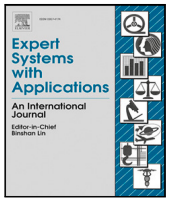
Expert Systems With Applications 238 (2024) 121698

Contents lists available at ScienceDirect

Expert Systems With Applications

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



Analyzing and forecasting service demands using human mobility data: A two-stage predictive framework with decomposition and multivariate analysis

Zhiyuan Wei, Sayanti Mukherjee *

Department of Industrial and Systems Engineering, University at Buffalo - The State University of New York, Buffalo NY 14260, USA

ARTICLE INFO

Keywords:
Critical facilities
Service demand forecasting
Spatiotemporal mobility patterns
Data decomposition
Multivariate time series

ABSTRACT

Accurate service demand forecasts at critical facilities are fundamental for efficiently managing resources and promptly providing essential services to people and community. However, it has received little attention in the literature, mainly due to the unavailability of granular data and the lack of sophisticated forecasting methods. To address this gap, we provide a new perspective on sensing service demands at critical facilities leveraging fine-grained human mobility data, and propose a novel data-driven framework to forecast mobility patterns at the neighborhood level. Specifically, we develop a two-stage forecasting scheme to manage large-scale and complex human movement information. The first stage is to decompose the large-scale mobility data into spatial and temporal patterns, whereas the second stage is to model complex temporal dynamics using multivariate time series analysis. The proposed framework is implemented using real human mobility data obtained from mobile phone users. The results show that our model demonstrates the best predictive performance for varying forecast horizons, when compared to multiple benchmark methods including traditionally-used statistical and deep learning models. We also performed model robustness checks, showing that the proposed model is robust in making short-term and long-term forecasts. The proposed predictive framework could help businesses and local governments accurately forecast service demands for critical facilities for better allocating their resources.

1. Introduction

Critical facilities play a pivotal role in providing essential services to the communities, ensuring that our society can function smoothly and effectively (Logan & Guikema, 2020). Examples of critical facilities include hospitals, pharmacies, gas stations, and grocery stores, where people visit these facilities on a regular basis to fulfill their basic needs. Effectively managing critical facilities, such as inventory control, often requires forecasting customers' demand for future services (Fildes, Ma, & Kolassa, 2022). To support such service demand forecasts, it is of critical importance to build effective information management systems to analyze demand patterns from historical data (Nguyen, Tran, Thomassey, & Hamad, 2021). This often requires both high-quality data to characterize the behavior of customers and advanced analytics tools to make accurate and timely forecasts (Petropoulos et al., 2022).

Traditionally, service demands at critical facilities are often characterized by various data, such as sales data (Bi, Adomavicius, Li, & Qu, 2022; Loureiro, Miguéis, & da Silva, 2018), economic data (Osadchiy, Gaur, & Seshadri, 2013), and social media data (Schaer, Kourentzes, & Fildes, 2019). However, these traditional data used for sensing and predicting service demands often fall short of capturing dynamic customers' behaviors, such as how far they travel and where they come from (Petropoulos et al., 2022). These behaviors are deemed to be fundamental for businesses to understand their catchment/service areas and improve service demand forecasts (Belavina, 2021; Waddington, Clarke, Clarke, & Newing, 2018).

Recently, the burgeoning of large-scale mobility sensing data obtained from mobile phone users opens unique opportunities for better capturing human movement patterns, which are valuable to reveal the behavior of customers in visiting critical facilities (Liu, Liu, Lu, Teng, Zhu, & Xiong, 2017). This is because individuals can "vote with their feet", suggesting that people travel to different service places that can better match their preferences and satisfy their needs (Tiebout, 1956). Analyzing human mobility patterns can offer instructive insights into understanding the demand behaviors of people in accessing critical facilities (Wei & Mukherjee, 2023). However, most of the previous studies use descriptive analysis to describe human mobility patterns at various locations including public parking (Nie et al., 2021) and recreational spaces (Marcelo et al., 2022). Descriptive analysis falls

* Corresponding author.
E-mail addresses: zwei7@buffalo.edu (Z. Wei), sayantim@buffalo.edu (S. Mukherjee).

https://doi.org/10.1016/j.eswa.2023.121698
Received 3 June 2023; Received in revised form 1 September 2023; Accepted 16 September 2023
Available online 27 September 2023

short of predicting future service demand patterns, which are crucial for resource prepositioning at critical facilities. Therefore, our study focuses on developing a novel predictive framework that utilizes human mobility data to forecast service demand patterns for critical facilities.

Despite the availability and accessibility of large amount of mobile phone data, predicting human mobility patterns is still challenging. This is because the people's movement behaviors involve complex spatiotemporal dependencies, and have nonlinear interactions with various contextual factors such as sociodemographics (Wang, Currim, & Ram, 2022). This complexity of mobility patterns requires more sophisticated predictive techniques (Cheng, Trepanier, & Sun, 2022). The traditional statistical methods such as autoregressive integrated moving average (ARIMA) often require the assumption of stationarity and linear functional form in the data, thus limiting their ability to analyze complex nonlinear time series patterns (Harvey, 1990). In contrast, conventional machine learning models (e.g., random forest, support vector regression) are flexible in handling high-dimensional and nonlinear data, but they often fall short of capturing temporal dependencies in the variable (Au, Choi, & Yu, 2008). To model time-dependent patterns, deep learning models including the Long Short-Term Memory (LSTM) are utilized to analyze dynamic patterns that vary over time. However, deep learning models are difficult to interpret (Herm, Heinrich, Wanner, & Janiesch, 2023), and may not be suitable to incorporate static contextual information such as sociodemographics in the analysis (Ray, Jank, Dutta, & Mullarkey, 2023). Some studies have developed hybrid models that combine the strength of various models in order to produce better forecast results (Punia & Shankar, 2022; Van Steenbergen & Mes, 2020). Still, the design of hybrid models needs to consider potential issues such as overfitting and model selection, which have received less attention in the literature. Therefore, to overcome the existing challenges in time series modeling to handle complex and high-dimensional data, we aim to develop a data-driven framework to provide accurate and timely forecasts by exploring human mobility patterns.

To the best of our knowledge, this is the first study exploring the use of human mobility data to analyze and forecast the service demands. The large-scale big data on human mobility patterns obtained from mobile phone data exhibit complex spatial and temporal dependencies, posing significant challenges to most traditional methods, mostly owing to higher computational cost and complexities. Traditional methods such as ARIMA and VAR often fall short of modeling complex relationships efficiently, leading to inaccurate forecasting results. To address these challenges, we propose a novel two-stage predictive framework to efficiently handle large-scale and complex time series human mobility data. Specifically, the first stage of the framework aims to reduce the size of mobility data by extracting the key spatial and temporal patterns leveraging decomposition method. However, the conventional matrix decomposition methods are not suitable for time-series forecasting. To overcome this limitation, we introduce the second stage of our framework to incorporate the time dynamics required for time-series prediction. We also illustrate the extension of the proposed framework by integrating various contextual variables (e.g., sociodemographics) of interest. In a nutshell, the contribution of this research to the extant literature is summarized as follows.

- This paper provides a new perspective to analyze and forecast service demands leveraging human mobility patterns. This point of view allows businesses and governments to identify dynamic service demands of people at a fine-grained spatial scale (i.e., neighborhood level), which may not be apparent from traditional data that focused on either the state or county level.
- We propose a novel two-stage forecasting framework that incorporates data decomposition and multivariate time series modeling techniques. This framework can capture both complex spatial and temporal dependencies of large-scale data, and flexibly incorporate contextual information, which hold promise for providing better forecast results.
- We perform a series of experiments, including sensitivity analysis and robustness checks, to demonstrate the capabilities of the proposed framework using the real-world mobile phone data. We also illustrate superior predictive performances of our model in comparison to a variety of benchmark methods including the traditionally-used statistical (e.g., ARIMA, VAR) and deep learning models (e.g., LSTM).

The remainder of the paper is organized as follows. Section 2 summarizes the related literature which motivates and supports the importance of this paper. Section 3 provides an overview of the design of our two-stage predictive framework. Section 4 presents a case study to illustrate the applicability of the framework. Section 5 provides numerical experiments to show the model performance. Discussions and conclusions of this study are summarized in Section 6 and Section 7 respectively.

2. Related work

In this section, we first highlight the gaps related to the various data used for service demand characterization. Thereafter, we present the various challenges associated with the different traditionally-used forecasting models for service demand prediction.

2.1. Service demand characterization

To accurately capture service demand patterns at critical facilities is a crucial step for service demand forecasting. Traditionally, there is a variety of data that can be leveraged to characterize service demand patterns. For example, sales data refers to information collected on the sales transactions of a business. It often exhibits strong temporal trends and seasonal variations, which can be used to understand the demand of customers in retail establishment (Choi, Yu, & Au, 2011; Vallés-Pérez et al., 2022). Waddington et al. (2018) explore the temporal fluctuations of store sales to understand the demand to grocery stores and highlight the needs of integrating the information across multiple stores to generate more accurate demand in a region. However, sales data is typically classified as proprietary commercial information, which is not publicly accessible (Fildes et al., 2022). In other words, sales data collected by one store may not be readily available to others, hindering the stakeholders such as local governments from making informed decisions at the regional level (Waddington et al., 2018). Another limitation of sales data is that it often results in an underestimation of service demand under certain scenarios such as out-of-stock situations (Chen & Ou, 2009).

Economic data or macroeconomic information is another type of data used to reveal service demand patterns, where a significant economic upswing (or downturn) is highly correlated with growth (or decline) in demand (Osadchiy et al., 2013; Sagaert, Aghezzaf, Kourentzes, & Desmet, 2018). Osadchiy et al. (2013) develop a statistical model to predict the sales at retailer stores based on financial indicators such as market return information. Suryani, Chou, and Chen (2010) consider economic conditions such as Gross Domestic Product (GDP) to capture the demand of people at airports, and to predict the future air travel demand. Since economic information is frequently gathered at coarse spatiotemporal scales, it is not adequate to reflect the nuanced and detailed customer's demand at the micro level, which is the key for inventory management (Fildes et al., 2022).

Social media such as Twitter and Meta is another data source to uncover service demands by providing information about customers' preferences, opinions, and emotions. Papanagnou and Matthews-Amune (2018) investigate the use of internet information including Google search intensity and YouTube watch video duration to understand the demand patterns to pharmacies and highlight the importance of text mining technique in demand forecasting. Even though the value of social media data has been pronounced in previous studies (Cui, Gallino,

Moreno, & Zhang, 2018; Terroso-Saenz, Flores, & Muñoz, 2022), there are also some concerns regarding its usage such as high bias and limited coverage, especially when users fail to provide accurate information in their social media (Seyedan & Mafakheri, 2020).

With the recent advancement of global positioning system (GPS) technology, human mobility data is becoming increasingly accessible and abundant, which has the potential to characterize the dynamic demand patterns of people at critical facilities (Liu et al., 2017). For example, an increased frequency of visits to grocery stores in the face of natural disaster often suggests a greater level of demand among people due to a panic buying behavior (Wei & Mukherjee, 2022). Nie et al. (2021) utilize human mobility data at parking facilities to understand the public parking demand at different time of day. Human mobility data overcomes the limitations of traditionally-used data in capturing customers' behaviors in access to critical facilities, and provides a new perspective to characterize the demand patterns of people (Marcelo et al., 2022). However, most studies only utilize descriptive analysis to gain insights on what has happened in the past, without focusing on what mobility patterns will occur in the future. Therefore, there is a need to conduct predictive analysis that can make accurate estimations about service demand forecasts for critical facilities in the near future.

2.2. Matrix decomposition

Matrix decomposition (or matrix factorization) refers to the transformation of a matrix into a canonical form, which has been widely applied in a variety of applications such as missing data imputation, feature extraction, and dimensionality reduction (Fanaee-T & Gama, 2016; Kolda & Bader, 2009). Tensor is a generalized matrix to represent multidimensional data. Growing attention has been paid to extracting latent patterns from human movements using decomposition techniques. For instance, Fan, Song, and Shibasaki (2014) use the tensor factorization method to decompose the population flow into various components representing basic mobility patterns for working, commuting and entertaining fields. Du, Zhou, Liu, Cui, and Xiong (2019) develop a framework based on tensor decomposition to identify underlying citywide transit service hot spots using public transit data. To address the missing data issues in human mobility patterns, Chen, He, and Sun (2019) present a Bayesian tensor decomposition approach for spatiotemporal traffic data imputation.

However, the standard matrix/tensor decomposition methods often fall short of capturing the dynamics of temporal patterns in the data, thus making them unsuitable for time series forecasting (Kargas et al., 2021). To facilitate times-series prediction, linear time series models are often incorporated into the matrix decomposition process to account for temporal evolution. For example, Yu, Rao, and Dhillon (2016) develop a temporal matrix factorization model, where autoregressive (AR) process is embedded to model latent temporal factors from the matrix decomposition. Chen and Sun (2021) present a Bayesian temporal factorization framework that combines vector autoregressive (VAR) process and matrix factorization for modeling spatiotemporal data with missing values. However, these studies apply the linear time series models to the matrix decomposition process, which can restrict the ability to handle complex nonlinear data. These challenges call for development of more sophisticated models to forecast large-scale complex time series data, which we will discuss in the next section.

2.3. Forecasting methods

In general, time series forecasting models could be broadly classified into three main categories: statistical time-series models, machine learning (including deep learning) methods, and hybrid methods (Punia & Shankar, 2022). In what follows, we discuss the strengths and weaknesses of these different types of models, highlighting the research gaps in their application for service demand forecasting, which serve as the motivation for our work.

A number of statistical models, such as autoregressive integrated moving average (ARIMA), seasonal ARIMA (SARIMA), and vector autoregression (VAR), have been utilized to forecast the service demand. For example, Papanagnou and Matthews-Amune (2018) use VAR model to predict medicine demands using external variables such as Google search intensity and YouTube video watch duration. Xu, Qi, and Hua (2010) apply ARIMA model to predict the demand for agricultural commodities at different whole foods markets. The simple statistical models often assume certain form of the relationship in the data (e.g., linear), which may limit the handling of complex nonlinear data with high frequency and volatility (Harvey, 1990).

In contrast, supervised machine learning models have been widely used to analyze nonlinear patterns, as they are flexible without the specification of certain functional form in the data (Cavalcante, Frazzon, Forcellini, & Ivanov, 2019). The benefit of using machine learning is highlighted in previous studies on demand forecasting (Carbonneau, Laframboise, & Vahidov, 2008; Loureiro et al., 2018). Ferreira, Lee, and Simchi-Levi (2016) illustrate the effectiveness of regression trees with bootstrap aggregation (a.k.a., bagging) in the prediction of both flash sales and new products. To further account for temporal patterns, deep learning models such as Long Short-Term Memory (LSTM) are often used to model complex time series patterns (Nguyen et al., 2021). Wang et al. (2022) propose a context-aware LSTM model to forecast time-dependent ridership demand using bus GPS trajectory and automatic fare collection (AFC) data. Zhu, Ninh, Zhao, and Liu (2021) consider the cross-series information for predicting pharmaceutical demands using recurrent neural network (RNN). However, training deep learning models can be computationally expensive as it requires to learn a large amount of model parameters, making it difficult to deploy these models in resource-constrained environments in a timely manner (Yang, Xue, Ding, Wu, & Gao, 2021).

Hybrid models that involve the combination of various models could jointly improve the overall forecast accuracy. Choi et al. (2011) develop a hybrid scheme for sales forecasting at apparel retail stores where the classic statistical method and wavelet transformation are combined to show the improved predictive accuracy. Arunraj and Ahrens (2015) introduce a hybrid model integrating SARIMA and quantile regression to forecast the daily demand in retail stores. Punia and Shankar (2022) present another hybrid model combining LSTM and random forest for demand forecasts of packaged food products considering both structured and unstructured data (Punia & Shankar, 2022). Hybrid models provide ways of incorporating the strengths of various models, but those studies simply assemble models that share same or similar functionality, without fully accounting for the issues of model selection and overfitting.

To summarize, most studies on demand forecasting for critical facilities utilize univariate models, which solely account for a single autoregressive variable in the analysis (Arunraj & Ahrens, 2015; Choi et al., 2011; Loureiro et al., 2018). The univariate models often fall short of capturing the interdependencies of multiple response variables, which may constrain the model's predictive power. On the contrary, the multivariate model that considers the joint effect of multiple response variables is found to have superior predictive performance over univariate models (Wei, Narin, & Mukherjee, 2022). However, little attention has been paid to exploring multivariate time series models that can allow the incorporation of multiple autoregressive patterns into the demand forecasting for critical facilities (Fildes et al., 2022). To overcome the above-mentioned challenges in time series forecasting and to further enhance predictive performance, this paper develops a novel two-stage forecasting framework by integrating data decomposition technique and multivariate time series modeling. Our framework is different from the most existing demand forecasting schemes in two ways. On the one hand, the proposed framework first extracts the main spatial and temporal patterns from human mobility data in the first stage, and then models the dynamics of temporal patterns for prediction in the second stage. The synthesis of these two stages

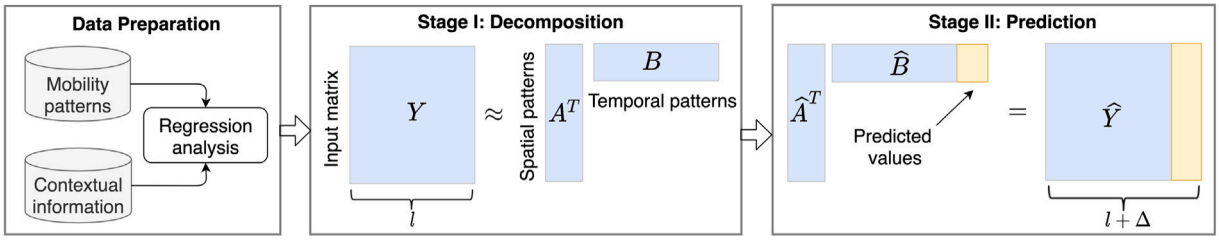


Fig. 1. Overview of the proposed two-stage forecasting framework with contextual information.

could efficiently handle large-scale and complex spatiotemporal data. On the other hand, our framework is flexible to incorporate contextual information, which is valuable for businesses to understand the key factors related to service demand patterns.

3. Methodological framework

In this section, we introduce the proposed two-stage forecasting framework. Stage I is described in Section 3.1, and Stage II is presented in Section 3.2. We also illustrate how our framework can be extended by incorporating contextual information in Section 3.3. The schematic representation of the proposed two-stage forecasting framework with contextual information is represented in Fig. 1.

3.1. Stage I: Decomposition

The first stage aims to discover the latent features in the data by decomposing the original data to a more compact form. Specifically, we use $\boldsymbol{Y} = \{y_{it}, \forall(i,t)\}$ to represent the service demands that vary across both neighborhoods ($i = 1,\dots,n$) and time ($t = 1,\dots,l$). The decomposition allows $\boldsymbol{Y} \in \mathbb{R}^{n\times l}$ to be factorized into neighborhood-specific latent factor $\boldsymbol{a}_i = [a_{i1}, a_{i2}, \dots, a_{ik}]^T$ and time-specific latent factor $\boldsymbol{b}_t = [b_{t1}, b_{t2}, \dots, b_{tk}]^T$. Mathematically, it can be written as

$$y_{it} \approx \boldsymbol{a}_i^T \boldsymbol{b}_t = \sum_{r=1}^{k} a_{ir} b_{tr}, \text{ for } i = 1,\dots,n,\ t = 1,\dots,l. \tag{1}$$

Here $\boldsymbol{a}_i \in \mathbb{R}^{k\times n}$, $\boldsymbol{b}_t \in \mathbb{R}^{k\times l}$, and $k < \min(n,l)$. The use of such a decomposition technique in Eq. (1) enjoys several benefits including feature extraction and data compression. On the one hand, the latent factors could reveal certain characteristics of original data across each dimension (Kolda & Bader, 2009). On the other hand, it can save computational expense by reducing original data (with the size of $n\times l$) to more compact data (with the size of $(n+l)\times k$), while still maintaining data fidelity. This can substantially improve computation efficiency, especially when dealing with a large-scale dataset (Chen, He, & Wang, 2018).

To find $\boldsymbol{A} = [\boldsymbol{a}_1,\dots,\boldsymbol{a}_n]$ and $\boldsymbol{B} = [\boldsymbol{b}_1,\dots,\boldsymbol{b}_l]$ to represent $\boldsymbol{Y}$ (i.e., $\boldsymbol{Y} \approx \boldsymbol{A}^T\boldsymbol{B}$), we formulate this as an optimization problem to minimize a loss function $\mathcal{L}(\boldsymbol{A},\boldsymbol{B})$, which is given by

$$\min_{\boldsymbol{A},\boldsymbol{B}} \quad \mathcal{L}(\boldsymbol{A},\boldsymbol{B}) = \underbrace{\sum_{(i,t)\in\Omega} \left(y_{it} - \boldsymbol{a}_i^T \boldsymbol{b}_t \right)^2}_{\text{reconstruction error}} + \underbrace{\lambda \left(\sum_{i=1}^{n} \left\|\boldsymbol{a}_i\right\|_2^2 + \sum_{t=1}^{l} \left\|\boldsymbol{b}_t\right\|_2^2 \right)}_{L_2 \text{ regularization}}. \tag{2}$$

The first term in $\mathcal{L}(\boldsymbol{A},\boldsymbol{B})$ describes the reconstruction error between the actual value and its estimation, while the second term is the L_2 regularization to improve the generalization performance of the model by avoiding overfitting. The parameter λ controls the extent of regularization, $\Omega = \{(i,t) : y_{it} \text{ is observed}\}$ represents the set of observed data, and $\|\cdot\|_2$ is the Frobenius norm defined as the square root of the sum of the element-wise absolute squares. Note that, $\mathcal{L}(\boldsymbol{A},\boldsymbol{B})$ is a non-convex function due to the $\boldsymbol{a}_i^T\boldsymbol{b}_t$ term, thus is a NP-hard to optimize (Hillar & Lim, 2013). To address this challenge, we apply the Alternating Least Squares (ALS) method, with the idea of iteratively optimizing one matrix at a time while holding the other matrices constant. In other words, it alternates between updating the neighborhood matrix and the time matrix, until convergence. Specifically, when fixing $\boldsymbol{b}_t$, we can optimize $\boldsymbol{a}_i$ through

$$\hat{\boldsymbol{a}}_i = \arg\min_{\boldsymbol{a}_i} \sum_{\{t:(i,t)\in\Omega\}} \left(y_{it} - \boldsymbol{a}_i^T \boldsymbol{b}_t \right)^2 + \lambda \|\boldsymbol{a}_i\|_2^2, \quad i = 1,\dots,n. \tag{3}$$

Similarly, when holding $\boldsymbol{a}_i$ as a constant, $\boldsymbol{b}_t$ can be optimized through

$$\hat{\boldsymbol{b}}_t = \arg\min_{\boldsymbol{b}_t} \sum_{\{i:(i,t)\in\Omega\}} \left(y_{it} - \boldsymbol{a}_i^T \boldsymbol{b}_t \right)^2 + \lambda \|\boldsymbol{b}_t\|_2^2, \quad t = 1,\dots,l. \tag{4}$$

From Eqs. (3) and (4), it can be observed that the algorithm solves a least squares problem to obtain the latent factors, which is a convex problem. We further derive the closed form solution to update $\hat{\boldsymbol{a}}_i$ and $\hat{\boldsymbol{b}}_j$ through the first-order conditions (i.e., setting the partial derivatives equal to zero). Then, we have

$$\hat{\boldsymbol{a}}_i = \left(\sum_{t:(i,t)\in\Omega} \boldsymbol{b}_t^T \boldsymbol{b}_t + \lambda I \right)^{-1} \sum_{t:(i,t)\in\Omega} y_{it} \boldsymbol{b}_t, \quad i = 1,\dots,n, \tag{5}$$

$$\hat{\boldsymbol{b}}_t = \left(\sum_{i:(i,t)\in\Omega} \boldsymbol{a}_i^T \boldsymbol{a}_i + \lambda I \right)^{-1} \sum_{i:(i,t)\in\Omega} y_{it} \boldsymbol{a}_i, \quad t = 1,\dots,l, \tag{6}$$

where I is the identify matrix that has dimension of $k \times k$.

3.2. Stage II: Prediction

In Stage I, the mobility patterns are decomposed into two latent factors (i.e., neighborhoods and time), where the time-specific factor signals the time only up to l. However, we are interested in predicting the future timestamps such as $l+1$. The matrix decomposition often falls short of predicting the future dynamics based on static latent factors (Yu et al., 2016). To address this problem, we further impose temporal dynamics on the latent time factor $\boldsymbol{b}_t$, so that it can evolve over time. Note that, we have a total of k different time series according to the matrix $\boldsymbol{B} \in \mathbb{R}^{k\times l}$. To fully utilize the interdependency of multiple time series, we leverage multivariate time series approach to simultaneously model various dependent variables co-evolving over time. This approach overcomes the assumption of univariate analysis that often assumes each time series pattern is independent (Tsay, 2013). Therefore, in this paper, we utilize dynamic mode decomposition (DMD) method for multivariate time series forecasting. There are several advantages of using the DMD method. On the one hand, it is a data-driven algorithm that does not require any prior knowledge of the data being analyzed, which is well-suited for nonlinear data (Yu et al., 2016). On the other hand, it works by extracting dominating dynamic modes from the data and predicting multiple time series simultaneously, which is ideal for multivariate analysis (Cheng et al., 2022). In what follows, we present the details of how we use DMD to predict the latent time factors in our study.

Given the matrix $\boldsymbol{B}$ obtained from Stage I, we first partition it into two matrices denoted by $\bar{\boldsymbol{B}} = [\boldsymbol{b}_1,\dots,\boldsymbol{b}_{l-1}] \in \mathbb{R}^{k\times(l-1)}$ and $\bar{\boldsymbol{B}}' = [\boldsymbol{b}_2,\dots,\boldsymbol{b}_l] \in \mathbb{R}^{k\times(l-1)}$ with the same size, where $\bar{\boldsymbol{B}}'$ is just shifted one time step ahead from $\bar{\boldsymbol{B}}$. The goal of DMD is to find a best fit operator

$\boldsymbol{D} \in \mathbb{R}^{k\times k}$ that can advance $\bar{\boldsymbol{B}}$ into $\bar{\boldsymbol{B}}'$, which is mathematically given by

$$\bar{\boldsymbol{B}}' \approx \boldsymbol{D}\bar{\boldsymbol{B}}. \tag{7}$$

Then, we compute a low-rank approximation to $\bar{\boldsymbol{B}}$ by applying the truncated singular value decomposition (SVD) method, while retaining the r largest singular values and vectors (Yu et al., 2016). This can be expressed as

$$\bar{\boldsymbol{B}} \approx \bar{\boldsymbol{B}}_r = \boldsymbol{U}_r\boldsymbol{\Sigma}_r\boldsymbol{V}_r^\top, \tag{8}$$

where $\boldsymbol{U}_r \in \mathbb{R}^{k\times r}$ and $\boldsymbol{V}_r \in \mathbb{R}^{(l-1)\times r}$ are unitary matrices, and $\boldsymbol{\Sigma}_r \in \mathbb{R}^{r\times r}$ is a diagonal matrix with r non-zero singular values. Based on Eqs. (7) and (8), we can solve $\boldsymbol{D}$ through the following equation

$$\boldsymbol{D} \approx \bar{\boldsymbol{B}}'\bar{\boldsymbol{B}}^+ \approx \bar{\boldsymbol{B}}'\bar{\boldsymbol{B}}_r^+ = \bar{\boldsymbol{B}}'(\boldsymbol{U}_r\boldsymbol{\Sigma}_r\boldsymbol{V}_r^\top)^+ = \bar{\boldsymbol{B}}'\boldsymbol{V}_r\boldsymbol{\Sigma}_r^{-1}\boldsymbol{U}_r^\top. \tag{9}$$

Note that $\bar{\boldsymbol{B}}^+$ denotes the Moore–Penrose inverse of a matrix $\bar{\boldsymbol{B}}$, whereas $\bar{\boldsymbol{B}}'$ is just a representation of a collection of time-specific factors.

Next, instead of computing the full matrix $\boldsymbol{D}$, which is often computationally expensive, the low-rank projection $\widetilde{\boldsymbol{D}} \in \mathbb{R}^{r\times r}$ is derived through the following equation:

$$\widetilde{\boldsymbol{D}} = \boldsymbol{U}_r^+\boldsymbol{D}\boldsymbol{U}_r \approx \boldsymbol{U}_r^+\bar{\boldsymbol{B}}'\boldsymbol{V}_r\boldsymbol{\Sigma}_r^{-1}\boldsymbol{U}_r^\top\boldsymbol{U}_r = \boldsymbol{U}_r^+\bar{\boldsymbol{B}}'\boldsymbol{V}_r\boldsymbol{\Sigma}_r^{-1}. \tag{10}$$

Here $\bar{\boldsymbol{B}}'$ is an output of Stage I, $\boldsymbol{U}_r$, $\boldsymbol{\Sigma}_r$, and $\boldsymbol{V}_r$ can be derived from Eq. (8). It can be noted that $\widetilde{\boldsymbol{D}}$ and $\boldsymbol{D}$ both have the same nonzero leading eigenvalues (Tu, Rowley, Luchtenburg, Brunton, & Kutz, 2014). Based on eigendecomposition, we have $\widetilde{\boldsymbol{D}}\boldsymbol{W} = \boldsymbol{W}\boldsymbol{\Lambda}$, where the columns of $\boldsymbol{W}$ are eigenvectors, and the entries of the diagonal matrix $\boldsymbol{\Lambda}$ are the corresponding eigenvalues. The DMD modes (i.e., the eigenvectors of $\boldsymbol{D}$) can be derived by $\boldsymbol{\Phi} = \bar{\boldsymbol{B}}'\boldsymbol{V}_r\boldsymbol{\Sigma}_r^{-1}\boldsymbol{W}$. That is, each column of $\boldsymbol{\Phi}$ is a DMD mode corresponding to a particular eigenvalue in $\boldsymbol{\Lambda}$. To this end, with the approximated eigenvalues and eigenvectors of $\boldsymbol{D}$ in hand, we can predict the next time step, which is analytically constructed as

$$\boldsymbol{b}_t = \boldsymbol{\Phi}\boldsymbol{\Lambda}^t\boldsymbol{b}_0, \text{ for } t = l+1, \dots. \tag{11}$$

The vector $\boldsymbol{b}_0 = \boldsymbol{\Phi}^{-1}\boldsymbol{b}_1$ represents the initial amplitude of each DMD mode. It is worth noting that Eq. (11) provides a closed form of prediction for latent time factors, so that no numerical iteration is needed. This can greatly reduce the computational expenses.

To summarize, based on the proposed two-stage forecasting framework, the predicted values across neighborhoods at the future time stamps are calculated through

$$\hat{y}_{it} = \hat{\boldsymbol{a}}_i^T\boldsymbol{b}_t, \text{ for } t = l+1, \dots, \tag{12}$$

where $\hat{\boldsymbol{a}}_i$ can be numerically estimated by Eq. (5), and $\boldsymbol{b}_t$ can be analytically obtained through Eq. (12).

3.3. Integrating contextual information

The proposed two-stage framework can be extended by integrating contextual information (e.g., sociodemographic backgrounds and weather conditions). To achieve this, we first denote $\boldsymbol{X}$ as a set of independent variables of interest. This can be, for example, $\boldsymbol{X} = \{\boldsymbol{x}^{\text{tavg}}, \boldsymbol{x}^{\text{pop}}\}$, where $\boldsymbol{x}^{\text{tavg}}$ indicates average temperature and $\boldsymbol{x}^{\text{pop}}$ is population. Then, we represent $g(\cdot)$ as the regression model of interest (e.g., linear model, regression tree) that can build the relationship between dependent variable $\boldsymbol{Y}$ and independent variables $\boldsymbol{X}$, i.e., $\boldsymbol{Y} = g(\boldsymbol{X}) + \boldsymbol{\epsilon}$. In this study, we select random forest as the regression model, as it has shown to be effective in regression analysis in terms of both predictability and interpretability (Breiman, 2001; Ferreira et al., 2016). With contextual information, we can apply our framework using residuals $\boldsymbol{\epsilon}$ as model inputs rather than $\boldsymbol{Y}$. Analyzing the residuals allows our model to capture the unexplained variability in the data, which could further improve model predictive performance.

To summarize, the key innovation of our framework lies in how these three components depicted in Fig. 1 (data preparation, decomposition, and prediction) are combined and integrated seamlessly to achieve better predictive performance from complex human mobility data. The algorithm for our proposed two-stage forecasting framework with contextual information is illustrated in Algorithm 1. This algorithm could benefit readers and practitioners by helping them implement our framework more effectively, as we have provided analytical solutions and numerical iterations for the equations in the algorithm to make their usage easier.

Algorithm 1 Two-stage forecasting framework with contextual information

Input: $\boldsymbol{Y}$, $\boldsymbol{X}$
Output: $\{g(\boldsymbol{X}) + \hat{e}_{it}\}$ for $i = 1, \cdots, n$ and $t = l+1$
Initialization: $\boldsymbol{a}_i$ and $\boldsymbol{b}_t$ with random values
Perform regression analysis $g(\cdot)$ and obtain residuals $\epsilon = \boldsymbol{Y} - g(\boldsymbol{X})$
$y_{it} \leftarrow e_{it}$, $\forall(i,t)$
repeat ▷ Stage I: decomposition
    Update $\hat{\boldsymbol{a}}_i$ using Equation (5), $i = 1, \cdots, n$
    Update $\hat{\boldsymbol{b}}_t$ using Equation (6), $t = 1, \cdots, l$
until convergence
Predict $\boldsymbol{b}_t$ for $t = l+1$ using Equation (11) ▷ Stage II: prediction
Calculate $\hat{y}_{it}$ for $t = l+1$ using Equation (12)
$\hat{e}_{it} \leftarrow \hat{y}_{it}$, $\forall(i,t)$

4. Experimental setting

In this section, we present how to demonstrate the applicability of our proposed framework using real-world human mobility data. An overview of the steps involved in the model implementation is shown in Fig. 2. Specifically, we begin with an introduction of human mobility data collected from SafeGraph company and other types of data in Section 4.1. Then, we provide a brief introduction to benchmark models that are used to compare our model in Section 4.2. Following that, the details on model implementation and performance evaluation are described in Section 4.3.

4.1. Data preparation

We apply our framework to analyze and predict the service demand of the gas stations in the Harris County, Texas, U.S. The county seat is the city of Houston, which is the largest city in Texas and fourth largest city in the U.S. The selection of gas station is due to two main reasons. First, gas stations are critical facilities that provide essential service for transportation on a daily basis. Second, accurate predictions of gasoline demand are critical for ensuring the continuity of energy supply, considering the fact that more than 90% of households in Harris County reported owning at least one vehicle (Houston State of Health, 2023).

4.1.1. Human mobility data

Human mobility data from mobile phone users in Harris County is obtained from SafeGraph, which is a company that gathers geospatial data from mobile phone users (SafeGraph, 2020). Specifically, two types of data are collected from SafeGraph. One is human mobility information, expressed in the CSV (Comma-Separated Values) format, including the daily number of people who left their neighborhood (where their homes are located) and the daily number of visitors at each service facility. Another type of data is the geographical information, which is stored in the JSON (JavaScript Object Notation) format and includes geographic coordinates for service locations and polygons for neighborhoods. The mobility data is anonymous and aggregated at each census block group (CBG) to protect individual privacy (SafeGraph, 2020). Here, the CBG is a geographic area defined by the U.S. Census Bureau that contains between 600 and 3000 people, and used to represent the residential neighborhood. The data also contains foot traffic

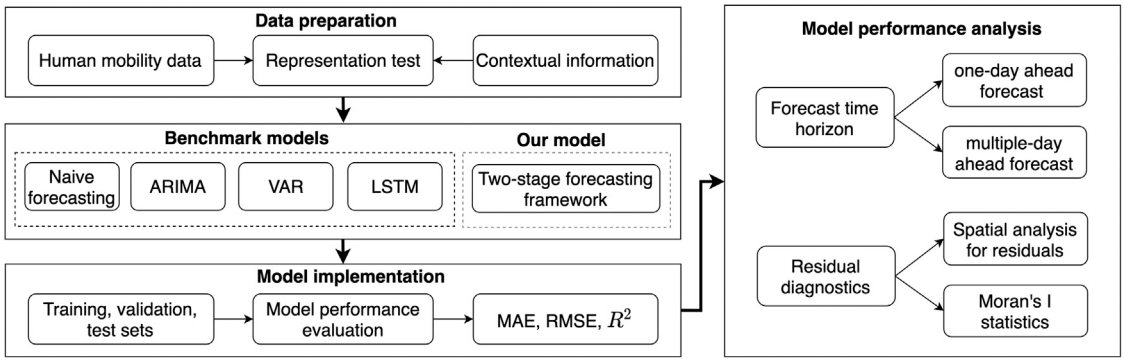


Fig. 2. An overview of the steps in model implementation.

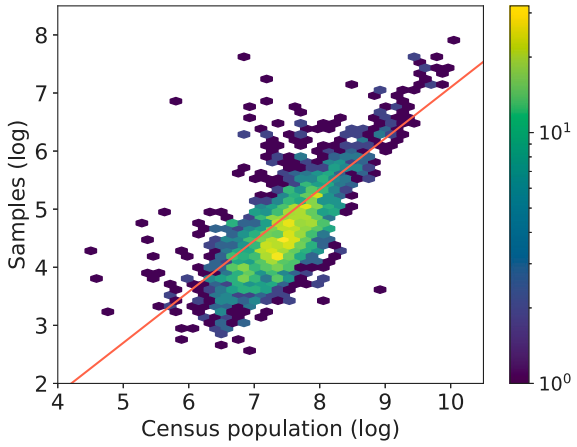


Fig. 3. Comparison between the sampled mobile phone users and the census population. Each hexagon is color-coded based on the number of neighborhoods within each bin.

information at various types of points-of-interest (POI) including retail stores and gas stations. Each POI is associated with a business type based on the North American Industry Classification System (NAICS) code. Based on the methodology developed in previous work to connect CBGs and POIs (Wei & Mukherjee, 2023), we obtained the daily human movement patterns to gas stations (NAICS code: 447110) at each CBG in our study region. In our analysis, the response variable y_{it} represents the daily number of customers visiting a gas station in each neighborhood. This response variable could signal the demand of people for gasoline service across varying neighborhoods. Daily mobility patterns to gas stations were recorded for each neighborhood for the first three months of 2021, totaling 12 weeks. In brief, our database covers 0.32 million mobile phone users, which account for around 7% of that census population, across 2142 CBGs for 12 weeks in our study region. Here, we assume that the CBG population remains constant during our analysis period of 12 weeks, which is a reasonable and practical assumption since the census population generally remains constant during a short- or medium-term period.

To ensure that mobility data being collected is representative, we implemented the representation test, where the sampled data is compared with the census data. First, a hexbin (a.k.a., hexagonal binning) plot is used to visualize the distribution of the sampled data to the census data using hexagonal bins. As shown in Fig. 3, each hexagon is color-coded based on the number of neighborhoods within each bin, where the lighter (darker) colors indicate more (less) neighborhoods in each bin. It can be observed that most hexagons are centered around the regression line, which suggests a strong positive correlation between the sampled data and the census population. Second, a quantitative analysis is performed by calculating the Pearson correlation coefficient (denoted as ρ). We confirmed that the sampled data is highly correlated with the census data ($\rho = 0.81$) across all neighborhoods, and this relationship is statistically significant ($p < 0.05$). The results of these analyses suggest that our dataset does not contain significant biases to the census population. This is also in line with the previous study that shows SafeGraph data is generally representative of the census data (Brelsford, Moehl, Weber, Sparks, Tuccillo, & Rose, 2022).

4.1.2. Contextual information

We then collected climate information and sociodemographic characteristics pertaining to our study region. The climate data are gathered from the National Oceanic and Atmospheric Administration (NOAA) (NOAA, 2021). The sociodemographic variables are obtained from the American Community Survey (ACS) 5-year estimates of the U.S. Census Bureau (US Census Bureau, 2020). We also introduced an indicator variable to represent the day of week. The description of the quantitative variables is exhibited in Table 1.

Table 1 also displays the Pearson's correlation coefficient (Pearson's ρ) and the corresponding p-value for each quantitative variable with respect to our response variable. All variables depicted in Table 1 are continuous and measured at the CBG level. Here, the p-value is a statistical measurement used to validate the significance of a hypothesis test. A low p-value (typically below a significance level of 0.05) indicates that we have enough evidence to reject the null hypothesis, indicating that there is a significant correlation between the independent variable and the dependent variable. We observe that all the variables are statistically significant based on the p-values, suggesting that the inclusion of these information may help understand and predict mobility patterns for the gas stations. For climate variables, mobility behaviors are negatively correlated with daily precipitation, wind speed, and air pressure, but positively linked to average air temperature. For sociodemographic variables, a higher proportion of white, older, and low-income people is linked to fewer mobility movements. On the contrary, the proportion of people who have not completed high school is positively associated with mobility patterns to gas stations, which may be attributed to work activities.

4.2. Benchmark models

Our proposed model is compared with four benchmark models including naive forecasting, autoregressive integrated moving average (ARIMA), vector autoregression (VAR), and long short-term memory (LSTM), which are widely used for time series forecasting in the literature (Ferreira et al., 2016; Nguyen et al., 2021; Xu et al., 2010). The naive forecast simply predicts the future value by taking the mean value of the historical data, which is a useful baseline to compare more complex forecasting methods. ARIMA and VAR are representative examples of parametric models, which often restricts the data being analyzed to follow a particular distribution or assumption (e.g., linearity). On the other hand, LSTM is representative example of non-parametric

Table 1
Description of quantitative variables from the data.

Variable	Description (Unit)	Pearson's ρ	p-value
tavg	The average air temperature (°C).	0.037	$p < 0.05$
prcp	The daily precipitation (mm).	−0.032	$p < 0.05$
wspd	The average wind speed (km/h).	−0.027	$p < 0.05$
pres	The average sea-level air pressure (hPa).	−0.018	$p < 0.05$
total population	Census population (count)	0.847	$p < 0.05$
white population	Proportion of population identified as white (%).	−0.052	$p < 0.05$
older population	Proportion of population aged 65 and older (%).	−0.201	$p < 0.05$
less than high school	Proportion of population have not completed high school (%).	0.08	$p < 0.05$
low income households	Proportion of population below $59,999 household income (%).	−0.043	$p < 0.05$

model without the specification of certain functional form in the data. A short description of these models is presented here. For more details on these time series models, readers may refer to the book (Box, Jenkins, Reinsel, & Ljung, 2015).

ARIMA is a univariate statistical model that is made up of three parts (Box et al., 2015). The first part is autoregressive that captures the linear relationship between an observation and previous lagged observations. The second part is integration that accounts for the rate of change of the growth/decline in the data. The last part is moving average that explores the lagged forecast errors.

VAR is a multivariate statistical model that is able to capture the joint behavior of multiple time series variables. It is an extension of the univariate autoregressive models. In the VAR model, a set of time series variables are modeled as linear functions of their historical values, as well as the historical observations of the other variables in the system. This allows for the modeling of interdependencies among the variables that are co-evolving over time (Box et al., 2015).

LSTM model is a machine learning model designed for learning long-term dependencies in time series data. LSTM introduces a memory cell with three gates (input gate, forget gate, and output gate) to control the flow of information (Hochreiter & Schmidhuber, 1997). Specifically, the input gate controls the weight of information (current input and previous state) to be added into the cell state. The function of the forget gate is to determine what information to discard from the cell state. Lastly, the output gate allows the model to selectively output information from the current cell state.

Note that, ARIMA is a univariate model and often deals with a single time series, whereas VAR and LSTM are multivariate models allowing for modeling multiple time series simultaneously. Thus, we apply ARIMA to model the mobility patterns in each neighborhood individually, while VAR and LSTM are implemented for all neighborhoods at once.

4.3. Implementation details

To implement all the models, we split the data into three sets – training, validation, and test sets – based on the chronological order to preserve temporal dependencies in the data. The training set is utilized to train a model, the validation set is applied to tune the model hyperparameters, and the test set is leveraged to evaluate the predictive capability of the model (Hastie, Tibshirani, Friedman, & Friedman, 2009). From the collected data, we use the first nine weeks as the training set, the week ten as the validation set, and the last two weeks as the test set. To identify the best configuration of the model, we adopt the grid search strategy, which is a widely used method for hyperparameter tuning (Hastie et al., 2009). The proposed model has three parameters to be optimized: (a) the size of lower-dimensional matrix k, (b) the regularization term λ, and (c) the number of singular values r. For each combination of model parameters, the model is trained on the training set and evaluated on the validation set. The optimal combination of hyperparameters that yields the best model performance is selected as the final model. Following that, the final model is evaluated on the test set. For model evaluation, three widely-adopted statistical metrics are utilized, namely, mean absolute error (MAE), root mean square error (RMSE), and R^2, which are given by

$$\text{MAE} = \frac{1}{|\Omega'|}\sum_{(i,t)\in\Omega'}|y_{it} - \hat{y}_{it}| \tag{13}$$

$$\text{RMSE} = \sqrt{\frac{1}{|\Omega'|}\sum_{(i,t)\in\Omega'}(y_{it} - \hat{y}_{it})^2}. \tag{14}$$

$$R^2 = 1 - \frac{\sum_{(i,t)\in\Omega'}(y_{it} - \hat{y}_{it})^2}{\sum_{(i,t)\in\Omega'}(y_{it} - \bar{y}_{it})^2} \tag{15}$$

Note that, y_{it} and $\hat{y}_{it}$ are the actual and predicted values, $\bar{y}_{it}$ represents the mean of the actual values. and $|\Omega'|$ is the cardinality of the corresponding set. The smaller values of MAE and RMSE, or the larger values of R^2, indicate the better predictive performance of the model.

5. Results

In this section, we discuss the results from our numerical experiments. The predictive performance of our proposed model and multiple benchmark models is reported in Section 5.1. Then, we present the sensitivity analysis and robustness checks of the proposed model in Section 5.2 and Section 5.3 respectively. All the numerical analyses are performed using Python 3.9.13 on a 64-bit Dell Precision 3650 Tower with an 11th Gen Intel(R) Core(TM) i7-11700K processor running at 3.60 GHz.

5.1. Model performance

Table 2 shows the predictive performance of models for one-day ahead forecast on the test set. The values in boldface indicate the best predictive performance. We observe that our model performs the best (with the smallest MAE and RMSE, and the largest R^2) in comparison with the benchmark models. Note that the MAE for our model is 33.8, meaning that the predicted number of people visiting the gas stations in each CBG is off by 33.8 on average from the observed number of people visiting the gas stations. It can also be observed that the Naive forecast performs the worst, since it only takes the mean of historical data for future forecasts. Compared with the linear models (i.e., ARIMA and VAR), our model can better capture the nonlinear patterns in the data. Specifically, our model is 20% and 9% better than ARIMA and VAR respectively in terms of predictive performance measured by MAE. Compared with the LSTM, the results show that our model is better suited for multivariate time series data, as it is capable of capturing the interdependencies among other variables. Our model is 10% better than LSTM with respect to predictive performance measured by MAE. In addition, we performed statistical t-test and showed that the average performance of the proposed model is statistically better than all the other models ($p < 0.05$).

We also plot the distribution of predictive errors on the test set for all the models, as shown in Fig. 4. The distribution of errors indicate the uncertainty in the model prediction. It can be observed that LSTM produces few outliers that are outside of the box plots. Naive forecast and ARIMA tend to have the larger spread of errors measured by the interquartile range, suggesting these models have a higher uncertainty

Table 2
Predictive performance comparisons between benchmark models and the proposed model in terms of MAE, RMSE, and R^2: mean (standard deviation).

	Naive	ARIMA	VAR	LSTM	Our model
MAE	49.17 (11.61)	41.19 (12.97)	36.96 (10.85)	37.76 (9.97)	**33.80 (9.00)**
RMSE	73.04 (19.42)	61.32 (18.84)	54.86 (15.57)	57.09 (14.24)	**50.86 (13.61)**
R^2	0.79 (0.11)	0.89 (0.09)	0.91 (0.06)	0.89 (0.1)	**0.93 (0.04)**

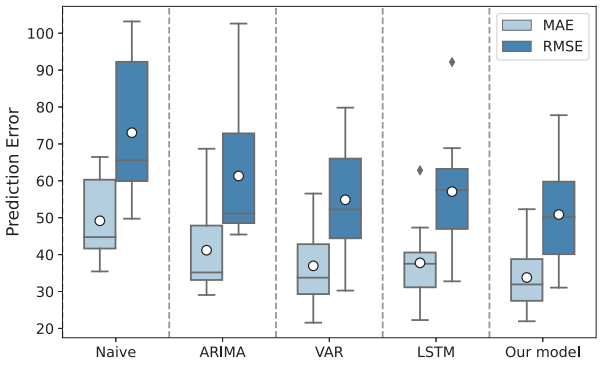


Fig. 4. Box plots for predictive performance of benchmark models and the proposed model. The hollow dots indicate the mean values.

in their predictions and are less stable in providing accurate predictions. On the other hand, our proposed model has the smallest variance for MAE, RMSE, and R^2. The low variance typically indicates that the model is more stable in predictions with less uncertainty, which is desirable for accurate and reliable demand forecasts.

To further check the indispensability of the matrix decomposition in the first stage, we implemented an ablation experiment by removing the first stage from our two-stage framework. Specifically, we only use the DMD in the second stage for time-series forecasting. The results from the DMD model show that the MAE, RMSE, and R^2 are 40.35, 58.74, and 0.9 respectively. In contrast, our two-stage forecasting framework, which integrates both matrix decomposition and the DMD model for time-series prediction, can achieve better forecasting results with 33.80, 50.86, and 0.93 for MAE, RMSE, and R^2 respectively. This confirms the necessity of including decomposition as the first stage to further enhance forecast results. As a result, the proposed two-stage forecasting framework and its various components should be implemented altogether to fully realize its potential for time-series service demand forecasting.

5.2. Sensitivity analysis

We performed sensitivity analysis to examine the performance of the proposed method with respect to changes in a model parameter, while holding the other parameters constant. Fig. 5 displays the one-way sensitivity analysis for three model parameters: k, λ, and r. The vertical axis indicates the predictive performance measured by MAE on the validation set. The optimal values of model parameters are marked as the vertical dashed lines (i.e., $k^* = 30$, $\lambda^* = 100$, $r^* = 20$), which are obtained from the grid search strategy.

We observe that the predictive error initially increases and subsequently decreases considerably, as the value of k increases. This indicates adding more latent factors could enhance the model performance, but this improvement becomes less significant as k grows. Similarly, it can be noticed that the predictive error of the model decreases rapidly as the value of λ rises to 100, and then stabilizes around 32.5 irrespective of the increase in λ. For parameter r, it can be seen that the predictive error first decreases and then increases, as the growing of r. When the value of r gets larger, the model tends to have a higher prediction error, suggesting that the model is prone to overfitting the data.

5.3. Robustness check

The purpose of robustness check is to assess the robustness of the model's performance to different settings, and to provide greater confidence in the reliability of the model's conclusions. We first evaluated the model predictive performance by adjusting the forecast horizon from the one-day ahead to ten-day ahead forecasts. Here, the forecast horizon refers to the length of time into the future for which a model is used to generate forecasts (Box et al., 2015). We choose the unit of forecast horizon as consecutive days, as it could help businesses make informed operational decisions (e.g., inventory management) and strategic planning (e.g., pricing).

Fig. 6 displays the model performance with the changing of forecast horizons. It can be seen that the predictive accuracy decreases (reflected by the increasing of MAE), as forecast horizon gets larger. This is because, in general, longer forecast horizons are associated with more uncertainty where potential changes may occur in the system over a longer period. But still, our proposed model outperforms other models for all forecast horizons, suggesting that the proposed forecasting framework can be well suited for different forecast horizons.

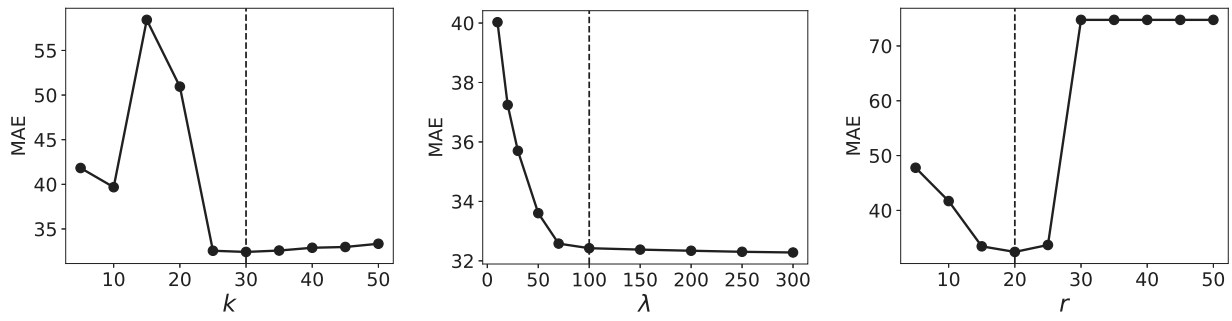
Next, to check whether the residuals of our model (defined as the difference between actual values and predicted values) exhibit spatial autocorrelation or not, we calculated the Moran's I statistic (Li, Calder, & Cressie, 2007). Specifically, Moran's I measures the degree of spatial similarity among residuals in the study region, which is given by

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{16}$$

Here, n is the total number of neighborhoods, w_{ij} is a measure of spatial proximity between neighborhoods i and j, x_i (x_j) is the residual of the neighborhoods i (j), and $\bar{x}$ is a mean of residuals. We calculated w_{ij} using the inverse of the Haversine distance between the centroids of neighborhoods. Note that, the value of Moran's I ranges between -1 and $+1$. In particular, the positive (negative) value indicates positive (negative) spatial autocorrelation, meaning that the residuals across neighboring locations tend to have similar (dissimilar) values. The value of 0 indicates no spatial autocorrelation, so that the residuals are independent across neighborhoods. The Moran's I for our model residuals is found to be 0.0004, suggesting that there is no spatial autocorrelation in the residuals. This satisfies the assumption of independence of the error terms in the model, which could produce unbiased estimates of model parameters (Li et al., 2007). In addition, Fig. 7 displays the spatial distribution of model residuals in our study region, where the coded colors are based on the quantiles of the residuals. The positive residuals marked as dark red indicate that predicted values are less than actual values, meaning that the model has underestimated the number of people visiting gas stations. The negative residuals marked as light red indicate that the model has overestimated the actual number of people in access to gas stations. Overall, we can observe that the patterns are random and do not exhibit clustered patterns, which confirms the absence of spatial autocorrelation in the residuals.

6. Discussion

This paper contributes to the existing literature related to demand forecasts by developing a novel two-stage predictive framework to provide accurate forecasts for service demands, leveraging large-scale



(a) k: size of latent factors (b) λ: regularization (c) r: number of singular values

Fig. 5. Sensitivity analysis for model parameters.

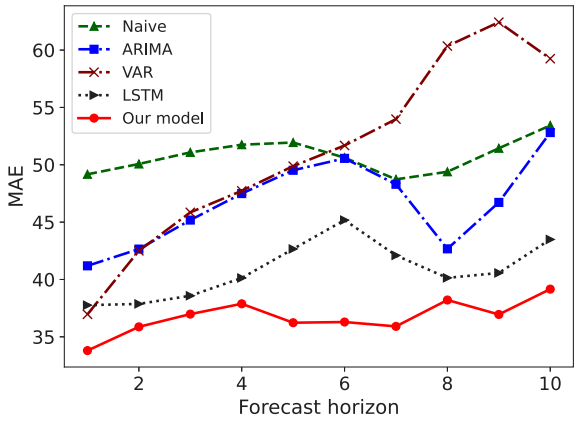


Fig. 6. Predictive performance over different forecast horizons.

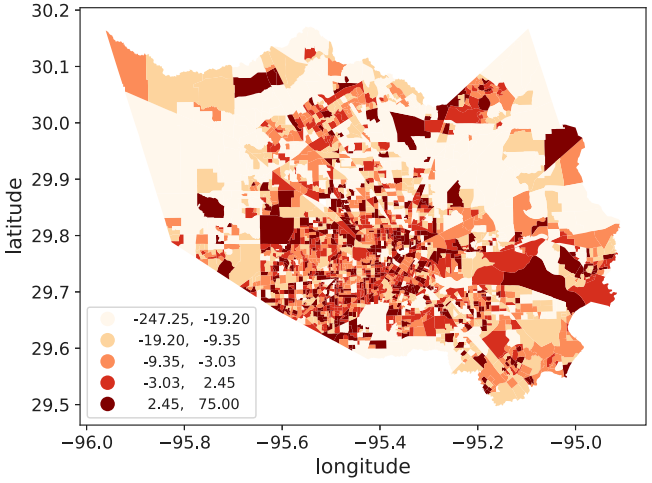


Fig. 7. Spatial distribution of the residuals of the proposed model. The legend indicates the quantiles of the residuals, which are defined as the differences between actual and predicted values of the proposed model. The darker (lighter) shades indicate the regions where the model underestimates (overestimates) the actual values.

and complex human mobility information. In a nutshell, the proposed framework has several benefits. On the one hand, the first stage involves reducing the input spatiotemporal data into smaller pieces while still maintaining data fidelity. This decomposition technique can extract the main characteristics (i.e., latent spatial and temporal factors) from the original data, thereby preventing the overfitting issue and mitigating computational costs. On the other hand, the second stage involves utilizing multivariate time series modeling to explore the complex spatiotemporal dependencies of the service demands. By taking into account the interdependencies of multiple time series patterns, the model's predictive performance could be further enhanced. The synergy of these two stages is capable of making an accurate and timely prediction by learning nonlinear data with complex spatiotemporal dependencies.

This work also has practical implications. The proposed forecasting framework can produce more accurate mobility forecasts, which could play a vital role in information management by enabling effective resource allocation, supporting planning and decision making, and optimizing supply chain operations. The improved predictive accuracy could significantly help decision makers to better understand the service demand of people in visiting critical facilities, which is the key for supply chain Marra, Ho, and Edwards (2012). For example, if the demand for a service is projected to increase in the area in the near future, businesses can utilize this forecast to expand their operations to meet the increased demand. Similarly, emergency planners would significantly benefit from more accurate forecasts, as it would allow them to pre-allocate relief resources such as foods and medicines to the areas with high demands, potentially saving lives in the face of natural disaster (Fuqua & Hespeler, 2022). In addition, sensitivity analysis is performed to examine the sensitivity of the model with respect to the change of model parameters. This could guide decision makers to select the optimal values in the model by trading off between accuracy and efficiency. For instance, exploring more latent factors in the data may result in a smaller reconstruction error, but may also require more computational resources. By and large, this work can be considered one of the key building blocks to help businesses shift from the traditional style of demand sensing to a more hybrid approach that utilizes advanced predictive analytics for demand forecasting.

7. Conclusions and future work

Understanding the service demand of people in visiting critical facilities at finer spatiotemporal resolutions plays a critical role in resource management. But, it is challenging to characterize people's service demand behaviors due to data unavailability and lack of sophisticated predictive techniques. To address these challenges, this study provides a new venue to characterize people's service demand to critical facilities by their mobility patterns. Specifically, we propose a two-stage forecasting framework that integrates data decomposition and multivariate time series analysis to provide better mobility forecasts to critical facilities at the neighborhood level. Our model demonstrates the superior performances over multiple benchmark models including ARIMA, VAR, and LSTM, in terms of predictive accuracy and stability. The proposed framework could help businesses harness the value of mobility information of customers to better analyze and predict service demand patterns, which could support informed decision-making in the allocation of resources.

In this study, we recognize certain limitations that could be studied in the future research. On the one hand, this paper collects the human mobility data in Houston, Texas, and provides a case study to forecast mobility patterns to gas stations. Future work could expand the use of the proposed framework to other regions and other critical facilities (or services) of interest, which may provide additional insights on the heterogeneities of the service demands in different areas. Comparing the demand patterns of neighborhoods with varying socioeconomic characteristics (e.g., low-income vs high-income) would help governments better allocate resources and funding to mitigate the existing socioeconomic disparities among population groups. On the other hand, even though human mobility data has merits to reveal people's dynamic movement behaviors at critical facilities at finer spatiotemporal resolutions, it also exhibits some limitations. Mobility data is often passively collected from mobile phone users, which may not reveal the personal information such as opinions and emotions regarding the services. Mobility data may not be fully observed for all segments of the population. Thus, future research could consider the integration of multiple sources of data such as human mobility data, social media data, and sales information to further explore the potential of enhancing predictive accuracy. The fusion of different data sources often requires the development of more advanced Artificial Intelligence (AI) technology for business organizations. Another potential direction for future work would be to further extend our framework by examining how comparative methods can affect the model performance through comparative analysis.

CRediT authorship contribution statement

Zhiyuan Wei: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Sayanti Mukherjee:** Conceptualization, Methodology, Writing – review & editing, Supervision.

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.

Funding

(1) National Science Foundation
Grant/Award Number: Award#2324616; Award#2308524
(2) Center for Geohazards Study, University at Buffalo (SUNY)

References

Arunraj, N. S., & Ahrens, D. (2015). A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. *International Journal of Production Economics*, *170*, 321–335.
Au, K.-F., Choi, T.-M., & Yu, Y. (2008). Fashion retail forecasting by evolutionary neural networks. *International Journal of Production Economics*, *114*(2), 615–630.
Belavina, E. (2021). Grocery store density and food waste. *Manufacturing & Service Operations Management*, *23*(1), 1–18.
Bi, X., Adomavicius, G., Li, W., & Qu, A. (2022). Improving sales forecasting accuracy: A tensor factorization approach with demand awareness. *INFORMS Journal on Computing*, *34*(3), 1644–1660.
Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.
Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.
Brelsford, C., Moehl, J., Weber, E., Sparks, K., Tuccillo, J. V., & Rose, A. (2022). Spatial and temporal characterization of activity in public space, 2019–2020. *Scientific Data*, *9*(1), 379.
Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. *European Journal of Operational Research*, *184*(3), 1140–1154.
Cavalcante, I. M., Frazzon, E. M., Forcellini, F. A., & Ivanov, D. (2019). A supervised machine learning approach to data-driven simulation of resilient supplier selection in digital manufacturing. *International Journal of Information Management*, *49*, 86–97.
Chen, X., He, Z., & Sun, L. (2019). A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation. *Transportation Research Part C: Emerging Technologies*, *98*, 73–84.
Chen, X., He, Z., & Wang, J. (2018). Spatial-temporal traffic speed patterns discovery and incomplete data recovery via SVD-combined tensor decomposition. *Transportation Research Part C: Emerging Technologies*, *86*, 59–77.
Chen, F., & Ou, T. (2009). Gray relation analysis and multilayer functional link network sales forecasting model for perishable food in convenience store. *Expert Systems with Applications*, *36*(3), 7054–7063.
Chen, X., & Sun, L. (2021). Bayesian temporal factorization for multidimensional time series prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 4659–4673.
Cheng, Z., Trepanier, M., & Sun, L. (2022). Real-time forecasting of metro origin-destination matrices with high-order weighted dynamic mode decomposition. *Transportation Science*, *56*(4), 904–918.
Choi, T.-M., Yu, Y., & Au, K.-F. (2011). A hybrid SARIMA wavelet transform method for sales forecasting. *Decision Support Systems*, *51*(1), 130–140.
Cui, R., Gallino, S., Moreno, A., & Zhang, D. J. (2018). The operational value of social media information. *Production and Operations Management*, *27*(10), 1749–1769.
Du, B., Zhou, W., Liu, C., Cui, Y., & Xiong, H. (2019). Transit pattern detection using tensor factorization. *INFORMS Journal on Computing*, *31*(2), 193–206.
Fan, Z., Song, X., & Shibasaki, R. (2014). CitySpectrum: A non-negative tensor factorization approach. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing* (pp. 213–223).
Fanaee-T, H., & Gama, J. (2016). Tensor-based anomaly detection: An interdisciplinary survey. *Knowledge-Based Systems*, *98*, 130–147.
Ferreira, K. J., Lee, B. H. A., & Simchi-Levi, D. (2016). Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management*, *18*(1), 69–88.
Fildes, R., Ma, S., & Kolassa, S. (2022). Retail forecasting: Research and practice. *International Journal of Forecasting*, *38*(4), 1283–1318.
Fuqua, D., & Hespeler, S. (2022). Commodity demand forecasting using modulated rank reduction for humanitarian logistics planning. *Expert Systems with Applications*, *206*, Article 117753.
Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.
Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction, vol. 2*. Springer.
Herm, L.-V., Heinrich, K., Wanner, J., & Janiesch, C. (2023). Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. *International Journal of Information Management*, *69*, Article 102538.
Hillar, C. J., & Lim, L.-H. (2013). Most tensor problems are NP-hard. *Journal of the ACM*, *60*(6), 1–39.
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.
Houston State of Health (2023). Households without a vehicle.
Kargas, N., Qian, C., Sidiropoulos, N. D., Xiao, C., Glass, L. M., & Sun, J. (2021). Stelar: spatio-temporal tensor factorization with latent epidemiological regularization. In *Proceedings of the AAAI conference on artificial intelligence, vol. 35, no. 6* (pp. 4830–4837).
Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, *51*(3), 455–500.
Li, H., Calder, C. A., & Cressie, N. (2007). Beyond Moran's I: testing for spatial dependence based on the spatial autoregressive model. *Geographical Analysis*, *39*(4), 357–375.
Liu, Y., Liu, C., Lu, X., Teng, M., Zhu, H., & Xiong, H. (2017). Point-of-interest demand modeling with human mobility patterns. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 947–955).
Logan, T. M., & Guikema, S. D. (2020). Reframing resilience: Equitable access to essential services. *Risk Analysis*, *40*(8), 1538–1553.
Loureiro, A. L., Miguéis, V. L., & da Silva, L. F. (2018). Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decision Support Systems*, *114*, 81–93.
Marcelo, G.-T., Constance, B., Joseph, M., Kay, A., David, Z., Maarten, V. S., et al. (2022). Do we have enough recreational spaces during pandemics? An answer based on the analysis of individual mobility patterns in Switzerland. *Landscape and Urban Planning*, *221*, Article 104373.
Marra, M., Ho, W., & Edwards, J. S. (2012). Supply chain knowledge management: A literature review. *Expert Systems with Applications*, *39*(5), 6103–6110.
Nguyen, H. D., Tran, K. P., Thomassey, S., & Hamad, M. (2021). Forecasting and anomaly detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management. *International Journal of Information Management*, *57*, Article 102282.

Nie, Y., Yang, W., Chen, Z., Lu, N., Huang, L., & Huang, H. (2021). Public curb parking demand estimation with poi distribution. *IEEE Transactions on Intelligent Transportation Systems*, *23*(5), 4614–4624.

NOAA (2021). National oceanic and atmospheric administration: Climate data online.

Osadchiy, N., Gaur, V., & Seshadri, S. (2013). Sales forecasting with financial indicators and experts' input. *Production and Operations Management*, *22*(5), 1056–1076.

Papanagnou, C. I., & Matthews-Amune, O. (2018). Coping with demand volatility in retail pharmacies with the aid of big data exploration. *Computers & Operations Research*, *98*, 343–354.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Taieb, S. B., et al. (2022). Forecasting: theory and practice. *International Journal of Forecasting*.

Punia, S., & Shankar, S. (2022). Predictive analytics for demand forecasting: A deep learning-based decision support system. *Knowledge-Based Systems*, *258*, Article 109956.

Ray, A., Jank, W., Dutta, K., & Mullarkey, M. (2023). An LSTM+ model for managing epidemics: Using population mobility and vulnerability for forecasting COVID-19 hospital admissions. *INFORMS Journal on Computing*.

SafeGraph (2020). Maintaining a high quality, global POI database is hard.

Sagaert, Y. R., Aghezzaf, E.-H., Kourentzes, N., & Desmet, B. (2018). Tactical sales forecasting using a very large set of macroeconomic indicators. *European Journal of Operational Research*, *264*(2), 558–569.

Schaer, O., Kourentzes, N., & Fildes, R. (2019). Demand forecasting with user-generated online information. *International Journal of Forecasting*, *35*(1), 197–212.

Seyedan, M., & Mafakheri, F. (2020). Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. *Journal of Big Data*, *7*(1), 1–22.

Suryani, E., Chou, S.-Y., & Chen, C.-H. (2010). Air passenger demand forecasting and passenger terminal capacity expansion: A system dynamics framework. *Expert Systems with Applications*, *37*(3), 2324–2339.

Terroso-Saenz, F., Flores, R., & Muñoz, A. (2022). Human mobility forecasting with region-based flows and geotagged Twitter data. *Expert Systems with Applications*, *203*, Article 117477.

Tiebout, C. M. (1956). A pure theory of local expenditures. *Journal of Political Economy*, *64*(5), 416–424.

Tsay, R. S. (2013). *Multivariate time series analysis: With R and financial applications*. John Wiley & Sons.

Tu, J. H., Rowley, C. W., Luchtenburg, D. M., Brunton, S. L., & Kutz, J. N. (2014). On dynamic mode decomposition: Theory and applications. *Journal of Computational Dynamics*, [ISSN: 2158-2491] *1*(2), 391–421. http://dx.doi.org/10.3934/jcd.2014.1.391.

US Census Bureau (2020). American community survey (ACS). URL https://www.census.gov/programs-surveys/acs.

Vallés-Pérez, I., Soria-Olivas, E., Martínez-Sober, M., Serrano-López, A. J., Gómez-Sanchís, J., & Mateo, F. (2022). Approaching sales forecasting using recurrent neural networks and transformers. *Expert Systems with Applications*, *201*, Article 116993.

Van Steenbergen, R., & Mes, M. R. (2020). Forecasting demand profiles of new products. *Decision Support Systems*, *139*, Article 113401.

Waddington, T. B., Clarke, G. P., Clarke, M., & Newing, A. (2018). Open all hours: Spatiotemporal fluctuations in UK grocery store sales and catchment area demand. *The International Review of Retail, Distribution and Consumer Research*, *28*(1), 1–26.

Wang, Y., Currim, F., & Ram, S. (2022). Deep learning of spatiotemporal patterns for urban mobility prediction using big data. *Information Systems Research*, *33*(2), 579–598.

Wei, Z., & Mukherjee, S. (2022). Mapping human mobility variation and identifying critical services during a disaster using dynamic mobility network. In *IIE annual conference. proceedings* (pp. 1–6). Institute of Industrial and Systems Engineers (IISE).

Wei, Z., & Mukherjee, S. (2023). Examining income segregation within activity spaces under natural disaster using dynamic mobility network. *Sustainable Cities and Society*, Article 104408.

Wei, Z., Narin, A. B., & Mukherjee, S. (2022). Multidimensional population health modeling: A data-driven multivariate statistical learning approach. *IEEE Access*, *10*, 22737–22755.

Xu, X., Qi, Y., & Hua, Z. (2010). Forecasting demand of commodities after natural disasters. *Expert Systems with Applications*, *37*(6), 4313–4317.

Yang, X., Xue, Q., Ding, M., Wu, J., & Gao, Z. (2021). Short-term prediction of passenger volume for urban rail systems: A deep learning approach based on smart-card data. *International Journal of Production Economics*, *231*, Article 107920.

Yu, H.-F., Rao, N., & Dhillon, I. S. (2016). Temporal regularized matrix factorization for high-dimensional time series prediction. *Advances in Neural Information Processing Systems*, *29*.

Zhu, X., Ninh, A., Zhao, H., & Liu, Z. (2021). Demand forecasting with supply-chain information and machine learning: Evidence in the pharmaceutical industry. *Production and Operations Management*, *30*(9), 3231–3252.