Geo-Tracking Consumers and its Privacy Trade-offs

March 28, 2024

Abstract

Can geo-tracking data allow firms to better predict consumers' future behaviors? If so, how might potential privacy regulations limit the usefulness of geo-tracking data for prediction? Using data with over 120 million driving instances for 38,980 app users, and their visits to 422 restaurants in Texas, the authors quantify the extent to which geo-tracking data allow restaurants to better predict the number of visits one week ahead. They show that geo-tracking data increase the performance of prediction models by 14.77% relative to models that use demographic, behavioral, and static home location information. Simulation exercises that limit *what* data are tracked and in *what form*, *where*, and *how frequently* these data are tracked show a decrease in the predictive performance of models that use geo-tracking data. However, the decrease varies by the type of restriction; regulations that restrict *what* data are geo-tracked (i.e., summaries of driving behaviors) and in *what form* (i.e., synthetic data generated with nearby users' data) result in the largest decreases in predictive performance (16.24% and 8.09%), while regulations that restrict *where* users are geo-tracked (i.e., within a few miles of a business location) and *how frequently* (i.e., at longer intervals) result in smaller decreases (3.56% and .77-2.46%, depending on the frequency). Importantly, models with restricted geo-tracking generally outperform models that do not use any geo-tracking information. These findings can assist managers and policymakers in assessing the risks and benefits associated with the use of geo-tracking data.

Keywords:
Mobile apps, geo-tracking, privacy, targeting, machine learning

INTRODUCTION

In recent years, most firms have become capable of tracking consumers' locations and movement patterns through their own or third-party mobile apps (Valentine-De et al. 2018). Firms use geo-tracking data to predict consumers' future actions and make strategic decisions (Sun et al. 2022). Many restaurants, for example, use geo-tracking data to predict future visits to their locations to improve their service and create local promotions (Dean 2023). Burger King implemented a well-known application of geo-tracking when it offered the Whopper burger for one cent to customers who ordered it through its app while located within 600 meters of a McDonald's (Clifford 2018).[1]

Despite the potential usefulness of geo-tracking data for firms, these data reveal sensitive personal information about consumers (Bleier, Goldfarb, and Tucker 2020; Choi, Jerath, and Sarvary 2023; Goldfarb and Tucker 2012). For example, Canadian coffee chain Tim Hortons evoked a "mass invasion of privacy" by geo-tracking its app users round-the-clock (Austen 2022). As a consequence, the use of geo-tracking data by firms has attracted legal and regulatory action (Binns et al. 2018; Tau 2023). Data broker Kochava was sued by the Federal Trade Commission (FTC) for selling consumers' geolocation data that made it possible to identify their visits to sensitive locations (FTC 2022).[2] Recent privacy regulations, such as the California Privacy Rights Act (CPRA), explicitly recognize consumer location data as personal and sensitive information. Taking these concerns into account, researchers have proposed data obfuscation approaches to make such data privacy-preserving while retaining their usefulness for firms (Macha et al. 2023).

Though geo-tracking has attracted the attention of consumers, firms, regulators, and researchers, it is not clear to what extent geo-tracking data allow firms to better predict future outcomes and how privacy regulations impact the usefulness of these data. In this context, our research addresses two objectives. First, we examine the extent to which geo-tracking

[1] Many third-party firms, such as Radar and Bluedot enable businesses to build their geo-tracking capabilities and constitute the growing multi-billion dollar location data ecosystem (Macha et al. 2023)

[2] FTC also charged data vendors InMarket and X-Mode for selling consumers' raw location data (FTC 2024).

data are useful when predicting app users' visits to a business location one week ahead, relative to not using any consumer data and using only demographic, behavioral, and static home location information. Second, we examine how restricting geo-tracking data under potential privacy regulations impacts the usefulness of these data for prediction.

To address our first research question, we identify an application of geo-tracking data in the restaurant industry. While the potential usefulness of geo-tracking data depends on specific business objectives, our application focuses on predicting the number of visits to a restaurant one week ahead using the previous week's geo-tracking data as input.[3] Our empirical context is ideal for this research because many restaurants have access to geo-tracking data through their own or third-party apps, and are interested in predicting the number of visits they expect each week (Oblander and McCarthy 2023). Improving the predictions of weekly visits can allow restaurants to prevent under- or over-committing resources, such as staff and inventory.

To address our second research question, we then examine how restricting geo-tracking data under potential privacy regulations impacts the predictive performance of models that use these data. Specifically, we simulate four types of regulations that restrict *what* geo-tracking data are tracked and in *what form*, *where*, and *how frequently* these data are tracked. We motivate and develop these simulations based on privacy regulations, industry practices, and recommendations from the data obfuscation literature.[4]

In our research, we use proprietary data from a Texas-based app that tracks individual-level driving. The purpose of the app is to encourage safe driving. To do this, the app rewards points for driving safely without using one's phone. The points can be redeemed as discounts at business locations, which are primarily restaurants that the app has partnered with. The app has over 200,000 users. For our research, we can access data from a random sample of 38,980 app users, including over 120 million driving instances for 60 weeks in 2018-2019.

[3]We identify the prediction of weekly visits as a relevant objective for restaurant managers through a set of structured interviews. See Web Appendix A for a summary of interview responses.

[4]We verify that consumers perceive our simulation exercises to be privacy-preserving by conducting a survey about their perceptions (Jerath and Miller 2024; Lin and Strulov-Shlain 2023). See Web Appendix B for details.

We focus on visits to 422 standalone restaurants in 40 cities in Texas where the app was present during this time. While we are limited to a subset of restaurants, our sample is representative and includes chains like McDonald's and independent stores like Ozona Grill.

Our empirical approach exploits the detailed trip-level information for each user to extract features that characterize users' driving trajectories and restaurant visits. We then aggregate the data to the restaurant-week level. We use these data to train and evaluate a machine learning (ML) model at the restaurant-week level and then predict the total number of visits from app users one week ahead. We quantify our model performance using the out-of-sample root mean squared error (RMSE). We also report the results from alternative models at the restaurant level and for alternative metrics, such as the mean absolute error (MAE).

Our analysis shows that geo-tracking data improve the predictive performance of our models by 14.77% relative to models that use demographic, behavioral, and static home location information, and by 22.27% relative to models without any consumer data (i.e., only restaurant and time-related information). We also find that using geo-tracking data reduces the likelihood of both over-predicting and under-predicting future visits, which may allow restaurants to reduce the likelihood of both wasting supplies and staff hours (because of over-prediction) and of impacting customers' experience (because of under-prediction).

After quantifying the usefulness of geo-tracking data for prediction, we turn to our simulations of privacy regulations. Not surprisingly, we find that all forms of restrictions on geo-tracking data reduce their predictive value, though the extent of the decrease varies significantly across the simulations. Among our simulations, we find that restricting *what* data are geo-tracked (i.e., summaries of driving behaviors) and in *what form* (i.e., synthetic data based on nearby users) results in the largest decreases in predictive performance (16.24% and 8.09%), while restricting *where* users are geo-tracked (i.e., within a few miles of a business location) and *how frequently* (i.e., at longer intervals) results in smaller decreases (3.56% and .77-2.46%, depending on the frequency). Importantly, models with restricted geo-tracking generally outperform models that do not use any geo-tracking data.

Our research contributes to the growing literature on the value of consumer data for firms and to the literature on privacy regulations and data governance. First, the existing work on the value of data primarily focuses on the value of online data rather than offline geo-tracking data (e.g., Berman and Israeli 2022; Korganbekova and Zuber 2023; Rafieian and Yoganarasimhan 2021; Wernerfelt et al. 2022; Yoganarasimhan 2020). In one exception that is closest to our research, Sun et al. (2022) use omnichannel data, including data on consumers' offline visits for predicting each consumer's online activity. In contrast, we focus on predicting offline visits using complete geo-tracking data and on quantifying the implications of restricting geo-tracking under potential privacy regulations. Second, the research on privacy in marketing examines the impact of regulations for online firms (e.g., Goldberg, Johnson, and Shriver 2024; Johnson et al. 2023; Johnson, Shriver, and Goldberg 2023; Miller and Skiera 2023; Peukert et al. 2022, Zhao, Yildirim, and Chintagunta 2021), consumer perceptions of privacy (e.g., Jerath and Miller 2024; Lin and Strulov-Shlain 2023), and data obfuscation schemes (e.g., Li et al. 2023; Macha et al. 2023; Tian, Turjeman, and Levy 2023). We adapt recommendations from this literature for offline geo-tracking data in our simulations. In this way, our research extends these papers and the ongoing policy debate about the governance of consumers' geo-tracking data (e.g., FTC 2024; Tau 2023).

Our research has several implications for firms and policymakers. First, we show that using geo-tracking data can improve a firm's ability to predict weekly visits by reducing both over- and under-prediction of visits. Based on their context and the costs associated with over- and under-predicting, firms can evaluate their decision to collect and use geo-tracking data. Second, we propose practical ways in which firms can protect consumers' geo-tracking data. Third, we identify ways of restricting data that still allow firms to get predictive value from them. Specifically, we show that firms can restrict *how often* and *where* they geo-track consumers with relatively little loss in predictive value compared with restricting *what* data and *in what form* data are tracked. Finally, our findings from the simulation exercises can allow policymakers to assess predictive losses from various types of regulatory simulations.

RELATED LITERATURE

Our research relates to two streams of literature on the value of consumer data for firms and on privacy regulations and data obfuscation schemes for data collection and use.

Value of Consumer Data for Firms

Research on the value of data quantifies how different types of consumer data contribute to marketing outcomes of interest in various settings and applications. Historically, the marketing literature has documented the usefulness of consumers' purchase histories relative to their demographic data for designing targeted pricing strategies (e.g., Acquisti and Varian 2005; Rossi, McCulloch, and Allenby 1996). With the growth in digital and mobile technologies, the sources and types of consumer data available to firms have expanded rapidly (Lamberton and Stephen 2016; Varga et al. 2024). However, most research in this domain focuses on consumers' online data and not their offline trajectories (e.g., Berman and Israeli 2022; Wernerfelt et al. 2022; Yoganarasimhan 2020).

The paper that comes closest to our research is Sun et al. (2022). Their research quantifies the usefulness of both online and offline data for predicting each consumer's online actions i.e., their likelihood of visiting, considering a purchase, and purchasing at multiple websites two weeks ahead. In their work, Sun et al. (2022) use data on past online and offline trajectories and show that omnichannel predictions outperform single-channel ones by 7.38%. In contrast, we focus on predicting the total visits by app users to a retail location each week. We also examine how the accuracy of these predictions is impacted by privacy-preserving restrictions imposed on the data that firms can use. Such an inquiry is important for physical retail locations that are interested in predicting weekly visits, but that may be subject to privacy regulations that restrict data tracking.

Overall, relative to the literature on the value of data, our research focuses on quantifying the predictive value of geo-tracking data beyond traditional metrics like demographics,

behavioral, and static location information, and evaluating how this predictive value might change under potential privacy restrictions imposed on geo-tracking data.

Data Governance, Data Obfuscation, and Privacy

The research on data governance, data obfuscation, and privacy examines regulations and proposes restrictions that can impact how firms use customer data (e.g., Johnson et al. 2023; Macha et al. 2023). We broadly categorize these restrictions into the following types: Restricting *what* data are tracked and in *what form*, *where*, and *how frequently*. Table 1 summarizes these restrictions with motivating examples and their application to geo-tracking.

Table 1: Overview of Privacy Restrictions and their Related Geo-Tracking Simulations

Type of restriction	Motivation	Potential geo-tracking simulation	Example research
What user data are tracked	GDPR protects user data that contain personally identifiable information. Google's Sandbox technology proposes anonymizing user browsing data within the Chrome browser instead of collecting raw browsing data using cookies.	User-level summarization: Use summaries of driving behaviors extracted from geo-tracking data.	Johnson, Shriver, and Goldberg (2023); Miller and Skiera (2023)
In what form data are tracked	The data obfuscation literature proposes anonymizing user data to prevent re-identification, e.g., not using uniquely identifying information, adding noise to the data while preserving its value.	Synthetic data generation: Replace the user's data with nearby users' (i.e., k-nearest neighbors') data.	Li et al. (2023); Macha et al. (2023)
Where users are tracked	Google allows ad targeting depending on a country's laws. Many data vendors only sell consumer data tracked within specific geofences.	Geographical restrictions: Geo-track users only if they were within certain distances of a location (e.g., one mile).	Danaher et al. (2015); Dubé et al. (2017)
How frequently users are tracked	Regulatory actions discourage the sale of consumers' raw location data streams and could deter firms from high-frequency tracking. Firms may elect to track lower-frequency data or data at only static locations to optimize storage costs.	Frequency restrictions: Geo-track trips at lower frequency intervals, at the start and end points of a trip, or for a randomly selected trip per week instead of constantly tracking all trips.	Kim, Bradlow, and Iyengar (2022); Trusov, Ma, and Jamal (2016)

Notes: GDPR = General Data Protection Regulation. In Web Appendix B, we report consumer surveys to verify that our geo-tracking simulations are perceived as privacy-preserving.

What user data are tracked. Most privacy regulations, such as the California Privacy Rights Act (CPRA), consider geo-tracking data as sensitive personal data. One way in which privacy regulations restrict geo-tracking is to completely ban any form of location tracking (Tau 2023). In some cases, regulatory action also prohibits selling consumers' sensitive location data (FTC 2024). By comparing the predictive performance of models with geo-tracking data and those without any consumer data (i.e., only restaurant and time-related information) and without any geo-tracking data (i.e., only customer demographics, behavioral information, and static location), our research addresses this possibility.

In practice, however, a complete ban is less likely. Instead, regulations typically restrict what user data can be tracked. For example, the General Data Protection Regulation (GDPR) requires anonymizing personally identifiable information (PII) about users (Wang, Jiang, and Yang 2023), which impacts digital firms, publishers, and web technology vendors (e.g., Johnson, Shriver, and Goldberg 2023; Miller and Skiera 2023). Similarly, Google's *Topics* API in its Privacy Sandbox hides the specific sites users visit and, instead, infers broad interest-based categories to serve relevant ads.[5] Individual-level data generally perform better for targeting advertisements to consumers relative to aggregate data (Danaher 2023). As such, data brokers commonly use algorithms to create individual user profiles by combining data from multiple sources (Lin and Misra 2022; Neumann, Tucker, and Whitfield 2021; Yan, Miller, and Skiera 2022).[6]

One way to make geo-tracking data privacy-safe is user-level summarization, i.e., to extract features from geo-tracking data that describe users' driving behaviors without recording latitude-longitude coordinates or sensitive home location data. We address this possibility in our counterfactual on user-level summarization.

In what form data are tracked. Since individual-level geo-tracking data pose privacy risks by revealing exact home locations and trajectories, extant literature has proposed location privacy-preservation mechanisms for data obfuscation (Jiang et al. 2021). These mechanisms

[5] See, for example, Google's policy "Topics: Relevant ads without cookies". Accessed on March 25th, 2024.

[6] In practice, some firms (e.g., Bridg) create privacy-safe profiles for retail media networks to reach their audience.

broadly seek to change the form in which data are tracked, e.g., by adding noise or dropping data, in order to preserve consumer privacy while maintaining the usefulness of the data. Macha et al. (2023), for example, quantify the risk that from a set of trajectories, a private feature, such as home location, may be identified. Then, they find user-specific subsets of trajectories that preserve the usefulness of the data to an advertiser while keeping the privacy feature hidden. Methods, such as k-anonymity, similarly seek to anonymize data with respect to the unique user records that identify it to prevent linking a user's identity to sensitive datasets (Li et al. 2023). Microsoft's proposed Ad Selection API for their Microsoft Edge browser has privacy protections built into it, including k-anonymity constraints.[7]

The application of data obfuscation methods to geo-tracking data is not trivial. Altering location data can prevent them from being usable for businesses. Approaches that do balance utility-privacy trade-offs tend to be computationally intensive and do not generally lend themselves to business applications (Cunha, Mendes, and Vilela 2021; Terrovitis et al. 2017). The high dimensionality of spatial data results in these methods often lacking interpretability for managerial applications (see, for example, the discussion in Macha et al. 2023).

Our approach to make geo-tracking data privacy-safe is to adapt the recommendations from the data obfuscation literature in a managerially relevant way for location data. One way of doing this is to not use each user's geo-tracking data at all and instead, use their nearest neighbors' averaged data to construct synthetic geo-tracking features. We address this possibility in our simulation exercise on synthetic data generation.

Where users are tracked. Most apps and firms rely on geo-tracking in confined geofenced areas for their marketing applications. The likelihood of shoppers redeeming mobile coupons for stores inside a shopping mall increases if they receive the coupons in close proximity to the focal stores (Danaher et al. 2015). Similarly, targeting based on real-time or historical consumer location within geofences is more effective for firms (Dubé et al. 2017). In practice, third-party data providers, such as Radar and BlueDot enable geofencing services for firms

[7] See, for example, Microsoft's announcement "New privacy-preserving ads API coming to Microsoft Edge." Accessed on March 14th, 2024.

in a way that allows them to observe shoppers in their vicinity. Similarly, location-specific laws can allow firms to target users in select locations only.[8]

One way to restrict geo-tracking data is to disallow the tracking of users that are beyond a certain distance of a store and only track users who enter the geofence. We address this possibility in our simulation exercise on geographical restrictions.

How frequently users are tracked. The frequency with which geo-tracking data can be collected may be consequential for consumer privacy as it is for firm decision-making (Kim, Bradlow, and Iyengar 2022). Temporal limitations imposed on individual-level web tracking abilities impact online businesses (Trusov, Ma, and Jamal 2016). Regulatory actions that discourage or prohibit the sale of consumers' raw location data streams could deter firms from high-frequency tracking to avoid litigation and reputational damage, in addition to other deterrents like the cost of storing high-frequency data (FTC 2024). In the context of geo-tracking, temporal restrictions would suggest tracking consumers' locations at longer intervals or at static locations near points-of-interest (POIs) only.

One way to restrict the geo-tracking data is to reduce the frequency with which these data are tracked. We consider this possibility in our simulation exercise on frequency restrictions.

Overall, recent privacy regulations, firms' own self-regulation practices, and the data obfuscation literature seek to protect consumers' data and privacy in various ways. We leverage a few key privacy measures and the extant research on privacy to derive and evaluate privacy-safe simulation exercises for geo-tracking data in our application. Through a set of consumer surveys, we provide some face validity that these simulation exercises are perceived as privacy-preserving by consumers (see Web Appendix B). Importantly, by design, each of our proposed geo-tracking simulations restricts consumer geo-tracking data in a different way, either completely transforming the data (e.g., *what* and *in what form* data are tracked) or reducing the scope and extent of tracking (e.g., *where* and *how frequently* data are tracked). In this way, our proposed simulations vary in how they protect consumer geo-tracking data.

[8] See, for example, Google's policy "Target ads to geographic locations." Accessed on March 25th, 2024.

DATA AND EMPIRICAL STRATEGY

Data Sources

Our primary data source is a safe-driving app based in Texas, U.S.[9] The app has over 200,000 users. For our research, we accessed a random sample of about 20% of the app users (i.e., 38,980). The purpose of the app is to encourage safe driving. The app can detect when a user is driving at a speed of over 10 miles per hour and if they are using their phone while driving. The app incentivizes safe driving by awarding users a fixed number of points for each mile driven when they do not use their phone while driving. The points can be redeemed at partnering firms, which are primarily restaurants. Although the app is unique in its safe driving aspect, it shares some commonalities with deals and delivery apps by offering information about local businesses. The nature of data collected by the app is also not unique, and most apps that access location tracking have similar data collection abilities, e.g., food delivery apps and navigation apps.

The app records a user's position (i.e., latitude and longitude) every three minutes once it detects that a trip has begun. The app uses this information to record locations and driving speed with date and timestamps. The app company gave us access to individual-level geo-tracking data for 60 weeks between September 2018 and October 2019, comprising over 120 million driving points. We use these data to identify driving trajectories and restaurant visits. We also have access to data on app users' demographics (e.g., age, gender, zip code) self-reported by users at the time of signing up for the app.

The second source of our data is Safegraph.[10] This dataset consists of the polygons (i.e., geometries) that identify each restaurant in Texas. We use Safegraph's geometry/polygon data to identify restaurant locations.

[9]The app is headquartered in College Station, Texas. During our sample period, the app had business partners and users in cities, such as Houston and San Antonio, among others.

[10]See https://www.safegraph.com/academics "SafeGraph is a data company that aggregates anonymized location data from numerous applications in order to provide insights about physical places, via the SafeGraph Community. To enhance privacy, SafeGraph excludes census block group information if fewer than five devices visited an establishment in a month from a census block group." We use Safegraph's geometry/polygon data only.

We supplement these data with additional data from Yelp to identify the category of each restaurant using Yelp's category tags (Klopack 2024).[11] We use these data when constructing our feature set relating to past user visits to a restaurant of the same category and brand as the target restaurant for which we are interested in making predictions.

Finally, we also access the American Community Survey (ACS) 2016 5-year estimate for the Census Block Groups (CBG) in our sample to generate demographic features, such as household income and education. These features are not directly recorded at the user level by the app but could contain important demographic information at the CBG level.[12]

Identifying Visits

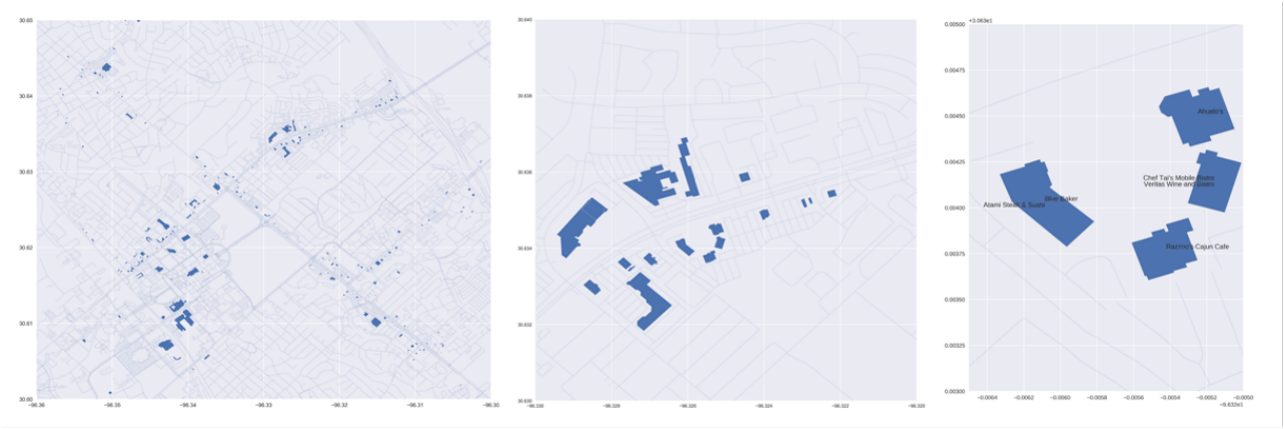
We identify restaurant visits from geo-tracking data by combining them with retail geometries in two steps. First, we identify instances when a driver is stationary. Based on the data, we operationalize "stops" as intervals of at least 10 but at most 120 minutes when the driver is not moving (Pappalardo and Simini 2018). Next, we classify a stop as a "visit" when the location coordinates lie within a polygon defined by the longitude-latitude pair of each vertex of the restaurant location. Examples of polygons of restaurants in Bryan/College Station are shaded in blue in the maps shown in Figure 1.

Identifying visits by overlaying polygons and driving trajectories may not capture a fraction of the true visits to restaurants that are adjacent or that are within another store (e.g., Subway within Walmart). For this reason, we only use standalone restaurants in our analyses. Similarly, if the last point recorded for a specific trip lies outside the polygon of a restaurant, we would not identify this point as a visit if the user entered the restaurant just after the last GPS point was recorded. In these cases, training the model with the resulting data would lead to under-performance relative to a model trained with true visit data.

[11] The Yelp category tags and the main category appear in Web Appendix C.

[12] The American Community Survey has been used extensively in academic research to extract demographic information at the Census Block Group and Census Tract levels. See, for example, Avenancio-León and Howard (2022); Bertrand, Kamenica, and Pan (2015); Chetty, Hendren, and Katz (2016); Klopack, Lewis, and Luco (2024); Landvoigt, Piazzesi, and Schneider (2015); Naik, Raskar, and Hidalgo (2016).

Figure 1: Restaurant Polygons in Our Data



Notes: The polygons are geometric boundaries available for each store using satellite imagery.

Sample Restaurants for Analysis

Our sample of restaurants for making predictions comprises 422 standalone restaurant locations. To arrive at this sample, we considered 10,582 standalone restaurant locations across the 40 cities in Texas in which the app was present during our data period. To avoid sparse outcomes, we chose the sample of 422 restaurants (about 4% of the total restaurants) that at least 10 app users visited in our data period. We verify that these locations are representative of the larger set of 10,582 standalone restaurants. Specifically, we examined the type of restaurant (chain vs. non-chain) and the distribution of restaurant brands and categories overall and in our sample. Our sample is similar to the broader set of restaurants in terms of the proportion of chain stores (i.e., 75.05% overall vs. 73.69% in our sample), the brand of the restaurant (i.e, top chains are fast food brands like McDonald's, Starbucks, and Sonic overall and in our sample), and the food category of the restaurant (i.e., top categories are American, Burger, and Latin American overall and in our sample).[13]

[13]Even though our 422 restaurants are a representative sample, in Web Appendix Tables D1-D4, we report the results for alternative samples of restaurants.

Aggregating Data to the Restaurant-week Level

We identify restaurant visits at the user-week level. However, restaurants are typically interested in predicting the total number of visits (see a summary of our interviews with managers in Web Appendix A). To transform user-week level data to the restaurant-week level, we take the sum of visits from the app users to a restaurant in a given week. We use this total number of visits as our target outcome of the prediction.

In addition to the total visits, we also provide a way to aggregate data for the input features (i.e., predictors like demographics, behavioral, and geo-tracking information). To do this, we identify the relevant population for each restaurant as users who were within 30 miles of the restaurant in the previous week, the maximum distance users travel to visit restaurants in our data.[14] We then aggregate the individual-level features of these users to the restaurant-week level. For example, if a Starbucks location has a subset of app users nearby in a given week, we average their user-week level data on each feature to create the data for that Starbucks location for that week. Note that in this approach, the relevant set of users for a restaurant varies by week depending on who was driving nearby.[15] Our models predict one week ahead to allow inter-temporal separation between the time period over which the feature set is constructed and the outcome (Lee, Yang, and Anderson 2021).

Information Sets for Predicting Aggregate Visits

In this section, we introduce four models that vary in the information set used as input and that allow us to quantify the relative predictive value of geo-tracking data. The models include a baseline model without any consumer data (i.e., only restaurant- and time-related features) and three models with different sets of consumer data. We report the information sets used in our prediction models in Table 2.

[14]In Web Appendix Table D5, we report the results for an alternative threshold of 17 miles. We use 17 as our alternative threshold because it is the mean distance users travel to visit a restaurant, conditional on visiting.

[15]In Web Appendix Table D6, we repeat this analysis by aggregating over users that live, rather than drive, within 30 miles of a restaurant.

Table 2: Information Sets for Predicting Restaurant Visits

Model	Feature set used as input
Baseline	Restaurant and time features
Model A	Demographics + behavioral information + home-zip code distance
Model B	Demographics + behavioral information + home-tracked distance
Model C	Demographics + behavioral information + home-tracked distance + other driving features

Notes: Baseline features include restaurant category, city, and season effects. *Demographics* include age, gender, and census-block level data on education, income, etc. but no location information. *Behavioral information* refers to consumers' number of past visits to the restaurant in the data period before the target week in the prediction model (e.g., 30 past weeks when predicting for week 31, 31 past weeks when predicting for week 32, and so on). *Home-zip code* distance refers to the distance between the centroid of the zip code (self-reported in the app) in which a customer resides and the target restaurant in the prediction model. *Home-tracked* distance refers to the distance between the latitude-longitude of the customer's home (recovered from geo-tracking data) and the target restaurant. Unlike static home locations, *other driving features* capture time-varying information recovered from geo-tracking data, e.g., trip distance, which is the minimum average distance between the user's trips and the target restaurant overall and for different time-of-day windows (e.g., 8 am to 12 pm, 12 pm to 4 pm, and so on).

Baseline Model. Restaurants can make a prediction about the total number of visits in a week without using any consumer-level information. Retailers commonly form an idea about the visits they expect based on their category, location, and the time of year, among other aggregate features (e.g., Varga et al. 2024). Therefore, we use this information in our baseline model with restaurant- and time-related features. Our baseline model allows us to benchmark the predictive performance of models with other sets of consumer-level predictors (e.g., demographics, behavioral, geo-tracking data).

Model A: Demographic and Behavioral Information with Home-Zip Code Distances. In our first specification, we include information that is commonly available to restaurants about their customers' demographics and past behavior relating to the restaurant. The demographic features include consumers' age, gender, and publicly-available ACS data that contain information about the population, race, employment, income, home-work commute, household size, and education at the census-block level. We also include behavioral information about the customers' number of past visits to the restaurant. Restaurant managers are likely able to access historical information on the total number of invoices, orders, or reservations for customers who use their app or reservation system. However, they cannot easily get this

information for competing restaurants or for other locations of their own parent brand in the case of chains (Weis 2023). Therefore, we include past visits to the target restaurant as behavioral information when predicting visits to that restaurant for a given week. However, we treat visits to other restaurants of the same brand or category as geo-tracking information.

Since behavioral information aims to capture consumers' past patronage of the restaurant, we reserve the first half of our data (i.e., weeks 1-30) to compute this information. We then make predictions for each week, starting week 31 till week 60. In this approach, past visits use the data from weeks 1-30 to predict for week 31, then weeks 1-31 to predict for week 32, and so on (consistent with Sun et al. 2022). Thus, behavioral information captures visits computed over an increasing number of weeks every week.

Finally, we also include information on consumers' home zip codes in Model A. Many restaurants do not observe their customers' home addresses. However, restaurants can often access consumers' home zip code information through their app or reservation system. We compute the home-zip code distance as the distance from the centroid of a zip code to the target restaurant for which we are predicting. We compute this distance for each restaurant in our data for the relevant set of consumers each week.

Model B: Demographic and Behavioral Information with Home-Tracked Distances. In this specification, we replace the home-zip code distance in Model A with home-tracked distance, i.e., the distance between the latitude-longitude coordinates of a consumer's home address and the target restaurant for which we are making the prediction. Home-tracked distance is derived from geo-tracking data and captures static information about a consumer's precise home relative to the restaurant. To compute home-tracked distances, we follow Pappalardo et al. (2022) and first use the latitude and longitude of geo-tracking data to identify the home location of each user. We then use the home location to calculate a user-restaurant specific "distance from home" measure (i.e., home-tracked distance to a restaurant). As with home-zip distance, we also compute the home-tracked distance for each restaurant in our data for the relevant set of consumers each week.

Model C: Demographic, Behavioral, and Complete Geo-Tracking Information. In this specification, we build on Model B's static home location and add time-varying features from geo-tracking data. Because any single geo-tracking metric can be fairly privacy-friendly relative to the full geo-tracking data and because, in practice, firms are likely to use multiple metrics, we extract several features. This task is not trivial. First, geo-tracking data have spatio-temporal richness but are also noisy. Second, using the entire trajectory of consumer movement as inputs in our prediction models is computationally intensive and not readily scalable. Finally, any information derived at the consumer-week level needs to be aggregated to the restaurant-week level (see Section on Aggregating Data to the Restaurant-week Level).

To extract useful information from geo-tracking, we focus on the following key features of these data. First, similar to the home-tracked distance, we compute a trip distance for each user-week as the minimum distance at which the user was from each restaurant in our data in the previous week before the week for which we are predicting visits. It provides additional information relative to home-tracked distance because it changes every week and considers a user's driving activity relative to the target restaurant. Second, since users' driving patterns vary depending on the time of day, we also compute the trip distance by time-of-day to each restaurant for different time-of-day windows (e.g., 8 am to 12 pm, 12 pm to 4 pm) in the previous week. Third, geo-tracking data can provide information to managers about their consumers' visits to other restaurants and not just their own restaurants. Therefore, we include geo-tracking features about the past number of visits and the recency of visit to any restaurant of the same category (i.e., number and recency of category visits) and the same brand (i.e., number and recency of brand visits) as the target restaurant for which we are predicting. Finally, we also include information about the previous week's visits to the target, same-brand, and same-category restaurant for those driving near the restaurant (i.e., within 30 miles). Geo-tracking data allows tracking this information not commonly available through point-of-sale terminals or reservation systems of a restaurant.

We report the summary statistics of the features at the restaurant-week level in Table 3.

Table 3: Information Sets for Prediction and Summary Statistics

Feature	Description	Mean
Target	No. of weekly visits by app users	2.35
Demographic/behavioral information		
Age	Average age of users	32.65
Gender (female)	Proportion of females	.47
Race	Proportion of white population in a census block	.72
Employment	Proportion of full-time employed people	.45
Commuters	Proportion of $\geq$ 16 year-olds commuting to work	.44
Family households	Proportion of households with $>$ 1 person in a census block	.24
Education	Proportion of $\geq$ 25 year-old with highschool diploma	.58
Income	Median household income ($) in the past 12 months	72,509
Home-zip code distance	Distance (in miles) between the centroid of the home-zip code and the restaurant	61.54
Past visits	No. of visits to the restaurant in the past	64.93
Geo-tracking information		
Home-tracked distance	Distance (in miles) between home coordinates inferred using geo-tracking data and the restaurant	27.03
Trip distance	Minimum distance (in miles) between trip coordinates and restaurant for the previous week	11.06
Trip distance - 0004	Trip distance for trips between midnight and 4 am	10.80
Trip distance - 0408	Trip distance for trips between 4 am and 8 am	10.76
Trip distance - 0812	Trip distance for trips between 8 am and 12 pm	11.02
Trip distance - 1216	Trip distance for trips between 12 pm and 4 pm	10.50
Trip distance - 1620	Trip distance for trips between 4 pm and 8 pm	10.71
Trip distance - 2000	Trip distance for trips between 8 pm and midnight	10.63
Past brand visits	No. of visits to same-brand restaurants in the past	1,383
Past category visits	No. of visits to same-category restaurants in the past	4,943
Previous week visits	No. of visits to the restaurant in the previous week if the individual was within 30 miles of the restaurant	1.58
Previous week brand visits	Previous week visits to same-brand restaurants	7.84
Previous week category visits	Previous week visits to same-category restaurants	28.24
Recency of past visit	Days since last visit to the restaurant	54.10
Recency of past brand visit	Days since last visit to same-brand restaurants	49.94
Recency of past category visit	Days since last visit to same-category restaurants	46.10

Notes: The statistics are reported for the 422 restaurants in our sample and their aggregated data over 5,951 observations at the restaurant-week level. Past period refers to rolling window of 30 or more weeks upto the prediction week and previous week refers to the week before the prediction week.

On average, a restaurant gets 2.35 visits from the app users in a week. In terms of demographics, the average age of users is 32.65, 47% are female, 72% are white, 45% are in full-time employment, 44% are commuters, 24% live in family households, 58% have at least a high school diploma, and the median household income is $72,509. Based on their home zip code, on average, users live 61.54 miles from the restaurant.[16] On average, a restaurant records 64.93 visits in the past from the users in our sample. Among the geo-tracking features, the home-tracked distance is 27.03 miles and the trip distance is 11.06 miles. Based on the time of day, the trip distance ranges from 10.50 miles during the 12pm to 4pm window and 11.02 miles from 8am to 12pm. On average, restaurants of the same brand as the target restaurant record 1,383 visits in the past weeks from the app users and those of the same category receive 4,943 visits. The target restaurant received 1.58 visits in the week before the week of prediction, same-brand restaurants received 7.84 visits, and the same-category restaurants received 28.24 visits. On average, the last consumer visit was 54 days ago to the target restaurant, 50 days ago to the same-brand restaurants, and 46 days ago to the same-category restaurants.

To illustrate the patterns captured by the geo-tracking data, the panel at the center of Figure 2 presents a two-dimensional projection of all the driving features in a multi-dimensional space of our entire feature set (see Web Appendix E for details). Each point in this projection represents a user. The driving features determine the shape of this visual, but the points in the figure are color-coded by one feature, i.e., the minimum distance between a user's trips to the target restaurant for a specific restaurant-week. The users represented by purple dots have lower trip distances, while those represented by pink dots have higher trip distances. To illustrate these patterns, we also show the heat maps of driving patterns for two users in different regions of the projections (Van der Maaten and Hinton 2008). The user from the purple region of the plot on the left, for example, has a lower driving distance and drove

[16] Several reasons may explain the disparity between the *home-zip code* and the *home-tracked* distance. For example, the app is headquartered in College Station, a college town. If students in College Station report their zip code in the app as that of their hometown rather than their local zip code, these distances will differ significantly.

mostly within Katy, Texas. In contrast, the user from the pink region of the plot on the right has a higher trip distance spanning Killeen, Austin, and College Station in a week.

Figure 2: Visualizing Geo-Tracking Data: Trip Distance

Machine Learning Framework and Data Splits

Once we generate the total number of visits by app users each week to our sample of restaurants and the input feature sets at the restaurant-week level, our final dataset for prediction has 5,951 restaurant-week level observations. This dataset represents an unbalanced panel of 422 restaurants observed over one or more of the 30 weeks for which we make predictions (i.e., the second half of the data). We split the data at random into 80% restaurant-weeks for training a model and use the remaining 20% restaurant-weeks as test data to evaluate the trained model. Thus, our main empirical strategy is to train one ML model at the restaurant-week level using the training data and evaluate and report the model's performance using the test data. The separation into training and test data ensures that the model is learning general patterns during training and is less likely to overfit to the same restaurant-weeks whose patterns it learns. We quantify the out-of-sample model performance for the test data using the root mean squared error (RMSE).

While our main empirical approach of estimating models with different information sets already benchmarks against a model with only restaurant- and time- features, we also report three alternative empirical approaches. First, we report alternative models that explicitly predict deviations in visits based on seasonality as the target rather than the total number of visits (see Web Appendix Table D7). Second, we report a model that splits the data by week within each restaurant to train and predict visits for that restaurant rather than at the restaurant-week level (see Web Appendix Table D8). Finally, we repeat our main analysis for an alternative outcome that scales up the total visits from app users to the entire population of the city in which the restaurant is located, i.e., target visits $\times$ ratio of the city's population to the number of app users in that city in our sample (see Web Appendix Table D9). Across these alternative approaches and outcomes, we find results that are consistent with those from our main model, although the magnitude varies.

Model Training

To predict visits to each restaurant one week ahead, we train one model for all the restaurant-weeks in our training data. We then evaluate the performance of the model on the test data.

Because our goal is to quantify the predictive value of geo-tracking data over and above demographic, behavioral, and static location information, we use various ML algorithms, such as the Least Absolute Shrinkage and Selection Operator (Lasso), Ridge Regression, Elastic Net Regression, and Boosted Regression Trees. Because Elastic Net offers better performance than Ridge and Lasso, and is more efficient than boosted trees for our data, we report this model in the main results. We report the alternative models in Web Appendix Tables D10, D11, and D12. The results from these models follow similar patterns as those from our main model.

Out-of-sample Evaluation

We evaluate the performance of our trained ML model by computing the root mean squared error (RMSE) associated with using each information set in Table 2. We report the average RMSE for each model for the test data.

The RMSE is the square root of the sum of squared differences between the predicted and observed number of visits for each restaurant and week combination rw in our data. It is computed as follows:

$$RMSE = \sqrt{\frac{1}{N_{test}} \sum_{rw=1}^{N_{test}} (y_{rw} - \hat{y}_{rw})^2}$$

We also report the mean absolute error (MAE) as an alternative performance metric. MAE is calculated by taking the average of the absolute differences between the predicted and observed number of visits for each restaurant-week. The formula for MAE is as follows:

$$MAE = \frac{1}{N_{test}} \sum_{rw=1}^{N_{test}} |y_{rw} - \hat{y}_{rw}|$$

Bootstrapping Procedure

We use a bootstrap procedure to evaluate the differences in predictive performance across models with different information sets. We implement this procedure by generating 2,000 bootstrap samples from the training and test data with replacement for each restaurant-week combination in our data. For each sample, we compute the performance metrics for the models under consideration. By doing this over the bootstrap samples, we can construct a distribution for each measure of interest and model under consideration. We then use the percentile method to construct 95% bootstrapped confidence intervals of each evaluation metric. Our bootstrapping procedure allows us to quantify the uncertainty in our estimates.

RESULTS: PREDICTIVE PERFORMANCE BY INFORMATION SET

In this section, we present the findings for our first research objective: examining the extent to which geo-tracking information improves the predictive performance of our models relative to consumers' demographics, behavioral, and static home location information.

Table 4 reports the predictive performance of each of our models corresponding to the information sets in Table 3. Specifically, we report the RMSE for each model, the difference in RMSE between models, and the bootstrapped confidence intervals computed using the test data at the restaurant-week level.

The results in Table 4 show that all the models with consumer data (i.e., Models A-C) perform better than the baseline model in terms of reducing the RMSE, our measure of predictive performance. Model C, which includes the geo-tracking data, performs better than all other models, including those that contain static location information. The RMSE for Model C is 22.27% lower than that of the baseline model, 14.73% lower than that of Model A with demographics, behavioral information, and home-zip code distance, and 14.77% lower than that of Model B with demographics, behavioral information, and home-tracked distance. Our bootstrapped confidence intervals further show that the differences in performance between Model C with geo-tracking data and all other models are statistically significant, though we also find that there is no significant difference between Models A and B.

While the performance metrics improve when we include geo-tracking data, the percentage improvement in the root mean squared metric does not easily lend itself to business objectives. To address this, we evaluate the performance of Model C separately for instances when the model over-predicts and for instances when it under-predicts and compare its performance to that of Model A, which does not use geo-tracking data. The idea behind this analysis is that over- and under-prediction may generate different costs to restaurants, and thus, it is relevant to determine whether Model C performs better than Model A across these instances or if it is better only in one of these instances. We present our findings in Figure 3, which reports the

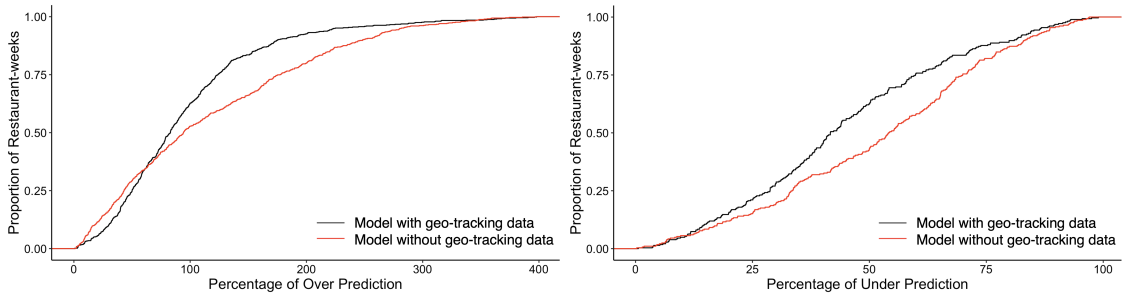
Table 4: Results: Predictive Performance by Information Set

Model	Mean RMSE	Difference from Baseline	Difference from Model A	Difference from Model B
Baseline	5.643 [5.589, 5.698]			
Model A: Home-zip distances	5.144 [5.097, 5.191]	.499 [.487, .512]		
Model B: Home-tracked distances	5.146 [5.088, 5.193]	.497 [.485, .510]	-.002 [-.003, -.002]	
Model C: Home-tracked, trip distances, and other driving	4.386 [4.341, 4.432]	1.257 [1.223, 1.291]	.758 [.730, .785]	.760 [.732, .787]

Notes: RMSE = Root mean squared error. In column "Mean RMSE," the confidence interval corresponds to that of the mean RMSE for that model. In all other columns, the confidence interval corresponds to that of the mean difference between two models. We implement this using 2,000 bootstrap replications for each model, as described in the text. N = 5,951 restaurant-weeks. Note that all feature sets include controls for seasons, cities, and categories. Results correspond to Elastic Net models.

empirical cumulative distribution function (CDF) of the absolute difference between actual and predicted visits, as a percentage of actual visits for models with geo-tracking data (Model C) and those without geo-tracking data (Model A). The empirical CDF of models without geo-tracking lies to the right of that of models with geo-tracking data, which suggests that models using geo-tracking predict visits better in both cases.

Figure 3: Actual vs. Predicted Visits under Models with and without Geo-tracking Data



Notes: The plots show the cumulative proportion of restaurant-weeks by the extent of over- and under-prediction of visits using models with and without geo-tracking data (Model C vs. A).

SIMULATION EXERCISES: RESULTS UNDER RESTRICTED GEO-TRACKING

In this section, we report our results for our second research question: how does restricting geo-tracking data under different types of privacy regulations impact the predictive performance of models that use these data? This question is important from a policy perspective because of the recent emergence of regulations restricting data tracking (FTC 2024; Klosowski 2021). If policymakers regulate consumer geo-tracking by requiring, for example, data summarization so firms cannot access coordinate-level data, or by imposing geographical restrictions on where users are geo-tracked, how would such restrictions impact the predictive performance of models that use these restricted data?

To answer this question, we outline an approach to quantify the extent to which restricting geo-tracking under alternative privacy regulations impacts predictive performance. In Table 1, we motivated four types of regulatory restrictions and how they can be applied to geo-tracking data. Next, we describe how we simulate each of these regulations in our setting and discuss how they impact predictions from our models relative to unrestricted geo-tracking.

User-level summarization. In the first category of simulations, we consider regulations requiring data to be anonymized with respect to the user that generates that data through summarization. We implement a version of this following Pappalardo and Simini (2018) and construct driving summary features for each user-week. These summary features contain aggregated information about a user's driving behavior in a week, such as the total distance traveled, entropy (i.e., variability of the locations visited relative to their past distribution of visited locations), time of driving, number of days driven, and number of trips each week. These features are independent of the target restaurant's location and capture general mobility patterns about a user in a privacy-preserving way (e.g., see consumer surveys in Web Appendix B). The technical details of computing these features and their summary statistics appear in Web Appendix E. Under this simulation, we train a model with the

baseline features, the demographic and behavioral information, and these summary features. Unlike Model C that uses home-tracked locations, trip-distances, and other geo-tracking features, this model uses aggregated driving summaries.

Synthetic data generation. In the second category of simulations, we consider regulations that require adding perturbation or noise to the data to generate synthetic data. We implement a version of this using a k-Nearest Neighbor approach at the user level, which allows us to ensure that users cannot be re-identified from these data. Specifically, we identify each app user's $k = 10$ nearest neighbors based on home locations and use the average of their data to replace the user's data. For example, to compute the focal user's home location, we take the centroid of the polygon formed by the home locations of these nearest neighbors. To compute variables such as the minimum distance between a trip and a restaurant, we consider the mean distance among the set of minimum distances of the nearest neighbors and the target restaurant. By following this approach, we ensure that the data that we use as input for our models cannot be used to re-identify specific users but is generated based on averages over the other users. Under this simulation, we train Model C with the new synthetic data generated for each user-week and aggregated to the restaurant-week level as in our main analysis.[17]

Geographical restrictions. In the third category of simulations, we explore how geographical restrictions impact prediction outcomes. We implement this simulation using geofences that restrict firms to observe only users who entered a certain radius (e.g., one mile) of their location. Any users that are outside this distance are, therefore, not observable to the firms and are excluded from our simulated prediction model. However, conditional on a user being within the one-mile geofence, the firm is able to observe their geo-tracking data and is able to recover, for example, true home locations. Under this simulation, we train Model C using these restricted geo-tracking data on only those users who entered the geofence.[18]

[17] To aggregate to the restaurant-week level, we consider users whose "synthetic" distance generated using the neighbors' data was within 30 miles of the restaurant in the week before the prediction week to mimic a situation in which the firm can only observe synthetic data and never access the focal user's original data.

[18] An alternative way of implementing geofences is to restrict both users and trajectories to be tracked only within

Frequency restrictions. In the fourth category of simulations, we consider regulations that restrict *how often* users may be tracked but still allow firms to use the data at the coordinate level. In practice, we consider regulations that record the data less often than what our focal app does. We implement two versions of this.

In the first version, we keep the first point of each trip but systematically drop the data within a trip. Specifically, we re-construct the geo-tracking data at lower frequencies, assuming they are collected at one-half and one-third frequency of the original three-minute interval with which our data provider records the data.[19] These exercises are meant to represent choices firms might make about temporal granularity when deciding how often to record data.

In the second version, we implement simulations that reduce geo-tracking frequency in ways that replicate static geo-tracking regulations. These exercises represent scenarios in which firms may track users at specific points of interest, e.g., at the start and end of their trips at places of business, recreation, and so on. In practice, we consider two implementations that differ along this dimension. In the first one, we record the first and last points of a trip and drop all other records. In the second one, we record data for one random trip per user-week, keep all records of it, and drop any other trips that week.

Under each simulation exercise, we use the restricted geo-tracking data for both training and test purposes. Even though we implement four types of privacy restrictions based on Table 1, it is important to note that these restrictions are qualitatively different in how and how much they may protect consumer privacy. While some restrictions completely prevent user identification, others may simply hide some features of their data.[20]

Next, we discuss the results of our simulations.

---

the geofence. However, doing so artificially forces the home-tracked distances to be within the geofence. Therefore, we prefer a more conservative approach of keeping all user data conditional on the user being within the geofence. We also consider the possibility that, in practice, businesses could use a different geofence distance around their store locations. We report alternative versions of this simulation with other radii in Web Appendix Table D13.

[19] In Web Appendix Table D13, we also implement a variation that collects data at one-tenth of the original frequency and one that keeps the overall *amount* of data the same as the 1/2 geo-tracking frequency but drops data randomly rather than systematically.

[20] In Web Appendix F, we provide examples from our data to illustrate how each simulation protects user identity.

Table 5: Results: Simulation Exercises with Varying Restrictions on Geo-Tracking

Simulation	Mean RMSE	Difference	Percentage
Complete geo-tracking (Model C in Table 4)	4.386 [4.341, 4.432]		
User-level summarization			
Summary features instead of geo-tracking	5.098 [5.050, 5.146]	.712 [.646, .777]	16.24%
Synthetic data generation			
K-nearest neighbors' data	4.741 [4.692, 4.791]	.355 [.342, .368]	8.09%
Geographical restrictions			
Geofenced users within 1 mile of restaurant	4.542 [4.494, 4.591]	.156 [.093, .220]	3.56%
Frequency restrictions			
Reduced frequency of geo-tracking			
1/2 frequency	4.420 [4.377, 4.463]	.034 [.004, .063]	.77%
$1/3^{rd}$ frequency	4.494 [4.448, 4.540]	.108 [.098, .117]	2.46%
Static geo-tracking			
First- and last- trip points only	4.492 [4.445, 4.538]	.106 [.100, .111]	2.42%
One trip per week at random only	4.485 [4.438, 4.532]	.099 [.038, .160]	2.26%

Notes: RMSE = Root mean squared error. The complete geo-tracking model includes demographics, behavioral, and geo-tracking data (i.e., Model C of Table 4). This Table reports the mean and bootstrapped confidence intervals (in square brackets) of the RMSE of each model and the difference in RMSE between the complete geo-tracking model and each simulation using the test data. Results correspond to Elastic Net models.

Results: Predictive Performance under Varying Restrictions on Geo-Tracking

Table 5 presents our findings for the simulation analysis. In each simulation, we report the results for Model C after re-training and evaluating it using restricted geo-tracking data.

Our main finding in this section is that all the restrictions that we evaluate result in a lower predictive performance than that of Model C with complete geo-tracking. However, the decrease in performance varies by the type of restriction. Importantly, we find that even

under various policy restrictions, models with geo-tracking information generally perform better than those that do not use geo-tracking information at all.

Next, we describe the results in detail. First, we find that regulations that use summaries of geo-tracking data at the user level result in the largest decrease in predictive performance relative to Model C. Specifically, the RMSE is 16.24% larger than the one in Table 4.

Second, we find that regulations that generate synthetic data using k-nearest neighbors (k-NN) also result in higher RMSE, which is 8.09% larger than Model C with complete geo-tracking presented in Table 4. While the use of synthetic data results in a significant decrease in predictive performance, this decrease is significantly smaller than the one associated with using summarization of geo-tracking data. By construction, the synthetic data approach introduces fewer changes to the data that our model can use as input than driving summaries, and our findings imply that this less restrictive approach preserves some of the predictive power that is lost under user-level summarization.[21]

Third, we find that regulations that restrict firms to observe data only for users who were within a mile of their locations result in a 3.56% decrease in predictive performance relative to Model C with complete geo-tracking data. This simulation shows that though geofences impose a significant restriction to firms that collect geo-tracking data, the restricted data are still useful when predicting customer visits relative to a context in which these data are not available at all or if they are restricted in other ways (e.g., limiting *what* data are tracked).

Fourth, we consider simulations that restrict the frequency with which geo-tracking data are recorded, including reduced frequency and static geo-tracking at specific times only. We find that using reduced-frequency data decreases the predictive performance by .77% and 2.46% under one-half and one-third tracking frequencies, while using static geo-tracking decreases the predictive performance by around 2% relative to complete geo-tracking.

Overall, our findings show that regulations that restrict the data that can be used to

[21] The main difference between user-level summarization and synthetic data generation is that the synthetic data approach keeps all the geo-tracking features but in a form that makes it harder to identify individual users, while user-level summarization replaces the sensitive geo-tracking features with driving summaries.

predict app users' visits reduce the performance of the model relative to when unrestricted geo-tracking data are used. Though the direction of effects is somewhat expected, the relative magnitude of prediction losses provides us with rich insights and shows that not all regulatory restrictions are equal in their implications for prediction. Specifically, we find that the largest reductions in predictive performance are associated with restrictions that fully transform the data. For example, user-level summarization, which seeks to prevent user identification, results in the largest decreases in predictive performance. On the other extreme, frequency restrictions, which retain coordinates in their raw form, reduce predictive performance the least. Between these extremes, we consider a number of regulations that vary in how they restrict geo-tracking data. Even in the most restrictive cases, the prediction model performs similarly to or better than when the models are restricted to not using geo-tracking data.

ROBUSTNESS CHECKS

Alternative Restaurants

In our main analysis, we report the results for a sample of 422 restaurants that at least 10 app users visited. Even though we verified that these 422 restaurants are representative of the broader set of restaurants, it is possible that the results are sensitive to the specific restaurant sample. Thus, in Web Appendix Table D1 and Table D2, we report the results for alternative samples of restaurants based on more stringent sample selection criteria of at least 15 and 25 users visiting those restaurants anytime in our data period. It is also possible that our findings are generated by the sample of restaurants that have relatively few visits and that our model is unable to predict similarly for more popular restaurants. To examine this, we report the results for the top restaurants by number of visits during our prediction period in Web Appendix Table D3 and Table D4. We find qualitatively similar results across all the samples, although the magnitude varies based on the number of visits in each sample.

Alternative Aggregation Thresholds

In our main analysis, we use a 30-mile distance threshold to identify a restaurant's potential customers and aggregate their demographic, behavioral, and geo-tracking features to predict the total visits to a restaurant each week. To test the robustness of this 30-mile threshold, which is based on the maximum distance consumers traveled to visit a restaurant in our data, we also report the results for two alternative analyses. In the first analysis, we use a threshold of 17 miles, which is the mean distance consumers travel to visit a restaurant in our data. The results are reported in Web Appendix Table D5. In the second analysis, instead of aggregating over consumers who drove within 30 miles of the restaurant in the previous week, we aggregate over consumers whose home-zip code distance is within 30 miles of the restaurant. The results are reported in Web Appendix Table D6. We find that Model C with the geo-tracking data outperforms the models without these data in both cases.

Alternative Model Setups

We test the robustness of our results to three different model setups based on alternative target and data splits. First, instead of predicting the total number of visits, we predict deviations in visits as our outcome since restaurant managers may be interested in weekly demand fluctuations. By comparing our models to a baseline, we already capture the predictive value of geo-tracking data to baseline predictions using restaurant- and time-features. However, we also directly model deviations in this alternative model setup. To generate the data on these deviations, we regress the total visits on season identifiers and compute the residuals. Next, we use these residuals as the target of our prediction models, excluding the top .05% observations with the highest residuals. The results appear in Web Appendix Table D7. Second, we train and report a "pooled" model. Instead of splitting the restaurant-week level data into training and test sets, in this approach, we take each restaurant and split the weeks into training and test, and stack them. To train our prediction model, we also include week fixed effects as predictors. The results appear in Web Appendix

Table D8. Finally, we also train a model using scaled visits i.e., the visits from app users scaled up to reflect the population of the city in which the restaurant is located. The results appear in Web Appendix Table D9. In each of the three setups, Model C with geo-tracking data outperforms the other models.

Alternative Machine Learning Models

While our main models use an Elastic Net Regression, we also train alternative ML models, such as Lasso, Ridge, and Boosted Regression Trees, to make sure the results are not unique to the model we use. Since our main interest is in comparing various information sets, we repeat the specifications reported in the main results in Table 4 and report their results for alternative models in Web Appendix Tables D10, D11, and D12. The findings are consistent with those reported in the main analysis.

Alternative Implementation of Simulation Exercises

In addition to our main simulation exercises, we implemented alternative versions using different parameters for the simulations. Specifically, we trained the geofence models using larger radii of two, five, and ten miles rather than the one-mile radius in our main analysis. Similarly, we extended the frequency simulation to one-tenth tracking and one-half random tracking (i.e., drop half of the driving instances at random, rather than at systematic intervals). We report the findings from these models in Web Appendix Table D13. We find support for the main result that the predictive performance of the models with these alternative thresholds is lower than that of models with complete geo-tracking.

Alternative Metrics

Our main results allow us to compare the performance of various information sets using RMSE as the performance metric. In addition, we also report the results in our main analysis for an alternative performance metric, i.e., the Mean Absolute Error (MAE). MAE tends

to be less sensitive to outliers than RMSE. The results for MAE appear in Web Appendix Table D14 and show patterns that are similar to the RMSE results.

Outlier Drivers

It is possible that some users in our data may be restaurant delivery drivers or commercial taxi drivers. Such drivers are likely to have higher than typical levels of driving distances. To make sure our prediction results are not driven by learning these outlier drivers' patterns, we drop any user with driving distances of more than the mean plus three times the standard deviation and train our prediction model after excluding them. The results appear in Web Appendix Table D15 and are similar as those we reported earlier.

CONCLUSION

In recent years, many firms have started collecting geo-tracking consumer data and using them to inform their marketing and operational decisions (Clifford 2018). However, geo-tracking evokes privacy concerns among app users and regulators. For example, companies like Tim Hortons have attracted public scrutiny due to their geo-tracking practices (Austen 2022). Many emerging privacy regulations, such as the California Privacy Rights Act, treat consumer locations as sensitive data and restrict geo-tracking.

In this research, we examined two questions: First, to what extent are geo-tracking data useful relative to not using consumer data and using only demographic, behavioral data, and static home location information for predicting visits to a business location? Second, how does restricting geo-tracking data under various privacy regulations impact the usefulness of these data for prediction?

We answered our research questions in the context of the restaurant industry using an application we identified through in-depth interviews with managers, i.e., predicting the total visits to a restaurant one week ahead (see Web Appendix A for managerial interviews). Specifically, we used proprietary data from a safe-driving app in Texas with 120 million

driving instances for 38,980 individual users to make predictions for 5,951 restaurant-weeks in our sample using a machine learning (ML) approach. From our interviews, we learned that predictions at a weekly frequency can allow managers to make better decisions about their marketing and operations. While the potential usefulness of geo-tracking data for businesses depends on the application under study, we focused on weekly prediction of visits. Importantly, we compared the performance of models with complete geo-tracking against those of models with restricted geo-tracking using simulation exercises that are motivated by privacy regulations, industry practices, recommendations from the data obfuscation literature, and a consumer survey we conducted (see Web Appendix B).

Our research has several key findings. First, we find that using geo-tracking data increases predictive performance by 14.77% when compared with using demographic, behavioral, and static home location information, and by 22.27% when considering only baseline models that do not include any consumer data. Second, models with geo-tracking data perform better both by reducing the extent of over- and under- prediction of visits. Third, imposing privacy restrictions that limit *what* geo-tracking data are tracked, in *what form*, *where*, and *how frequently* reduces the usefulness of geo-tracking for prediction by 1%-16% relative to complete geo-tracking. The extent of the decrease in predictive performance depends on the type of regulation. Specifically, regulations that restrict *what* data are geo-tracked (i.e., summaries of driving behaviors) and in *what form* (i.e., synthetic data generated with nearby users' data) are associated with the largest decreases in predictive performance (16.24% and 8.09% respectively). We also identify restrictions that have a smaller relative impact on the predictive performance, such as limiting *how frequently* users are tracked (.77-2.46%, depending on the frequency) or *where* they are tracked (3.56%). Finally and importantly, models that use restricted geo-tracking still generally outperform models that do not use geo-tracking information.

Managerial Implications

Our results have several implications for firms and policymakers. First, our finding that geo-tracking data allow firms to better predict the total visits from users can help managers plan their marketing and operations. Our interviews with restaurant managers revealed that they plan each week's resource commitments in advance (e.g., decisions to hire part-time additional staff). For example, the manager of Culvers' and Red Lobster said that: "Knowing how many customers to expect the following week can reduce cost, increase profits, and help manage [staff and inventory] without waste." By improving the prediction of total visits, we show that geo-tracking data can help managers better plan ahead.

Second, we find that using geo-tracking data for prediction reduces both over- and under-prediction of visits by app users. The better performance of our model in both cases allows firms to evaluate, based on the cost of over- and under-prediction, their decision to collect and use geo-tracking data for prediction applications.

Third, our additional analyses show that geo-tracking data can be useful for within-restaurant predictions over weeks and for predicting deviations from typically expected visits. These findings imply that restaurants can better plan for fluctuations in expected visits if they have access to geo-tracking data about their consumers. Restaurants that experience more fluctuations in their traffic could consider leveraging such data for their planning and decisions.

Fourth, in our simulation exercises, we propose practical ways in which firms can protect and restrict consumers' geo-tracking data, and show that limiting *what* data are tracked and *in what form* results in the highest predictive losses for firms in our application. However, other restrictions, such as *where* data are tracked and *how frequently*, still allow firms to derive relatively higher predictive value from geo-tracking, which implies that firms are better off restricting geo-tracking in these ways when possible. Our findings can also be useful when firms are purchasing geo-tracking data from third party vendors and need to make decisions about what data and how much data to access. Reduced frequency of tracking, for example, can allow firms to save on data storage and server costs relative to complete tracking. More

importantly, our findings imply that there are ways in which firms can restrict geo-tracking data in privacy-safe ways (while still getting predictive value from them), which can allow them to potentially mitigate the risk of litigation and reputational damage, given recent regulatory action against the mishandling of consumer geolocation data streams (FTC 2024).

Finally, our simulation exercises are meaningful for regulators who often evaluate various types of restrictions to impose because they can better quantify the relative predictive losses of each in the context of our application.

Limitations and Future Research

Our research has limitations that future research can address. First, our data come from the users of a single app. Even though we demonstrate our findings for scaled outcomes at the city-population level, we do not observe the overall population of visitors to a restaurant and can only analyze the total visits from app users. If managers were interested in predicting aggregate demand to their location from all customers and not just app users, they should interpret our outcome measure with some degree of finite-sample measurement error. While both the app users and the restaurants in our sample are representative of the general population, we are limited to our sample of restaurants and users, so there could be value in expanding our analysis beyond our sample. Second, many restaurants use geo-tracking data to make real-time predictions. Our ML framework predicts visits one week ahead using past weeks' data. Future research may address real-time predictions under privacy restrictions for other relevant applications, e.g., for targeting mobile coupons. Third, we acknowledge that the potential usefulness of geo-tracking data for businesses depends on the application under study. Our paper can speak to one application, which uses geo-tracking information to predict visits to a restaurant in a week. If data are available, future research can extend the usefulness of geo-tracking in other contexts, for other outcomes (e.g., revenues and profits), and for other applications of interest. Given the heightened concerns with using personal consumer data, firms may also prefer to use such data in

situations where it advances consumer interest, which is another potential avenue for future research (Soleymanian, Weinberg, and Zhu 2019). Fourth, while our simulated regulations are generally perceived as privacy-preserving by consumers we surveyed, future research can better elicit consumers' willingness to share data under different geo-tracking data-sharing scenarios since consumers' stated and actual preferences for privacy may vary (Adjerid, Peer, and Acquisti 2018). Finally, while we examine simulated regulations drawn from the current policy landscape and privacy literature, there may be other regulations of interest we cannot study. For example, most app users in our setting are adults, so we cannot comment on privacy protection for children (e.g., Johnson et al. 2023). Similarly, app users in our setting have opted-in to be tracked, so we cannot comment on consenting vs. non-consenting users (e.g., Lin and Strulov-Shlain 2023). Future research, if data are available, can specifically examine these additional important regulations.

REFERENCES

Acquisti, Alessandro and Hal R Varian (2005), "Conditioning Prices on Purchase History," *Marketing Science*, 24 (3), 367–381.

Adjerid, Idris, Eyal Peer, and Alessandro Acquisti (2018), "Beyond the Privacy Paradox," *MIS quarterly*, 42 (2), 465–488.

Austen, Ian (2022), "'A Mass Invasion of Privacy' but No Penalties for Tim Hortons," *New York Times* http://tinyurl.com/33x6ujz5.

Avenancio-León, Carlos F and Troup Howard (2022), "The Assessment Gap: Racial Inequalities in Property Taxation," *The Quarterly Journal of Economics*, 137 (3), 1383–1434.

Berman, Ron and Ayelet Israeli (2022), "The Value of Descriptive Analytics: Evidence from Online Retailers," *Marketing Science*, 41 (6), 1074–1096.

Bertrand, Marianne, Emir Kamenica, and Jessica Pan (2015), "Gender Identity and Relative Income within Households," *The Quarterly Journal of Economics*, 130 (2), 571–614.

Binns, Reuben, Ulrik Lyngs, Max Van Kleek, Jun Zhao, Timothy Libert, and Nigel Shadbolt (2018), "Third Party Tracking in the Mobile Ecosystem," *Proceedings of the 10th ACM Conference on Web Science*, pages 23–31.

Bleier, Alexander, Avi Goldfarb, and Catherine Tucker (2020), "Consumer Privacy and the Future of Data-based Innovation and Marketing," *International Journal of Research in Marketing*, 37 (3), 466–480.

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz (2016), "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment," *American Economic Review*, 106 (4), 855–902.

Choi, W Jason, Kinshuk Jerath, and Miklos Sarvary (2023), "Consumer Privacy Choices and (Un) Targeted Advertising along the Purchase Journey," *Journal of Marketing Research*, 60 (5), 889–907.

Clifford, Catherine (2018), "You can now Head to McDonald's to get a Burger King Whopper for 1 cent — Here's how," *CNBC* http://tinyurl.com/4zv3hbep.

Cunha, Mariana, Ricardo Mendes, and João P Vilela (2021), "A Survey of Privacy-preserving Mechanisms for Heterogeneous Data Types," *Computer Science Review*, 41, 100403.

Danaher, Peter J (2023), "Optimal Microtargeting of Advertising," *Journal of Marketing Research*, 60 (3), 564–584.

Danaher, Peter J, Michael S Smith, Kulan Ranasinghe, and Tracey S Danaher (2015), "Where, When, and How Long: Factors that Influence the Redemption of Mobile Phone Coupons," *Journal of Marketing Research*, 52 (5), 710–725.

Dean, Grace (2023), "McDonald's and Chick-fil-A are Tracking Your Location to Make Sure Your Fries are Perfectly Crispy When you Come Collect Your Mobile Order," *Business Insider* http://tinyurl.com/4zkr622h.

Dubé, Jean-Pierre, Zheng Fang, Nathan Fong, and Xueming Luo (2017), "Competitive Price Targeting with Smartphone Coupons," *Marketing Science*, 36 (6), 944–975.

FTC (2022), "FTC Sues Kochava for Selling Data that Tracks People at Reproductive Health Clinics, Places of Worship, and Other Sensitive Locations," *Federal Trade Commission* http://tinyurl.com/pwbedwke.

FTC (2024), "FTC Order Will Ban InMarket from Selling Precise Consumer Location Data," *Federal Trade Commission* https://tinyurl.com/3mh4nzw7.

Goldberg, Samuel G, Garrett A Johnson, and Scott K Shriver (2024), "Regulating Privacy Online: An Economic Evaluation of the GDPR," *American Economic Journal: Economic Policy*, 16 (1), 325–358.

Goldfarb, Avi and Catherine Tucker (2012), "Shifts in Privacy Concerns," *The American Economic Review*, 102 (3), 349–353.

Gonzalez, Marta C, Cesar A Hidalgo, and Albert-Laszlo Barabasi (2008), "Understanding Individual Human Mobility Patterns," *Nature*, 453 (7196), 779–782.

Hoteit, Sahar, Stefano Secci, Stanislav Sobolevsky, Carlo Ratti, and Guy Pujolle (2014), "Estimating Human Trajectories and Hotspots through Mobile Phone Data," *Computer Networks*, 64, 296–307.

Jerath, Kinshuk and Klaus M Miller (2024), "Using the Dual-Privacy Framework to Understand Consumers' Perceived Privacy Violations Under Different Firm Practices in Online Advertising," *Working paper*.

Jiang, Hongbo, Jie Li, Ping Zhao, Fanzi Zeng, Zhu Xiao, and Arun Iyengar (2021), "Location Privacy-preserving Mechanisms in Location-based Services: A Comprehensive Survey," *ACM Computing Surveys (CSUR)*, 54 (1), 1–36.

Johnson, Garrett, Tesary Lin, James C Cooper, and Liang Zhong (2023), "COPPAcalypse? The Youtube Settlement's Impact on Kids Content," *Working paper.*

Johnson, Garrett, Scott Shriver, and Samuel Goldberg (2023), "Privacy & Market Concentration: Intended & Unintended Consequences of the GDPR," *Management Science*, 69 (10), 5695–6415.

Kim, Mingyung, Eric T Bradlow, and Raghuram Iyengar (2022), "Selecting Data Granularity and Model Specification using the Scaled Power Likelihood with Multiple Weights," *Marketing Science*, 41 (4), 848–866.

Klopack, Ben (2024), "One Size Fits All? The Value of Standardized Retail Chains," *RAND Journal of Economics*, 55 (1), 55–86.

Klopack, Ben, Eric Lewis, and Fernando Luco (2024), "Economic Consequences of Natural Disasters: The Impact of Hurricane Harvey on Local Retail and Consumer Welfare," *Working Paper*.

Klosowski, Thorin (2021), "The State of Consumer Data Privacy Laws in the US (And Why it Matters)," http://tinyurl.com/ypjhssxj.

Korganbekova, Malika and Cole Zuber (2023), "Balancing User Privacy and Personalization," *Working Paper*.

Lamberton, Cait and Andrew T Stephen (2016), "A Thematic Exploration of Digital, Social Media, and Mobile Marketing: Research Evolution from 2000 to 2015 and an Agenda for Future Inquiry," *Journal of Marketing*, 80 (6), 146–172.

Landvoigt, Tim, Monika Piazzesi, and Martin Schneider (2015), "The Housing Market(s) of San Diego," *American Economic Review*, 105 (4), 1371–1407.

Lee, Jung Youn, Joonhyuk Yang, and Eric T. Anderson (2021), "Buying and Payment Habits: Using Grocery Data to Predict Credit Card Payments," *Working paper*.

Li, Shaobo, Matthew J. Schneider, Yan Yu, and Sachin Gupta (2023), "Reidentification Risk in Panel Data: Protecting for K-Anonymity," *Information Systems Research*, 34 (3), 811–1319.

Lin, Tesary and Sanjog Misra (2022), "Frontiers: The Identity Fragmentation Bias," *Marketing Science*, 41 (3), 433–440.

Lin, Tesary and Avner Strulov-Shlain (2023), "Choice Architecture, Privacy Valuations, and Selection Bias in Consumer Data," *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2023-58).

Macha, Meghanath, Natasha Zhang Foutz, Beibei Li, and Anindya Ghose (2023), "Personalized Privacy Preservation in Consumer Mobile Trajectories," *Information Systems Research*, forthcoming.

Miller, Klaus M. and Bernd Skiera (2023), "Economic Loss of Cookie Lifetime Restrictions," *International Journal of Research in Marketing*, Forthcoming.

Naik, Nikhil, Ramesh Raskar, and César A. Hidalgo (2016), "Cities Are Physical Too: Using Computer Vision to Measure the Quality and Impact of Urban Appearance," *American Economic Review*, 106 (5), 128–32.

Neumann, Nico, Catherine E. Tucker, and Timothy Whitfield (2021), "Frontiers: How Effective is Third-party Consumer Profiling? Evidence from Field Studies," *Marketing Science*, 38 (6), 918–926.

Oblander, Shin and Daniel McCarthy (2023), "Frontiers: Estimating the Long-Term Impact of Major Events on Consumption Patterns: Evidence from COVID-19," *Marketing Science*, 42 (5), 839–1028.

Pappalardo, Luca and Filippo Simini (2018), "Data-driven Generation of Spatio-temporal Routines in Human Mobility," *Data Mining and Knowledge Discovery*, 32 (3), 787–829.

Pappalardo, Luca, Filippo Simini, Gianni Barlacchi, and Roberto Pellungrini (2022), "scikit-mobility: A Python Library for the Analysis, Generation, and Risk Assessment of Mobility Data," *Journal of Statistical Software*, 103 (1), 1–38.

Peukert, Christian, Stefan Bechtold, Michail Batikas, and Tobias Kretschmer (2022), "Regulatory Spillovers and Data Governance: Evidence from the GDPR," *Marketing Science*, 41 (4), 746–768.

Rafieian, Omid and Hema Yoganarasimhan (2021), "Targeting and Privacy in Mobile Advertising," *Marketing Science*, 40 (2), 193–218.

Rossi, Peter E., Robert E. McCulloch, and Greg M. Allenby (1996), "The Value of Purchase History Data in Target Marketing," *Marketing Science*, 15 (4), 321–340.

Smith, H Jeff, Sandra J Milberg, and Sandra J Burke (1996), "Information Privacy: Measuring Individuals' Concerns about Organizational Practices," *MIS Quarterly*, pages 167–196.

Soleymanian, Miremad, Charles B Weinberg, and Ting Zhu (2019), "Sensor Data and Behavioral Tracking: Does Usage-based Auto Insurance Benefit Drivers?," *Marketing Science*, 38 (1), 21–43.

Song, Chaoming, Zehui Qu, Nicholas Blumm, and Albert-László Barabási (2010), "Limits of Predictability in Human Mobility," *Science*, 327 (5968), 1018–1021.

Sun, Chenshuo, Panagiotis Adamopoulos, Anindya Ghose, and Xueming Luo (2022), "Predicting Stages in Omnichannel Path to Purchase: A Deep Learning Model," *Information Systems Research*, 33 (2), 429–445.

Tau, Byron (2023), "Selling Your Cellphone Location Data Might Soon Be Banned in U.S. for First Time," *Wall Street Journal* http://tinyurl.com/bdsa8khx.

Terrovitis, Manolis, Giorgos Poulis, Nikos Mamoulis, and Spiros Skiadopoulos (2017), "Local Suppression and Splitting Techniques for Privacy Preserving Publication of Trajectories," *IEEE Transactions on Knowledge and Data Engineering*, 29 (7), 1466–1479.

Tian, Longxiu, Dana Turjeman, and Samuel Levy (2023), "Privacy Preserving Data Fusion," *Working paper*.

Trusov, Michael, Liye Ma, and Zainab Jamal (2016), "Crumbs of the Cookie: User Profiling in Customer-base Analysis and Behavioral Targeting," *Marketing Science*, 35 (3), 405–426.

Valentine-De, Jennifer, Natasha Singer, Michael H Keller, and Aaron Krolik (2018), "Your Apps Know Where You Were Last Night, and They're Not Keeping it Secret," *New York Times* http://tinyurl.com/2ucxjh9f.

Van der Maaten, Laurens and Geoffrey Hinton (2008), "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, 9 (11), 2579–2605.

Varga, Marton, Anita Tusche, Paulo Albuquerque, Nadine Gier, Bernd Weber, and Hilke Plassmann (2024), "Does fMRI Data Improve Predictions of Product Adoption by Store Managers and Sales to Consumers for Consumer-Packaged Goods?," *Working Paper*.

Wang, Pengyuan, Li Jiang, and Jian Yang (2023), "EXPRESS: The Early Impact of GDPR Compliance on Display Advertising: The Case of an Ad Publisher," *Journal of Marketing Research*, forthcoming.

Wang, Yanwen, Chunhua Wu, and Ting Zhu (2019), "Mobile Hailing Technology and Taxi Driving Behaviors," *Marketing Science*, 38 (5), 734–755.

Weis, Ashley (2023), "What Franchisors Should Know About Data Privacy Compliance in 2023," *Franchise Wire* https://tinyurl.com/mwrvd5c8.

Wernerfelt, Nils, Anna Tuchman, Bradley Shapiro, and Robert Moakler (2022), "Estimating the Value of Offsite Data to Advertisers on Meta," *Working Paper*.

Yan, Shunyao, Klaus M Miller, and Bernd Skiera (2022), "How Does the Adoption of Ad Blockers Affect News Consumption?," *Journal of Marketing Research*, 59 (5), 1002–1018.

Yoganarasimhan, Hema (2020), "Search Personalization using Machine Learning," *Management Science*, 66 (3), 1045–1070.

Zhao, Yu, Pinar Yildirim, and Pradeep K. Chintagunta (2021), "Privacy Regulations and Online Search Friction: Evidence from GDPR," *Working paper*.

WEB APPENDIX A

INTERVIEWS OF RESTAURANT MANAGERS AND OWNERS

Table A1: Managerial Responses on How Predictions of Consumer Visits May be Useful, the Consumer Information they Collect/Wish to Collect, and at what Intervals

Manager Role	**Experience**	**Gist of Interview Responses**
Manager Jimmy John's (JJ)	5 years	Predicting consumer visits every week would help us with staffing and figuring out if we're overstaffed or understaffed. We can't fill staffing gaps immediately, so we receive staff applications on a weekly basis and look at their availability for the week. Customer demographics are important and will also tell us where our customers reside. Our app asks for an address for delivery, so we know how close customers live to JJ's store. Knowing consumers' location and visit patterns to other restaurants could help us target ads. I wouldn't prefer to keep data for more than a few weeks or six months at most to not alarm customers.
Manager Asian restaurant in Chicago	2-3 years	At our small Chinese restaurant in Chicago, we collect demand data and customer visits. A prediction algorithm that can tell us how likely a customer is to visit us each week would be helpful. This would help us to know who is coming and how often, which would allow us to plan and manage our resources effectively. Knowing what kind of food our customers like based on their past visits and/or to other restaurants also helps us make better decisions.
Manager and Chef Culvers' and Red Lobster	3+ years	It [predicting customer visits] would reduce costs greatly. If you can predict how many people are likely to visit and maybe even which meals they will need, you can plan which ingredients to order, or how much staff to hire. In my experience, you'd have half of the staff fulltime and half of the staff part-time. Schedules get set typically on Fri/Sat for the following week. Knowing how many customers to expect the following week can reduce cost, increase profits, and help manage all of those without waste. It would also be useful to see how many are local vs. out-of-town customers. Are they mostly going to be coming in for mornings or afternoons, then you could have specials around those meals and times instead of all day deals. Or even a special deal of the week.
Food Administrator University Dining Services	27+ years	If a prediction algorithm used data specific to my restaurant to make predictions about my consumers' visits, I would use it to make decisions. It would be important to ensure that the algorithm considers factors, such as the demographics and needs of our customers. However, I would not rely solely on the algorithm as it is important to stay in touch with what customers want. In terms of the frequency at which I would like to predict consumer visits, I would say weekly would be the most useful. This would help me to tailor our offerings and marketing strategies to better meet the needs of our customers.

Manager Role	Experience	Gist of Interview Responses
Owner, Three fine-dining restaurants in Atlanta	20+ years	I'd love to see the data on who comes back to our restaurant for repeat service. If I knew how many customers (and who) is coming, I can prepare my restaurant and work towards satisfying them, e.g., if I have mostly female or non-binary customers, or where they come from, maybe their relationship status. More detailed information about our customers is incredibly useful when it comes to satisfying them and growing our business. However, we also need to be careful not to overload ourselves by trying to predict customer visits every single day or by collecting too much data about them. It might be better to focus on weekly predictions, so that we can plan and make sure we are ready to provide the best possible experience for our customers. It can also help us plan ingredients, staffing, and store locations i.e., where to locate based on what kind of demographics and competitors are there and what information we have on them.
Manager Chilli's	18 months	The data I'd really be interested in is repeat customers. POS systems don't really track that and not everyone has rewards programs (or joins one). It would also be useful to know how many people are likely to come in and overlap that with community events they plan to ahead. Demographics about customers can also help look at trends about who they are, whether they are visiting if they are more proximate, how far they'd travel for a restaurant etc. Monthly predictions are fine, but for some decisions, weekly or more regular data could also help especially for scheduling. For example, if you know Thursday is going to be busy for family customers, hopefully you can increase your customer accounts too, you can move people in and out faster. If they can come and go quickly, it will improve the customer experience.
Hostess Restaurant in Chicago	2 years	A prediction algorithm that can tell me how likely a customer is to visit my restaurant would be very helpful. I could use this information to adjust staffing schedules, move tables around, and plan weekly specials. I could also use it to determine when to promote happy hour specials and other promotions on social media. I would like to be able to predict consumer visits on a weekly basis. This is because most decisions, including staffing, scheduling, and social media marketing as well as ordering supplies are planned on a weekly basis.
Manager Crosby's Kitchen in Chicago	15+ years	The most important decisions I make are hiring and customer service. Predicting consumer visits can help because the problem is that sometimes it's slow and sometimes it's busy. I don't like to over or understaff. So I'd like to know the predictions a week in advance. Sometimes also a month in advance because other people I work with have lives outside the restaurant, but usually a week is fine.

Notes: The interviews were conducted after securing Institutional Review Boards (IRB) approval. Managers were recruited through a research database at a large public university. The only inclusion criterion was some experience in the restaurant industry in the U.S. Each interview lasted about 30 minutes via zoom. We include restaurant name and location only if the managers agreed to share it.

WEB APPENDIX B

CONSUMER SURVEYS

In this section, we report the results of a survey we conducted to ask consumers about their privacy perceptions towards geo-tracking. While our privacy restrictions under the simulation exercises are primarily motivated by the current regulatory environment, industry practices, and recommendations from the data obfuscation literature, we further wanted to validate that the simulation exercises are privacy preserving from the perspective of end consumers who are subject to such tracking.

We recruited survey participants through a research database at a large public university in the U.S. after securing IRB approvals. The survey was administered using a Qualtrics link. Participation in the survey was voluntary. Participants were offered a Target gift card worth five dollars for completing the survey. Upon accessing Qualtrics, the participants were presented the following information:

*"Imagine you are using an app on your phone that gives you points that you can redeem at local restaurants and businesses. Read the scenarios below about the kind of data the app tracks from your usage, then indicate the extent to which you agree with each statement that follows: <**insert scenario**>*

We presented the scenarios listed in Table B1 in a random order. After each scenario was presented, we asked survey participants to indicate their privacy perceptions for that scenario across five dimensions adapted from the privacy literature (e.g., Smith, Milberg, and Burke 1996).

Table B1: Scenarios

Text of Scenarios in the Survey
Complete tracking: The app collects data about your geo-location constantly after it has detected that driving has started.
Summary features: The app collects data about your driving summary (e.g., total distance you drive, time of day when you drive) rather than geo-location coordinates.
Synthetic data: The app collects data about your geo-location coordinates, but adds noise to your data (e.g., using similar users' data) so that your exact address is not identified.
Geofence: The app collects data if you were near a local restaurant.
Frequency: The app collects data about your geo-location coordinates with a frequency of every half hour (i.e., 30 minutes) after it has detected that driving has started.
Static first- and last- points: The app collects data about your geo-location coordinates at the start and end of your trip.
Static random trip per week: The app collects data about your geo-location coordinates for one trip at random each week.

The privacy dimensions were:

1. It bothers me that the app collects these data.
2. I'm concerned that the app is collecting too much information about me.

3. I am concerned about my privacy and how the app might use my data.
4. It bothers me to give my information to this app.
5. I will stop using this app in the future.

Participants were asked to rank each of the five statements for each scenario they were presented on a scale of 1 (strongly disagree) to 5 (strongly agree).

Results of the Survey

Overall, 191 participants completed the survey. Of the participants, 60% were female. The average age was 44 years. Across the five privacy measures, participants' privacy concerns under complete geo-tracking had an average score of 4.18 out of 5. Relative to complete geo-tracking, user-level summarization had an average score of 3.17 ($p < .001$), synthetic data generation had an average score of 3.51 ($p < .001$), and geographic restrictions had an average score of 3.22 ($p < .001$). Among frequency restrictions, the half-hour interval of tracking had an average score of 3.73 ($p < .001$), the static tracking of first- and last-trip points only had an average score of 3.77 ($p < .001$), and one trip per week at random had an average score of 3.46 ($p < .001$). We report the mean rating for each question and averages across the five questions in the survey for each scenario in Table B2.

Table B2: Results: Mean Rating for Each Survey Question and Overall

Simulation	Q1	Q2	Q3	Q4	Q5	Mean
Complete geo-tracking	4.20	4.27	4.32	4.24	3.88	4.18
User-level summarization						
Summary features instead of geo-tracking	3.11	3.18	3.43	3.19	2.92	3.17
Synthetic data						
Add noise to the data	3.53	3.58	3.68	3.54	3.24	3.51
Geographical restrictions						
Geofenced users when near restaurant	3.21	3.20	3.45	3.25	2.99	3.22
Frequency restrictions						
Reduced frequency of geo-tracking						
Reduced frequency	3.71	3.78	3.93	3.76	3.49	3.73
Static geo-tracking						
First- and last- trip points only	3.78	3.87	4.04	3.81	3.34	3.77
One trip per week at random only	3.46	3.43	3.74	3.49	3.19	3.46

Notes: N = 191. Survey statements Q1-Q5 were as follows: Q1. It bothers me that the app collects these data. Q2. I'm concerned that the app is collecting too much information about me. Q3. I am concerned about my privacy and how the app might use my data. Q4. It bothers me to give my information to this app. Q5. I will stop using this app in the future. Participants were asked to rank each statement from 1 (strongly disagree) to 5 (strongly agree). Reduced frequency refers to half hour.

WEB APPENDIX C

CATEGORIZATION OF RESTAURANTS

Table C1: Categories of Restaurants using Yelp API

	Categories	Description	Popular tags
1	American	Restaurants serving American cuisine, but excluding restaurants specializing in burgers and sandwiches, and excluding restaurants that were also tagged as another type.	American (Traditional), American (new), breakfast, brunch, chicken wings, diners
2	Asian	Restaurants specializing in cuisines from south Asian, east Asian, and southeast Asian countries, as well as pacific islands.	Chinese, Japanese, sushi bars, Asian fusion, Thai, Indian, Hawaiian
3	Burgers	Restaurants with tag "Burgers".	Burgers, hot dogs, sports bars, Steakhouses
4	Coffee	Restaurants with tag "Coffee".	Tea, coffee, café
5	Dessert	Restaurants with tag "Dessert".	Dessert, frozen yogurt
6	European	Restaurants specializing in Italian, French, or other European cuisines, except for restaurants also tagged "Pizza".	Italian, French, Irish, Wine Bars, Noodles, Mediterranean
7	Latin American	Restaurants specializing in cuisines from south and central America and the Caribbean.	Mexicana, Tex-Mex, Latín American, seafood, Caribbean, Cuban
8	Pizza	Restaurants with tag "Pizza".	Pizza, Italian, salad
9	Sandwiches	Restaurants with tag "Sandwiches", "Deli", or "Cheesesteaks".	Sandwiches, deli, cheesesteaks
10	Other	Restaurants not tagged as any of the above categories.	

WEB APPENDIX D

ROBUSTNESS CHECKS AND ADDITIONAL ANALYSES

In this section, we present the robustness checks for alternative sample of restaurants (Tables D1-D4), alternative aggregation approaches (Table D5 and Table D6), alternative model setups (Tables D7- D9), alternative ML models (Tables D10- D12), alternative ways of implementing the simulation exercises (Table D13), alternative metrics (Table D14), and outlier drivers (Table D15).

Table D1: Results: Predictive Performance of Elastic Net Regression by Information Set for Restaurants with At Least 15 Users Visiting

Model	Mean RMSE	Difference from Baseline	Difference from Model A	Difference from Model B
Baseline	6.191 [6.130, 6.252]			
Model A: Home-zip distances	5.671 [5.618, 5.724]	.520 [.506, .534]		
Model B: Home-tracked distances	5.673 [5.620, 5.726]	.518 [.504, .532]	-.0028 [-.0032, -.0024]	
Model C: Home-tracked, trip distances, and other driving	4.828 [4.776, 4.880]	1.363 [1.325, 1.401]	.843 [.811, .874]	.845 [.814, .876]

Notes: RMSE = Root mean squared error. In column "Mean RMSE," the confidence interval corresponds to that of the mean RMSE for that model. In all other columns, the confidence interval corresponds to that of the mean difference between two models. We implement this using 2,000 bootstrap replications for each model, as described in the main text. N = 4,756 for 299 restaurants across 30 prediction weeks. The restaurants are selected based on at least 15 app users visiting anytime in the data period. Note that all feature sets include the baseline feature set of seasons, city, and category.

Table D2: Results: Predictive Performance of Elastic Net Regression by Information Set for Restaurants with At Least 25 Users Visiting

Model	Mean RMSE	Difference from Baseline	Difference from Model A	Difference from Model B
Baseline	7.164 [7.087, 7.241]			
Model A: Home-zip distances	6.602 [6.536, 6.668]	.562 [.544, .580]		
Model B: Home-tracked distances	6.603 [6.537, 6.669]	.561 [.543, .579]	-.0007 [-.0017, .0003]	
Model C: Home-tracked, trip distances, and other driving	5.683 [5.619, 5.747]	1.481 [1.431, 1.531]	.919 [.879, .960]	.920 [.880, .961]

Notes: RMSE = Root mean squared error. In column "mean RMSE," the confidence interval corresponds to that of the mean RMSE for that model. In all other columns, the confidence interval corresponds to that of the mean difference between two models. We implement this using 2,000 bootstrap replications for each model, as described in the main text. N = 2,855 for 142 restaurants across 30 prediction weeks. The restaurants are selected based on at least 25 app users visiting anytime in the data period. Note that all feature sets include the baseline feature set of seasons, city, and category.

Table D3: Results: Predictive Performance of Elastic Net Regression by Information Set for Top 25 Restaurants by Visits

Model	Mean RMSE	Difference from Baseline	Difference from Model A	Difference from Model B
Baseline	15.235 [15.076, 15.394]			
Model A: Home-zip distances	14.974 [14.826, 15.122]	.261 [.238, .286]		
Model B: Home-tracked distances	14.961 [14.813, 15.109]	.274 [.251, .299]	.013 [.006, .020]	
Model C: Home-tracked, trip distances, and other driving	13.580 [13.431, 13.721]	1.655 [1.569, 1.743]	1.394 [1.315, 1.473]	1.381 [1.303, 1.459]

Notes: RMSE = Root mean squared error. In column "mean RMSE," the confidence interval corresponds to that of the mean RMSE for that model. In all other columns, the confidence interval corresponds to that of the mean difference between two models. We implement this using 2,000 bootstrap replications for each model, as described in the main text. N = 567 restaurant-weeks for a sample of 25 restaurants with the highest number visits in the prediction period. Note that all feature sets include the baseline feature set of seasons, city, and category.

Table D4: Results: Predictive Performance of Elastic Net Regression by Information Set for Top 50 Restaurants by Visits

Model	Mean RMSE	Difference from Baseline	Difference from Model A	Difference from Model B
Baseline	11.258 [11.139, 11.377]			
Model A: Home-zip distances	10.859 [10.756, 10.962]	.399 [.373, .425]		
Model B: Home-tracked distances	10.868 [10.764, 10.972]	.390 [.364, .416]	-.009 [-.012, -.006]	
Model C: Home-tracked, trip distances, and other driving	9.529 [9.426, 9.632]	1.729 [1.657, 1.801]	1.330 [1.272, 1.388]	1.339 [1.281, 1.397]

Notes: RMSE = Root mean squared error. In column "mean RMSE," the confidence interval corresponds to that of the mean RMSE for that model. In all other columns, the confidence interval corresponds to that of the mean difference between two models. We implement this using 2,000 bootstrap replications for each model, as described in the main text. N = 1,152 restaurant-weeks for a sample of 50 restaurants with the highest number visits in the prediction period. Note that all feature sets include the baseline feature set of seasons, city, and category.

Table D5: Results: Predictive Performance of Elastic Net Regression by Information Set for Alternative Aggregation Threshold of 17 Miles

Model	Mean RMSE	Difference from Baseline	Difference from Model A	Difference from Model B
Baseline	5.643 [5.589, 5.698]			
Model A: Home-zip distances	5.147 [5.100, 5.194]	.496 [.485, .509]		
Model B: Home-tracked distances	5.146 [5.099, 5.193]	.497 [.486, .510]	.0011 [.0008, .0014]	
Model C: Home-tracked, trip distances, and other driving	4.383 [4.337, 4.429]	1.260 [1.227, 1.295]	.764 [.736, .791]	.763 [.735, .791]

Notes: RMSE = Root mean squared error. In column "Mean RMSE," the confidence interval corresponds to that of the mean RMSE for that model. In all other columns, the confidence interval corresponds to that of the mean difference between two models. We implement this using 2,000 bootstrap replications for each model, as described in the main text. N = 5,951 restaurant-weeks. Note that all feature sets include the baseline feature set of seasons, city, and category. Compared to the main model that aggregates over users within 30 miles (i.e., the maximum distance consumers travel to visit a restaurant in our data) of a restaurant, we use the alternative threshold of 17 miles (i.e., the mean distance consumers travel to visit a restaurant in our data) for this analysis.

Table D6: Results: Predictive Performance of Elastic Net Regression by Information Set for Aggregation over Users based on Home-zip Distance

Model	Mean RMSE	Difference from Baseline	Difference from Model A	Difference from Model B
Baseline	5.643 [5.589, 5.698]			
Model A: Home-zip distances	5.145 [5.098, 5.192]	.498 [.487, .511]		
Model B: Home-tracked distances	5.144 [5.097, 5.191]	.499 [.488, .512]	.0007 [.0003, .0013]	
Model C: Home-tracked, trip distances, and other driving	4.491 [4.446, 4.536]	1.152 [1.125, 1.181]	.654 [.632, .676]	.653 [.631, .675]

Notes: RMSE = Root mean squared error. In column "Mean RMSE," the confidence interval corresponds to that of the mean RMSE for that model. In all other columns, the confidence interval corresponds to that of the mean difference between two models. We implement this using 2,000 bootstrap replications for each model, as described in the main text. N = 5,951 restaurant-weeks. Note that all feature sets include the baseline feature set of seasons, city, and category. Compared to the main model that aggregates over users within 30 miles (i.e., the maximum distance consumers travel to visit a restaurant in our data) of a restaurant, we aggregate over users whose home-zip code distance from the restaurant is within 30 miles for this analysis.

Table D7: Results: Predictive Performance of Elastic Net Regression by Information Set for the Deviation Model

Model	Mean RMSE	Difference from Baseline	Difference from Model A	Difference from Model B
Baseline	2.381 [2.351, 2.411]			
Model A: Home-zip distances	2.357 [2.331, 2.383]	.024 [.019, .029]		
Model B: Home-tracked distances	2.357 [2.331, 2.383]	.024 [.019, .029]	.000 [.000, .000]	
Model C: Home-tracked, trip distances, and other driving	2.271 [2.247, 2.295]	.110 [.100, .120]	.086 [.078, .094]	.086 [.078, .094]

Notes: RMSE = Root mean squared error. In column "Mean RMSE," the confidence interval corresponds to that of the mean RMSE for that model. In all other columns, the confidence interval corresponds to that of the mean difference between two models. We implement this using 2,000 bootstrap replications for each model, as described in the main text. N = 5,921 restaurant-weeks excluding restaurant-weeks with top .05% residuals (i.e., very high residuals) based on seasonality. Models A and B have the same mean RMSE upto three decimal places.

Table D8: Results: Predictive Performance of Elastic Net Regression by Information Set for a "Pooled" Model within Restaurants

Model	Mean RMSE	Difference from Baseline	Difference from Model A	Difference from Model B
Baseline	5.664 [5.624, 5.704]			
Model A: Home-zip distances	5.151 [5.117, 5.185]	.513 [.503, .523]		
Model B: Home-tracked distances	5.151 [5.117, 5.185]	.513 [.503, .523]	.000 [.000, .000]	
Model C: Home-tracked, trip distances, and other driving	4.422 [4.389, 4.455]	1.242 [1.216, 1.268]	.729 [.708, .750]	.729 [.708, .750]

Notes: RMSE = Root mean squared error. In column "Mean RMSE," the confidence interval corresponds to that of the mean RMSE for that model. In all other columns, the confidence interval corresponds to that of the mean difference between two models. We implement this using 2,000 bootstrap replications for each model, as described in the main text. N = 5,951 restaurant-weeks. Model A and B have the same results upto three decimal places. Note that all feature sets include controls for seasons, city, and category as well as a week identifier. Compared to the main analysis with restaurant-week level splits, the data in this analysis is split within restaurant by week i.e., training data contain subset of weeks for a restaurant and the test data contain the remaining weeks for the same restaurant. However, we use the week prior to the prediction week to construct the input features, consistent with the main model.

Table D9: Results: Predictive Performance of Elastic Net Regression by Information Set for Visits Scaled by the Proportion of App Users in a City

Model	Mean RMSE	Difference from Baseline	Difference from Model A	Difference from Model B
Baseline	1,217.77 [1,208.10, 1,227.45]			
Model A: Home-zip distances	1,109.20 [1,099.74, 1,118.66]	108.57 [106.65, 110.49]		
Model B: Home-tracked distances	1,105.88 [1,096.53, 1,115.22]	111.89 [109.97, 113.82]	3.32 [2.90, 3.75]	
Model C: Home-tracked, trip distances, and other driving	1,069.69 [1,061.11, 1,078.27]	148.08 [145.48, 150.68]	39.51 [37.76, 41.26]	36.19 [34.54, 37.83]

Notes: RMSE = Root mean squared error. In column "Mean RMSE," the confidence interval corresponds to that of the mean RMSE for that model. In all other columns, the confidence interval corresponds to that of the mean difference between two models. We implement this using 2,000 bootstrap replications for each model, as described in the main text. N = 5,951 restaurant-weeks. Model A and B have the same results up to three decimal places. Note that all feature sets include the baseline feature set of seasons, city, and category. This model uses target visits in each restaurant-week scaled to the proportion of app users i.e., number of app users divided by the total population of the city in which the restaurant is located.

Table D10: Results: Predictive Performance of Ridge Regression by Information Set

Model	Mean RMSE	Difference from Baseline	Difference from Model A	Difference from Model B
Baseline	5.643 [5.588, 5.698]			
Model A: Home-zip distances	5.136 [5.088, 5.184]	.507 [.495, .519]		
Model B: Home-tracked distances	5.139 [5.091, 5.187]	.504 [.492, .516]	-.0024 [-.0026, -.0021]	
Model C: Home-tracked, trip distances, and other driving	4.392 [4.347, 4.437]	1.251 [1.221, 1.281]	.744 [.720, .768]	.747 [.723, .771]

Notes: RMSE = Root mean squared error. In column "Mean RMSE," the confidence interval corresponds to that of the mean RMSE for that model. In all other columns, the confidence interval corresponds to that of the mean difference between two models. We implement this using 2,000 bootstrap replications for each model, as described in the main text. N = 5,951 restaurant-weeks. Note that all feature sets include the baseline feature set of seasons, city, and category.

Table D11: Results: Predictive Performance of Lasso Regression by Information Set

Model	Mean RMSE	Difference from Baseline	Difference from Model A	Difference from Model B
Baseline	5.644 [5.589, 5.699]			
Model A: Home-zip distances	5.144 [5.097, 5.191]	.500 [.487, .513]		
Model B: Home-tracked distances	5.147 [5.100, 5.194]	.497 [.484, .510]	-.0021 [-.0025, -.0017]	
Model C: Home-tracked, trip distances, and other driving	4.384 [4.338, 4.430]	1.260 [1.221, 1.294]	.760 [.735, .791]	.763 [.735, .791]

Notes: RMSE = Root mean squared error. In column "Mean RMSE," the confidence interval corresponds to that of the mean RMSE for that model. In all other columns, the confidence interval corresponds to that of the mean difference between two models. We implement this using 2,000 bootstrap replications for each model, as described in the main text. N = 5,951 restaurant-weeks. Note that all feature sets include the baseline feature set of seasons, city, and category.

Table D12: Results: Predictive Performance of Boosted Regression Trees by Information Set

Model	Mean RMSE	Difference from Baseline	Difference from Model A	Difference from Model B
Baseline	5.577 [5.521, 5.633]			
Model A: Home-zip distances	5.014 [4.963, 5.065]	.563 [.554, .572]		
Model B: Home-tracked distances	5.081 [5.028, 5.134]	.496 [.488, .504]	-.067 [-.070, -.064]	
Model C: Home-tracked, trip distances, and other driving	4.371 [4.325, 4.417]	1.206 [1.190, 1.222]	.643 [.632, .654]	.710 [.699, .721]

Notes: RMSE = Root mean squared error. In column "Mean RMSE," the confidence interval corresponds to that of the mean RMSE for that model. In all other columns, the confidence interval corresponds to that of the mean difference between two models. We implement this using 2,000 bootstrap replications for each model, as described in the main text. N = 5,951 restaurant-weeks. Note that all feature sets include the baseline feature set of seasons, city, and category.

Table D13: Results: Predictive Performance of Elastic Net under Alternative Implementation of Simulation Exercises

Simulation	Mean RMSE	Difference
Complete geo-tracking (Model C in Table 4)	4.386	
	[4.341, 4.432]	
Alternative geographical restrictions		
Geofenced users within 2 miles of restaurant	4.444	.058
	[4.398, 4.491]	[.047, .069]
Geofenced users within 5 miles of restaurant	4.526	.139
	[4.479, 4.573]	[.129, .150]
Geofenced users within 10 miles of restaurant	4.594	.207
	[4.546, 4.642]	[.198, .217]
Alternative frequency restrictions		
1/2 frequency at random	4.493	.107
	[4.447, 4.540]	[.102, .112]
$1/10^{th}$ frequency	4.462	.075
	[4.417, 4.506]	[.057, .094]

Notes: RMSE = Root mean squared error. The complete geo-tracking model includes demographics, behavioral, and geo-tracking data (i.e., Model C of Table 4). The table reports the mean and bootstrapped confidence intervals (in square brackets) of the RMSE of each model and the difference in the RMSE between the complete geo-tracking model and each simulation using the test data, similar to the results reported in Table 5.

Table D14: Results: Predictive Performance of Elastic Net Regression by Information Set using Mean Absolute Error (MAE) Metric

Model	Mean MAE	Difference from Baseline	Difference from Model A	Difference from Model B
Baseline	1.818 [1.813, 1.823]			
Model A: Home-zip distances	1.672 [1.667, 1.677]	.146 [.142, .150]		
Model B: Home-tracked distances	1.672 [1.667, 1.677]	.146 [.142, .150]	.000 [.000, .000]	
Model C: Home-tracked, trip distances, and other driving	1.307 [1.303, 1.311]	.511 [.507, .515]	.365 [.361, .369]	.365 [.361, .369]

Notes: MAE = Mean absolute error. In column "Mean MAE," the confidence interval corresponds to that of the mean MAE for that model. In all other columns, the confidence interval corresponds to that of the mean difference between two models. We implement this using 2,000 bootstrap replications for each model, as described in the main text. N = 5,951 restaurant-weeks. Models A and B have the same mean MAE upto three decimal places. Note that all feature sets include the baseline feature set of seasons, city, and category.

Table D15: Results: Predictive Performance of Elastic Net Regression by Information Set after Excluding Outlier Drivers

Model	Mean RMSE	Difference from Baseline	Difference from Model A	Difference from Model B
Baseline	5.470 [5.417, 5.523]			
Model A: Home-zip distances	4.988 [4.942, 5.034]	.482 [.470, .491]		
Model B: Home-tracked distances	4.990 [4.945, 5.037]	.480 [.467, .494]	-.0022 [-.0025, -.0019]	
Model C: Home-tracked, trip distances, and other driving	4.289 [4.243, 4.335]	1.181 [1.148, 1.214]	.699 [.672, .726]	.701 [.675, .729]

Notes: RMSE = Root mean squared error. In column "Mean RMSE," the confidence interval corresponds to that of the mean RMSE for that model. In all other columns, the confidence interval corresponds to that of the mean difference between two models. We implement this using 2,000 bootstrap replications for each model, as described in the main text. N = 5,938 after dropping outlier drivers with more than mean plus three times standard deviation of driving distances. Note that all feature sets include the baseline feature set of seasons, city, and category.

WEB APPENDIX E

USER-LEVEL SUMMARIZATION OF GEO-TRACKING DATA: TECHNICAL DETAILS

We follow Pappalardo and Simini (2018) to extract mobility patterns for the users in our data from their geo-coordinates. Because coordinate-level data may be considered sensitive, and policymakers may restrict its use, in our simulations, we explore how summarizing these raw data at the user level without revealing exact locations might perform when used as inputs in our analyses. Because human mobility follows remarkably consistent patterns (Gonzalez, Hidalgo, and Barabasi 2008), using aggregated data rather than the geo-tracking data does not mean necessarily mean that our predictive performance will be hurt. To implement this approach, we rely on the DIary-based TRAjectory Simulator (DITRAS) framework (Pappalardo and Simini 2018). This framework separates the temporal characteristics of human mobility from its spatial characteristics. It turns mobility data into a diary generator represented as a Markov model, which we can use to generate features of interest.

Next, we describe the features that we compute at the user-week level to summarize geo-tracking data, following Pappalardo and Simini (2018)'s feature set. The first set of features relates to the randomness of consumers' driving trajectories. This is important because as the randomness of driving behavior increases, the likelihood of an algorithm being able to learn from past information to predict future visits decreases. Consumers whose driving patterns have a lower degree of randomness may be driving similar routes, e.g., they may have the same commute from home to work. As a result, they may be exposed to the same set of restaurants along their route. However, those with a higher degree of randomness, e.g., who may be driving out of town more frequently, may be exposed to a different set of routes and restaurants more. Thus, capturing the randomness in driving patterns can be informative of visitation decisions.

We use three measures of randomness: random, uncorrelated, and real entropy based on the mobility literature. Entropy is the informational value of past driving behavior when trying to predict future behavior (Pappalardo and Simini 2018). Random entropy measures the uncertainty of an individual's next location, assuming that this individual's movement is completely random among N possible locations (Wang, Wu, and Zhu 2019). Uncorrelated entropy captures the heterogeneity of locations visited by the user. Real entropy additionally accounts for the order in which different locations are visited by users and their time spent at each location, thus capturing the user's full spatiotemporal mobility (Song et al. 2010).

The second set of features computed from geo-tracking data relates to how much the app users drive. We take this into account by computing the radius of gyration, which is the characteristic distance traveled by the driver (Gonzalez, Hidalgo, and Barabasi 2008). The radius of gyration allows us to identify how far consumers typically drive, thus providing useful information related to each consumer. In addition to trajectory-based characteristics, we also capture the overall number of days on which a consumer drives, how many coordinates lie along their trip routes, and the maximum distance they traveled away from the focal location where they spend the most time during the training period.

Finally, we focus on the specific characteristics of each trip, for example, the number of quick stops (of $\leq$ 10 minutes) and long stops (of $\geq$ 60 minutes, Hoteit et al. 2014). We

also include a measure of the proportion of trips completed by a user in various windows of time to account for specific time-of-day effects, and the average number of trips per week (Pappalardo and Simini 2018). The summary statistics of these features at the user-week level appear in Table E1.

Next, we describe the technical details of computing these features. Human mobility tends to display a great degree of spatial and temporal regularity. Driving points that follow a spatial distribution of displacements over all the users can be well approximated by a truncated power-law with random walk pattern of step size Δ_r (Gonzalez, Hidalgo, and Barabasi 2008).

$$\Pr(\Delta_r) = (\Delta_r + \Delta_{r_0})^{-\beta} \exp(\frac{-\Delta_r}{k}) \tag{1}$$

where $\beta = 1.75 \pm .15$, $\Delta_{r_0} = 1.5$km, and cutoff distance of $k|_{D1} \approx 400$km and $k|_{D2} \approx 80$km.

Table E1: Summary Statistics of DITRAS Features of Geo-Tracking Data

Feature	Description	Mean
Random entropy	Variability of a user's visited locations if each location is visited with equal probability	5.916
Uncorrelated entropy	Variability of a user's visited locations based on probabilities of past visits	.99
Real entropy	Variability of a user's visited locations based on probabilities and order of past visits	5.802
Radius of gyration (miles)	Characteristic distance traveled by a user	17.69
Unique days	Average no. of unique days of driving	3.87
Locations	Average no. of unique points in a user's trip trajectory	84.21
Max distance (miles)	Maximum distance traveled by users from their home	51.44
Short stops	No. of stops of $\leq$10 minutes	15.37
Short stops at restaurants	No. of stops at restaurants for $\leq$10 minutes	2.26
Short stops at unique restaurants	No. of stops at unique restaurants for $\leq$10 minutes	1.77
Long stops	No. of stops for $\geq$60 minutes	8.32
Long stops at restaurants	No. of stops at restaurants for $\geq$60 minutes	.94
Long stops at unique restaurants	No. of stops at a unique restaurants for $\geq$60 minutes	.72
Morning driving	Proportion of trips in the morning (before 11am)	.33
Afternoon driving	Proportion of trips in the afternoon (11am to 5pm)	.33
Evening driving	Proportion of trips in the evening (after 5pm)	.34
Trip frequency	Average number of trips by a user	12.57

Notes: The summary geo-tracking features are computed for all users using their raw geo-coordinates each week. The reported numbers are aggregated over 5,951 restaurant-weeks for our sample of 422 restaurants. The number of stops at unique restaurants are computed as the count of unique restaurants a consumer stops at, e.g., if a consumer stops at twice at the same Pizza Hut location and once at a Starbucks location, the number of stops at unique restaurants will be two. DITRAS = DIary-based TRAjectory Simulator.

1. Radius of gyration:

   By this formulation, human motion follows a truncated Levy flight random walk with a probability distribution that is heavy-tailed. We can recover the radius of gyration, the characteristic distance travelled by user a when observed up to time t, as follows:

$$r_g^a(t) = \sqrt{\frac{1}{n_c^a} \sum_{i=1}^{n_c^a} \left(\overrightarrow{r_i^a} - \overrightarrow{r_{cm}^a} \right)^2} \tag{2}$$

   where $\overrightarrow{r_i^a}$ represents the $i = 1, 2, .., n_c^a(t)$ positions recorded by user a and $\overrightarrow{r_{cm}^a}$ is the center of mass of the trajectory.

2. Entropy:

   Entropy is a measure of variability in a users' mobility. We compute three types of entropy: random, uncorreleted and real entropy (Song et al. 2010).

   (a) Random entropy captures the degree of predictability of the user's whereabouts if each location is visited with equal probability.

$$S_i^{rand} = \log_2 N_i \tag{3}$$

   where N_i is the number of distinct locations visited by user i,

   (b) Uncorrelated entropy captures the degree of predictability of the user's whereabouts taking into account past visitation patterns.

$$S_i^{unc} = -\sum_{j=1}^{N_i} p\left(j\right) \log_2 p_i\left(j\right) \tag{4}$$

   where $p\left(j\right)$ is the historical probability that location j was visited by user i characterizing the heterogeneity of visit patterns.

   (c) Real entropy captures the degree of predictability of the user's whereabouts taking into account past visitation patterns as well as the order in which a user visits a location. It captures the full spatiotemporal order in a user's mobility pattern.

$$S_i^{real} = -\sum_{T_i^{'} \subset T_i} P\left(T_i^{'}\right) \log_2 \left[P\left(T_i^{'}\right) \right] \tag{5}$$

   where $P\left(T_i^{'}\right)$ is the probability of finding a particular time-ordered sub sequence and $T_i^{'}$in the trajectory T_i.

$$T_i = \{X_1, X_2, X_3, .., X_L\} \tag{6}$$

   which denotes the sequence of locations at which user i was observed at each time interval.

WEB APPENDIX F

PRIVACY-PRESERVATION UNDER SIMULATION EXERCISES

In this section, we provide examples of how each of our simulation exercises might protect individual user data. There are two main ways in which users can be identified from our data. First, geo-tracking data can be used to infer precise home locations (i.e., latitude and longitude), which can then be linked to other datasets uniquely (e.g., property records). Second, geo-tracking data can also generate unique individual records (i.e., trajectories of places visited), so even if true home locations are not observed, the uniqueness of records can still serve as individual quasi-identifiers (e.g., Li et al. 2023). Next we discuss these issues for each of our simulation exercises.

User-level summarization

Our first simulation exercise that limits *what* user data are tracked completely replaces geo-tracking data, including home-tracked locations, with summaries of driving behaviors. In this way, it prevents user home locations from being identified or used at all in the models. It also replaces unique user records with driving behaviors that tend to overlap between users, e.g., two users with different commutes may have similar driving distances each week. In this way, the user-level summarization of driving behaviors protects individual identities.

Synthetic data generation

Our second simulation exercise that limits *what form* user data are tracked in generates synthetic data i.e., replaces a user's geo-tracking data, including home-tracked locations, with those of their k-Nearest Neighbors (KNN).

In Figure F1, we illustrate the shift in home location using a sample user from our data. The Figure shows that KNN shifts the true home location (blue dot) to further away (green dot), making it difficult to identify the user based on their true home.

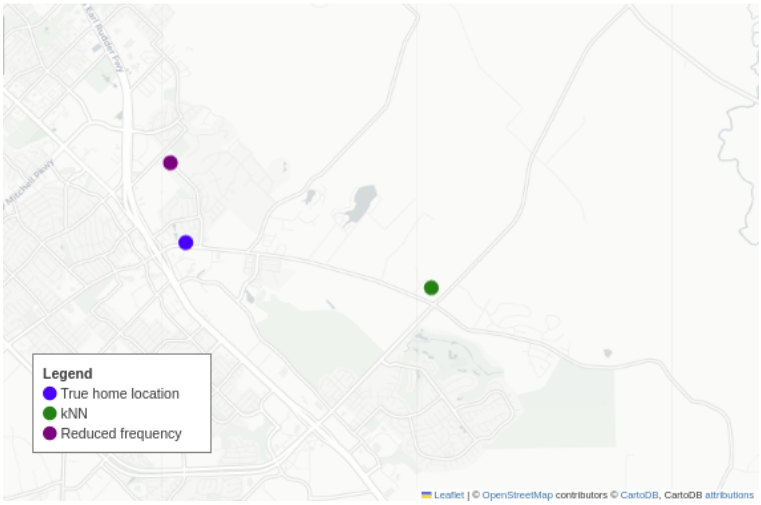


Figure F1: Example of User's Home Location: True vs. Simulated

By construction, synthetic data generated from other users' data can also duplicate records when users share the same set of neighbors. This can prevent users from uniquely being identified both by changing their data completely and by potentially creating the same set of data for two or more users.

Geographical restrictions

Our third simulation exercise that limits *where* users are tracked only keeps user data if they were within a mile of the restaurant in the previous week i.e., completely protects some users if they were outside the geofence.

Figure F2 illustrates these patterns for a sample of users from our data based on their distance to one restaurant in a given week (Chick-fil-A in Houston in Week 32 of our data). The figure plots the home locations of all users who, when geo-tracking data are unrestricted, are in the relevant category for this restaurant. The Figure classifies these users into two categories. Users whose home location is plotted in green are users who were not within the geofence, and users whose home location is plotted in blue are users who were in the geofence. In this exercise, all the data of users reported in blue are available, but none of the data of those users reported in green are available as inputs for Model C under geofencing restrictions.

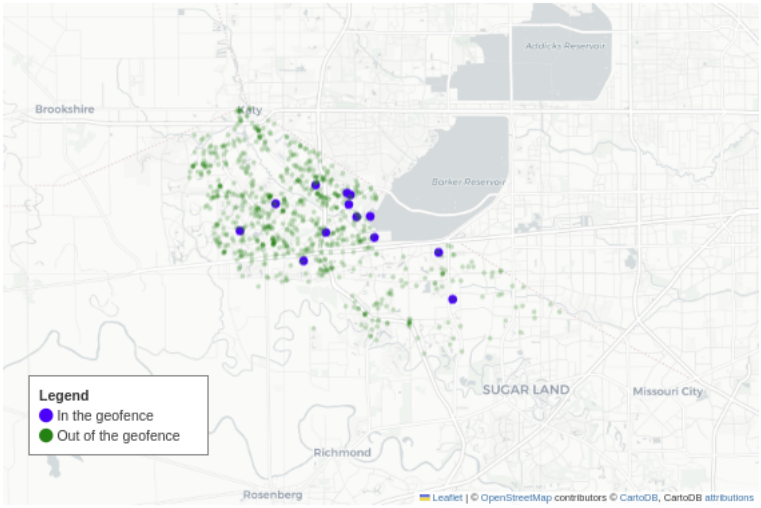


Figure F2: Example of Users Protected (vs. Not) Under Geofencing Simulations

Frequency restrictions

Our fourth simulation exercise limits *how frequently* users are tracked by reducing the frequency of tracking, e.g., one-third the frequency of original tracking.

In Figure F1, we illustrate the shift in home location using a sample user from our data. The Figure shows that reduced frequency tracking shifts the true home location (blue dot) to further away (purple dot), making it difficult to identify the user based on their true home. In this example, the shift induced by the KNN synthetic data is much larger than that induced under reduced frequency of tracking.

While the frequency restrictions hide a user's true home location, it is possible that they may still uniquely identify them using their records of past trajectories. In this way, frequency restrictions could be less privacy-preserving than some of the other simulations we discuss that completely transform the data and do not keep any raw coordinates.