Contents lists available at ScienceDirect

Journal of Business Research

journal homepage: www.elsevier.com/locate/jbusres

Predicting mobility using limited data during early stages of a pandemic

Michael T. Lash [a], S. Sajeesh [b,*], Ozgur M. Araz [b]

[a] School of Business, University of Kansas, Lawrence, KS 66045, United States
[b] College of Business, University of Nebraska - Lincoln, Lincoln, NE 68588, United States

ARTICLE INFO

Keywords:
Risk perception
Retail activity
Mobility
COVID-19
Multicollinearity
Hill-climbing algorithm

ABSTRACT

The COVID-19 pandemic has changed consumer behavior substantially. In this study, we explore the drivers of consumer mobility in several metropolitan areas in the United States under the perceived risks of COVID-19. We capture multiple dimensions of perceived risk using local and national cases and death counts of COVID-19, along with real-time Google Trends data for personal protective equipment (PPE). While Google Trends data are popular inputs in many studies, the risk of multicollinearity escalates with the addition of more relevant terms. Therefore, multicollinearity-alleviating methods are needed to appropriately leverage information provided by Google Trends data. We develop and utilize a novel optimization scheme to induce linear models containing strictly significant covariates and minimal multicollinearity. We find that there are a variety of unique factors that drive mobility in different geographic locations, as well as several factors that are common to all locations.

1. Introduction

Market disruptions are defined as profound changes in the business landscape that force organizations and supply chains to undergo significant transformations instead of incremental changes (Edelman & Heller, 2014). Such disruptions may impact the interactions among market participants. A technology-driven market disruption, such as a groundbreaking invention, could be welfare-enhancing since both the firm that is disrupting the market and consumers could benefit at the expense of the firm's competitors. In contrast, with a negative market-level shock, all parties could be worse off. Therefore, it becomes essential not only to study trends and the extent of changes in consumer behavior but also to predict future behavior in the presence of such shocks (Sheth, 2020). The global pandemic caused by the novel Coronavirus disease 2019 (COVID-19) provides a crucial backdrop to studying these phenomena.[1]

It is clear that the COVID-19 pandemic has also had a significant impact on the economy (Donthu & Gustafsson, 2020). Restrictions such as travel bans, cancellation of social events (concerts, sporting events, etc.), closure of non-essential businesses, and "stay-at-home" orders to mitigate the virus's transmission have influenced consumers' mobility patterns and reduced shopper traffic and supply disruptions in many industries (Kumar et al., 2019). These interventions, in turn, have negatively impacted the profitability of brick-and-mortar stores (Pantano et al., 2020). Thus, it has become vital to study the drivers of consumer mobility in such environments since retailers of essential and non-essential goods face contrasting demand shifts due to changes in consumer mobility (Roggeveen & Sethuraman, 2020). A better understanding of consumer mobility patterns can help managers optimize staffing, inventory, and in-store advertising (Sundararaj, 2017).

In addition to externally imposed restrictions, consumer mobility could be shaped by two important aspects. First, the specific value of pandemic health metrics that consumers observe in their local vicinity relative to the broader market could affect consumer mobility. The daily number of cases and deaths are typically used metrics to track the coronavirus pandemic (Lehmann, 2020). Second, consumers' risk perception of disease transmission could influence their willingness to adopt preventative health behaviors, such as avoiding the extent to which they travel. In contrast to the stated risk perception in consumer surveys, one could use consumers' online search patterns of scarce

* Corresponding author.
E-mail addresses: michael.lash@ku.edu (M.T. Lash), ssajeesh2@unl.edu (S. Sajeesh), oaraz2@unl.edu (O.M. Araz).

[1] It is well known now that Coronavirus Disease 2019 (COVID-19) is caused by a virus strain, SARS-CoV-2 (CDC, 2020a). It was first detected in December 2019 in Wuhan, China, and shortly after, it has spread to nearly every country. By early April 2020, more than 1.5 million people had been infected with COVID-19, and it had caused over eighty-eight thousand deaths worldwide (Johns Hopkins, 2020). The primary mode of transmission is through respiratory droplets expelled from the mouth or nose of an infected person and inhaled by a healthy person (CDC, 2020b).

https://doi.org/10.1016/j.jbusres.2022.113413
Received 5 January 2022; Received in revised form 24 October 2022; Accepted 28 October 2022
Available online 6 January 2023
0148-2963/© 2022 Elsevier Inc. All rights reserved.

pandemic-related paraphernalia to measure intrinsic risk perception.

Further, consumer behavior changes may vary based on the social and demographic characteristics of different communities and the rate and extent of transmission in the early stages of the pandemic. Thus, it may be essential to study consumer mobility patterns in distinct geographical areas to understand commonalities and differences in the underlying drivers of consumer mobility, specifically retail mobility (Clemons, 2008). In this paper, we study the effect of pandemic health metric information and consumers' perceived risk due to the pandemic on consumers' mobility decisions in three distinct geographic locations. We use real-time data from Google Mobility, Google Trends, and Twitter to develop a novel prediction algorithm for consumers' mobility patterns. While Google Trends data have been popular input in several studies to measure public interest, the risk of multicollinearity escalates with the addition of more relevant terms in such studies. Therefore, effective methods must be developed and implemented to maximize the benefit of the information from such data sources. With the developed novel algorithm, our methodology effectively handles multicollinearity and lags in covariates while ensuring that all model-included covariates are statistically significant. These considerations allow the results to be readily interpretable. Consumer mobility is defined as the aggregated, and anonymized movement pattern observed among consumers based on their mobile device location. Further, retail mobility refers specifically to consumer movement trends for retail locations such as grocery stores, restaurants, and shopping centers.

Based on the above discussion, we address the following research questions in this study:

(1) Are the risk severity perception indicators (i.e., Google Trends search data) affected by risk susceptibility measure (i.e., pandemic health impact metric)?
(2) How is consumer mobility affected by the two distinct dimensions of risk perceptions: susceptibility and severity?
(3) Given the correlated nature of Google Trends search data, how do we construct a parsimonious model to predict retail mobility?
(4) How do the risk perception dimensions differentially impact the various mobility activities in different metro areas?

We make the following three main contributions. First, several studies (e.g., Persson et al., 2021) have focused on analyzing how consumer mobility drives disease transmission and pandemic spread. In contrast, we focus on the reverse problem – how pandemic health metrics drive consumers' risk perception and subsequently impact retail mobility. Second, guided by the integrated framework on how newer models can aid decision-making in retailing and related supply chains (Bradlow et al., 2017), we develop a novel steepest ascent, steepest descent hill-climbing algorithm that generates linear predictive models of consumer mobility patterns. The algorithm ensures that minimal multicollinearity exists among the predictors and that all predictors used in the model are statistically significant. These considerations allow us to readily interpret which defined covariates have a bearing on the retail phenomena of interest. Furthermore, our developed modeling framework can also be adapted to study future disruptive events. Finally, our study broadens our understanding of the underlying drivers of consumer mobility during the initial phase of the COVID-19 pandemic – we find estimates of the relative effects of local vs global pandemic health metrics and consumer risk perceptions factors on consumer mobility. For instance, we find that in certain locations, searches for hand sanitizer, disinfectant, and masks are associated with decreases in retail mobility, but those different locations respond to different searches (e.g., disinfectant and hand sanitizer searches in Houston and hand sanitizer and mask searches in Omaha). Simultaneously, different locations seem to place different emphases on local vs national COVID-19 cases and deaths in the context of retail mobility. For instance, all locations responded to increases in either local cases or deaths by increasing retail mobility (the result is statistically significant in all cases), suggesting that individuals were likely to withhold their shopping tendencies when the pandemic was milder but were only willing to withhold such behavior for so long, which incidentally is when the virus had spread further, and COVID-19 cases and deaths had increased. Our comprehensive set of results can be useful to managers and policymakers in designing the appropriate supply chain and marketing strategies that are influenced by consumer mobility factors.

2. Literature review

We first discuss the key papers that form the building block for our empirical methodology. Subsequently, we review the literature on the applications of similar datasets and highlight our relative contributions.

2.1. Review of methods literature

A key aspect of this study is the introduction of predictive models containing minimal multicollinearity, using only statistically significant covariates. While we ultimately propose a model with these desired properties, our work builds on several studies in the extant literature we discuss below.

First, a well-established measure of multicollinearity is the variance inflation factor (VIF) (Hair Jr et al., 2016; Sheather, 2009; Chennamaneni et al., 2016; Weisberg, 2005). According to (Hair Jr et al., 2016), multicollinearity is present when the VIF of a covariate is greater than four. Other works have suggested that a VIF of 5 (Sheather, 2009) or even 10 (Weisberg, 2005) is acceptable. We will make use of this measure when developing our predictive model. Second, there are a variety of well-known variable selection procedures. Among these, sequential forward selection (SFS) (Grechanovsky & Pinsker, 1995; Lash et al., 2017, 2019), also referred to as stepwise selection (Zhang, 2016), is particularly relevant to this work since we incorporate this as part of our proposed method. SFS works by iteratively selecting the most favorable covariate according to some criteria or metric (typically, predictive performance improvement) and retaining the selected variable in the model. The process is repeated until either there are no more covariates to select or the addition of a variable produces a worse model (according to the adopted criteria). As mentioned, however, these methods tend to focus on predictive performance (Marcano-Cedeño et al., 2010; Ververidis & Kotropoulos, 2005; Cotter et al., 1999; Peduzzi et al., 1980; Hastie et al., 2020), rather than multicollinearity reduction or covariate statistical significance. SFS is a part of our methodological innovation in this space.

A procedure similar to SFS is backward variable selection (BVS), otherwise called backward elimination. When BVS is used, a model is initially induced on the full set of covariates. Subsequently, according to some adopted goodness-measuring metrics, the least favorable covariate is removed. The procedure then repeats until only desirable covariates (according to the chosen metric) remain. However, these methods, like SFS, tend to focus on predictive performance improvement as their criteria (Nguyen et al., 2014; Meyer et al., 2010), as do the very few algorithms that employ both SFS and BVS (Mao, 2004; Kano & Harada, 2000). Several select works, however, have focused on the methodology that sequentially selects based on p-values (Lash et al., 2017, 2019) rather than on definitive predictive performance-improving selection criteria.

Both SFS and BVS procedures belong to a broader class of optimization methods referred to as hill-climbing or local search optimization methodologies. Local search methods can be used to solve a wide variety of problems, including binary optimization (Bertsimas et al., 2013) and domain-specific problems, such as examination timetabling (Caramia et al., 2008). Hill-climbing algorithms, specifically, have been used in a variety of applications as well, including wind turbine optimization (Karabacak et al., 2019), reservoir optimization (Alsukni et al., 2019), and a variety of discrete optimization problems (Vaughan et al., 2005). Our novel hill climbing algorithm builds on the existing literature and

Table 1
Comparison with past literature by type of optimization method and optimization objective.

		Objective			
		Predictive Performance	P-Value	VIF	All Three
Optimization Method	SFS	Zhang, 2016; Marcano-Cedeño et al., 2010; Ververidis & Kotropoulos, 2005; Cotter et al., 1999; Peduzzi et al., 1980; Hastie et al., 2020	Grechanovsky & Pinsker, 1995; Lash et al., 2017; Lash et al., 2019	Dupuis and Maria-Pia, 2013	–
	BVS	Nguyen et al., 2014; Meyer et al., 2010	–	–	–
	Both	Mao, 2004; Kano & Harada, 2000	–	–	Our method

incorporates multiple criteria, as illustrated in Table 1 below:

2.2. *Review of data and data applications literature*

From a data perspective, web search engines' trend data have been used for different purposes, including surveillance or demand prediction of products and services (Du et al., 2015; Rivera, 2016). Laato et al. (2020) investigated consumers' purchasing behavior during the early stages of the COVID-19 pandemic, when fear of consumer market disruptions affected multiple sectors and supply chains. More recently, Keane & Neal (2021) used Google Trends and search data to model consumer panic during the COVID-19 pandemic and highlight how local and national governmental pandemic policies impact consumer panic across countries. Simionescu & Raišienė (2021) also used Google Trends data to study unemployment trends during the COVID- 19 pandemic in the EU new member states. Search engine data and trend indicators have also been used in supply chain and operations literature. Boone et al. (2018) used Google Trends data to improve sales forecasts and reduce out-of-sample forecast errors. Fritzsch et al. (2020) also used Google Trends to forecast sales using weekly data at the product level. Google Trends information has also been used to measure consumer buzz and forecast movie revenues (France et al., 2021). Our analysis builds on this stream of research and uses google search data (as a measure of consumer risk perception) to predict consumer mobility.

In addition, Google Trends data have been widely used to inform epidemiological studies. Ahmad et al. (2020) presented a study to assess the predictability of COVID-19 incidence using Google Trends data on internet search interest of certain gastrointestinal symptoms and terms. The study stated that these internet search data could be useful for predicting COVID-19 cases in the United States. Similarly, Asseo et al. (2020) used taste and smell loss-related search terms to track the cases in the United States and Italy and discussed the benefits and limitations of using Google Trends data in disease surveillance. Internet search data are also useful for improving forecasts for certain infectious disease activities, for improving surveillance, and supporting real-time decisions. In earlier studies, Google Trends data have been used in forecasting influenza activity and predicting related outcomes (Araz et al., 2014; Yang et al., 2015). Kandula et al. (2019) used Google Trends data to forecast influenza-associated hospitalization in the United States and suggested that these web-search data can provide important real-time information and improve the accuracy of forecasts for hospitalizations.

Prior research has also highlighted how consumers' risk perception influences their use of public transportation, such as ride-sharing services or metro trains. (e.g., Wang et al., 2019; Basu & Ferreira, 2021; Garaus & Garaus, 2021). Chernozhukov et al. (2021) use Google Mobility data to measure the impact of the social distancing policies during the COVID-19 pandemic. Using geo-spatial analyses, a large telecommunication data set is also used to monitor human mobility during the COVID-19 pandemic (Persson et al., 2021). We complement this stream of research by using the developed predictive model to forecast consumer mobility in three US cities using publicly available data from Google that reflects "real-time" consumer trends. Such data is available for a majority of regions/cities across the globe, and our algorithm could be used to forecast retail mobility, given the data.

With rare events, organizations may not have relevant prior experiences to draw inferences and aid decision-making. The Covid-19 pandemic represents such a rare event. Responding to rare events requires organizations to mobilize and adjust existing resources quickly as well as develop new capabilities (Henningsson et al., 2021). Further, rare events (such as the Covid-19 pandemic) are usually characterized by less available data (Oehmen et al., 2020). For example, in the case of the Covid-19 pandemic, retailers may not have accurate epidemiological data on transmission and may have to use other heuristics, such as consumer mobility patterns, to optimize retail decisions. During the early days of the COVID-19 pandemic, due to potential and actual social distancing policies, including mandated lockdowns or potential lockdowns, consumer trends were disrupted, which also affected retail activities globally (OECD, 2020). As the epidemic growth showed geographic variation, some states, e.g., New York (NY), observed early surges in hospitalizations and early long-term state-wide lockdowns. Meanwhile, other states, e.g., Nebraska (NE) and Texas (TX), observed increased cases and deaths sometime later than NY. These geographic variations in cases and deaths also produced varying demands on certain items, such as hand sanitizer and masks, as well as varying mobility patterns, including retail activity patterns.

Our work contributes to and broadens the extant literature by focusing on how COVID-19 risk perceptions impact retail mobility (rather than on how retail mobility impacts the spread of COVID-19) through the development of a novel hill-climbing heuristic that produces linear models containing minimal multicollinearity and strictly significant covariates, and by analyzing the factors that drive consumer mobility during the initial phase of the COVID-19 pandemic.

More broadly, recent technological advances are transforming retailing (Shankar et al., 2020), and understanding the drivers of consumer mobility data can enable organizations to further optimize technology-driven tools to manage market disruptions. Our method is scalable to include more risk perception indicators to predict retail mobility, which could be a starting point toward developing actionable retailing strategies. Although we do not focus on the impact of retail mobility on specific retail strategies, we believe that the outcomes from our predictive model could be used as strategic inputs to optimize several retail decisions, such as staffing, inventory, and location-based online and in-store advertising decisions.

3. Data and methodology

This section presents the data and the methodology used to predict mobility as a function of consumers' perceived risk. Prior research has identified that perceived risk plays a vital role in consumer behavior. Based on the stay-at-home order issued by the state and county governments, one significant behavioral change relates to their mobility decisions. Therefore, due to individuals' perceived risk, individuals may restrict or change their mobility for different activities. We first describe the risk perception dimensions and then discuss data sources that map onto these dimensions.

Risk perception is broadly defined as evaluating the subjective probability of a negative outcome and its consequences (Sjöberg et al., 2004; Menon et al., 2008). Prior research suggests that risk perceptions have two dimensions, susceptibility and severity (El-Toukhy, 2015).

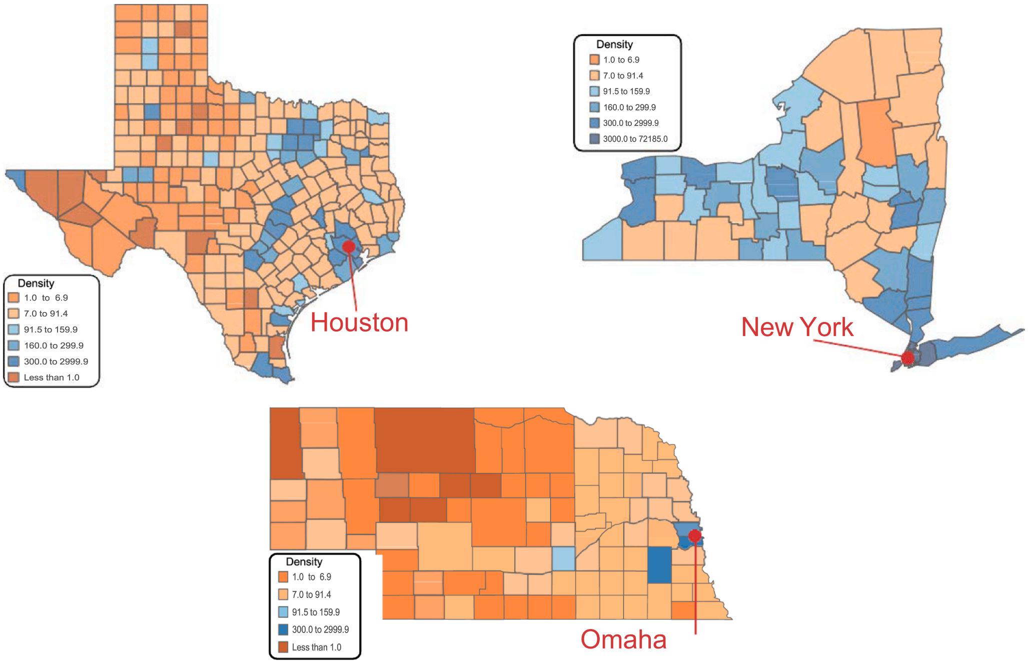


Fig. 1. The three states and coinciding cities considered by our study with population density by county indicated with coloring. Blue indicates counties with higher population density and orange lower population density. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Susceptibility refers to the likelihood of experiencing a health risk, whereas severity refers to the seriousness of the risk (Brewer et al., 2007).

The local and national cases and death data due to the pandemic publicized by the media and captured in the Pandemic Impact metrics provide consumers with a measure of the pandemic's prevalence in their local communities, impacting their risk perceptions in terms of their susceptibility to the pandemic. Further, such risk assessment by consumers affects their decision to search for information (Maser & Weiermair, 1998). Information search for pandemic-related paraphernalia captures the risk perception in terms of severity. Moreover, mortality data may also directly affect consumers' perceptions of the severity of the risk associated with the pandemic. Therefore, in our empirical analysis, we incorporate the effect of both these dimensions of risk perception on mobility.

Since Google Trends data present timely web-search information on certain items in different locations, the search trends on masks, hand sanitizers, and disinfectants can reflect the perceived COVID-19 risk of individuals in these locations. These trends are impacted by reported local COVID-19 death and case data and nationwide COVID-19 case and death data shared in the national news. Further, Google Mobility data present real-time information about these activities in various locations. While the COVID-19-related epidemiological data can affect these mobility activities, they can also be affected by individuals' risk

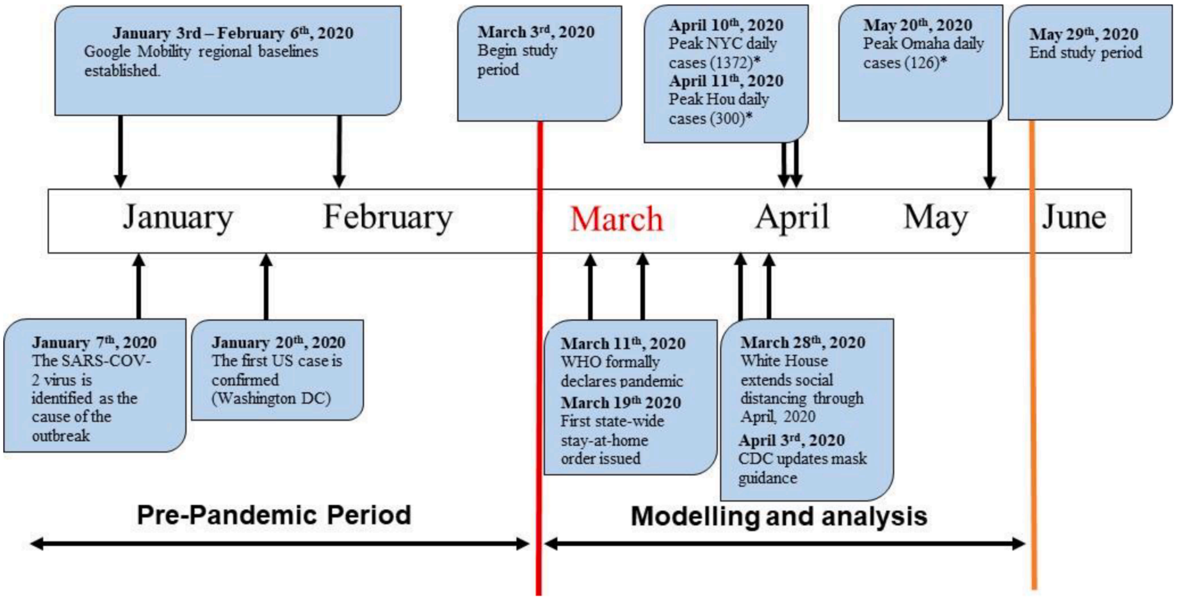

January 3rd – February 6th, 2020
Google Mobility regional baselines established.
March 3rd, 2020
Begin study period
April 10th, 2020
Peak NYC daily cases (1372)*
April 11th, 2020
Peak Hou daily cases (300)*
May 20th, 2020
Peak Omaha daily cases (126)*
May 29th, 2020
End study period
January
February
March
April
May
June
January 7th, 2020
The SARS-COV-2 virus is identified as the cause of the outbreak
January 20th, 2020
The first US case is confirmed (Washington DC)
March 11th, 2020
WHO formally declares pandemic
March 19th 2020
First state-wide stay-at-home order issued
March 28th, 2020
White House extends social distancing through April, 2020
April 3rd, 2020
CDC updates mask guidance
Pre-Pandemic Period
Modelling and analysis

Fig. 2. A timeline showing several relevant COVID milestones, our study period, and the period in which Google Mobility established baselines from which relative COVID mobility is measured. * Indicates peak cases within the sample period (March 3 - May 29, 2020).

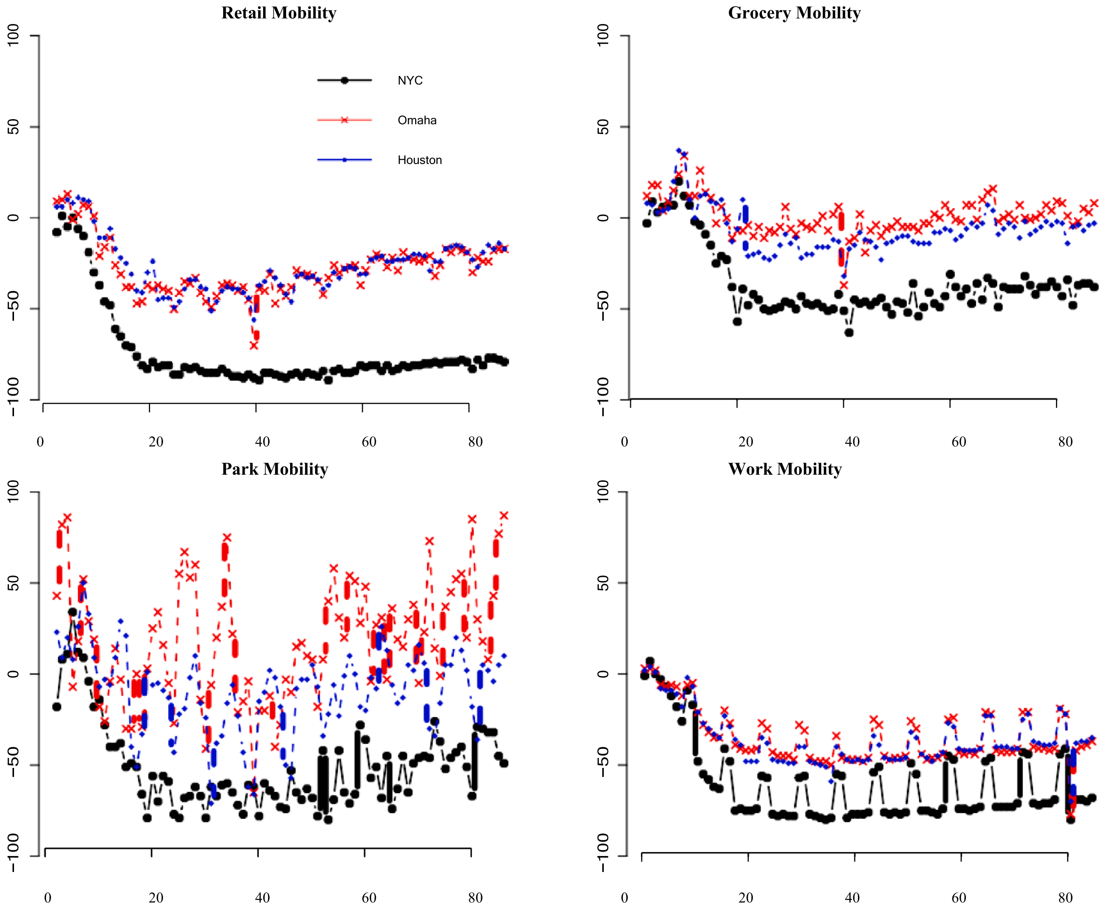


Fig. 3. Google Mobility data across three cities from March 3rd to May 29th, 2020.

perceptions inferred from the Google Trends data we gathered on specific personal protection equipment(s) (PPEs).

3.1. Data

The COVID-19 pandemic has continuously evolved over time. As such, it was necessary to truncate data collection efforts and proceed with our analysis. Here, we explain the time frame and locations selected for inclusion in this study and discuss the investigated Google Trends search terms.

3.1.1. Choice of locations

One of the objectives of our study is to understand the relative impact of local vs global cases of COVID-19 and other perception-of-risk measures on consumer mobility in different geographic locations. Therefore, we decided to focus on three major cities in the US: New York City (NYC), Houston, and Omaha. These cities differ in terms of their population density, socioeconomic characteristics, and the transmission dynamics of COVID-19 during the initial stages of the pandemic. In NYC, there was a surge in the number of cases and deaths due to COVID-19 starting in March 2020, whereas the Houston area was not impacted until much later in the summer. Omaha is a relatively smaller city (although still an established metropolitan area) in the Midwest and is geographically separated from both NYC and Houston. All these dynamics and features of the metro areas allowed us to study the differential impact of local and national pandemic health metric information. All these metro areas are presented in Fig. 1, with population densities across the considered states mapped at the county level. These maps demonstrate that these locations are either the most populated or one of the most populated metro areas in their corresponding state.

3.1.2. Time frame

We collect data for the 12-week period from 3rd March 2020 to 29th May 2020 for our analysis. This time frame is chosen because the World Health Organization (WHO) classified the COVID-19 outbreak as a pandemic on 11th March 2020. This led all state and county governments to issue a stay-at-home order. Around the third week of May 2020, there was a shift in government policy, allowing businesses to reopen in a phased manner, albeit with restrictions. Furthermore, the stay-at-home order was modified to a “safer at home” order.

To further clarify our study period, the temporal evolution of the COVID-19 pandemic, and the period in which Google Mobility baselines were established, we provide Fig. 2 below.

WHO and the Center for Disease Control (CDC) assess risk and preparedness for a pandemic on a continuum of four pandemic phases (alert, pandemic, transition, and interpandemic phases) (CDC 2016, WHO 2010). The time frame of our study corresponds to the alert and initial pandemic phases. During such a period, there is limited data about the accurate estimates of the virus transmissibility and severity, which are crucial parameters in understanding and predicting the course of a pandemic in epidemiology. Thus, we use available real-time data as proxy measures of consumers’ perceptions of risk severity and susceptibility to understand drivers of consumer mobility. The outcomes from our predictive model could be used as strategic inputs to improve several retail decisions even when there is limited data about virus transmission and severity.

3.1.3. Search items used in the measurement of risk

We chose the following three search terms as a measurement of consumers’ risk perception: hand sanitizer, masks, and disinfectant. These items were chosen as they were common virus-preventative

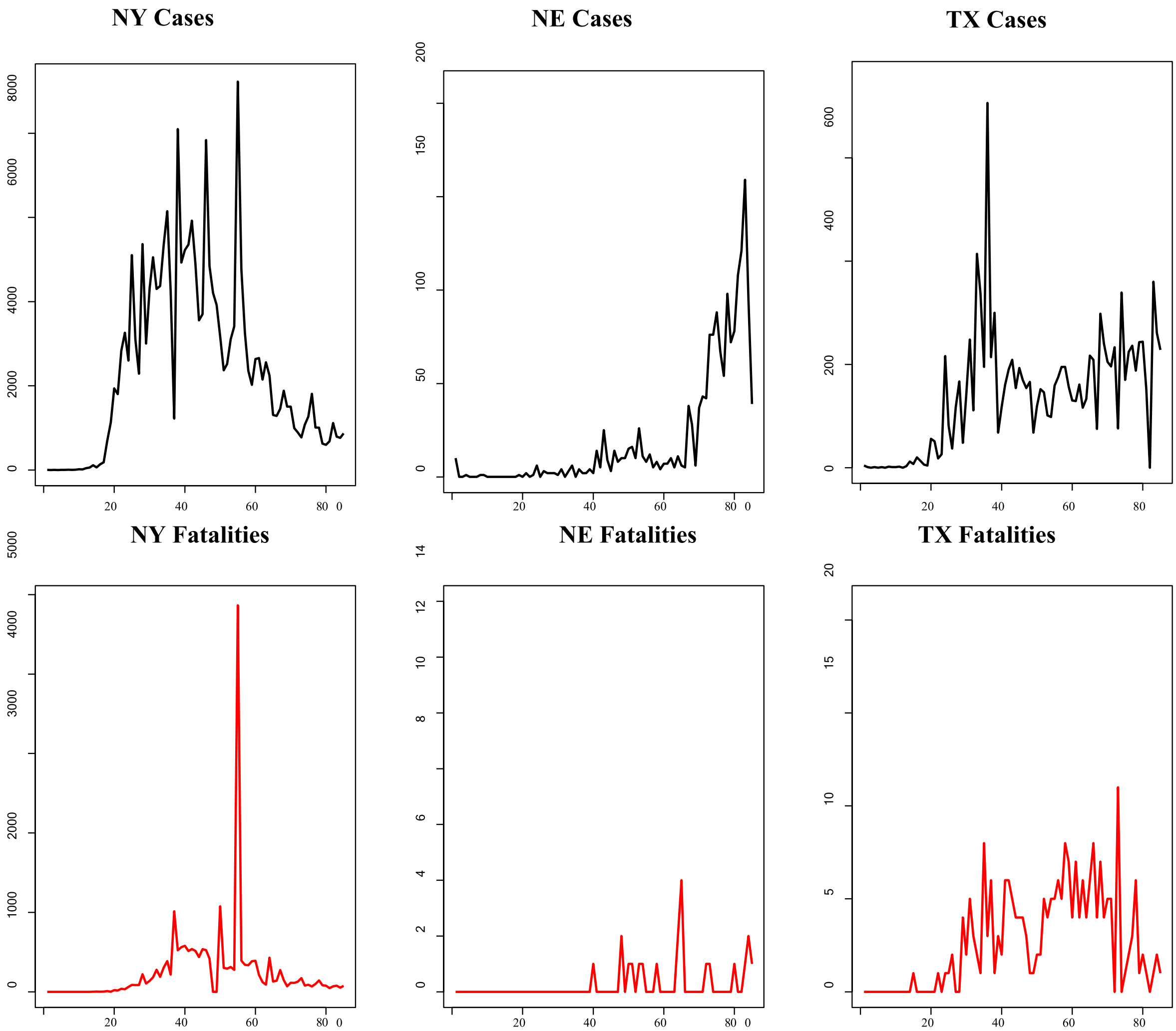


Fig. 4. Cases and Fatalities by city from March 3rd to May 29th, 2020.

measures presented to the public at the time of the study. In fact, there was a documented shortage of these items during the period of our analysis. Although we restricted our analysis to these search terms, our model can easily be scaled to incorporate additional google search terms. Further, the predictive accuracy of our parsimonious model is sufficiently high, as detailed in the Results section.

3.2. Variable descriptions

In this section, we describe the dependent and independent variables used in our econometric models. These models aim to determine the factors that significantly impact consumer mobility activities as a consequence of the COVID-19 pandemic. We combine four main sources of data: 1) Google Trends data, 2) Google Mobility data, 3) Data on mortality and number of infections due to COVID-19, and 4) Twitter data (tweets) relating to the COVID-19 pandemic.

3.2.1. Dependent variables

Google Mobility Data: Google mobility data is generated by aggregating information from users (who have agreed to share their location information) in conjunction with Google Maps.[2] Further, the mobility information is grouped into six broad categories that have comparable features from a social distancing guidance perspective. The six categories are: (i) retail & recreation, (ii) grocery & pharmacy, (iii) parks, (iv) transit.

stations, (v) workplaces, and (vi) residential places. These data have been constructed by comparing visits and lengths of stays at certain places relative to a baseline (Google Mobility, 2021). The retail & recreation cate- gory provides data on mobility trends for places such as restaurants, cafes, and shopping centers. Grocery & pharmacy category provides data on mobility trends for sites considered to be essential trips, including grocery markets, drug stores, and pharmacies. Similar subcategories of related locations are grouped within parks, transit stations, workplaces, and residential places (Google Mobility, 2021). The use of such types of consumer mobility data is also in vogue in the extant literature (e.g., Persson et al., 2021).

Note that the Google mobility data compare mobility for the reported date to a baseline day. The baseline day represents a normal value for that day of the week calculated as the median value from the 5-week period, Jan 3 - Feb 6, 2020. Thus, consistent with our data source, we are interested in predicting retail mobility relative to the baseline. Fig. 3 shows the time-series representation of the Google mobility values for the various categories across three cities.

Our analysis primarily focuses on predicting retail mobility, representing more leisurely retail activities, and grocery and pharmacy mobility, representing retail engagements that are more out of necessity. Fig. 3 highlights differences in mobility types across cities – the differences arise due to differences in availability of public transportation, the density of retail outlets, stage of the pandemic as well as differences in

[2] More details can be found at https://www.google.com/covid19/mobility/.

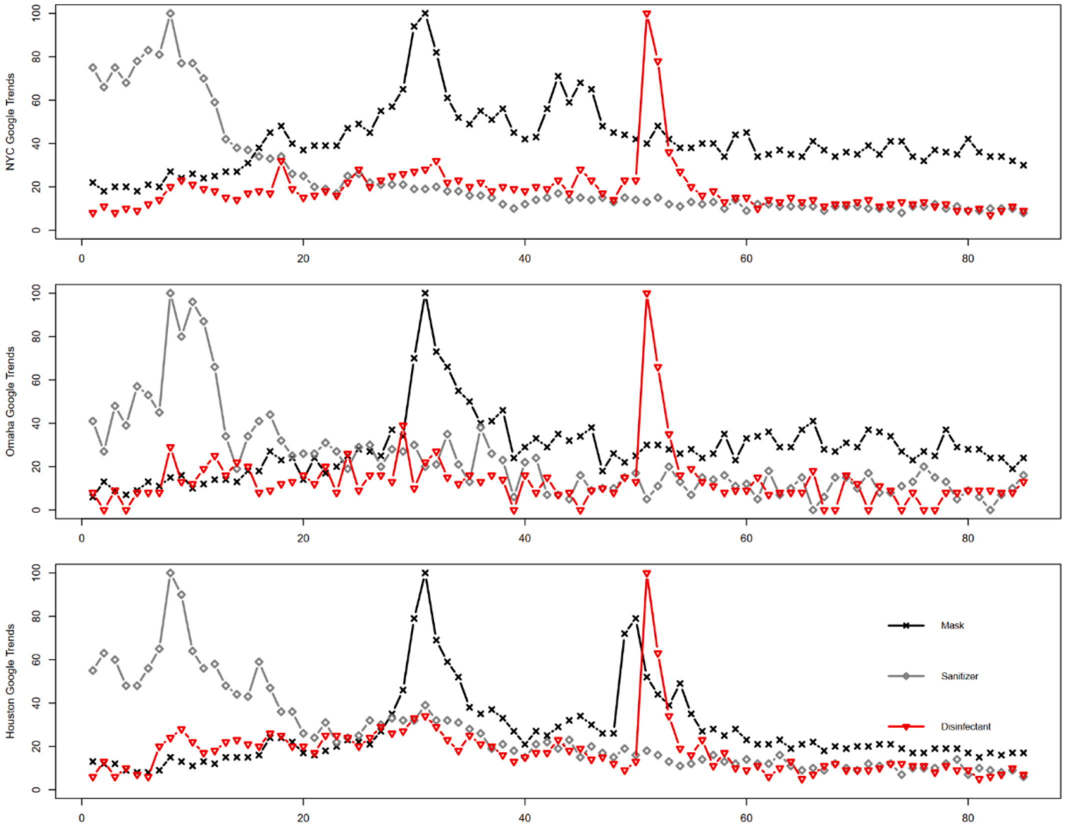


Fig. 5. Google Trends values by city from March 3rd to May 29th, 2020.

consumer density. Among the three cities under analysis, the effect of the pandemic was observed to a larger extent in New York City initially. This substantially impacted retail mobility in NYC in the time period under consideration. Further, as Fig. 1 points out, the consumer density is much higher in New York City relative to Houston and Omaha, which impacts retail mobility.

3.2.2. Independent variables

COVID-19 Cases and Mortality: We incorporate location-specific COVID-19 case and death counts into our analyses, as well as counts for the US as a whole. Such data represent immediate indicators of COVID- 19 severity and are likely the factors driving Google searches for masks, hand sanitizer, and disinfectants. Although case fatality rates and the prevalence are epidemiologically better indicators of the severity and the transmissibility of the virus spreading in communities, at the time of the study, news outlets were presenting both the number of deaths and the confirmed cases separately, which were driving the public perception about the risk associated with COVID-19. Therefore, for each state and city considered in the study, daily cases and death counts are obtained from New York Times data repository and used.[3] Fig. 4 shows the COVID-19 daily cases and deaths at the three locations of interest as a time series.

Google Trends: Google Trends provides a comparison of the search volume of different queries over time. Google assigns a popularity index ranging from 0 to 100 to keyword searches in which 100 represents the maximum search interest for the selected location and time (Google Trends, 2021). Fig. 5 represents the Google Trends values for our focal search terms in the three cities as a time-series. The data represent the relative search interest for that specific term in that specific geographic region as a proportion of all searches on all topics in Google. We build on the literature on the use of Google Trends data for pandemic-related studies (e.g., (Ahmad et al., 2020), (Asseo et al., 2020)). However, in contrast to extant literature that studied the effect of Google Trends on cases/incidence, we use the search trends data to measure consumers' risk perception and analyze its impact on retail mobility.

COVID-19 Tweets: Twitter provides a platform by which individuals can express their opinions to the public through small, 280-word posts referred to as "tweets". Tweets that are specific to the COVID-19 pandemic provide a means of assessing public engagement and awareness related to the spread of the disease. Since retail mobility may be closely related to the public's broader awareness of the pandemic, we propose to use COVID-19-specific tweets to measure engagement and awareness. Furthermore, social media-derived variables have also proven successful in other domains (Hu et al., 2019; Barrett & Orlikowski, 2021; Zhang & Ram, 2020). Our tweet dataset, which is publicly available (Chen et al., 2020), consists of approximately 60 million tweets for the period under consideration (March-May 2020). Of these 60 million tweets, 20,000 have location-specific information, and of these 20,000 tweets, 1,800 are specific to Nebraska, Texas, and New York states. Since there are relatively few tweets specific to the states of the cities in our study under consideration, we elect to use these statewide counts rather than further refine the geographic scope of eligible tweets.

3.3. Modeling approach

In this study, we want to quantify the impact of each variable on the phenomena of interest. To do so, we propose to construct ordinary least squares (OLS) regression models and then analyze the sign and magnitude of the coefficients to determine each variable's impact on the phenomena under consideration.

We construct two sets of ordinary least squares (OLS) predictive

[3] More information can be found at https://github.com/nytimes/covid-19-data.

models:

A1. Models to predict each of the Google Trends variables: mask, hand sanitizer, and disinfectant search popularity.
A2. Models to predict each of the Google Mobility variables: retail mobility and grocery and pharmacy store mobility.

Additionally, the models we construct are location-specific. In other words, we construct models for Houston, New York, and Omaha independent of one another in order to assess and compare the factors that drive the phenomena of interest in a location-specific manner. Therefore, we construct and analyze five different models – three Google Trends models and two Google Mobility models – for each location, thus totaling 15 predictive models overall.

3.3.1. *Empirical models*

Based on our variables described in Section 3.2, the Google Trends OLS models can be expressed as:

$$gt_t^{(k)} = \beta_0^{(k)} + \sum_{l=1}^{T}\Big[\beta^{(k)}_{1_{(t-l)}} Cases^{(k)}_{(t-l)} + \beta^{(k)}_{2_{(t-l)}} Deaths^{(k)}_{(t-l)} + \beta_{3_{(t-l)}} USCases_{(t-l)} + \beta_{4_{(t-l)}} USDeaths_{(t-l)}\Big], \quad (1)$$

where the β terms are coefficients in the model and $l = 1, \ldots, T$ represents the number of days preceding the observation of each independent variable gt. We refer to $l = 1, \ldots, T$ as the *lag time* of each variable and each collection of variables $var^{(l)}$: $l = 1, \ldots, T$ as a so-called "variable group". Intuitively, there is likely a high degree of correlation between the observed value of each variable from one day to the next (i.e., among the variables in a variable group). For instance, the number of US COVID-19 cases one day (i.e., $l = 1$) prior is likely highly correlated with the number of US COVID-19 cases two days (i.e., $l = 2$) prior. The presence of multicollinearity presents a problem if we wish to interpret the coefficients as an indication of the effect of the predictors on the predicted Google Trends variable. We will discuss and alleviate this issue in the next section. Note that $gt \in \{mask, hand\ sanitizer, disinfectant\}$ and $k \in \{Houston, NYC, Omaha\}$.

We also build and analyze OLS models that predict two Google Mobility variables: retail mobility and grocery and pharmacy mobility. We can express the location-specific models as:

$$gm_t^{(k)} = \beta_0^{(k)} + \sum_{l=1}^{T}\Bigg[\Bigg(\sum_{gt\in GT}\beta^{(k)}_{gt_l} gt_l^{(k)}\Bigg) + \Bigg(\sum_{gm'\in GM'}\beta^{(k)}_{gm'_l} gm'^{(k)}_t\Bigg) + \beta^{(k)}_{1_{(t-l)}} Cases^{(k)}_{(t-l)} + \beta^{(k)}_{2_{(t-l)}} Deaths^{(k)}_{(t-l)} + \beta_{3_{(t-l)}} USCases_{(t-l)} + \beta_{4_{(t-l)}} USDeaths_{(t-l)} + \beta^{(k)}_{5_{(t-l)}} Tweets^{(k)}_{(t-l)}\Bigg], \quad (2)$$

where gm $\in$ {retail, grocery and pharmacy}, gm′ $\in$ GM′ = {park, transit, workplace, residential}, and GT = {mask, hand sanitizer, disinfectant}. Again, note the presence of temporal lag l = 1, …, T.

As we have mentioned, in both the models captured by Equations (1) and (2), there is likely to be a high degree of multicollinearity both among variables in the same variable group and among variables in other groups (see the Appendix for location-specific correlation matrices). This is problematic since we wish to interpret the sign and magnitude of the coefficients as indicators of the effects of the predictors on the predicted variable of interest, as well as understand when such variables have the highest degree of impact. Therefore, in the next section, we propose a method to induce OLS models that contain minimal multicollinearity and only statistically significant covariates, thus providing the desired interpretability.

Table 2
Grocery and pharmacy mobility sequential forward selection (SFS) results obtained for each of our three selected locations. The covariates with a VIF > *4* are highlighted with the VIF values in red. The coefficients that are flipped due to multicollinearity, as compared to the results obtained using our method (Table 9), are highlighted in blue.

City	Variables	Lag	Coeff	VIF
Houston	Transit Mobility	1	35.662***	5.195
	Local Cases	2	-12.054***	1.593
	Res Mobility	2	-10.441*	2.364
	Hand San	1	-19.408***	3.275
	Park Mobility	3	-12.171**	2.034
	US Deaths	6	9.831**	3.274
	C19 Tweets	2	-6.221**	1.181
	Work Mobility	6	23.985***	4.577
NY	Res Mobility	1	8.354***	1.716
	Local Deaths	7	-9.383*	1.128
	Work Mobility	7	11.499**	4.116
	Local Cases	1	-11.488***	2.083
	Transit Mobility	3	58.422***	4.68
	Mask	7	-13.415***	2.383
	US Cases	7	20.6***	2.75
Omaha	Transit Mobility	4	16.121**	2.832
	Res Mobility	6	-11.426**	2.509
	Local Cases	3	7.691*	1.108
	US Cases	6	35.341***	10.451
	Mask	7	-24.315***	1.783
	Local Deaths	1	11.85**	1.148
	Work Mobility	3	21.997***	3.052
	US Deaths	6	-17.011**	7.337

3.3.2. *The multicollinearity problem*

To explore and illustrate the problem of multicollinearity in the context of our problem setting, we initially adopt a sequential forward selection (SFS) procedure that selects the covariate with the smallest p-value at each iteration for inclusion in the final model. Note that we will adopt this method as a sub-procedure in our proposed method in the next subsection and provide an algorithmic sketch in the Appendix section – Algorithm B.1.

We apply SFS to our Grocery and Pharmacy Mobility prediction problem for each of our three selected locations and present the results in Table 2 below. The results of Table 2 show the benefits of SFS – the models are relatively sparse with strictly significant coefficients at the $\rho \leq 0.05$ level. However, the result also highlights the issue of multicollinearity, even when such a method is adopted. As we discussed in the related works section, when the VIF of a covariate is greater than four, multicollinearity is present (Hair Jr et al., 2016). The covariates in Table 2 with VIF $\geq$ 4.0 are highlighted in grey, and the VIF values are highlighted in red. As shown, a certain degree of multicollinearity is still present in each location-specific model.

The problem with multicollinearity is that it affects model coefficient estimates, making interpretation and assessment of these coefficients with respect to the phenomena of interest unreliable, even flipping the sign of effected coefficients. We can see these effects in each of the model results disclosed in Table 2. Those coefficients that have a different sign than those uncovered by our method (disclosed in the next sub-section) are highlighted in blue (interested readers can compare these results to the results of our proposed model, found in Table 9).

Here we can see that even though the p-values are significant, the signs of these coefficients are still flipped. Therefore, if we are to reliably interpret the coefficients of the models in our problem setting, additional innovation is needed.

3.3.3. *Sequential forward p-value Selection, Backward VIF Elimination*

As we illustrated in the previous sub-section, multicollinearity is present in our expressed models, even when accepted procedures, such as SFS, are adopted. This is an issue since we wish to interpret the sign and magnitude of the coefficients in our models as indicators of the predicted outcomes. The presence of multicollinearity prevents this, however, oftentimes inverting the sign of collinear variables, skewing the magnitude of the coefficients, and producing insignificant p-values.

Simultaneously, we also wish to determine which lagged variables significantly affect the outcome of interest and to keep only these variables in our models. By keeping only these lagged variables, we can not only comment on which of our defined variables have a bearing on the various outcomes of interest but also *when* these variables have a significant impact on such outcomes.

To summarize, we want our OLS models to produce three benefits:

B1. Determine the optimal *lag time* between the predictors and the dependent variable.
B2. Models with *only* statistically significant predictors (variables).
B3. Models based on variables that have minimal multicollinearity.

To achieve the above desirables, we begin by formulating an optimization problem that transforms (B1), (B2), and (B3), above, into a mathematical formulation we can optimize over:

$$\begin{aligned}&\min_{X'} \boldsymbol{I}^{\intercal}(\boldsymbol{y}-(\boldsymbol{X}'\widehat{\boldsymbol{\beta}}))^{2}\frac{1}{n}\\&s.t.\rho\left(\widehat{\beta}_{j}\right)\leq\alpha\forall_{j}\\&VIF\left(\boldsymbol{X}'_{j}\right)\leq\gamma\forall_{j}\end{aligned} \tag{3}$$

where $\widehat{\beta}=(\mathbf{X}'^{,\top}\mathbf{X}')^{-1}\mathbf{X}'^{\top}\mathbf{y}$, which is the closed form solution of OLS, $\mathbf{I}$ is an identity vector, $\mathbf{X}'\in \mathrm{R}^{n\times p\prime}$ is a design matrix, where $p'\leq p$, with $\mathbf{X}\in \mathrm{R}^{n\times p}$ being the original design matrix. In other words, $\mathbf{X}'$ is a design matrix that contains fewer (or the same) number of features as the original design matrix $\mathbf{X}$. The function $\rho(\cdot)$ determines the p-value of an inputted coefficient βj. α is the largest p-value of a particular coefficient allowed. The function $V\ IF\ (\cdot)$ determines the variance inflation factor (VIF) of each column j of a design matrix $\mathbf{X}'$. γ is the largest VIF value allowed. Briefly put, Equation (3) optimizes over the features that are included in the design matrix $\mathbf{X}'$ in a manner that:

C1. Produces a model that most minimizes predictive error.
C2. Ensures that all coefficients are statistically significant at the α level or better via the $\rho\ (\cdot)$ function. We select $\alpha = 0.05$ when running our experiments by convention.
C3. Ensures that all predictors have minimal multicollinearity via the $V\ IF\ (\cdot)$ function and corresponding cutoff γ. A VIF cutoff of 4 is the most conservative (i.e., smallest) VIF cutoff value we found in the relevant literature (Hair Jr et al., 2016), and therefore adopt this value in our experiments (i.e., $\gamma = 4$).

To solve the optimization model of Equation (3), we propose a steepest-ascent, steepest-descent hill-climbing algorithm we refer to as *Sequential Forward p-value Selection, Backward VIF Elimination* (SFPS-BVE). Our SFPS-BVE algorithm is provided and discussed in Appendix B. In short, our SFPS-BVE algorithm consists of two sub-procedures: Sequential Forward p-value Selection (SFPS) and Sequential Backward VIF Elimination (SBVE), performed in that order. The SFPS procedure is designed to produce an OLS model consisting of only statistically significant p-values, which addresses ($C2$) above. Our p-value criterion is in contrast with previous forward variable selection algorithms which tend to focus on predictive performance (Marcano-Cedeño et al., 2010; Ververidis & Kotropoulos, 2005; Cotter et al., 1999; Peduzzi et al., 1980; Hastie et al., 2020). The SBVE procedure is subsequently applied to ensure that none of the covariates exhibit any multicollinearity, which addresses (C3), above. Here, it is also worth noting that backward variable selection algorithms also tend to focus on predictive performance improvement (Nguyen et al., 2014; Meyer et al., 2010), as do the very few algorithms that employ both forward and backward selection, e.g., (Mao, 2004; Kano & Harada, 2000).

The closed form solution to the OLS procedure ensures that the obtained model most minimizes predictive error using the currently selected set of covariates; therefore, (C1), above, is also addressed. Thus, our proposed SFPS-BVE algorithm is able to address all of the above-enumerated desirables, enabling us to reliably interpret the sign and magnitude of the coefficients of the obtained OLS models as indicators of each variable's effect on the predicted phenomena of interest. We provide and discuss this algorithm in Appendix B.

3.4. *Analysis and model evaluation*

In this section, we first elaborate on our proposed method of analysis, followed by a brief discussion of how the lagged variables are instantiated in our datasets. Finally, we discuss the model evaluation.

We conduct all of our experiments using our proposed *Sequential Forward p-value Selection-Backward VIF Elimination* method elaborated in Section 4 with $\alpha = 0.05$ and $\gamma = 4$. In particular, we apply SFPS-BVE and then analyze the coefficients, p-values, and VIF values of all model-selected features to assess the impact of each feature on each dependent phenomena of interest.

As mentioned before, there are two categories of dependent variables: the three Google Trends variables, which include hand sanitizer, mask, and disinfectant searches, and the two Google Mobility variables, of which we will analyze retail mobility and grocery and pharmacy mobility. We will use the model disclosed by Equation (1) (Section 3.3.1) to analyze the three Google Trends variables and the model disclosed by Equation (2) (also Section 3.3.1) to analyze the two Google Mobility variables. For each of the dependent variables, a model is constructed for each of the three locations considered in this study – Houston, NYC, and Omaha – for a total of 15 models.

In order to analyze the different variables using the models proposed in Section 3.3, we apply a *lag* to all predictors (independent variables) used to assess a particular phenomenon of interest. Our selected lag duration, expressed in both models (Equations (1) and (2)) using T, is seven days – i.e., $T = 7$ – thus producing seven independent variables in each so-called variable group. To be more concrete, we observe a particular phenomenon of interest y at t – i.e., yt – and then associate each corresponding predictor observation at $\mathrm{l} = 1, 2, \ldots, 7$ days preceding the observation of yt to create each of our datasets. We believe a seven-day lag period is a reasonable length of time in which a particular predictor may continue having an impact on each of the dependent variables.

For our Google Trends prediction models, we will apply SFPS-BVE to the entire dataset, consisting of 84 days, without setting aside any days for testing. We do this since we are predominantly interested in analyzing the impact of local and nationwide COVID-19 data on the three Google Trends variables rather than quantifying the predictive capacity of such variables. When analyzing Google Mobility data, however, we set aside the last seven days of our study period (May 23rd - May 29th) to also comment on the predictive capacity of our obtained models. We elect to use a test set for this analysis since these variables are the primary focus of this study, and we wish to assess the predictive and extrapolatory capacity of the obtained models.

4. Results

In this section, we present our results, beginning with Google Trends and followed by Google Mobility. As mentioned, our results are provided in terms of the coefficients, p-values, and VIF values of the variables selected by our method. In each of the provided coefficient results tables, we indicate a p-value significance of ≤ 0.001 with * * *, p-value significance of $0.001 < \rho \leq 0.01$ with **, and p-value significance of $0.01 < \rho \leq 0.05$ with *. Additionally, we also provide the lag (days) of each

Table 3
Masks.

City	Variables	Lag (days)	Coefficients	VIF
Houston	US Cases	1	53.324***	1.71
	Loc Cases	2	−28.462*	1.512
	Loc Deaths	6	−23.808***	1.272
NY	US Cases	1	37.281***	2.357
	NY Loc Cases	2	21.829**	1.811
	US Deaths	7	−27.169***	1.703
Omaha	US Cases	1	46.212***	1.688
	US Deaths	7	−20.267***	1.688

Table 4
Hand Sanitizer.

City	Variables	Lag (days)	Coefficients	VIF
Houston	US Deaths	1	−21.823***	1.43
	Loc Deaths	7	−26.778***	1.43
New York	US Cases	2	−55.503***	3.109
	Loc Cases	5	27.941***	2.447
	US Deaths	7	−10.679*	1.685
Omaha	US Cases	4	−42.974***	1.025
	Loc Cases	5	−19.861**	1.025

selected covariate, which tells us when a particular variable has an impact on the corresponding phenomena of interest.

4.1. *Predicting and analyzing Google Trends data*

We first present the results of applying our method to the three Google Trends variables: hand sanitizer, masks, and disinfectant.

4.1.1. *Masks*

Table 3 presents the results of the mask search popularity model in each location. First, we can see that our method has worked as designed. This observation is also replicated across all models and presented results: all p-values are significant at the 0.05 level or better, and all VIF values are less than 4. Second, our results show a mix of factors both positively and negatively contributing to searches for masks. Interestingly, we observe that US Cases positively contribute to mask searches in all three cities, whereas US Deaths negatively contribute to mask searches in NY and Omaha. This may be attributable to the emphasis placed on “Stop the Spread” campaigns implemented by organizations such as the CDC (Centers for Disease Control and Prevention, 2020) and broadcast by national news sources, e.g., CNN (CNN, 2020). News agencies have frequently discussed the rising number of COVID-19 cases and that the spread of the disease can be curbed through the implementation of social distancing measures and wearing masks. Such discussions have been prevalent in both the mainstream and research circles alike (Betsch et al., 2020).

A plausible explanation of why local cases negatively impact mask search in Houston can be understood by noting that in the initial stages of the pandemic, the national media focused primarily on the US cases as a whole rather than local cases. Further, there were differences in local attitudes toward the use of masks. In addition, as we can see in Table 3, “Local Cases” is not as significant as “US cases” in predicting mask searches. Also, in the time period under consideration, the magnitude of cases was much higher than the deaths. This could explain why “US Deaths” have a negative impact on mask searches in NY and Omaha.

A key component to these increase/decrease observations is also the timing (lag). In all instances, more recent (i.e., fewer lagged days) variables positively influence searches for masks, while the less recent (more lagged days) variables negatively influence searches. This suggests that individuals in all three cities are aware of the most recent figures surrounding the spread of COVID-19 and are responding by seeking out means of protecting themselves – e.g., through the use of masks.

Table 5
Disinfectant.

City	Variables	Lag (days)	Coefficients	VIF
Houston	US Cases	1	25.1***	1.698
	Loc Deaths	3	−27.08***	1.308
	Loc Cases	5	−27.883**	1.472
New York	US Deaths	5	−16.827**	1.705
	Loc Cases	5	36.481***	1.705
Omaha	NA	NA	NA	NA

4.1.2. *Hand sanitizer*

Table 4 discloses the results of predicting hand sanitizer search by each location that we consider. Here, we observe that there is a mix of variables that contribute to and detract from hand sanitizer searches and that the contribution of these variables is unique to each city. Interestingly, the significant factors in the Houston and Omaha models all contribute to decreases in hand sanitizer searches. This seems a bit counter-intuitive since we might suppose that as local/national cases/deaths increase, searches for hand sanitizer would also increase. However, in examining Fig. 3, we can see that cases do not spike until late in the study period in Omaha, and deaths never spike in either city, which may contribute to individuals acting in a seemingly counter-intuitive manner. In New York, nationwide increases in cases and deaths detract from hand sanitizer searches, while local cases positively influence searches. This finding suggests that New Yorkers are more responsive to local news and reporting surrounding the development of the COVID-19 pandemic than nationwide reporting.

Comparing Tables 3 and 4, we observe that “US Deaths” positively impacts mask search, whereas it negatively impacts hand sanitizer search. This is not surprising given that mask and hand sanitizer search seems to be negatively correlated in the data - see correlation data in Appendix A. Further, in the early stages of the pandemic, the media emphasized masks much more than hand sanitizers which could have played a role.

4.1.3. *Disinfectant*

Table 5 discloses the *disinfectant* results by city. Again, a mix of factors contributes positively and negatively to disinfectant searches. Curiously, no statistically significant variables were found for the city of Omaha. This finding suggests that Omaha residents were not concerned with the purchase of disinfectants in response to the COVID-19 pandemic during the time frame considered in this study. However, we did observe that Omaha residents positively responded by searching for masks and hand sanitizer. Thus, this result is particularly interesting. On the other hand, both Houston and New York models produced statistically significant coefficients. In Houston, both local cases and deaths decreased searches for disinfectants, while US Cases increased searches. This again suggests that nationwide COVID-19 reporting may have a positive impact on preventative measures, such as the use of disinfectants. However, the New York result exhibits the opposite characteristics, with local cases positively contributing to disinfectant searches and US Deaths negatively contributing, again suggesting that individuals here may be more responsive to local COVID-19 coverage and reporting.

Collectively, Houston is most responsive to nationwide COVID-19 cases, with US Cases positively contributing to searches for hand sanitizer, masks, and disinfectants. This may be due to the time period we consider where the spread of the virus was ramping up, but deaths were relatively low. On the other hand, New York is most responsive to local COVID-19 figures, with Local Cases positively contributing to searches for hand sanitizer, masks, and disinfectants. These findings suggest that there are regional differences in the type of information that drives behavior, i.e., nationwide information in the case of Houston and local information in the case of New York.

Table 6
Training and testing RMSE for the Retail mobility predictive models, comparing our method to SFS. Testing period: May 23rd - May 29th (inclusive). Bold indicates a lower RMSE as compared to the other model's respective training/testing result.

	SFS		SFPS-BVIF	
City	**Train**	**Test**	**Train**	**Test**
Houston	3.507	8.15	3.507	8.15
NY	**1.915**	11.136	4.9	**8.794**
Omaha	5.264	9.184	5.264	9.184

Omaha, meanwhile, is a mixed bag with nationwide case-reporting positively contributing to searches for masks only and no information significantly contributing to searches for disinfectant; both local and nationwide information were found to negatively contribute to searches for hand sanitizer. This finding may suggest that individuals in Omaha were responsive to campaigns encouraging the use of a mask.

4.2. *Predicting and analyzing consumer mobility*

In this section, we examine the results of the Google Mobility models, beginning with retail mobility and followed by grocery and pharmacy mobility. Two results are presented with each type of mobility: predictive performance results and analytical results. The predictive performance results show the training and testing performance of each model. As we have mentioned, the test set consists of the last seven days of our data – May 23rd to May 29th (inclusive).

4.2.1. *Retail mobility*

Prior to analyzing the retail mobility models, we first report and discuss each model's predictive performance, also comparing the predictive performance of our method against sequential forward selection (SFS). Table 6 discloses the predictive performance results for the retail mobility models by city, with the results reported in terms of root mean squared error (RMSE).

We first note that the results of our method are comparable to that of SFS, with the NY model even outperforming SFS on the test set. This is particularly encouraging since our method is more "strict" in terms of the included covariates (e.g., allowing more multicollinearity may, at times, produce a more accurate model, but at the expense of interpretability). Furthermore, the training and testing RMSE of all models are fairly reasonable.

Further examining the results of the models produced using our model, we see that the Houston model has the lowest training and testing RMSE, followed by New York and then Omaha. The results indicate that the models are comparable in terms of predictive performance.

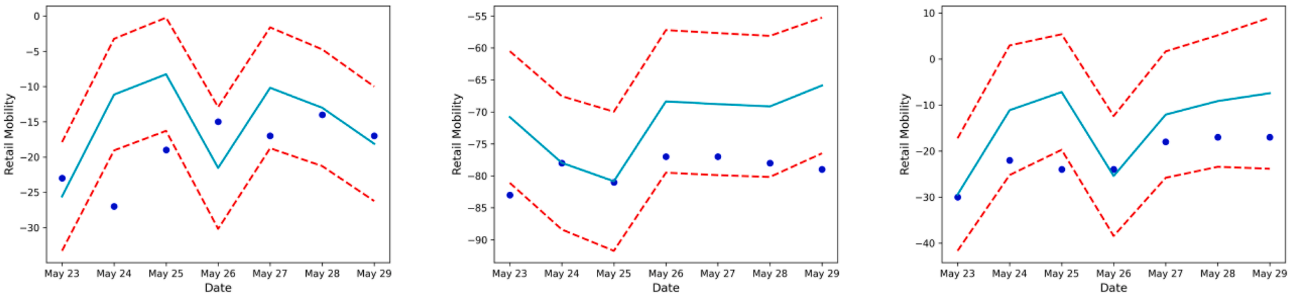
To further assess the quality of the retail mobility models obtained using our method, we create plots of the predicted vs actual retail mobility values by date, also showing the 95 % upper and lower confidence bounds on the predictions. Fig. 6 depicts these results. The results in Fig. 6 show that the majority of the observations (i.e., actual retail mobility values) are within the 95 % predictive confidence bounds, demonstrating that the induced models and subsequently performed model analysis are trustworthy. The Houston and New York models each have two points slightly outside the lower confidence bound, and Omaha has one. Each model also has several predictions (cyan line) that are very nearly spot on, again reinforcing the reliability of the results obtained from these models.

Table 7 discloses the *retail mobility* model results by city. There are a variety of factors that were found to both increase and decrease retail mobility in each city. Some factors are common to all cities, such as residential mobility, which intuitively decreases retail mobility in all three locations. In other words, as individuals are forced or elect to stay home, they tend to also not commute to retail locations. On the other hand, we also observe that many factors are unique to each city, with some factors even influencing retail mobility in opposite ways from one city to the next, such as parks mobility. This finding suggests that each location experiences and views the COVID-19 pandemic uniquely when it comes to engaging in retail behavior.

To be more concrete, we can see that residential mobility, disinfectant searches, and hand sanitizer searches all contribute to decreasing retail mobility in the city of Houston. These findings suggest that individuals living in Houston temper their retail behavior downward following proactive engagement with COVID-19 preventive measures. The disinfectant and hand sanitizer search results suggest that as

Table 7
Retail Mobility.

City	Variables	Lag (days)	Coefficients	VIF
Houston	Transit Mobility	1	24.139***	3.057
	Local Cases	1	6.93*	1.48
	Residential Mobility	2	−28.184***	2.403
	Disinfectant	3	−8.66*	1.213
	Hand Sanitizer	4	−20.445***	3.105
	Parks Mobility	7	13.174***	1.732
	US Deaths	7	5.104*	2.292
	Residential Mobility	1	−7.902**	1.46
NY	Local Deaths	3	2.436*	1.049
	Parks Mobility	3	28.72***	1.86
	Workplace Mobility	7	12.009***	2.16
	Parks Mobility	1	−9.261*	1.682
Omaha	Transit Mobility	1	27.894***	1.839
	Residential Mobility	2	−24.568***	1.421
	Local Cases	3	13.73***	1.275
	Hand Sanitizer	4	−12.4*	2.504
	US Cases	6	9.579*	3.278
	Mask	7	−22.569***	1.941



(a) Houston. (b) New York. (c) Omaha.

Fig. 6. Prediction vs actual retail mobility for the testing period May 23rd - May 29th (inclusive) with 95% prediction confidence bounds. Blue dots indicate observed retail mobility values, the cyan line indicates retail mobility predictions, and the red line indicates upper and lower 95% prediction confidence. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 8
Training and testing RMSE for the grocery and pharmacy mobility predictive models, comparing our method to SFS. Testing period: May 23rd - May 29th (inclusive). Bold indicates a lower RMSE as compared to the other model's respective training/testing result.

	SFS		SFPS-BVIF	
City	Train	Test	Train	Test
Houston	**4.627**	13.026	6.8	**9.426**
NY	**3.705**	**6.184**	6.883	8.4
Omaha	**4.794**	**29.089**	5.921	29.53

Table 9
Grocery and Pharmacy Mobility.

City	Variables	Lag (days)	Coefficients	VIF
Houston	Hand Sanitizer	1	2.465*	2.466
	Local Cases	2	−12.52*	1.593
	C19 Tweets	2	−3.315*	1.14
	Residential Mobility	2	−31.744***	1.351
	Parks Mobility	3	11.299*	1.176
	US Cases	6	5.039*	2.891
NY	Residential Mobility	1	−0.643*	1.555
	Local Cases	1	−6.008*	2.037
	Mask	7	−26.973***	2.173
	Local Deaths	7	−5.185*	1.121
	US Cases	7	14.739**	2.643
	Workplace Mobility	7	39.713***	2.389
Omaha	Local Deaths	1	13.18*	1.144
	Local Cases	3	8.384*	1.107
	Workplace Mobility	3	8.909*	1.992
	Transit Mobility	4	10.651*	2.757
	Residential Mobility	6	−14.326**	2.47
	US Deaths	6	7.675*	1.799
	Mask	7	−18.18**	1.695

individuals protect themselves from exposure to COVID-19 using chemical means, they begin to also implement social protocols by decreasing retail mobility.

Simultaneously, transit mobility, local cases, parks mobility, and US deaths positively contribute to retail mobility in Houston. Transit and park mobility show that if individuals in Houston are already mobile – i.e., visiting parks and generally commuting – they are also inclined to go shopping. The positive association of retail mobility with local cases and US deaths shows that the rise in local and national COVID-19 numbers has not been detrimental to individuals' willingness to engage in commerce.

In the New York model, we observe that residential mobility is the sole factor leading to decreases in retail mobility. While this finding shows that as more New Yorkers stayed at home, a greater decrease in retail mobility is observed. It is also worth noting that New York implemented strict lockdown procedures from March 22nd to May 7th, with a four-phase reopening plan beginning thereafter (New York State, 2020b; Silverstein, 2020; New York State, 2020a; Gold & Stevens, 2020). Therefore, the lack of shopping availability is likely the biggest factor driving a decrease in retail mobility in the state of New York: if businesses simply are not open, individuals cannot engage in commerce in the first place.

In New York, local deaths, park mobility, and workplace mobility were all found to be indicative of increased retail mobility. This suggests that more external engagements, such as working or going to the park, also lead to more retail behavior for individuals in New York. The park mobility finding for New York is opposite that of Omaha, suggesting that individuals in Omaha visit parks as an alternative to shopping. These orthogonal findings again suggest that retail behavior is driven and affected by different factors in different locations.

In addition to park mobility, residential mobility, hand sanitizer searches, and mask searches also indicate lower retail mobility. Searches for these items as indicators of decreased retail mobility suggest that individuals in Omaha are also practicing social distancing as a means of curbing the spread of the virus after seeking out protective measures, such as masks and hand sanitizer, even though fewer cases and deaths were observed during this time period (in Omaha).

Factors that contribute to increased retail mobility in Omaha include transit mobility, local cases, and US cases. While transit mobility is not surprising – after all, if folks are using transportation, they are likely using.

such means to engage in some level of commercial activity – local and US cases are. Therefore, we speculate that individuals in Omaha are responsive to discussions of preventive measures, such as using masks and hand sanitizer, but are not responsive to COVID-19 case increases either locally or nationally.

4.2.2. *Grocery and pharmacy mobility*

We now turn to examine the grocery and pharmacy mobility models. Prior to analyzing these models, we first quantify their predictive capabilities as compared to SFS, the results of which are in Table 8. As with retail mobility, we set aside the last seven days in our study period – May 23rd - May 29th – as a hold-out test set and assessed the models using RMSE.

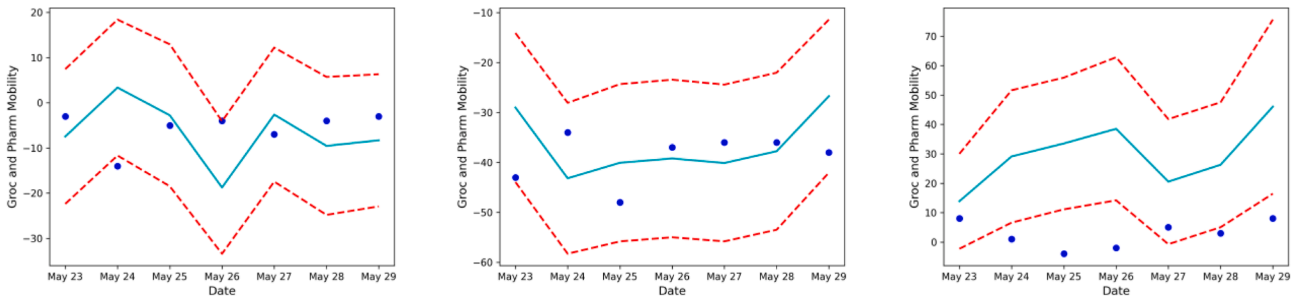
First, we see that the SFS models perform better than the models obtained using our method for NY and Omaha – our method performs better on the test set for Houston. We do, however, note that the models are all still fairly comparable. Again, this highlights the trade-off between model interpretability and, at times, predictive performance; the NY and Omaha SFS models are slightly more accurate but contain multicollinearity (as shown by Table 2) and, therefore, are not reliably interpretable.

Examining the results of our models, the training and testing RMSE for Houston and New York are very comparable to one another and to their retail mobility counterparts. On the other hand, the Omaha model has the lowest training RMSE but substantially higher testing (out-of-sample) RMSE (this is also the case with the SFS-obtained model). Since Omaha did not implement any type of mobility inhibiting protocols, such as lockdowns or non-essential worker stay-at-home orders, and grocery and pharmacy mobility serve to meet basic needs (i.e., food and medicine), this type of mobility may not have been drastically affected by any of our defined predictors.

To further assess the ability of our models to make out-of-sample predictions, we create visualizations for each, as we did with retail mobility, showing predicted vs actual grocery and pharmacy mobility values, along with 95 % prediction confidence bounds. These results are shown in Fig. 7. All but one grocery and pharmacy mobility value falls within the 95 % confidence bounds for the Houston model, and all values fall within the bounds for New York. All but two values fall outside the confidence bounds for the Omaha model, which is expected provided the testing (out-of-sample) RMSE. As we mentioned, Omaha did not implement any type of mobility inhibiting protocols, such as lockdowns or non-essential worker stay-at-home orders, and grocery and pharmacy mobility serve to meet basic needs (i.e., food and medicine). Hence, this type of mobility may not have been drastically affected and is, therefore, harder to predict using our defined covariates. However, it should be noted that the training (in-sample) RMSE is very reasonable, so interpretation of the results for the period before May 23rd should not be a problem for Omaha.

Table 9 discloses the *grocery and pharmacy mobility* model results.[4] As

[4] As an additional robustness check, instead of using local and U.S. cases and deaths as distinct, independent variables, we use the local and national death rates (where rate = deaths/cases) which captures the severity of the risk. We present the empirical estimates of this augmented model in Appendix C and observe that the pattern of results in the additional analysis is consistent with our main model estimates (Table 7 and Table 9). We thank an anonymous reviewer for suggesting this robustness check.



(a) Houston. (b) New York. (c) Omaha.

Fig. 7. Estimates vs actual grocery and pharmacy mobility for the testing period May 23rd - May 29th (inclusive) with 95% prediction confidence bounds. Blue dots indicate observed grocery and pharmacy mobility values, the cyan line indicates retail mobility predictions, and the red line indicates upper and lower 95% prediction confidence. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

with retail mobility, a variety of both unique and overlapping factors contribute to grocery and pharmacy mobility in each city. As a general observation, fewer “more” significant factors were found to be indicative of this type of mobility. This finding is not entirely unsurprising since food and medicine constitute basic needs and, pandemic or not, people still need these basic goods to survive.

Nevertheless, our considered variables provide clear indications of grocery and pharmacy mobility drivers. In Houston, searches for hand sanitizer, parks mobility, and the US cases all positively contributed to grocery and pharmacy mobility. Searches for hand sanitizer the preceding day are unsurprising since a grocery store is where one would procure such items. Parks mobility also makes sense since individuals who are willing to be mobile are also likely willing to visit the grocery store, as opposed to having groceries delivered or picked up using a pickup service, neighbors, etc. Curiously, increases in US COVID-19 cases also increase grocery and pharmacy mobility. This observation is tempered by the fact that, as local COVID-19 cases increase, grocery and pharmacy mobility decreases. It is worth pointing out that US Cases six days preceding increase mobility, and local cases one-day prior decrease grocery mobility. Therefore, taken in conjunction with one another, we may suppose that individuals in Houston are aware of the local figures surrounding COVID-19 and are responding accordingly. The simultaneity of US cases may actually suggest that Houstonians delayed going to the grocery store when cases were high a week earlier in the hopes that later, the spread would go down; this is an encouraging finding.

Other factors that lead to decreases in mobility in Houston include COVID-19 tweets and residential mobility. The COVID-19 tweets finding suggests that individuals in Houston pay attention to social media and/or conscientiously respond to the pandemic by expressing themselves on social media. Interestingly, all factors that produce decreases in grocery and pharmacy mobility occur during the preceding two days, suggesting that these effects are relatively immediate.

In New York, only two factors were found to increase grocery and pharmacy mobility, both of which were observed seven days prior: US cases and workplace mobility. Workplace mobility is unsurprising since one might imagine that individuals who are already mobile for work might stop at the grocery on the way home, for instance. It may seem surprising, however, that the optimal lag is seven days prior, although when we consider the frequency with which individuals visit the grocery store – typically weekly (Yoo et al., 2006) – the finding is once again meaningful.

As with Houston, increases in COVID-19 cases lead to increases in grocery and pharmacy mobility-seven days prior to increases in grocery and pharmacy mobility, while local cases one day prior lead to decreases in mobility, likely for the same reasons. Other factors contributing to decreases in grocery and pharmacy mobility in New York include residential mobility, mask searches, and local deaths, all of which make sense in the same contexts as previously discussed for other locations.

In Omaha, residential mobility and mask searches contribute to decreases in grocery and pharmacy mobility. In contrast, local deaths, local cases, workplace mobility, transit mobility, and US deaths all lead to increases. Interestingly, local deaths and cases and US deaths all lead to increases in grocery and pharmacy mobility at varying times (one, three, and six days, respectively). As we have seen, COVID-19 was not widespread in Omaha until later in our study period. These findings may suggest that individuals were aware of this initially but continued “business as usual” even as the virus began spreading throughout the community.

5. Managerial implications

The pandemic has substantially impacted the retail industry, especially during the early phase when governments strictly implemented social distancing policies. These social distancing policies have caused significant challenges in global supply chains leading to widespread business disruptions. Therefore, developing accurate forecasts has become more vital than ever, especially since retailers require such predictions to be more robust to economic disruptions caused by public health or other types of disasters. For example, natural disasters can challenge retail inventory management strategies due to a sudden shift in demand for certain products in different geographies, and real-time forecasts can be used effectively to allocate these critical items among stores (Morrice et al., 2016).

While disasters such as hurricanes can have a more predictable impact on demand for certain products, as presented in Morrice et al. (2016), businesses need to find more reliable data sources that can support forecasting efforts to mitigate the impact of disruptions on supply chains (Sharma et al., 2020). For example, Google Trends data can provide near-real-time information about consumer trends and perceived risks in different geographies, which have also been used for forecasting geo-spatial demand on certain products, e.g., see Nikolopoulos et al. (2020); Fritzsch et al. (2020); Boone et al. (2018). New data sources, such as Google Mobility, can also improve forecasting in the retail industry, as they also provide timely retail mobility data. However, there is a need to develop effective methodologies to accurately incorporate optimal lags in predictive variables, improving forecast accuracy in the retail industry.

Over the past several decades, complex forecasting methods have been developed to improve sales forecasts in the retail industry. As Ma & Fildes (2021) have stated, retail forecasting has focused on sales forecasting. However, no method dominates for all types of products and time periods. Therefore, the performance of the predictions is hugely dependent on the data available, products under study, and the geographic trends in which the demand emerges. In addition to testing various techniques to find the most accurate forecasting method specific to the product and the business, newly emerging data sources can definitely provide value in studying the retail industry’s demand and sales phenomenon. Brea et al. (2020) underscore the significance of

Table A1
Correlation matrix of Houston covariates.

	Mask	Hand	Disinf	L C	L D	US C	US DC19 Tweets	
Mask	1	–	–	–	–	–	–	–
Hand Sanitizer	−0.17	1	–	–	–	–	–	–
Disinfectant	0.4	0.16	1	–	–	–	–	–
Loc Cases	0.28	−0.53	−0.06	1	–	–	–	–
Loc Deaths	0.22	−0.51	−0.04	0.34	1	–	–	–
US Cases	0.62	−0.73	0.2	0.64	0.6	1	–	–
US Deaths	0.48	−0.64	0.06	0.59	0.63	0.86	1	–
C19 Tweets	−0.09	−0.26	−0.02	0.12	−0.1	0.09	0.1	1

these new data sources during uncertain times, especially when a pandemic hits. Incorporating these data (e.g., mobility and, more specifically, retail mobility data) requires a systematic approach to analyzing the timing of information and the optimal lags with the outcomes of interest. As demonstrated in this study, machine learning and forecasting techniques used in the retail industry can address this challenge.

Recent research (e.g., Sharma et al., 2020) has also urged the need to leverage technology and utilize more reliable data sources to improve forecasting efforts to mitigate the impact of disruptions on supply chains and business operations. Our paper highlights how aggregate-level, real-time data sources such as Google Trends search and mobility data could be employed to predict retail mobility. Our method is scalable to include more risk perception indicators to predict retail mobility, which could be a starting point toward developing actionable retailing strategies. Although we do not focus on the impact of retail mobility on specific retail strategies, we believe that the outcomes from our predictive model could be used as strategic inputs to improve several retail decisions, such as staffing, inventory, and in-store advertising decisions.

6. Conclusions

In this work, we examine the factors driving retail, grocery, and pharmacy mobility through the lens of the COVID-19 pandemic of 2020. We also analyze pandemic-associated online search behavior, such as online searches for hand sanitizer, masks, and disinfectants, and subsequently utilize such factors to also examine the aforementioned mobility phenomenon. Our analysis was conducted on data from three geographically dispersed locations in the United States: Houston (TX), New York City (NY), and Omaha (NE). To conduct our analyses, we built predictive models elicited from a novel steepest ascent, steepest descent hill-climbing algorithm that produced models containing only statistically significant coefficients with minimal multicollinearity. Our findings suggest that there are a variety of unique factors that contribute to and drive consumer behavior in each location, as well as several factors that are common to all locations. Furthermore, we find that different types of consumer engagements – i.e., retail vs grocery and pharmacy consumers – respond to different factors, which makes sense: grocery and pharmacy visits address basic human needs, while other types of retail engagements are more leisurely.

Results suggest that retail mobility can be predicted by risk-indicating search terms in Houston, while transit mobility, local cases, parks mobility, and US deaths also positively contribute to retail mobility. In New York City, local deaths, park mobility, and workplace mobility were all found to be indicative of increased retail mobility: i.e., more external engagements led to more retail behavior for individuals during the early phase of the pandemic. In Omaha, in addition to park mobility, residential mobility, hand sanitizer searches, and mask searches were also significant predictors of retail mobility. Therefore, given our approach to addressing multicollinearity and lags among the variables, real-time prediction of retail activity can be achieved with some level of accuracy using the search engine trends data in conjunction with the proper terms.

This study focused on the initial phase of the pandemic, during which a great deal of uncertainty regarding the severity and the transmissibility of the virus strain existed. In addition, there was a significant level of variation in local public health policies following the initial nationwide lockdowns. Since we focused on various factors contributing to the perceived risk in different locations and predicting retail mobility, our study can be extended with more recent data. Furthermore, the approach presented in this paper using the developed algorithms can be applied to more recently available data for understanding individuals' mobility patterns during different phases of the pandemic. Lastly, the methodology can also be used in predicting retail mobility and other mobility activities, using real-time data sources, during other disruptive events that may shake the global economy.

CRediT authorship contribution statement

Michael T. Lash: Data curation, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **S. Sajeesh:** Conceptualization, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Ozgur M. Araz:** Conceptualization, Data curation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Table A2
Correlation matrix of NY covariates.

	Mask	Hand	Disinf	L C	L D	US C	US DC19 Tweets	
Mask	1	–	–	–	–	–	–	–
Hand Sanitizer	−0.46	1	–	–	–	–	–	–
Disinfectant	0.33	−0.09	1	–	–	–	–	–
Loc Cases	0.64	−0.47	0.25	1	–	–	–	–
Loc Deaths	0.14	−0.22	0.09	0.54	1	–	–	–
US Cases	0.66	−0.83	0.32	0.66	0.29	1	–	–
US Deaths	0.46	−0.68	0.21	0.65	0.4	0.86	1	–
C19 Tweets	−0.06	−0.34	−0.04	0.04	0.14	0.17	0.21	1

Table A3
Correlation matrix of Omaha covariates.

	Mask	Hand	Disinf	L C	L D	US C	US DC19 Tweets	
Mask	1	–	–	–	–	–	–	–
Hand Sanitizer	−0.33	1	–	–	–	–	–	–
Disinfectant	0.1	0.04	1	–	–	–	–	–
Loc Cases	−0.02	−0.38	−0.19	1	–	–	–	–
Loc Deaths	0.02	−0.22	0.05	0.19	1	–	–	–
US Cases	0.67	−0.69	0.18	0.12	0.21	1	–	–
US Deaths	0.49	−0.6	0.08	0.02	0.21	0.86	1	–
C19 Tweets	−0.07	−0.1	0.05	−0.02	−0.09	0	0.03	1

Appendix A

The correlation matrices of non-lagged covariates by location are presented in Tables A1 (Houston), A2 (NY), and A3 (Omaha) below.

Appendix B

To discuss our SFPS-BVE algorithm, we begin by providing and discussing the SFPS component, followed by the SBVE component. We then provide the full SFPS-BVE algorithm. Algorithm B.1 discloses our SFPS procedure:

```
Algorithm B.1 Sequential Forward p-value Selection $SFPS(X, y, \mathscr{F})$
Require:$X, y, \mathscr{F}$
 1: Initialize $X^{(0)} \leftarrow 1, \mathscr{F}^{(0)} \leftarrow [Const], i \leftarrow 0, term \leftarrow False$
 2: while $term = False$ do
 3:   for $j \leftarrow 1, \cdots, |\mathscr{F}|$ do
 4:     $\boldsymbol{\beta}^{(j)}, \beta^{(j)} \leftarrow OLS(X^{(i)}$â§°$X_j, y)$
 5:   end for
 6:     $j^* \leftarrow \text{argmin}_j \left\{ \rho\left(\beta^{(j)}\right) | \rho\left(\boldsymbol{\beta}^{(j)}\right) \le \alpha : j = 1, \cdots, |\mathscr{F}| \right\}$
 7:   if $j^* \neq NULL$ then
 8:     $X^{(i+1)} \leftarrow X^{(i+1)}$â§°$X_{j^*}$
 9:     $\mathscr{F}^{(i+1)} \leftarrow \mathscr{F}^{(i)}$â§°$\mathscr{F}_j$
10:      $X \leftarrow X - group(j^*)$
11:      $\mathscr{F} \leftarrow \mathscr{F} - group(j^*)$
12:   else
13:      $term \leftarrow True$
14:   end if
15:      $i \leftarrow i + 1$
16: end while
17:$X^* \leftarrow X^{(i)}, \mathscr{F}^* \leftarrow \mathscr{F}^{(i)}$
Ensure:$X^*, \mathscr{F}^*$
```

The Sequential Forward p-value Selection (SFPS) algorithm begins by initializing several variables before executing the optimization procedure (Line 1). These variables include an initial design matrix $\mathbf{X}^{(0)}$, containing only a constant for the OLS offset term, a vector $\mathrm{F}^{(0)}$ for recording added feature names, an iteration counter *i*, and a Boolean variable *term* indicating whether the termination criteria has been satisfied. Next, the algorithm begins iterating until the termination criteria is satisfied (Line 2). During each iteration, OLS models are constructed on each remaining variable independently ($\mathbf{X}j$), along with any variables that have been added ($\mathbf{X}^{(i)}$) (Lines 3–5). Therefore, during the first iteration ($i = 0$), there are *p* models initially constructed, where each model is constructed on a single variable and an offset term only. Note that ++ indicates concatenation. In Line 4, we store the full model $\beta^{(j)}$ and separately replicate and store the β coefficient value corresponding to the particular variable selected at iteration *j* as $\beta^{(j)}$ for convenience purposes.

In Line 6, we find the index *j** of the variable that produced a model with the lowest p-value such that all p-values in the model are statistically significant at the α level or better. If the *j** returned from Line 6 is not null (i.e., the Line 6 conditions are met) (Line 7), then Lines 8–10 are executed. In Line 8, the $\mathbf{X}j$ vector is added to $\mathbf{X}^{(i+1)}$ and then deleted from **X**, along with the other variables belonging to the same variable group (i.e., the lagged variables with the same "meaning"), in Line 10. Likewise, in Line 9, the name of the selected variable F*j* is added to $\mathrm{F}^{(i+1)}$ and removed from F, along with the names of the variables belonging to the same lag group, on Line 11.

On the other hand, if the *j** returned from Line 6 is null (Line 12), then we set the *term* to True, and the optimization procedure terminates. Once this occurs, $\mathbf{X}^{(i)}$ and $\mathrm{F}^{(i)}$ become the optimized design matrices and feature set (Line 17) and are returned by the procedure.

Following the application of the SFPS procedure, the Sequential Backward VIF Elimination (SBVE) procedure is applied to ensure that minimal multicollinearity remains among the selected covariates of the SFPS procedure. This procedure is provided by Algorithm B.2:

Table C1
Retail Mobility results.

City	Variables	Lag (days)	Coefficients	VIF
Houston	Transit Mobility	1	21.157***	3.05
	Residential Mobility	2	−27.283***	2.34
	Disinfectant	3	−9.454*	1.2
	Hand Sanitizer	4	−19.964***	3.28
	Parks Mobility	7	11.177***	1.78
	US Rate	7	7.392**	2.02
NYC	Residential Mobility	1	−7.805**	1.45
	Loc Rate	3	1.162*	1.13
	Parks Mobility	3	28.387***	1.84
	Workplace Mobility	7	12.287***	2.33
Omaha	Workplace Mobility	1	10.155**	1.55
	Residential Mobility	2	−18.042***	1.79
	Disinfectant	3	−10.112*	1.05
	Hand Sanitizer	6	−13.244**	1.8
	Mask	7	−26.808***	1.43
	US Rate	7	20.126***	1.71
	Parks Mobility	7	7.15*	1.28

Table C2
Grocery and Pharmacy Mobility results.

City	Variables	Lag (days)	Coefficients	VIF
Houston	Hand Sanitizer	1	−3.923*	2.611
	Residential Mobility	2	−18.656***	1.663
	C19 Tweets	2	−6.954**	1.142
	Parks Mobility	3	−7.123*	1.91
	Workplace Mobility	6	34.376***	3.694
	US Rate	7	13.23***	1.659
NYC	Local Rate	2	−3.708*	1.234
	C19 Tweets	3	−6.525*	1.07
	US Rate	3	16.645***	1.637
	Residential Mobility	4	−18.775***	1.693
	Workplace Mobility	6	34.925***	1.637
	Mask	7	−13.345**	1.871
Omaha	US Rate	1	16.122***	1.79
	Local Rate	1	17.207***	1.124
	Workplace Mobility	3	13.367***	2.288
	Transit Mobility	4	12.99*	2.619
	Residential Mobility	6	−9.303*	2.188
	Parks Mobility	6	7.285*	1.165
	Mask	7	−22.631***	1.64

Algorithm B.2 Sequential Backward VIF Elimination $\boldsymbol{SBVE}(\boldsymbol{X}, \mathscr{F})$

```
Require: $X, \mathscr{F}$
1: Initialize $X^{(0)} \leftarrow 1, \mathscr{F}^{(0)} \leftarrow \mathscr{F}, i \leftarrow 0, term \leftarrow False$
2: while $term = False$ do
3:    for $j \leftarrow 1, \cdots, |\mathscr{F}^{(i)}|$ do
4:        $VIF_j \leftarrow VIF\left(X_j^{(i)}\right)$
5:    end for
6:    $j^* \leftarrow \text{argmin}_j \{VIF_j | VIF_j > \gamma : j = 1, \cdots, |\mathscr{F}^{(i)}|\}$
7:    if $j^* \neq NULL$ then
8:        $X^{(i+1)} \leftarrow X^{(i)} - X_{j^*}^{(i)}$
9:        $\mathscr{F}^{(i+1)} \leftarrow \mathscr{F}^{(i)} - \mathscr{F}_j^{(i)}$
10:   else
11:       $term \leftarrow True$
12:   end if
13:       $i \leftarrow i + 1$
14: end while
15: $X^* \leftarrow X^{(i)}, \mathscr{F}^* \leftarrow \mathscr{F}^{(i)}$
Ensure: $X^*, \mathscr{F}^*$
```

SBVE begins by initializing several variables (Line 1): $\mathbf{X}^{(0)}$ is initialized to $\mathbf{X}$,$\mathrm{F}^{(0)}$ is initialized to F, *i* is an iteration counter initialized to 0, and the Boolean variable *term* indicates whether the termination criteria have been satisfied, is initialized to *False*. Iteration over the optimization procedure begins on Line 2. At each iteration, the VIF is calculated for each variable (Lines 3–5). Then, the index *j** corresponding to the variable with the largest VIF, as long as the VIF is larger than γ, is selected (Line 6). If *j** is not null (i.e., a variable with a VIF larger than γ was found) (Line 7), the *j*th variable (Line 8) and feature (Line 9) are removed. Otherwise, the algorithm's termination criterion is satisfied, and the *term* is set equal to true (Line 11). Once the termination criterion is satisfied, the current design matrix and vector of feature names are considered optimized (Line 15) and are returned by the algorithm.

After applying both the SFPS and SBVE procedures, a new OLS model is induced on the "optimized" design matrix to create the final model for analysis. For the sake of clarity, we provide the full SFPS-BVE procedure in Algorithm B.3:

Application of our proposed SFPS-BVE procedure produces an OLS model consisting of statistically significant covariates with minimal multicollinearity, thus allowing us to reliably interpret the sign and magnitude of the coefficients as indicators of the dependent phenomena of interest.

Algorithm B.3 Sequential Forward p-value Selection-Backward VIF Elimination $\boldsymbol{SFPS-BVE}(\boldsymbol{X}, \boldsymbol{y}, \mathscr{F})$

```
Require:$\boldsymbol{X}, \boldsymbol{y}, \mathscr{F}$
  1:$\boldsymbol{X}', \mathscr{F}' \leftarrow SFPS(\boldsymbol{X}, \boldsymbol{y}, \mathscr{F})$
  2:$\boldsymbol{X}^*, \mathscr{F}^* \leftarrow SBVE(\boldsymbol{X}', \mathscr{F}')$
  3:$\boldsymbol{\beta}^* \leftarrow OLS(\boldsymbol{X}^*, \boldsymbol{y})$
Ensure:$\boldsymbol{X}^*, \mathscr{F}^*, \boldsymbol{\beta}^*$
```

Appendix C

The additional analysis results using local and national *death rates* for retail mobility (Table C1) and grocery and pharmacy mobility (Table C2) are presented below.

References

Ahmad, Imama, Flanagan, Ryan, & Staller, Kyle. 2020. Increased internet search interest for GI symptoms may predict COVID-19 cases in US hospitals. *Clinical Gastroenterology and Hepatology*, **18**(12), 2833–2834.E3.

Alsukni, E., Arabeyyat, O. S., Awadallah, M. A., Alsamarraie, L., Abu-Doush, I., & Al-Betar, M. A. (2019). Multiple-reservoir scheduling using β-hill climbing algorithm. *Journal of Intelligent Systems, 28*(4), 559–570.

Araz, OM, Bentley, D, & Muelleman, RL. 2014. Using Google Flu Trends data in forecasting influenza-like– illness related ED visits in Omaha, Nebraska. *American Journal of Emergency Medicine*, **32**(9), 1016–1023.

Asseo, Kim, Fierro, Fabrizio, Slavutsky, Yuli, Frasnelli, Johannes, & Niv, Masha Y. 2020. Utility and limi- tations of Google searches for tracking disease: the case of taste and smell loss as markers for COVID-19. *medRxiv*.

Barrett, M., & Orlikowski, W. (2021). Scale matters: Doing practice-based studies of contemporary digital phenomena. *MIS Quarterly, 45*(1), 467–472.

Basu, R., & Ferreira, J. (2021). Sustainable mobility in auto-dominated Metro Boston: Challenges and opportunities post-COVID-19. *Transport Policy, 103*, 197–210.

Bertsimas, D., Iancu, D. A., & Katz, D. (2013). A new local search algorithm for binary optimiza- tion. *INFORMS Journal on Computing, 25*(2), 208–221.

Betsch, C., Korn, L., Sprengholz, P., Felgendreff, L., Eitze, S., Schmid, P., & Böhm, R. (2020). Social and behavioral consequences of mask policies during the COVID-19 pandemic. *Pro- ceedings of the National Academy of Sciences, 117*(36), 21851–21853.

Boone, T., Ganeshan, R., Hicks, R.L., & N.R., Sanders. 2018. Can Google Trends Improve Your Sales Forecast? *Production and Operations Management*, **27**(10), 1770–1774.

Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The role of big data and predictive analytics in retailing. *Journal of Retailing, 93*(1), 79–95.

Brea, C., Bicanic, S., Li, Y. N., & Bhardwaj, S. (2020). Predicting consumer demand in an unpredictable world. *Harvard Business Review*, 1–7.

Brewer, N. T., Chapman, G. B., Gibbons, F. X., Gerrard, M., McCaul, K. D., & Weinstein, N. D. (2007). Meta-analysis of the relationship between risk perception and health behavior: The example of vaccination. *Health psychology, 26*(2), 136.

Caramia, M., Dell'Olmo, P., & Italiano, G. F. (2008). Novel local-search-based approaches to university examination timetabling. *INFORMS Journal on Computing, 20*(1), 86–99.

CDC 2016. The Continuum of Pandemic Phases. https://www.cdc.gov/flu/pandemic-resources/planning-preparedness/global-planning-508.html. [Accessed: July 14, 2022].

CDC, Centers for Disease Control *and* Prevention. 2020a. *Coronavirus disease 2019 (COVID-19) how COVID- 19 spreads*. https://www.cdc.gov/coronavirus/2019-ncov/about/ transmission.html. Date retrieved: March 13, 2020.

CDC, Centers for Disease Control *and* Prevention. 2020b. *Coronavirus disease 2019 (COVID-19) situation summary*. https://www.cdc.gov/cor onavirus/2019-nCoV/summary.html. Date retrieved: March 13, 2020.

Centers for Disease Control and Prevention. 2020. *Stop the Spread*. https://www.cdc.gov/coronavirus/ 2019- ncov/communication/stop-the-spread.html. Accessed: 2020-12-29.

Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking Social Media Discourse About the COVID- 19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health Surveill, 6*(2), e19273.

Chennamaneni, P. R., Echambadi, R., Hess, J. D., & Syam, N. (2016). Diagnosing harmful collinearity in moderated regressions: A roadmap. *International Journal of Research in Marketing, 33*(1), 172–182.

Chernozhukov, V., Kasahara, H., & Schrimpf, P. (2021). Causal impact of masks, policies, behavior on early covid-19 pandemic in the US. *Journal of econometrics, 220*(1), 23–62.

Clemons, E. K. (2008). How information changes consumer behavior and how consumer behavior determines corporate strategy. *Journal of Management Information Systems, 25*(2), 13–40.

CNN. 2020. *Three simple acts can stop Covid-19 outbreaks, study finds*. https://www.cnn.com/2020/07/21/health/covid-19-three-things-will-stop-it-wellness/index.html. Ac- cessed: 2020-12-29.

Cotter, S. F., Rao, B. D., Kreutz-Delgado, K., & Adler, J. (1999). Forward sequential algorithms for best basis selection. *IEE Proceedings-Vision, Image and Signal Processing, 146*(5), 235–244.

Donthu, N., & Gustafsson, A. (2020). Effects of COVID-19 on business and research. *Journal of Business Research, 117*, 284–289.

Du, R. Y., Hu, Y.e., & Sina, D. (2015). Leveraging trends in online searches for product features in market response modeling. *Journal of Marketing, 79*(1), 29–43.

Edelman, D., & Heller, J. (2014). Marketing disruption: Five blind spots on the road to marketing's potential. https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/marketing-disruption-five-blind-spots-on-the-road-to-marketing039s-potential [Accessed: January 13, 2021].

El-Toukhy, S. (2015). Parsing susceptibility and severity dimensions of health risk perceptions. *Journal of health communication, 20*(5), 499–511.

France, S. L., Shi, Y., & Kazandjian, B. (2021). Web Trends: A valuable tool for business research. *Journal of Business Research, 132*, 666–679.

Fritzsch, B., Kai, W., Philipp, S., & Ullmann, G. (2020). Can google trends improve sales forecasts on a product level? *Applied Economics Letters, 27*(17), 1409–1414.

Garaus, M., & Garaus, C. (2021). The Impact of the Covid-19 Pandemic on Consumers' Intention to Use Shared-Mobility Services in German Cities. *Frontiers in Psychology, 12*, 367.

Gold, M. G., & Stevens, M. (2020). What Restrictions on Reopening Remain in New York? *New York Times*.

Google Mobility. 2021. *Community Mobility Reports*. https://www.google.com/covid19/mobility/?hl=en. [Last accessed 02-January-2021].

Google Trends. 2021. Exploring What the World is Searching. https://trends.google.com/trends/?geo=US [Accessed: January 26, 2021].

Grechanovsky, E., & Pinsker, I. (1995). Conditional p-values for the F-statistic in a forward selection procedure. *Computational statistics & data analysis, 20*(3), 239–263.

Hair Jr, Joseph F, Hult, G Tomas M, Ringle, Christian, & Sarstedt, Marko. 2016. *A primer on partial least squares structural equation modeling (PLS-SEM)*. Sage publications.

Hastie, T., Tibshirani, R., Tibshirani, R., et al. (2020). Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. *Statistical Science, 35*(4), 579–592.

Henningsson, S., Kettinger, W. J., Zhang, C., & Vaidyanathan, N. (2021). Transformative rare events: Leveraging digital affordance actualization. *European Journal of Information Systems, 30*(2), 137–156.

Hu, Y., Xu, A., Hong, Y., Gal, D., Sinha, V., & Akkiraju, R. (2019). Generating business intelligence through social media analytics: Measuring brand personality with consumer-, employee-, and firm-generated content. *Journal of Management Information Systems, 36*(3), 893–930.

Kandula, S., Pei, S., & Shaman, J. (2019). Improved forecasts of influenza-associated hospital- ization rates with Google search trends. *J. R. Soc. Interface, 16*(20190080), 1–11.

Kano, Y., & Harada, A. (2000). Stepwise variable selection in factor analysis. *Psychometrika, 65*(1), 7–22.

Karabacak, M., Fernández-Ramírez, L. M., Kamal, T., & Kamal, S. (2019). A new hill climbing maximum power tracking control for wind turbines with inertial effect compensation. *IEEE Transactions on Industrial Electronics, 66*(11), 8545–8556.

Keane, M., & Neal, T. (2021). Consumer panic in the COVID-19 pandemic. *Journal of econometrics*.

Kumar, A., Mehra, A., & Kumar, S. (2019). Why do stores drive online sales? Evidence of underlying mechanisms from a multichannel retailer. *Information Systems Research, 30*(1), 319–338.

Laato, S, AKMN, Islam, Farooq, A, & A, Dhir. 2020. Unusual purchasing behavior during the early stages of the COVID-19 pandemic. *Journal of Retailing and Consumer Services*, **57**, 102224.

Lash, M. T., Slater, J., Polgreen, P. M., & Segre, A. M. (2017). A large-scale exploration of factors affecting hand hygiene compliance using linear predictive models. In *Of: 2017 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 66–73). IEEE.

Lash, M. T., Slater, J., Polgreen, P. M., & Segre, A. M. (2019). 21 Million Opportunities: A 19 Facility Investigation of Factors Affecting Hand-Hygiene Compliance via Linear Predictive Models. *Journal of Healthcare Informatics Research, 3*(4), 393–413.

Lehmann, Christine. 2020. Many Metrics to Measure COVID-19, Which Are Best? https://www.webmd.com/lung/news/20200922/many-metrics-to-measure-covid-19-which-are-best [Ac- cessed: January 13, 2021].

Ma, S., & Fildes, R. (2021). Retail sales forecasting with meta-learning. *European Journal of Operational Research, 288*(1), 111–128.

Mao, K. Z. (2004). Orthogonal forward selection and backward elimination algorithms for feature subset selection. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 34*(1), 629–634.

Marcano-Cedeño, A., Quintanilla-Domínguez, J., Cortina-Januchs, M. G., & Andina, D. (2010). Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network. In *Of: IECON 2010–36th annual conference on IEEE industrial electronics society* (pp. 2845–2850). IEEE.

Maser, B., & Weiermair, K. (1998). Travel decision-making: From the vantage point of perceived risk and information preferences. *Journal of Travel & Tourism Marketing, 7* (4), 107–121.

Menon, G., Raghubir, P., & Agrawal, N. (2008). *Health risk perceptions and consumer psychology [in:] Handbook of consumer psychology*.

Meyer, P., Marbach, D., Roy, S., & Kellis, M. (2010). Information-Theoretic Inference of Gene Networks Using Backward Elimination. *Pages 700–705 of: BioComp*.

Morrice, D. J., Cronin, P., Tanrisever, F., & Butler, J. C. (2016). Supporting hurricane inventory management decisions with consumer demand estimates. *Journal of Operations Management, 45*, 86–100.

New York State. 2020a. *Amid Ongoing COVID-19 Pandemic, Governor Cuomo Announces' NYS on PAUSE' Extended until May 15*. https://www.governor.ny.gov/news/amid-ongoing-covid-19-pandemic-governor- cuomo-announces-nys-pause-extended-until-may-15. Accessed: 2020-12-29.

New York State. 2020b. *New York State on Pause*. https://coronavirus.health.ny.gov/new-york-state-pause. Accessed: 2020-12-29.

Nguyen, H. B., Xue, B., Liu, I., & Zhang, M. (2014). Filter based backward elimination in wrapper based PSO for feature selection in classification. *Pages 3111–3118 of: 2014 IEEE Congress on Evolutionary Computation (CEC)*. IEEE.

Nikolopoulos, K., Punia, S., Schafers, A., Tsinopoulos, C., & Vasilakis, C. (2020). Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain distruptions and governmental decisions. *European Journal of Operational Research*.

OECD. 2020. *COVID-19 and the retail sector: impact and policy responses*. https://www.oecd.org/coronavirus. [Last accessed 02-January-2021].

Oehmen, J., Locatelli, G., Wied, M., & Willumsen, P. (2020). Risk, uncertainty, ignorance, and myopia: Their managerial implications for B2B firms. *Industrial Marketing Management, 88*, 330–338.

Pantano, E., Pizzi, G., Scarpi, D., & Dennis, C. (2020). Competing during a pandemic? Retailers' ups and downs during the COVID-19 outbreak. *Journal of Business Research, 116*, 209–213.

Peduzzi, P. N., Hardy, R. J., & Holford, T. R. (1980). A stepwise variable selection procedure for nonlinear regression models. *Biometrics*, 511–516.

Persson, J., Parie, J. F., & Feuerriegel, S. (2021). Monitoring the COVID-19 epidemic with nationwide telecommunication data. *Proceedings of the National Academy of Sciences of the United States of America (forthcoming)*.

Rivera, R. (2016). A dynamic linear model to forecast hotel registrations in Puerto Rico using Google Trends data. *Tourism Management, 57*, 12–20.

Roggeveen, A. L., & Sethuraman, R. (2020). How the COVID-19 Pandemic May Change the World of Retailing. *Journal of Retailing, 96*(2), 169–171.

Shankar, V., Kalyanam, K., Setia, P., Golmohammadi, A., Tirunillai, S., Douglass, T., Hennessey, J., Bull, J.S., & Waddoups, R. (2020). How Technology is Changing Retail. *Journal of Retailing*.

Sharma, A., Adhikary, A., & Borah, S. B. (2020). Covid-19 s impact on supply chain decisions: Strategic insights from NASDAQ 100 firms using Twitter data. *Journal of Business Research, 117*, 443–449.

Sheather, S. (2009). *A modern approach to regression with R*. Springer Science & Business Media.

Sheth, J. (2020). Impact of Covid-19 on consumer behavior: Will the old habits return or die? *Journal of Business Research, 117*, 280–283.

Silverstein, J. (2020). New York City to close all theaters and shift restaurants to take-out and delivery only due to coronavirus. *CBS*.

Simionescu, M., & Raišienė, A. G. (2021). A bridge between sentiment indicators: What does Google Trends tell us about COVID-19 pandemic and employment expectations in the EU new member states? *Technological Forecasting and Social Change, 173*, Article 121170.

Sjöberg, L., Moen, B.-E., & Rundmo, T. (2004). Explaining risk perception. *An Evaluation of the Psychometric Paradigm in Risk Perception Research, 10*(2), 665.

Sundararaj, R. G. (2017). Why Mobility Data in the Retail Industry is Important. https://datafloq.com/read/mobility-data-in-the-retail-industry/3162 [Accessed: January 13, 2021].

Vaughan, D. E., Jacobson, S. H., Hall, S. N., & McLay, L. A. (2005). Simultaneous generalized hill-climbing algorithms for addressing sets of discrete optimization problems. *INFORMS Journal on Computing, 17*(4), 438–450.

Ververidis, Dimitrios, & Kotropoulos, Constantine. 2005. Sequential forward feature selection with low computational cost. *Pages 1–4 of: 2005 13th European Signal Processing Conference*. IEEE.

Wang, Y., Gu, J., Wang, S., & Wang, J. (2019). Understanding consumers' willingness to use ride-sharing services: The roles of perceived value and perceived risk. *Transportation Research Part C: Emerging Technologies, 105*, 504–519.

Weisberg, S. (2005). *Applied linear regression* (Vol. 528). John Wiley & Sons.

WHO (2010). Pandemic Influenza Preparedness and Response. https://www.ncbi.nlm.nih.gov/books/NBK143062/pdf/Bookshelf_NBK143062.pdf. [Accessed: July 14, 2022].

Yang, S., Santillana, M., & Kou, S. C. (2015). Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences, 112*(47), 14473–14478.

Yoo, S., Baranowski, T., Missaghian, M., Baranowski, J., Cullen, K., Fisher, J. O., Watson, K., Zakeri, I. F., & Nicklas, T. (2006). Food-purchasing patterns for home: A grocery store-intercept survey. *Public Health Nutrition, 9*(3), 384–393.

Zhang, W., & Ram, S. D. (2020). A Comprehensive Analysis of Triggers and Risk Factors for Asthma Based on Machine Learning and Large Heterogeneous Data Sources. *MIS Quarterly, 44*(1).

Zhang, Z. (2016). Variable selection with stepwise and best subset approaches. *Annals of Translational Medicine, 4*(7).

Michael Lash is an assistant professor in the Business Analytics Area at the University of Kansas School of Business. His research interests are broadly in the areas of data mining, machine learning, and business analytics. Before joining the University of Kansas School of Business, he served as a visiting assistant professor at the University of Iowa Tippie College of Business Business Analytics Department. He received his PhD in Computer Science from the University of Iowa in 2018. His work has appeared in a variety of peer-reviewed journals including Expert Systems with Applications (ESWA), the Journal of Management Information Systems (JMIS), and the International Journal of Data Mining and Bioinformatics (IJDMB), among others. His work has also appeared in numerous rigorously peer-reviewed conferences including the SIAM International Conference on Data Mining (SDM), the IEEE International Conference on Health Informatics (ICHI), and the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), among others.

S. Sajeesh is an Associate Professor of Marketing at the College of Business, University of Nebraska-Lincoln. Previously, he was a faculty member at Baruch College, City University of New York. He received his doctorate in Marketing from the Wharton School of the University of Pennsylvania. His research interests are in retailing, emerging markets, product differentiation and marketing strategy. More specifically, his research focuses on understanding firms' marketing strategy using quantitative models under two broad contexts: (1) influence of consumer behavior characteristics (e.g., variety seeking, reference dependence, consumption externality) on optimal firm strategies (positioning and pricing strategies), and (2) influence of market and firm characteristics on firms' marketing strategies (e.g. entry strategies) as well as implications for social welfare and public policy. His research has appeared in some of the premier journals in marketing and business such as *Marketing Science, Management Science, Production and Operations Management, International Journal of Research in Marketing,* and *Decision Sciences*.

Dr Özgür Araz is a Professor in the Supply Chain Management and Analytics Department. His research interests include systems simulation, business analytics, healthcare operations and public health informatics. His research had been supported by NIH, Veterans Engineering Resource Center (VERC), HDR company, Boys Town of Nebraska, Nebraska Medicine and the University of Nebraska. Before joining the College of Business at UNL, he served as a faculty member of the College of Public Health at the University of Nebraska Medical Center (UNMC). He received his Ph.D. in Industrial Engineering from the Ira A. Fulton Schools of Engineering at Arizona State University and was a postdoctoral research fellow at the Center for Computational Biology and Bioinformatics of The University of Texas at Austin. Dr Araz's published research appeared in peer review journals including Production and Operations Management, Decision Sciences, INFORMS Journal on Computing, Annals of Operations Research, Decision Support Systems, International Journal of Production Economics, IIE Transactions on Healthcare Systems Engineering, in addition to high impact medical and health care journals such as American Journal of Public Health, Emerging Infectious Diseases, Obesity, Medical Care, American Journal of Emergency Medicine, among many others. He is a member of INFORMS, POMS (Production and Operations Management Society), Decision Sciences Institute, System Dynamics Society and HIMSS (Healthcare Information Management Systems Society). He is an editorial advisory board member of the Transportation Research Part -E and also serves as associate editor for Decision Sciences and IISE Transactions on Healthcare Systems Engineering journals. He is the Public Health Informatics Area Editor for the Health Systems journal. He is also a faculty fellow of Nebraska Governace and Technology Center - and Daugherty Water for Food Global Institute.