Annals of Operations Research (2024) 339:297–328
https://doi.org/10.1007/s10479-022-04838-6

ORIGINAL RESEARCH

Deep-learning model using hybrid adaptive trend estimated series for modelling and forecasting sales

Md. Iftekharul Alam Efat[1] · Petr Hajek[2] · Mohammad Zoynul Abedin[3] · Rahat Uddin Azad[1] · Md. Al Jaber[1] · Shuvra Aditya[1] · Mohammad Kabir Hassan[4]

Accepted: 14 June 2022 / Published online: 1 July 2022

Abstract
Existing sales forecasting models are not comprehensive and flexible enough to consider dynamic changes and nonlinearities in sales time-series at the store and product levels. To capture different big data characteristics in sales forecasting data, such as seasonal and trend variations, this study develops a hybrid model combining adaptive trend estimated series (ATES) with a deep neural network model. ATES is first used to model seasonal effects and incorporate holiday, weekend, and marketing effects on sales. The deep neural network model is then proposed to model residuals by capturing complex high-level spatiotemporal features from the data. The proposed hybrid model is equipped with a feature-extraction component that automatically detects the patterns and trends in time-series, which makes the forecasting model robust against noise and time-series length. To validate the proposed hybrid model, a large volume of sales data is processed with a three-dimensional data model to effectively support business decisions at the product-specific store level. To demonstrate the effectiveness of the proposed model, a comparative analysis is performed with several state-of-the-art sales forecasting methods. Here, we show that the proposed hybrid model outperforms existing models for forecasting horizons ranging from one to 12 months.

Keywords Machine learning · Sales forecasting · Big data · Regression model · Deep learning

✉ Mohammad Zoynul Abedin
m.abedin@tees.ac.uk

Md. Iftekharul Alam Efat
iftekhar.efat@gmail.com

Petr Hajek
petr.hajek@upce.cz

[1] Institute of Information Technology, Noakhali Science and Technology University, Noakhali, Bangladesh

[2] Science and Research Centre, Faculty of Economics and Administration, University of Pardubice, Studentska 84, 53210 Pardubice, Czech Republic

[3] Department of Finance, Performance and Marketing, Teesside University International Business School, Teesside University, Middlesbrough, Tees Valley TS1 3BX, UK

[4] Department of Economics and Finance, University of New Orleans, New Orleans, LA 70148, USA

1 Introduction

Sales forecasting is a prediction process based on previous sales history, purchasing behaviour, promotional activities, and expected market conditions (Ma et al., 2016). Proper sales forecasting outlines demand for a product and also helps in decreasing the costs of storing inventory; at the same time, it reduces the probability that stock-outs and long customer wait times will occur (Pavlyshenko, 2018). This has a direct effect on a company's profits. Timely and accurate sales forecasts are crucial in bridging the gap between supply and demand (Gahirwal, 2013). Sales forecasting helps companies define their plans for sales and operations, and the improved management performance positively affects investors and results in increased company value. Poor sales forecasting may lead to insufficient or overstocked inventories and the failure to satisfy customer needs (Chen & Lu, 2017).

Earlier research has reported several challenges in sales forecasting. An appropriate forecasting methodology may be associated with incorrect assumptions, leading to poor sales forecasts. Notably, although the shallow neural network has been viewed as a promising sales forecasting model, it has insufficient capacity for capturing seasonal and trend variations from raw datasets (Zhang & Qi, 2005). In addition, discontinuous and unstable previous data can lead to defective predictions. Indeed, forecasts for individual items usually have short and sparse life-cycles, which prevents estimations of important features such as seasonality (Jha et al., 2015). Moreover, markets are so dynamic that new products, promotional campaigns, and changes in public sentiment make it difficult to make decisions based only on the modelling of previous data (Chu & Zhang, 2003; Disney et al., 2021). Other factors should also be considered, such as weather conditions, significant events, holidays, and socio-economic conditions, which can cause multiple seasonal cycles, a high volatility in sales, and high dimensionality in data (Ganesan et al., 2019; Thomassey, 2010; Eachempati et al., 2022).

Generally, poor sales forecasting is due to three main reasons. First, an inappropriate forecasting model may be chosen; it may not be able to capture the trends and seasonality aspects of sales time series properly. Forecasting sales with such a model may also lead to stock-outs or overstocks of products (Bose, 2009). Second, sales forecasts may be based on erroneous assumptions, such as estimated predictive density and relationships among lagged values and error terms (Kolassa, 2016). Finally, unless relevant features that affect future sales are selected, forecasts may be too slack, especially in the early stages of a new product launch (Liu et al., 2013). Therefore, the first and mandatory step for proper sales forecasting is to extract the features that are strongly related to the demand for items in the dataset, particularly the trends and seasonality of product sales (Zhang & Qi, 2005).

In addition to the above issues with data, the main challenge for developing sales forecasting models with big data is to effectively model scarce and skewed data at the store and item levels (Ma & Fildes, 2021; Ulrich et al., 2021). Indeed, the huge amounts of data collected using technologies such as point-of-scale (POS) and Internet of Things (IoT) allow companies to better understand customer behaviour and improve sales forecasting performance (Boone et al., 2019; Ren et al., 2020). However, at the same time, the characteristics of big data, namely, their large volume, variety, velocity, and veracity, pose several challenging problems for sales forecasting methods. More precisely, locally optimal solutions may have a negative effect on traditional statistical and machine learning methods when one is learning how to use large-scale models; they also require a high-performance computing capacity (Zhang et al., 2018). A large velocity of sales among items and stores also requires that the models provide real-time sales forecasts. Therefore, great complexity in computations is another issue to overcome.

Numerous statistical and machine learning methods use time-series analysis to generate forecasting models (Box et al., 2015). These models make use of historical demand patterns as a baseline to predict future demands (Ma et al., 2016; Zhao and Wang, 2017).

Statistical methods used previously for sales forecasting include traditional regression or time-series models, such as the single exponential smoothing (SES) (Taylor, 2011; Teunter et al., 2011; Li & Lim, 2018), autoregressive integrated moving average (ARIMA) (Ramos et al., 2015), and Holt-Winters exponential smoothing (HWES) methods (Gelper et al., 2010). These models were based on modelling trends and seasonalities in univariate time-series. To improve the performance of these traditional statistical models, the effects of additional important drivers, such as the marketing strategy, were later incorporated (Ali & Pinar, 2016; Huang et al., 2019).

Most recently, category- and store-specific seasonality and marketing effects were combined with adaptive sales figures to obtain an interpretable retail forecasting model that considered random disturbances as well as the bias introduced by model regularization (Ali & Gürlek, 2020). However, these traditional statistical models rely on pooling observations across items and stores; as a result, distorted estimates can occur when one is considering diverse items and stores and big data. In order to consider the periodicities and holiday, weekend, and marketing effects for specific items and stores, here we propose the adaptive trend estimated series (ATES) model, which uses the data polarity spectrum for specific items and stores, discovers underlying periodicities using the spectral density function, and uses Box–Cox transformation to model changes in the sales time-series.

Machine learning models, such as artificial neural networks and extreme learning machines, have been proved to outperform the above statistical methods by building generic model structures that can easily handle complex nonlinear patterns in data (Sun et al., 2008; Loureiro et al., 2018; Kharfan et al., 2021; Li et al., 2021). Notably, deep neural networks are reported to be effective in capturing more complex features and multilevel representations from sales forecasting datasets. Specifically, recurrent neural networks such as long short-term memory (LSTM) (Weng et al., 2019) and gated recurrent unit (GRU) neural networks (Noh et al., 2020) automatically capture high-level temporal features from big data to accurately forecast sales. These deep neural networks were specifically designed to handle sequential data with patterns over time.

However, the current sales forecasting datasets not only contain sequential data but also other data components, such as information about stores and items. To make use of this additional information, a spatiotemporal matrix should be produced. Recurrent neural networks on their own cannot effectively handle such datasets. The convolution neural network (CNN) is ideal for processing such data because it enables the capturing of both scale-invariant features and local trend features (Wu et al., 2020; Ma et al., 2016; Pan & Zhou, 2020). We here propose a novel GRU–CNN–LSTM architecture to take advantage of the above deep neural network models. The combination of CNN and GRU is used to extract complex high-level spatiotemporal features from big data, and LSTMs are then trained using these features to perform multistep-ahead forecasting.

To effectively capture different patterns in the underlying big data, here we first use the ATES model to capture linear patterns. Then, the extracted linear component (residues of the ATES model) is added to the nonlinear GRU–CNN–LSTM model. Using a hybrid forecasting model is considered a good strategy because it has the capabilities to capture linear and nonlinear patterns in the data (Zhang, 2003). Moreover, recent research suggests that such hybrid architectures provide forecasts that are more accurate than those estimated by either pure statistical or machine learning models (Smyl, 2020).

In summary, the contributions of this study are twofold:

- A novel ATES + GRU–CNN–LSTM sales forecasting hybrid model is proposed that is specifically designed for big data. The proposed model effectively uses the diverse spatiotemporal characteristics of items and stores present in big data. It automatically learns the three-dimensional features in the time-series data for each item and store and incorporates the effect of product promotion.
- A large benchmark dataset is used for a wide range of items and stores to demonstrate that the proposed hybrid model can achieve better forecasting results than state-of-the-art sales forecasting methods.

The rest of this paper is structured as follows. Section 2 introduces a brief review of the literature on the state-of-the-art methods used for sales forecasting. Section 3 presents techniques for data pre-processing and modelling. Section 4 explains in detail the proposed forecasting models. The empirical results for sales forecasting with big data are shown in Sect. 5. Section 6 presents discussions of the empirical results. Finally, Sect. 7 offers conclusions reached by the study and outlines future research directions.

2 Literature review

Accurate sales forecasting has been increasingly recognized as a major concern for companies in the supply chain because markets are becoming more competitive and companies need to be more cost-efficient to achieve a competitive advantage. Production schedules need to be precisely planned because changes in production cause higher costs. Further, many internal and external factors may have a significant impact on sales, but their effects are usually difficult to anticipate. Sales forecasting helps companies spot and analyze those factors and, thus, gives a company time to make decisions and effectively plan its sales and operations.

Sales forecasting methods can be categorized as qualitative or quantitative. Qualitative methods tend to be expensive and biased owing to the forecaster's limited capacity for processing complex business information. For sales forecasting with big data, quantitative methods are more appropriate because they make use of the information stored in a company's databases to produce frequent forecasts for a large number of items and stores. There are two main categories of quantitative forecasting methods, i.e., statistical time-series methods and machine learning methods. Statistical time-series methods mainly rely on a trend analysis of sales data to generate forecasts. Machine learning methods usually rely on a supervised mode; they minimize the forecasting error for historical training data by learning an accurate data representation from multiple inputs. The output is a prediction based on out-of-sample observations. Here, we review a large and growing body of studies that have investigated various quantitative methods for sales forecasting.

2.1 Sales forecasting using statistical time-series methods

In the past, the conventional method of forecasting sales was via statistical time-series methods. In the last decade, many studies have focused on statistical time-series methods, including linear models (using linear functional form), such as ARIMA (Ramos et al., 2015), seasonal ARIMA (SARIMA) (Arunraj & Ahrens, 2015), and HWES (Gelper et al., 2010), and nonlinear models, such as generalized autoregressive conditional heteroscedasticity (GARCH) (Chen & Ou, 2011) and the Markov regime-switching model (Choi et al., 2011). The forecasts of these methods were sometimes combined to obtain more accurate predictions (Misiorek et al., 2006). In particular, ARIMA and its variations were widely applied in the sales fore-

casting domain (Box et al., 2015). Recently, several hybrid methods have been developed that combine ARIMA with nonlinear methods to improve performance. For example, a blended probabilistic model was proposed that expanded the ARIMA and XGBoost algorithms to predict store sales; it outperformed the individual ARIMA and XGBoost models (Pavlyshenko, 2016). Similar findings were observed for the combination of ARIMA with artificial neural networks (Zhang, 2003).

By extending the exponential smoothing formulation, Taylor developed five univariate exponentially weighted methods for forecasting intraday time-series sales data to deal with both intraweek and intraday seasonal cycles (Taylor, 2010). The methods make use of discount weighted regression, time-varying splines, and singular value decomposition techniques. However, these methods may be prone to judgmental biases in producing clusters of intraweek periods within the seasonal exponential smoothing and a selection of knots in the regression splines.

When considering trends in consumer behaviour, as well as seasonal and holiday effects on time-series data, Taylor and Letham (2018) introduced a modular regression algorithm named Prophet, by which the unpredicted nature in time-series data can be calculated. Domain experts can easily adjust the Prophet model's parameters. Later, Navratil and Kolkova (2019) used the Prophet algorithm to identify seasonal trends in revenue development in the e-commerce industry. To account for the seasonal trend, holidays and seasonal changes were adjusted before fitting the model. Indeed, the Prophet model allows for the adjustments needed for scalable forecasting. However, the selection of relevant model components and the manual performance control require an analyst's experience and domain knowledge.

Scalable forecasting has become increasingly important since the emergence of big data analytics as a tool to model sales data features. Notably, it has become easier for modellers to account for the outliers, trends, seasonality, or stationarity in sales time-series with big data. However, to make the most of the potential of big data, it is important to learn from large-scale datasets. Harsoor and Patil (2015) used the Hadoop MapReduce algorithm to handle large volumes of sales data and execute batch processing by item representation. Afterward, Tehrani and Ahrens (2016) exploited the K-means clustering algorithm to pinpoint homogeneous groups of items in big data forecasting.

Ma et al. (2016) developed a LASSO (least absolute shrinkage and selection operator) regression model equipped with Granger causality testing and a feature extraction component to overcome the problem of high dimensionality incurred by using sales history, promotional information, and time event information. This method improved the accuracy of the baseline model by 12.6%. However, this approach is accurate and reliable only when a sufficiently long time-series is available. Moreover, this method is computationally efficient at the cost of discarding potentially useful factors. Similarly, Sagaert et al. (2018b) used multiple LASSO models to perform multistep-ahead sales forecasts. LASSO produced reliable forecasts in the case of high dimensional problems.

In summary, previous studies found that larger datasets, in particular those with an increasing number of observations per item, decreased forecasting errors. In addition, a larger diversity in stores and items resulted in a more reliable and robust forecasting performance. Despite the fact that statistical time-series methods are popular in the related literature, these forecasting models may have issues. First, the feature engineering is manual (except in LASSO) and to a large degree relies on the experience of the model builder. Second, these methods can be compromised when extrapolating the latent patterns present in big data. More precisely, linear models cannot capture common real-world features, such as occasional outlying samples or asymmetric life-cycles (Xia & Wong, 2014). Modelling complex relationships between time-series data and other determinants of sales is difficult in linear

models. The nonlinear forecasting model, on the other hand, assumes that the underlying process of generating time-series data is constant; this is not realistic, however, in the changing real-world business environment. Sales forecasting data that are nonrepetitive limit the reliable identification of model parameters. Therefore, machine learning models have emerged to solve the above problems.

2.2 Sales forecasting using neural network models

To address the above-mentioned inherent nonlinear features in sales forecasting data, machine learning models have been developed over the last decade.

The first applications of machine learning methods in sales forecasting used feed-forward neural networks and extreme learning machines (Sun et al., 2008). In this way, the high volatility of demand was effectively modelled. To improve the generalization performance of neural networks, heuristic algorithms were later introduced that fine-tuned the parameters of the sales forecasting systems (Wong & Guo, 2010), further, hybrid approaches were developed that considered relation analysis in time-series to support purchasing decisions (Chen & Ou, 2011).

To overcome the problem of diverse patterns in the groups of product items, clustering methods were used to feed the machine learning methods (e.g., support vector regression and extreme learning machine) (Chen & Lu, 2017). The curse of the dimensionality problem related to the large number of features was tackled using advanced feature selection methods (Jiménez et al., 2017). Another problem of neural network-based models is their lack of interpretability, which was addressed by machine learning models with uniform explanations produced on the level of the model and examples to enable what-if analyses (Bohanec et al., 2017). Alternatively, fuzzy sets were incorporated into the neural network model to obtain a fuzzy neural network model capable of modelling the effects of promotional marketing activities in big data-driven sales forecasting (Kumar et al., 2020). However, the problem of overfitting individual forecasting models was overcome by combining multiple diverse predictors in an ensemble forecasting model (Pavlyshenko, 2019; Ji et al., 2019).

To consider the large and highly diverse number of features affecting future sales, Loureiro et al. (2018) evaluated the performance of deep neural networks for the first time. Deep feed-forward neural networks were shown to be superior to shallow neural networks and other traditional machine learning methods (e.g., decision trees, random forest, support vector machines), in particular, for the large number of stores evaluated. This is attributed to the capacity of these deep neural networks for capturing complex multilevel features from large and high-dimensional sales forecasting datasets. The main limitation of deep feed-forward neural networks is that temporal features cannot be taken into account. Therefore, LSTM and GRU recurrent neural networks were recently used to model long and short high-level temporal patterns in sales time-series (Weng et al., 2019; Noh et al., 2020). GRU represents an enhanced (but simplified) version of LSTM because of its ability to solve the vanishing gradient problem.

Although recurrent neural networks are effective in modelling temporal components in the data, important sales determinants are present in other feature sources. Most importantly, it is necessary to consider the three-dimensional nature of the sales data, including not only the temporal component but also the dimensions related to stores and items. Therefore, spatiotemporal models based on CNN have attracted increased attention (Wu et al., 2020; Ma & Fildes, 2021; Pan & Zhou, 2020). Two CNNs were combined in a meta-learning framework proposed by (Ma & Fildes, 2021). The first CNN was used to extract sales data; the other

Table 1 State-of-the-art statistical time-series methods and neural networks for sales forecasting

Authors	Area	Method
Harsoor and Patil (2015)	Walmart stores	HWES & Hadoop
Tehrani and Ahrens (2016)	Fashion	k-Means Clustering
Ma et al. (2016)	Retail	LASSO
Zhao and Wang (2017)	E-commerce	CNN
Kechyn et al. (2018)	Grocery	WaveNet CNN
Sagaert et al. (2018b)	Tire industry	LASSO, Linear Regression
Loureiro et al. (2018)	Retail	Deep NN
Liang et al. (2019)	Product marketing	XGboost & LightGBM
Ganesan et al. (2019)	Food	Dynamic ANN
Pavlyshenko (2019)	Rossmann Store	Stacking (NN, LASSO, RF, ARIMA, ExtraTrees)
Kraus et al. (2020)	Pharmacies	Deep-embedded LSTM
Pan and Zhou (2020)	E-commerce	CNN
Ma and Fildes (2021)	Retail	CNN-based meta-learner
this study	Grocery	ATES + GRU–CNN–LSTM

CNN captured the high-level features from other influential factors. Dilated causal convolutions were used in the WaveNet CNN model (Kechyn et al., 2018). Thus, a computationally effective version of CNN was obtained that can be applied to large-volume datasets.

The advantages of recurrent neural networks and CNN were used in a GRU–CNN hybrid forecasting model (Wu et al., 2020). GRU was used to extract temporal features from the time sequence of sales, and the feature vector of other variables was utilized in the CNN module. However, GRU suffers from its gradient explosion problem originating from the nonlinear dynamics representing sequential data.

To overcome the above problems, here we propose to use the GRU–CNN component to extract the spatiotemporal features in the first phase and then apply LSTM for multistep-ahead forecasting. Thus, the number of parameters in LSTM is reduced, and, at the same time, the risk of the vanishing gradient is reduced; further, the blended GRU–CNN–LSTM architecture is more robust to noise in the data than is LSTM.

An overview of the state-of-the-art statistical time-series methods and machine learning methods is presented in Table 1.

3 Data modelling

To forecast sales with big data, we obtained a large-scale real-world dataset from Kaggle, which has details on sales of different products in different stores (Kaggle, 2018). Corporación Favorita, a large Ecuadorian grocery retailer that operates hundreds of supermarkets, maintained this dataset, which comprises roughly 125 million data samples over five years from 2013 to 2017 or an average of 70,000 per day. Each record has the date, store number, item number, unit sales, and on-promotion data, as shown in Table 2.

The dataset includes retail sales for 4,100 items and 54 stores. Each item in the dataset had its own characteristic patterns that could best be modelled independently. Grocery sales forecasting is a challenging problem for companies. Overforecasting may cause grocers to

Table 2 Data description

Attribute name	Description
id	Sale id
date	Date on which sales was done
store_nbr	Store id from which item was sold
item_nbr	Item id which was sold
unit_sales	Unit of item that was sold
on_promotion	Information of whether item was promoted or not

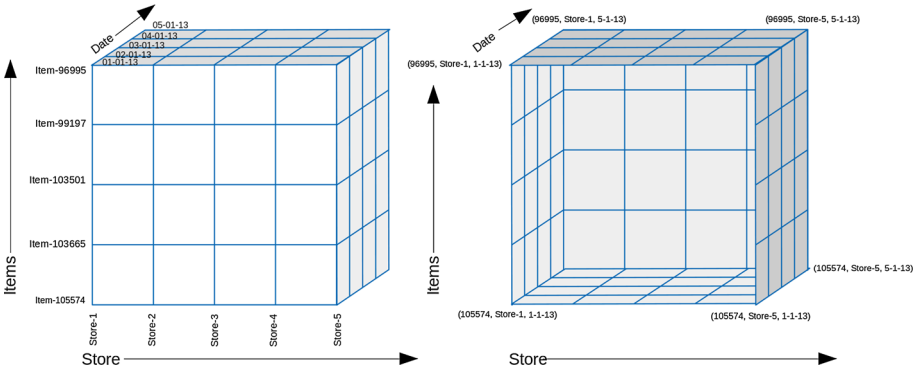


Fig. 1 Visualization of the three-dimensional data model for product-specific store-level sales forecasting. Date denotes time dimension; item represents product dimension (item); store stands for geographic dimension. This model is more realistic and enables better decisions at the store level, but additional noise must be modelled at the disaggregate level

become overstocked with perishable goods, whereas underforecasting may cause popular items to sell out quickly and potential profits to be lost. Moreover, retailers often have to provide customers with new stores and unique products to meet their needs and adapt to their seasonal tastes (Kechyn et al., 2018).

After data collection, data cleaning was performed by detecting and discarding corrupted, incorrect, and incomplete samples. Figure 1 presents a three-dimensional model by which big data can be easily evaluated and manipulated. In this model, the X, Y, and Z axes denote stores from different cities, unique items, and dates in the time span, respectively. To demonstrate the model efficiency, Fig. 2 shows the XY, YZ, and ZX plots displayed in two dimensions. Specifically, Fig. 2a describes the relation between item and store (city), Fig. 2b the relation between date and item, and Fig. 2c the relation between date and store.

To effectively model these relations, three different two-dimensional models must be implemented; this is less effective, however, particularly when dealing with big data. The three-dimensional model makes data retrieval and manipulation easier and also allows for the visualization of the dynamics of product sales and promotional, weekend, or holiday effects. This data representation also helps in the processing of data, overcoming the time-consuming performance issues of big data. Further, this data model can be fitted easily as an input to deep neural network models, thereby becoming part of the proposed GRU–CNN–LSTM architecture.

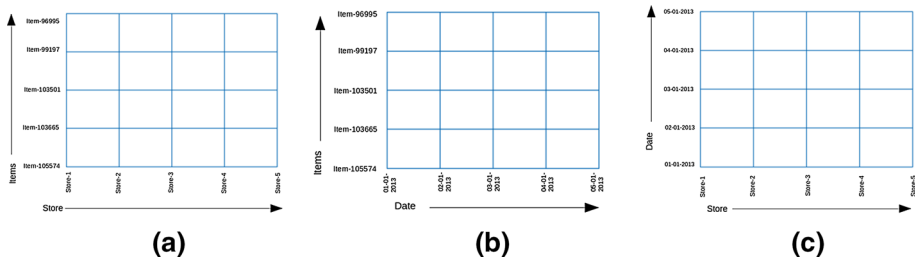


Fig. 2 Two-dimensional data models at the aggregate level: **a** aggregation of product data at store level; **b** product-level forecasting (without considering geographic conditions); and **c** store-level forecasting (not product-specific)

4 Proposed method

Sales forecasting is a traditional persuasive application of time-series methods to trend, seasonal, and cyclical factors. Nevertheless, forecasting for such time-series is challenging because of the difficulty in predicting demand, which depends on many factors. Therefore, state-of-the-art statistical time-series methods such as ARIMA, SARIMA, HWES, and PROPHET (Taylor & Letham, 2018) still perform satisfactorily on this prediction task (Liang et al., 2019). However, sales prediction is considered a regression problem as well as time-series problem due to the repetitive nature of past data. This motivates researchers to implement supervised machine learning algorithms, which often produce better results (Pavlyshenko, 2018). In this study, therefore, we propose a hybrid model specifically designed for sales forecasting, i.e., the ATES + GRU–CNN–LSTM model.

4.1 Adaptive trend estimated series (ATES)

Traditional statistical time-series methods treat sales data as multivariate panel data with shorter time-series, assuming cross-sectional variations because data come from multiple stores within a single retail chain (Ali & Gürlek, 2020). Sales S_{ijt} for the ith item and jth store at time t are considered a function of confounding variables, usually represented by the vector of seasonality variables, trend patterns, item-store fixed effects, and marketing variables.

Here, we propose to use the ATES model to deal with different seasonal effects present in business time-series. Inspired by the Prophet forecasting model, the seasonal effects are modelled in an additive manner, so that the newly added components can reveal new sources of seasonality. In addition to the flexibility in seasonality effects, the ATES model incorporates the effects of holidays and promotions, which are important determinants of retails sales.

In the proposed ATES model, the traditional multivariate regression model is combined with the spectral analysis of multivariate time-series, which helps researchers to understand the dynamics in the serially correlated data using simple nonparametric models. Spectral analysis has recently received considerable attention in big data analytics because it can be effectively applied to compare large sets of time series data (Caiado et al., 2020; von Sachs, 2020; Škare & Porada-Rochoń, 2020). The computational efficiency is achieved by condensing relevant information rather than using all of the information of the time-series. More precisely, the ATES model was outlined to consider the following data features: Box–Cox transformation, polarity spectrum, and trend periodogram.

4.1.1 Growth factor

The Box–Cox power transformation is a class of transformations used to simplify the structure of the forecasting model and obtain normal errors with constant variance. Previous empirical studies have shown that this transformation of time-series significantly improves the forecasting accuracy in the retail sales domain (Proietti & Lütkepohl, 2013). In the Box–Cox transformation, the appropriate value of parameter λ is identified, which indicates the power to which the time-series is raised:

$$G_t = \begin{cases} log(s_t) & \text{if } \lambda = 0 \\ \frac{s_t^{\lambda}-1}{\lambda} & \text{otherwise} \end{cases} \tag{1}$$

where s_t denotes the sales value at time t, and the value of parameter λ varies from -5 to $+5$. To illustrate, if $\lambda = 1$, then $G_t = s_t - 1$; therefore, the data are moved downward, but there is no transformation in the nature of the time-series (original scale is preserved). For all other values of λ, the time-series will change its shape to approximate normal distribution. Also note that, for $\lambda = 0$, we obtain the logarithmic transformation.

4.1.2 Data polarity spectrum

The time-series sales data have a habit of jumps in the sales of a store or product monthly, weekly, or after a certain period of time. This occurs frequently and depends on the human purchasing behaviour or promotional activity. Sometimes, similar patterns can be observed in other products or stores, thus indicating the nature of sales data as a polar spectrum. To illustrate, the data polarity spectrum is shown in Fig. 3 for monthly and weekly polarity.

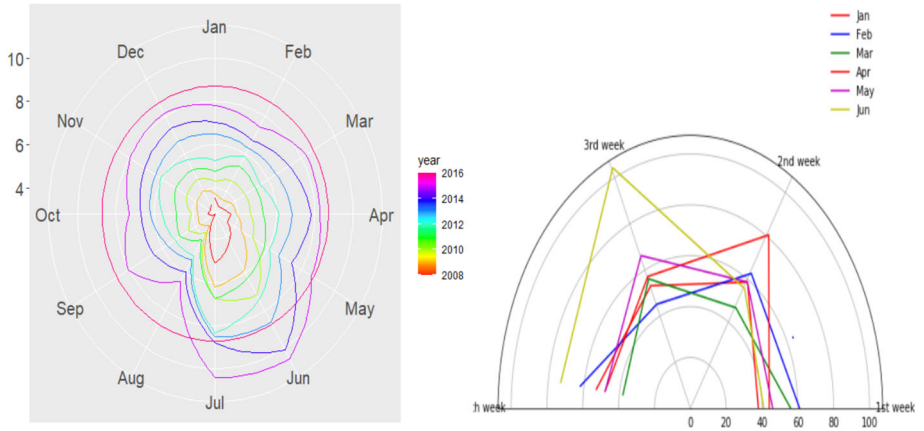
To detect the polarity spectrum, we considered the periodic moment, amplitude (height of the underlying function) with the phase (start point of the function), and a covariance stationary process. Random variation in the time-series allows the amplitude and phase to vary. Let $\{S_t\}$ be a sequence of scalar random variables, where $t = 0, \pm1, \pm2, ...$, and $E(S_t) = \mu, Cov(S_t, S_{t-k}) = \lambda_k$. Here, S_t is observed at ordered intervals such as a week or month. Also, the first and second moments of the procedure do not rely on time, that is, μ and λ_k are time-invariant. Therefore, the simple stochastic process of the polarity spectrum can be represented as:

$$S_t = \alpha cos(\omega t) + \delta sin(\omega t), \tag{2}$$

where α and δ are normally distributed random variables and ω is the angular frequency index (measured in cycles per unit time). Note that the amplitude is calculated as $A = sqrt(\alpha^2+\delta^2)$ and the phase is defined as $\phi = tan^{-1}(-\delta/\alpha)$.

4.1.3 Trend periodogram

Spectral analysis also allows us to examine underlying periodicities in the data. To investigate the periodic component of the time-series data, a periodogram is used to show the intensities (amplitude) against the data frequencies or periods (frequency characteristics). The periodogram is traditionally used to estimate the spectral density of the time-series. Considering n observations of the time-series, the trend periodogram is defined as the function of intensities $I(f_i)$ at frequency $f_i = 1/n$ or period $p_i = n/i$, that is, the ith harmonic of the primary frequency $1/n$ (Iwok, 2016). This actually determines the periodicity or seasonality



(a) Monthly Data Polarity Plot **(b)** Weekly Data Polarity Plot

Fig. 3 Polarity effect of monthly and weekly sales data

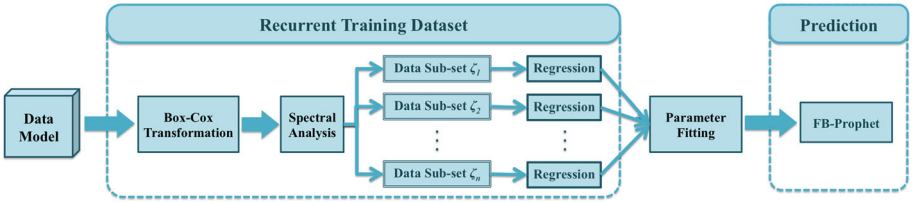


Fig. 4 Process flow of ATES model

of the time series, which also indicates the maximum peak of each period. Hence, the trend periodogram can be expressed as:

$$P_i = \sum_{i=1}^{\frac{n}{2}} I\,(f_i) \tag{3}$$

4.1.4 ATES model

This forecasting model is designed so that the periodic and polar characteristics are calculated on the transformed data to split the data into various parts; then, the forecasting model is fine-tuned and applied for each part separately, as illustrated in Fig. 4.

First, for each time unit t, the transformed value G_t is calculated using the power parameter λ determined for training data. The transformed time-series is divided into subsets as follows:

$$G_T = \{\{G_1, G_2, \ldots, G_x\}, \{G_{x+1}, G_{x+2}, \ldots, G_{x^1}\}, \ldots, \{G_{x^n+1}, G_{x^n+2}, \ldots, G_t\}\}. \tag{4}$$

Then, Eq. 4 is used to calculate the polarity spectrum for each subset of G_T, which actually determines whether the subset has the polarity nature or not. If

$$G_T^D = \{\{G_1, G_2, \ldots, G_d\}, \{G_{d+1}, G_{d+2}, \ldots, G_{d^1}\}, \ldots, \{G_{d^n+1}, G_{d^n+2}, \ldots, G_t\}\}. \tag{5}$$

Similarly, G_T is examined through the trend periodogram function, and each subset of G_T is restructured accordingly. Thus, the harmonic cycle nature of the training data is identified, which produces the periodogram set as follows:

$$G_T^P = \{\{G_1, G_2, \ldots, G_p\}, \{G_{p+1}, G_{p+2}, \ldots, G_{p^1}\}, \ldots, \{G_{p^n+1}, G_{p^n+2}, \ldots, G_t\}\} \quad (6)$$

Finally, the training data are split into n parts, noted as $\zeta = \{\zeta_1, \zeta_2, \ldots, \zeta_n\}$, using the following equation, where the nonintersect values are considered as noise or added to the next subset using the λ scaling value:

$$\zeta = G_T^D \cap G_T^P. \quad (7)$$

The kth subset is considered a "season" of time-series, where $\zeta_k = \{\{D_k, S_k\}, \{D_{k+1}, S_{k+1}\}, \ldots, \{D_{k+x}, S_{k+x}\}\}$, where D_k denotes the day for the kth subset, k=1,2,…,n.

Now, for each ζ_k, the Prophet decomposable time-series model is applied that includes the holiday, weekend, and promotional effects (Ali & Gürlek, 2020; Efat et al., 2018), as follows:

$$G_t(\zeta_k) = g_{tk} + s_{tk} + h_{tk} + p_{tk} + \epsilon_{tk}, \quad (8)$$

where G_t is the transformed sales amount in time t, g_{tk} denotes the trend function representing nonperiodic changes in the time-series, s_{tk} represents the periodic changes (seasonality), h_{tk} and p_{tk} represent the irregular effects of holidays and promotional variables, respectively, and ϵ_{tk} is the adjustable error value. For each ζ_k, the next subset is forecast, and the forecasting error E_k is calculated, which actually updates the error parameter ϵ_{tk}. Next, the ζ_k and ζ_{k+1} subset data are fed to training data to calculate the forecasting error and tune the remaining parameters in Eq. 8. This iterative process will produce forecast values, as presented in Algorithm 1.

Algorithm 1 ATES model for sales forecasting

Input: n parts of training data $\zeta = \{\zeta_1, \zeta_2, \ldots, \zeta_n\}$
Output: $forecasts \quad F = \{\{D_{t+1}, S_1\}, \{D_{t+2}, S_2\} \ldots, \{D_{t+y}, S_y\}\}$
Begin
$N \leftarrow \{1, 2, \ldots, n\}$
for each $N_i \in N$ **do**
$\zeta^T \leftarrow \zeta^T \cup \zeta_k$
$s_{tk} \leftarrow regression(\zeta_k)$
$g_{tk} \leftarrow \left(\frac{s_x^k}{s_1^k}\right)^{1/\sum x} - 1$
$h_{tk} \leftarrow Z_{ht} \times \kappa_h$
$p_{tk} \leftarrow Z_{pt} \times \kappa_p$
$F_k \leftarrow \{\{d_{k+1}, s_{k+1}\}, \ldots, \{d_{k+x}, s_{k+x}\}\}$
$F_c \leftarrow \zeta_{k+l}$
$\mathcal{E}_k \leftarrow \frac{2 \times SE_k}{\{d_{k+x+1}, s_{k+x+1}\}} \times 100\%$ $where$ $SE_k \leftarrow \frac{\sqrt{\frac{\sum_{j=k+1}^{x}(F_k - F_c)^2}{\sum_{j=k+1}^{x} x}}}{\sum_{j=k+1}^{x-1} x}$
$\xi_{k+1} \leftarrow paramFitting(\zeta_{k+1}, g_k)$
end for
$\xi \leftarrow median\{\xi_k, \xi_{k+1}, \ldots, \xi_n\}$
$M^c \leftarrow train(\zeta, \xi)$
$F \leftarrow \{\{D_{t+1}, S_1\}, \ldots, \{D_{t+y}, S_y\}\}$
End

Compared with traditional time-series models, the Prophet model does not explicitly account for the temporal features in the data. In addition, the model fits fast; it is also robust to missing values, is modular, has easily interpretable parameters, and accommodates multiple seasonality aspects. The modules of the Prophet model are defined as follows.

A nonlinear growth function g_{tk} is assumed:

$$g_{tk} = \left(\frac{S_x^k}{S_1^k}\right)^{1/\sum x} - 1. \tag{9}$$

The smooth seasonal effect is approximated using the Fourier series as follows:

$$s_t = \sum_{n=1}^{N} \left(a_n cos\left(\frac{2\pi nt}{P}\right) + b_n sin\left(\frac{2\pi nt}{P}\right)\right), \tag{10}$$

where P is the regular period, n represents the weight (number of cycles), and a_n and b_n are Fourier coefficients. The choice of n is automated using the Akaike information criterion because increasing N may result in model overfitting. This component allows for modelling seasonal patterns in the sales time-series.

The effects of holidays and promotional variables are defined as follows:

$$h_t = Z_{ht} \times \kappa_h, \tag{11}$$

$$p_t = Z_{pt} \times \kappa_p, \tag{12}$$

where Z_{ht} and Z_{pt} are vectors of dummy variables representing holiday and on-promotion dates, respectively, and κ_h and κ_p are parameters indicating a change in the forecast; priors $\kappa_h, \kappa_p \sim$ Normal (0, 5) are used (Taylor & Letham, 2018).

4.2 GRU–CNN–LSTM architecture

Inspired by the GRU–CNN and CNN–LSTM hybrid forecasting models (Wu et al., 2020; Kim & Cho, 2019), in which GRU (LSTM) captures temporal features of sales time-series and CNN is used to extract features from the spatiotemporal matrix characterizing the remaining input variables, here we propose a blended GRU–CNN–LSTM architecture, as shown in Fig. 5. By using LSTM for multistep-ahead forecasting in the second stage, the proposed model overcomes the gradient explosion problem of the GRU–CNN model and, at the same time, improves the LSTM robustness and learning effectivity by reducing its number of parameters. Here, the CNN feature extraction module works on high-dimensional data using a spatiotemporal matrix.

To reduce the dimensionality of the sales time-series, characteristic-based clustering was performed. That is, global characteristics were extracted from the data and clusters were formed based on these characteristics, rather than on the original time-series data. This not only makes the subsequent forecasts more effective but makes the forecasting model more robust to noise and missing data (Wang et al., 2009). Using this dimensionality reduction process, arbitrarily long time-series can be clustered. This is also useful for retail sales because time-series of different lengths are usually available for individual stores. Indeed, the structural-based clustering is reportedly more robust against time-series length and noise than dimensionality reduction methods that use point-level data (Wang et al., 2006). Structural characteristics of time-series, namely, the trend, seasonality, and kurtosis, were investigated using the characteristic-based clustering methods. We found that the selection of kurtosis

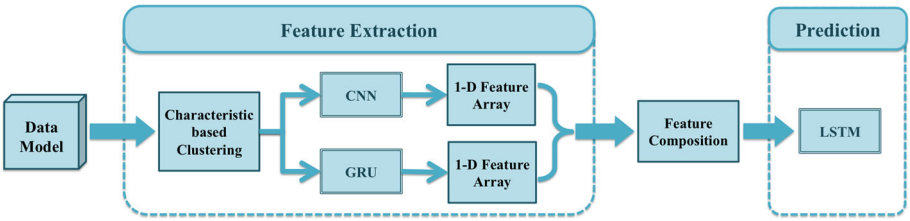


Fig. 5 Forecasting pipeline for blended GRU–CNN–LSTM architecture

(heavy-tails) generates high-quality hierarchical clustering, indicating a "peaked" distribution of the data. To identify peaks in the data, therefore, we used the excess kurtosis K, defined as follows:

$$K = \frac{1}{n\sigma^4}\sum_{t=1}^{n}(S_t - \overline{S_t})^4 - 3, \tag{13}$$

where σ is the standard deviation and n is the number of data samples. Where heavy-tails (upward or downward) were detected, we split the data into C parts. Thus, a limited number of data patterns is obtained that represent the relevant characteristics of the time-series, and, at the same time, the model becomes less sensitive to noise (Wang et al., 2006). The spatiotemporal matrix can be expressed as:

$$X = \begin{bmatrix} X_1(1)\, X_1(2) \cdots X_1(n) \\ X_2(1)\, X_2(2) \cdots X_2(n) \\ \vdots \qquad \ddots \quad \vdots \\ X_k(1)\, X_k(2) \cdots X_k(n) \end{bmatrix}$$

where n represents the nth time sequence of each kth data part.

The GRU feature extraction module preserves temporal information in the hidden layers and consists of two gates, i.e., the reset gate and update gate. The optimal time lag is selected using the reset gate. In recurrent neural networks, hidden state outputs h_t at time t are calculated based on the input time series x_t and the hidden state h_{t-1}. The recurrent neural network model can be represented as follows:

$$h_t = g_1(W_{hx}h_{t-1} + W_{hh}x_{t-1} + b_h), \tag{14}$$

$$\hat{y}_t = g_2(W_{hy}h_t + b_y), \tag{15}$$

where $\hat{y}_t$ denotes the output at time t; W_{hx}, W_{hh}, and W_{hy} represent the weight matrices of hidden layers; g_1 and g_2 are nonlinear activation functions; and b_h and b_y are bias vectors. GRU is then defined as follows:

$$u_t = \sigma(W_{ux}x_t + W_{uc}c_{t-1} + b_u), \tag{16}$$

$$r_t = \sigma(W_{rx}x_t + W_{rc}c_{t-1} + b_r), \tag{17}$$

$$\hat{c_t} = tanh(W_{cx}h_t + W_{rc}(r_t \otimes c_{t-1}) + b_c), \tag{18}$$

$$c_t = (1 - u_t) \otimes c_{t-1} + u_t \otimes \hat{c_t}, \tag{19}$$

where W_u, W_r, and W_c denote, respectively, the weight matrices of the update gate, reset gate, and candidate activation vector $\hat{c_t}$; c_t is the cell state vector; σ is the sigmoid activation function; $\otimes$ stands for the scalar product; and b_u, b_r, and b_c are bias vectors.

The LSTM structure is similar to that of GRU but consists of four gates; the input, input modulation, forget, and output gates:

$$i_t = \sigma(W_{ix}x_t + W_{hh}h_{t-1} + W_{ic}c_{t-1} + b_i), \tag{20}$$

$$f_t = \sigma(W_{fx}x_t + W_{hh}h_{t-1} + W_{fc}c_{t-1} + b_f), \tag{21}$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes tanh(W_{cx}x_t + W_{hh}h_{t-1} + W_{cc}c_{t-1} + b_c), \tag{22}$$

$$o_t = \sigma(W_{ox}x_t + W_{hh}h_{t-1} + W_{oc}c_{t-1} + b_o), \tag{23}$$

$$h_t = o_t \otimes tanh(c_t), \tag{24}$$

where W_i, W_f, and W_o are, respectively, the weight matrices of the input, forget, and output gate.

To build the GRU–CNN–LSTM architecture, at first the GRU and CNN are trained on the same training data separately as presented in Fig. 6. In the CNN module, the two-dimensional spatiotemporal matrices are stacked into three-dimensional matrix blocks. The res-convolution blocks perform convolutional operations to extract high-level features, and pooling layers are used to condense the features and produce the flattened (one-dimensional) data representation. Next, the GRU and CNN one-dimensional flattened data are merged accordingly as inputs for LSTM, where the merged data are divided into smaller datasets representing lags used for multistep-ahead forecasts. Each forecasting model consists of two LSTM layers and one dense layer to consolidate the forecasted output.

4.3 Hybrid model

The combination of the ATES and GRU–CNN–LSTM models allowed us to capture different aspects in the underlying data. Hybrid models integrating linear and nonlinear components have proved to an effective strategy for time-series forecasting (Zhang, 2003; do Nascimento Camelo et al., 2018; Smyl, 2020). However, the sales time-series data are assumed to comprise a linear component L_t and a nonlinear component N_t (Zhang, 2003). First, ATES is used to model the linear component, which leads to residuals containing only the nonlinear component. Second, the GRU–CNN–LSTM model is employed to model the nonlinear residuals from the ATES model. Then, the hybrid forecast F_t can be obtained as follows:

$$F_t = L_t(ATES_{forecast}) + N_t(GRU - CNN - LSTM_{forecast/ATES}) \tag{25}$$

where L_t denotes the linear component captured by applying the ATES model to the data, and N_t denotes the nonlinear component captured using the blended GRU–CNN–LSTM architecture. First, the proposed ATES model calculates the periodic and polar characteristics on the transformed data. Then, the ATES model is fine-tuned on training data, thus accommodating multiple seasonality aspects. From this component, the residuals R_t at time t are calculated that contain only the nonlinear relationship as follows:

$$R_t = F_t - \hat{L}_t, \tag{26}$$

where $\hat{L}_t$ is the forecasted value from ATES, and residuals R_t are modelled using the nonlinear component N_t (i.e., using the GRU–CNN–LSTM model). The combined forecast $\hat{y}_t$ is then obtained as follows:

$$\hat{y}_t = \hat{L}_t + \hat{N}_t, \tag{27}$$

where $\hat{N}_t$ is the forecasted value from GRU–CNN–LSTM.

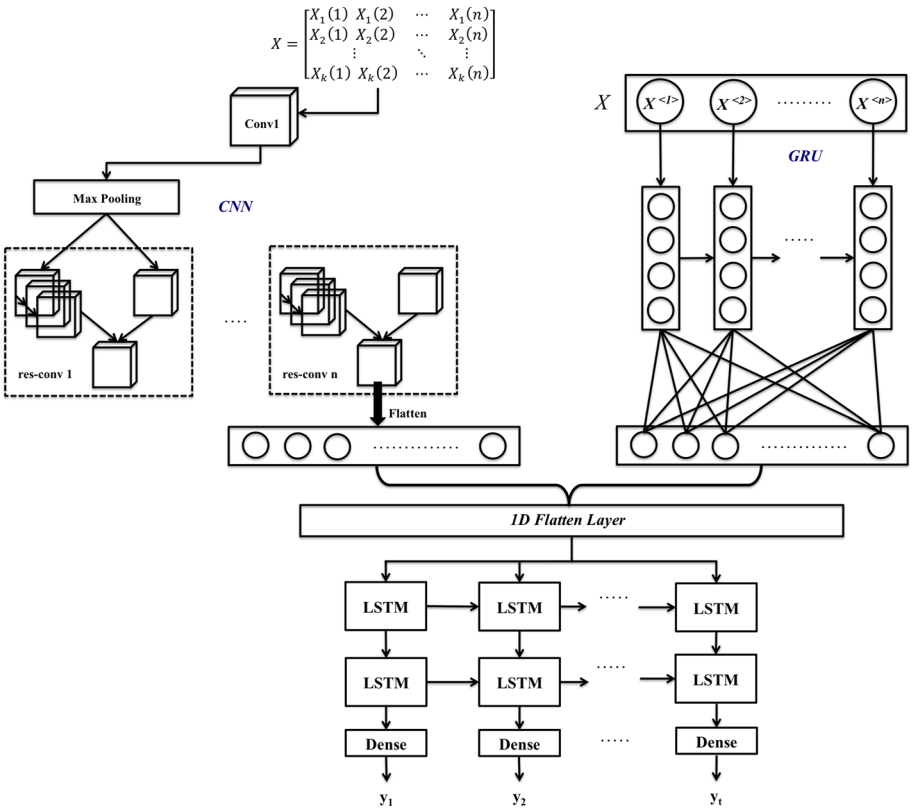


Fig. 6 GRU–CNN–LSTM architecture

However, residual analysis alone is inadequate to detect any nonlinear patterns in the data; therefore, we model the residuals using our blended GRU–CNN–LSTM model. Our blended GRU–CNN–LSTM architecture is a feed-forward network that operates on the deseasonalized and normalized data to provide multiple-step-ahead forecasts.

In this section, we outlined three methods for sales forecasting with big data. The ATES model utilizes the traditional multivariate regression model and augments it with the spectral analysis of multivariate time-series. The GRU–CNN–LSTM model was designed to capture high-level spatiotemporal features in the data. Finally, the hybrid model integrates both models, so that the linear and nonlinear components of the data can be effectively modelled.

5 Empirical result

5.1 Experimental setup

The proposed sales forecasting models were evaluated on the large real-world dataset introduced in Sect. 3. To assess the performance of the proposed models, each store within the dataset was given an individual forecast, and the performance measures were calculated as the average over all the stores in the dataset. The data for years 2013–2016 were used for

training; the data for the year 2017 were used for testing. To obtain the optimum values of the model hyperparameters, a tenfold cross-validation was applied on the training data. Hereinafter, we report the errors obtained for the testing data only.

The proposed models were deployed in Python. More precisely, the Python Data Analysis Library (Pandas) was used to process the data, and the Python scikit-learn package was used to implement the proposed methods. The Python code for Prophet is available at https://github.com/facebook/prophet.

To demonstrate the effectiveness of ATES, five other statistical time-series models were selected from those most effective at forecasting sales in earlier research: ARIMA (Ramos et al., 2015), SARIMA (Arunraj & Ahrens, 2015), SES (Taylor, 2011), HWES (Gelper et al., 2010), and PROPHET (Taylor & Letham, 2018). The parameters of the statistical time series models were set as follows:

- ARIMA—numbers of auto-regressive terms, nonseasonal differences, and lagged forecast errors $p, d, q = \{0, 1, 2, 3\}$
- HWES—seasonal and trend factors with a season cycle length of 28
- PROPHET—daily frequency and seasonality with periods = 365
- SES—span = 28, smoothing factor $\alpha = 2/(span + 1)$
- SARIMA—$p, d, q = \{0, 1, 2, 3\}$, seasonal order = 12

For the purpose of comparison with the proposed GRU–CNN–LSTM model, we used the traditional shallow, artificial feed-forward neural network (ANN) (Sun et al., 2008), a deep feed-forward neural network (DNN) (Loureiro et al., 2018), recurrent neural networks (GRU (Noh et al., 2020) and LSTM (Weng et al., 2019)), CNN (Pan & Zhou, 2020), and CNN + LSTM (Kim & Cho, 2019). The settings of the hyperparameters for the tested neural network models came from the extensive set of experiments; the optimal combinations, as obtained using the grid search procedure, are presented in Table 3.

Stochastic gradient descent with an Adam optimizer with MSE loss function was used to train the neural network-based models. Rectified linear unit (ReLU) was employed as hidden activation functions. The output shape of (33,1) was used to obtain predictions for the next 30 days and three, six, and 12 months. We also tested different numbers of hidden layers, sizes of windows and batches, epochs and learning rates, as indicated in Table 3.

5.2 Forecasting performance evaluation

To evaluate the forecasting performance of the proposed models, different point forecast accuracy measures were used, namely, the root mean square error (RMSE), mean absolute percentage error (MAPE), and percentage forecast error (PFE). Unlike the absolute error measures, minimizing the RMSE yields unbiased point forecasts, but it is sensitive to the intermittent demand items associated with high forecast errors (Kolassa, 2016). Therefore, we also employed MAPE and PFE to demonstrate the forecasting performance in terms of percentage. These measures are unit-free; therefore, they can be used to compare the forecasting performances across different datasets. The conventional measures do not provide a problem context. That is why PFE was used; it is a forecasting alternative that offers a high level of forecasting certainty (almost 95%), allowing us to say that the forecast will be within PFE% of the actual value (Klimberg and Ratick, 2000). The error measures are defined as follows:

$$RMSE = \sqrt{\frac{\sum \left(S_t - \hat{S}_t\right)^2}{n}}, \tag{28}$$

Table 3 Parameters of neural network-based forecasting models

Parameters	ANN	CNN	DNN	GRU	LSTM	CNN + LSTM	GRU–CNN–LSTM
Hidden layers	2	7	3	3	2	5	7
Window size	1	24	20	20	30	25	25
Batch size	30	30	30	40	16	30	32
No. of epochs	100	200	100	200	70	90	400
Input shape	(239,1)	(220, 24,1)	(220,20)	(240,1,20)	(220,20,1)	(187, 2, 18,1)	(187, 2, 18,1)
Learning rate	0.1	0.1	0.1	0.1	0.001	0.01	0.01

Table 4 Forecasting error performance for statistical time-series methods

Time dimension	ARIMA	HWES	PROPHET	SES	SARIMA	ATES
RMSE						
One month	1.1622	1.8134	1.0493	2.0516	2.0501	0.6185
One quarter	1.9369	3.1048	1.9826	3.0917	3.1025	1.7642
Half year	3.3461	4.8744	3.3650	5.8629	5.3274	3.9617
One year	4.9628	6.3890	5.3586	7.0677	6.0118	5.2242
MAPE						
One month	11.37%	14.88%	8.51%	15.84%	15.18%	4.62%
One quarter	14.65%	19.16%	11.39%	19.39%	18.93%	7.24%
Half year	21.37%	24.77%	13.05%	22.64%	21.02%	13.41%
One year	25.48%	27.87%	15.51%	25.42%	22.28%	15.36%
PFE						
One month	3.76%	4.39%	2.78%	8.56%	7.92%	2.14%
One quarter	4.56%	7.82%	4.92%	10.22%	11.66%	4.37%
Half year	5.83%	13.76%	5.26%	14.78%	14.42%	6.23%
One year	6.29%	16.27%	6.23%	17.03%	18.18%	8.45%

$$MAPE = \frac{\sum \frac{|S_t - \hat{S}_t|}{S_t}}{n} \times 100\%, \tag{29}$$

$$PFE = \frac{2 \times s_e}{\hat{S}_{t+1}} \times 100\%, \tag{30}$$

where n is the number of periods predicted, S_t and $\hat{S}_t$ are the actual and forecasted values in time t, respectively, and s_e is the standard error.

5.3 Experimental results

In the first set of experiments on the large benchmark dataset, we compared the performance of ATES with several state-of-the-art statistical time-series models. To evaluate the robustness of the forecasting models, forecasts were produced of multistep-ahead predictions, namely, 1-, 3-, 6-, and 12-month forecasting horizons. This is also important from the perspective of business sales and operations planning (Berry et al., 2020).

From Table 4, it can be observed that ATES consistently outperformed the compared models for the one- and three-month-ahead forecasting horizons. However, for the medium-term predictions of 6 and 12 months, the traditional ARIMA and Prophet methods performed better. This can be attributed to adaptive changes in the ATES model parameters based on seasonal or periodic effects. For sales forecasting with big data, this flexibility is even higher than for Prophet, which results in an error decrease for shorter forecasting horizons but may lead to uncertain intervals for longer projections (Taylor & Letham, 2018). For a better illustration, the one-month-ahead forecasts for a random store and particular product are presented in Fig. 7. Specifically, it can be seen that ATES is particularly effective in capturing unexpectedly high peaks of sales. This in turn may cause wide uncertainty intervals for longer forecasting horizons.

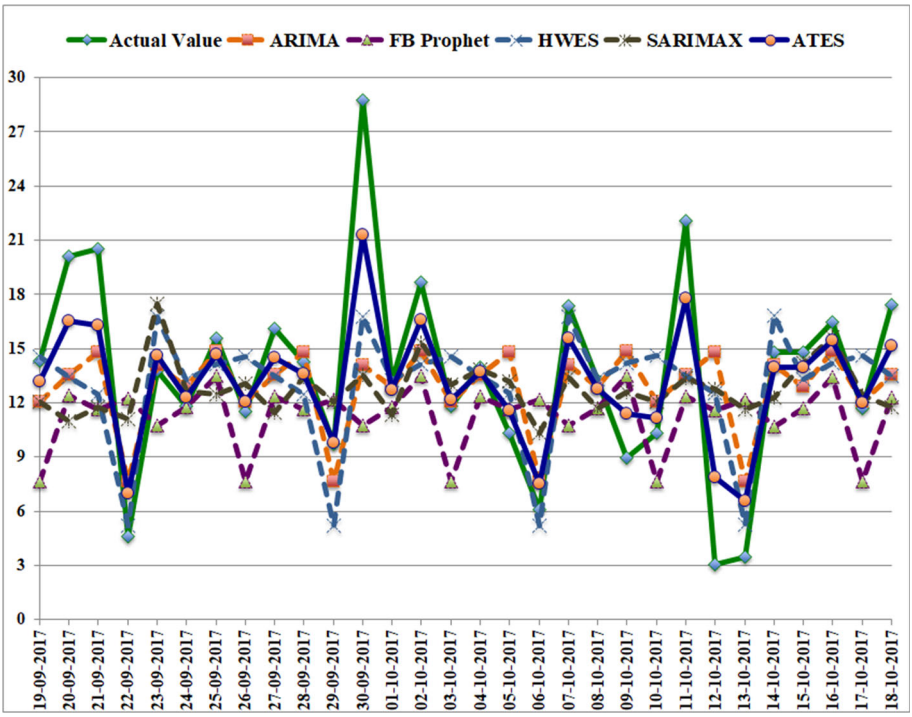


Fig. 7 Sales forecasts of statistical time-series methods for a selected item-store samples

In the next run of experiments, we employed the neural network models for sales forecasting. Table 5 shows that the proposed blended GRU–CNN–LSTM approach performed best in terms of most forecasting performance dimensions. When comparing the recurrent neural networks, GRU seemed superior to LSTM, indicating the vanishing gradient problem in LSTM leading to model overfitting. It must be admitted that this comparison failed to identify a clear benefit of GRU–CNN–LSTM over a longer forecasting horizon of six and 12 months. This suggests that characteristic-based clustering is particularly effective for short forecasting horizons; for medium-term predictions based on deep neural networks, this method appears to be less effective. In other words, for longer prediction periods, the original time-series data can also be used as inputs of recurrent neural networks to achieve high accuracy.

The results suggest that the proposed GRU–CNN–LSTM model overcomes this drawback of LSTM. This can be partly explained by a loss of information caused by the attening of the spatiotemporal matrices into one-dimensional time-series data in ANN, DNN, GRU, and LSTM. On the one hand, the results for CNN show that this neural network model was not effective in capturing the high-level temporal features in the data. On the other hand, CNN could be combined with recurrent neural networks to extract additional features from the data and thus produce more accurate forecasts. When compared with the results for the statistical time-series methods, the proposed GRU–CNN–LSTM model achieved remarkable forecasting errors for longer forecasting horizons. This was due to its capacity for recurrent predictions, capturing both short- and long-term spatiotemporal patterns in the data. As illustrated in Fig. 8, the proposed GRU–CNN–LSTM model was also effective in capturing sales peaks in the short term.

Table 5 Forecasting error performance for neural network models

Time dimension	ANN	CNN	DNN	LSTM	GRU	CNN+LSTM	GRU–CNN–LSTM
RMSE							
One month	1.3127	1.4451	1.0679	1.1421	1.0395	1.2833	0.9614
One quarter	2.4127	2.8964	1.8727	2.0716	1.5189	1.6185	1.4584
Half year	4.7519	4.5148	2.2459	3.7811	2.6958	2.7862	2.2183
One year	5.2274	5.6367	3.8391	5.2198	4.0294	3.9739	4.4719
MAPE							
One month	16.72%	14.78%	9.73%	17.58%	12.38%	12.92%	6.44%
One quarter	19.46%	18.61%	11.34%	19.03%	16.54%	14.81%	9.05%
Half year	21.03%	20.05%	14.28%	22.38%	18.89%	16.29%	14.17%
One year	22.59%	23.81%	22.07%	23.67%	20.76%	17.02%	17.44%
PFE							
One month	4.88%	8.94%	4.69%	6.02%	5.27%	5.48%	4.69%
One quarter	5.46%	12.37%	6.19%	6.47%	6.26%	7.44%	5.34%
Half year	7.39%	15.62%	11.44%	7.02%	10.68%	9.29%	7.17%
One year	8.12%	17.69%	13.15%	7.15%	12.69%	13.27%	7.41%

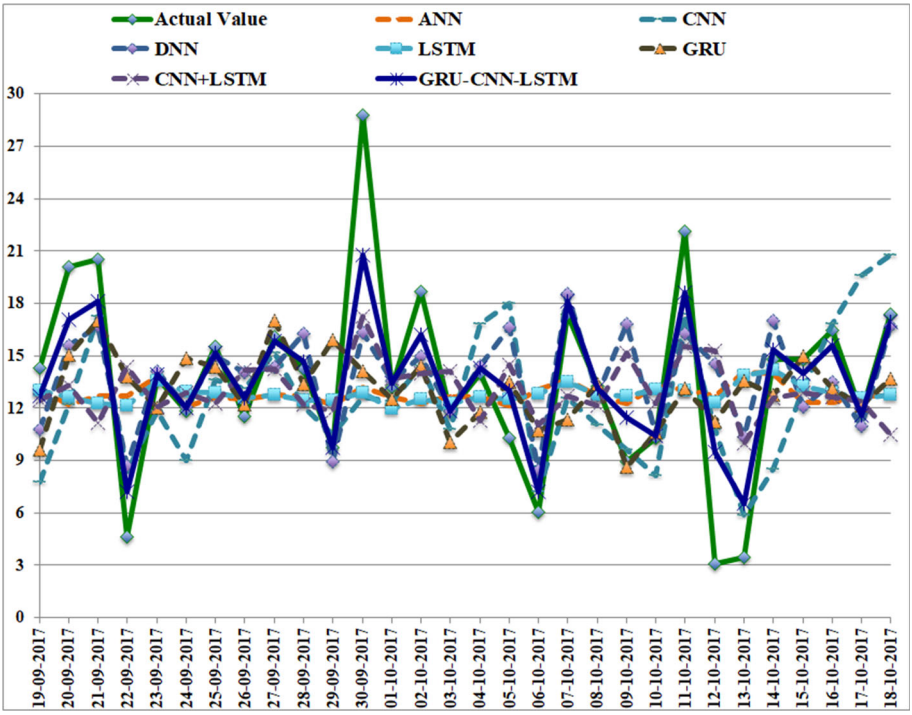


Fig. 8 Sales forecasts of neural network models for a selected item-store sample

Table 6 Forecasting error performance for hybrid time-series/neural network models

Time dimension	ATES +CNN	ATES + DNN	ATES + GRU	ATES + GRU–CNN–LSTM
RMSE				
One month	0.5849	0.4937	0.5298	**0.3166**
One quarter	1.2297	1.2124	1.0151	**0.9476**
Half year	1.9849	1.4126	2.1826	**1.0237**
One year	4.0217	3.4265	4.6392	**2.9274**
MAPE				
One month	3.89%	3.73%	4.26%	**2.57%**
One quarter	7.62%	6.94%	6.21%	**6.03%**
Half year	10.14%	11.81%	12.79%	**9.84%**
One year	15.24%	14.76%	15.07%	**12.61%**
PFE				
One month	2.07%	1.14%	1.62%	**1.03%**
One quarter	3.41%	3.39%	3.94%	**2.48%**
Half year	4.63%	5.44%	5.27%	**3.88%**
One year	8.39%	6.51%	7.91%	**6.02%**

In the final run of experiments, the hybrid ATES + GRU–CNN–LSTM model was compared with its hybrid counterparts, namely, ATES + CNN, ATES + DNN, and ATES + GRU. The compared hybrid models were trained in the same manner as ATES + GRU–CNN–LSTM. Table 6 shows that the proposed ATES + GRU–CNN–LSTM model outperformed its hybrid competitors in terms of all evaluation measures irrespective of time period. In addition, our hybrid model performed substantially better than single ATES (Table 4) and GRU–CNN–LSTM (Table 5), indicating that the underlying sales data comprised of linear and nonlinear components in the short- and long-term. This finding was confirmed by the good forecasting performance of the remaining hybrid models. Figure 9 shows that the hybrid models are particularly accurate in predicting the short-term sales peaks.

5.4 Significance tests

Choosing the appropriate forecasting model is challenging when one considers different accuracy measures. Moreover, comparisons among the forecasting models should be based on reliable error estimates. This is why we calculated the MAPE for each store in the data, which in turn allowed us to compare the methods statistically. To perform statistical comparisons, we used the standard Wilcoxon nonparametric signed-rank test as recommended in previous research because it is less restrictive to the nature of the error and less susceptible to outliers (Flores, 1989; Lu et al., 2012). The null hypothesis is that there is no significant difference between two samples.

The results of the Wilcoxon signed-rank test in Tables 7, 8, and 9 are based on the MAPE values. The p-values indicate the statistically different performance for the ATES, GRU–CNN–LSTM, and hybrid ATES + GRU–CNN–LSTM models when compared with the statistical time-series methods, neural networks, and hybrid models, respectively.

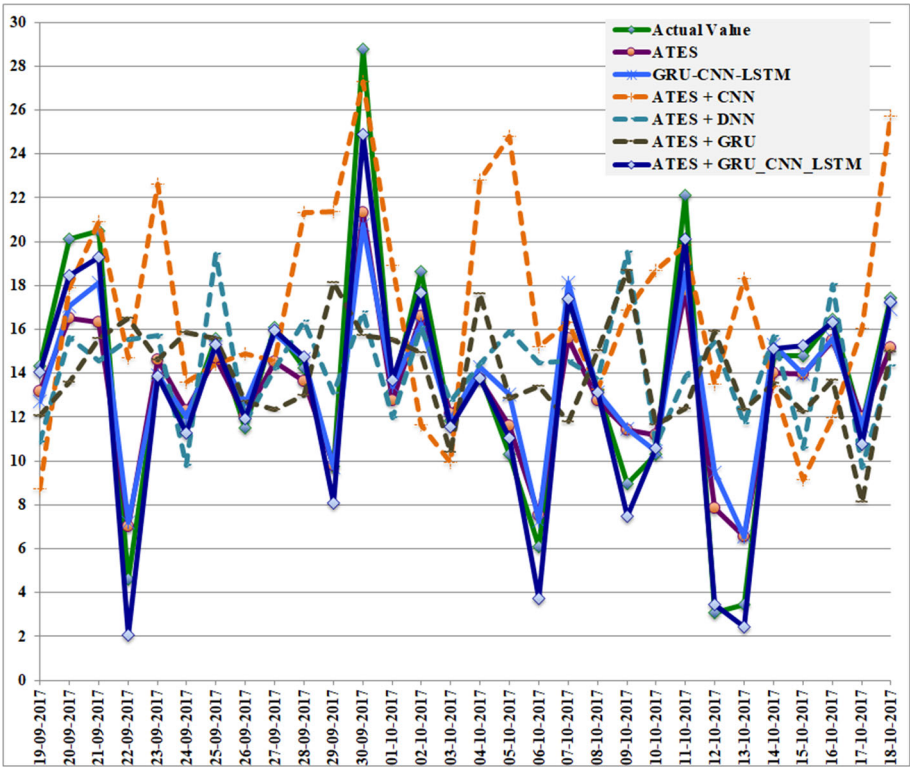


Fig. 9 Sales forecasts of hybrid time-series/neural network models for a selected item-store sample

Table 7 Results of Wilcoxon signed-rank test between the ATES model and five competing statistical time-series methods

		ARIMA	HWES	PROPHET	SES	SARIMA
ATES model	*Z*-statistic	− 15.76	− 15.27	− 15.65	− 15.58	− 15.11
	p value	5.50e−56	1.24e−52	3.18e−55	9.62e−55	1.32e−51

Table 8 Results of Wilcoxon signed-rank test between the GRU–CNN–LSTM model and six competing neural network models

		ANN	CNN	DNN	LSTM	GRU	CNN+LSTM
GRU–CNN–LSTM	*Z*-statistic	− 4.64	− 4.22	− 5.35	− 4.82	− 4.30	− 5.39
	p value	3.525e−6	2.356e−5	7.65e−8	1.413e−6	1.67e−5	7.015e−8

5.5 Robustness evaluation

To produce a robust and unbiased estimate of model prediction error, a common approach is to create multiple training/testing splits and average the errors over all the splits (Ferreira et al., 2017). Therefore, we split the data into multiple training/testing sets using different relative ratios. Specifically, in addition to the 80% relative ratio applied above, we considered

Table 9 Results of Wilcoxon signed-rank test between the hybrid ATES + GRU–CNN–LSTM model and three baseline hybrid time-series/neural network models

		ATES+CNN	ATES+DNN	ATES+GRU
ATES + GRU–CNN–LSTM	*Z*-statistic	−4.51	−4.76	−4.76
	p value	6.33e−6	1.92e−6	1.92e−6

Table 10 Robustness evaluation of statistical time-series methods for different proportions of training data

Relative ratio %	ARIMA (%)	HWES (%)	PROPHET (%)	SES (%)	SARIMA (%)	ATES (%)
80	11.37	14.88	8.51	15.84	15.18	**4.62**
70	12.56	18.87	10.33	23.27	17.89	**6.34**
60	18.34	22.77	14.67	27.54	19.02	**9.41**

Only the best results are in bold

Table 11 Robustness evaluation of neural network models for different proportions of training data

Relative ratio %	ANN (%)	CNN (%)	DNN	LSTM (%)	GRU (%)	CNN + LSTM (%)	GRU–CNN–LSTM (%)
80	16.72	14.78	9.73	17.58	12.38	12.92	**6.44**
70	17.26	14.59	10.44	19.33	16.54	13.81	**8.67**
60	20.14	26.81	15.32	21.11	18.32	16.15	**10.34**

Only the best results are in bold

Table 12 Robustness evaluation of hybrid time-series/neural network models for different proportions of training data

Relative ratio %	ATES + CNN (%)	ATES + DNN (%)	ATES+GRU (%)	ATES + GRU–CNN–LSTM (%)
80	3.89	3.73	4.26	**2.57**
70	6.01	5.47	6.29	**3.43**
60%	9.39	8.61	8.48	**6.24**

Only the best results are in bold

two other relative ratios, i.e., 60% and 70%. In other words, the first three (three and a half) years were used for training and the last two (one and a half) years were used for testing for the 60% (70%) ratio. The one-month-ahead forecasting horizon is evaluated in Tables 10, 11, and 12.

Table 10 shows that the ATES model consistently outperformed the compared statistical time-series methods irrespective of the relative ratio of training data. Similar evidence can be observed for the GRU–CNN–LSTM and ATES + GRU–CNN–LSTM models in Tables 11 and 12, respectively. Notably, the forecasting performance substantially dropped for the 60% relative ratio, indicating that there were insufficient training samples to properly capture the temporal patterns in the data.

To evaluate the robustness of the proposed model to different markets, we used the sales dataset for 45 Walmart stores.[1] The task was to predict department-wide sales for each store. The dataset contains sales from May 2, 2010, to November 1, 2012. Similar to the grocery sales dataset, each record contained the date, store number, department number, item number,

[1] https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data.

weekly sales for the given department, and whether the week is a holiday week. Overall, the dataset encompassed 421,570 data samples (the data for years 2010–2011 were used for training; the data for 2012 were used for testing).

The results for the Walmart dataset are presented in "Appendix 1", showing the performance of statistical time-series methods, neural network models, and hybrid models. When comparing the performance of statistical time-series and neural network models, the results suggest that ATES performed best for the shortest forecasting horizon, while the GRU–CNN–LSTM neural network model was more effective for the remaining time dimensions, thus confirming the results obtained for the grocery sales dataset. In addition, the hybrid time-series model ATES + GRU–CNN–LSTM provided further improvement of the forecasting performance in terms of most performance measures, confirming the effectiveness of the proposed hybrid model.

5.6 Comparison with existing models

It should be noted that the benchmarked dataset has been used in several related studies to date. Among the first studies, Chu et al. (2018) proposed a model called "substituting a subtree with an approximate subprogram" (SAS-GP), which used an effective global optimization of genetic programming. The model was applied to one-day-ahead predictions based on sales on seven previous days, and it significantly outperformed other evolutionary-based prediction methods. The WaveNet CNN model (Kechyn et al., 2018) was proposed to make use of dilated causal convolutions that are reportedly faster to train than recurrent neural networks. Feature maps in CNN were replaced with unsampled filters to avoid increasing receptive fields. As a result, the model performed excellently for the relatively short forecasting horizon of 16 days. An attention-based architecture called the temporal fusion transformer (TFT) was proposed by Lim et al. (2021) for multistep-ahead predictions, providing the forecaster with interpretable temporal dynamics. The time-varying relationship was considered in the used recurrent neural network, and the feature selection component was integrated into the model to detect relevant inputs. The one-month-ahead prediction of sales was performed using 90 days of past information to demonstrate that the TFT outperformed the ARIMA, deep-state space models, and autoregressive recurrent networks.

To demonstrate the effectiveness of the proposed model, the performance of the ATES + GRU–CNN–LSTM hybrid model was compared with the results of other models that have been applied to the grocery sales dataset in previous studies. The results obtained in the previous studies are presented in Table 13, which shows that these studies differ in the forecasting horizon and the performance measures used for the models. When considering forecasting models using the same performance measures as our study, our hybrid model outperforms these models in terms of both RMSE and MAPE. Specifically, the proposed model achieved a reduction in RMSE error by almost half compared with the SAS-GP (Chu et al., 2018) and CatBoost models (Ding e al., 2020). Similarly, there was a considerable reduction in MAPE (by almost 15%) compared with the RNN model augmented with a recalibration component (Kuleshov et la., 2018). This validates the effectiveness of the proposed model compared with existing machine learning and deep learning neural network models.

5.7 Summary of experimental results

The results of this study show which forecasting methods perform best in forecasting sales with big data. The adaptive and flexible ATES model, which integrates seasonal and periodic

Table 13 Comparison with existing models for the grocery sales dataset

Authors	Forecasting horizon	Method	Performance
Chu et al. (2018)	1-day	SAS-GP	RMSE = 0.631
Kechyn et al. (2018)	16-days	WaveNet	NWRMSLE = 0.578
Kuleshov et al. (2018)	1-month	RNN+recalibration	MAPE = 17.3%
Vairagade et al. (2019)	1-day	RF	MAE = 0.022, MSE = 0.001
Weng et al. (2019)	1-day	LightGBM + LSTM	NWRMSLE = 0.504
Ding et al. (2020)	1-day	LR, SVR, CatBoost	RMSE = 1.120(LR), RMSE = 0.782(SVR), RMSE = 0.605(CatBoost)
Lim et al. (2021)	1-month	TFT	P50QL = 0.147
Sprangers et al. (2022)	1-month	BiTCN	sMAPE = 0.674, NRMSE = 1.317
Paria et al. (2022)	1-month	HRDNN	sMAPE = 0.208
this study	1-month	ATES + GRU–CNN–LSTM	RMSE = 0.317, MAPE = 2.57%, PFE = 1.03%

BiTCN bidirectional temporal convolutional network, *HRDNN* hierarchical regularized DNN, LR - linear regression, *NRMSE* normalized RMSE, *NWRMSLE* normalized weighted root mean squared logarithmic error, *P50QL* P50 quantile loss, *RF* random forest, *SAS-GP* substituting a subtree with an approximate subprogram genetic programming, *sMAPE* symmetric MAPE, *SVR* support vector regression

effects performed significantly better than the competing models based on statistical time-series methods. Likewise, the proposed GRU–CNN–LSTM deep learning-based model was superior to the compared neural network models. However, the performance of the proposed models differed significantly for the forecasting horizons examined. The ATES additive model proved to be more accurate for one-month-ahead and three-month-ahead forecasting horizons but was not adequate for more complex and challenging longer forecasting horizons. Further improvement in forecasting performance could be achieved by combining ATES and GRU–CNN–LSTM.

6 Discussion

As noted in the literature review, forecasting sales with big data requires capturing complex multilevel features from the data. Compared with other machine learning methods, deep neural networks have proven to be the most effective for this challenging task (Loureiro et al., 2018). However, the results of previous studies are inconclusive in this regard. Vairagade et al. (2019) compared the random forest and neural networks to show that the random forest produced a lower MAPE and MSE. However, this finding turned out to be valid only for smaller datasets, as only 10-most-purchased items and a relatively short time-series were used for forecasting. Moreover, only the traditional feed-forward neural network was used for comparison. This study yielded results that corroborate the findings of a great deal of the previous studies in sales forecasting (Loureiro et al., 2018; Pan & Zhou, 2020; Ma & Fildes, 2021), suggesting that deep learning neural networks outperform statistical time-series methods. Given the results of the deep neural network comparisons, the capacity of the GRU model to solve the vanishing gradient problem proved to be particularly important in sales forecasting. However, this aspect of deep neural networks has proven to be only valid for

longer forecasting horizons, indicating that the effects of seasonality and holidays/promotions are crucial for shorter-term predictions of retail sales, which is consistent with earlier research (Arunraj & Ahrens, 2015; Ma et al., 2016).

Significant improvement in forecasting performance across forecasting horizons was achieved by combining the ATES and GRU–CNN–LSTM models. This finding is in agreement with Li et al. (2018), who showed the superiority of modelling the linear component and nonlinear residuals in sales data using a combined model over individual linear and nonlinear forecasting models. These results are also consistent with those of other studies (Arunraj & Ahrens, 2015) and suggest that, in the presence of strong seasonality, traditional and hybrid sales forecasting models perform better than neural network models.

7 Conclusions

This study set out to propose a sales forecasting hybrid model capturing both linear and nonlinear patterns in big data. To deal with the size of the data, data were split based on spectral characteristics in ATES, and kurtosis was used to produce data parts in the CNN module of the GRU–CNN–LSTM architecture. In addition, compared with models used in previous studies for sales forecasting, the models proposed in this study were intended for a more challenging task of multistep-ahead predictions ranging from 1 to 12 months. The present study confirms previous findings highlighting a role for hybrid forecasting models and contributes additional evidence that underlined the importance of (1) flexible additive modelling of seasonal effects in the linear component, and (2) spatiotemporal representation of items and stores in the nonlinear component of sales data. For the latter component of the hybrid model, high-level spatiotemporal features obtained from CNN and GRU were combined as flattened data and used by the LSTM component to forecast the sales data. One of the most interesting finding to emerge from this study is that the proposed hybrid ATES + GRU–CNN–LSTM model proved to be robust across forecasting horizons. Overall, using the hybrid model, it was possible to reduce the MAPE value by 2.05% for 1-month-ahead and by 2.75% for 12-month-ahead forecasts of grocery sales relative to the compared statistical time-series and neural network models.

The contribution of this study to the current literature lies in the development of a hybrid modelling approach to sales forecasting that is robust to forecasting horizons. To find a cutting-edge solution for multistep-ahead sales forecasting, we have devised a multi-stage methodology that additively models the big data characteristics in sales forecasting data. The evidence from this study suggests that the flexible linear component is perfectly adequate for short-term sales forecasts, while for accurate long-term forecasts it needs to be combined with a hybrid architecture based on deep learning. With reasonably accurate forecasting results up to one year in advance, the proposed solution can be expected to improve inventory management performance and adjust operational decisions especially in promotional marketing. Another important implication is that the underlying big data should be stored and accessed using a three-dimensional data model for effective product-specific store-level sales forecasting.

Finally, a number of limitations need to be noted regarding the present study. One limitation is that the weights of individual items were not considered. Indeed, giving a larger weight to perishable items is recommended in future studies. For that purpose, the loss function of the forecasting model must be modified. The proposed hybrid model can also be improved by using additional determinants of sales relevant for a given industry. One advantage of

our models is their easy modularity, which allows analysts to add relevant factors, such as macroeconomic indicators and government policies (Sagaert et al., 2018a). A further study could also assess the performance of spatiotemporal transformer-based deep learning models (Lim et al., 2021) in modelling the nonlinear residuals. For future research directions, we also recommend the application of the proposed forecasting models in different business domains with similar data characteristics such as e-commerce or energy forecasting as well as in other domains (e.g., mobile networks and air quality forecasting).

Acknowledgements This article was directed by Software Evaluation and Re-Engineering Research Lab (SERER Lab) and supported by the scientific research project of the Czech Sciences Foundation Grant No. 19-15498S.

Appendix 1: Forecasting error performance for the Walmart sales dataset

Statistical time-series methods

Time dimension	ARIMA	HWES	PROPHET	SES	SARIMA	ATES
RMSE						
One month	0.1072	0.1630	0.2135	0.1678	0.1147	**0.0824**
One quarter	0.2833	0.3345	0.3342	0.8877	0.2567	**0.2100**
Half year	0.3976	0.5782	0.4123	1.2340	0.4123	**0.3321**
One year	0.7560	0.8876	0.5678	1.3456	0.6318	**0.4567**
MAPE						
One month	14.58%	18.96%	23.78%	17.90%	13.59%	**8.62%**
One quarter	18.56%	24.80%	25.23%	28.31%	21.93%	**17.24%**
Half year	26.79%	31.21%	28.35%	47.10%	25.02%	**23.77%**
One year	41.35%	44.87%	31.41%	55.42%	32.28%	**27.66%**
PFE						
One Month	4.66%	4.99%	5.67%	4.96%	4.92%	**4.23%**
One quarter	8.56%	8.92%	9.22%	10.12%	8.26%	**7.38%**
Half year	11.23%	11.76%	11.22%	16.78%	11.42%	**10.26%**
One year	13.44%	13.77%	13.23%	20.03%	13.18%	**11.23%**

Neural network models							
Time dimension	ANN	CNN	DNN	LSTM	GRU	CNN + LSTM	GRU–CNN–LSTM
RMSE							
One month	0.2105	0.1194	0.1044	0.1223	0.1072	0.1361	**0.0918**
One quarter	0.3012	0.1988	0.2109	0.2467	0.2234	0.2123	**0.1876**
Half year	0.3987	0.3184	0.2908	0.3123	0.2987	0.2911	**0.2843**
One year	0.6881	0.4765	0.4908	0.5467	0.4654	0.4678	**0.4123**
MAPE							
One month	24.58%	15.11%	13.23%	13.79%	14.17%	14.09%	**12.11%**
One quarter	30.21%	24.45%	23.78%	21.90%	19.54%	20.10%	**17.00%**
Half year	41.89%	31.22%	29.09%	33.89%	28.22%	29.21%	**23.22%**
One year	49.56%	41.22%	39.54%	40.22%	38.78%	39.22%	**31.87%**
PFE							
One month	4.90%	4.60%	4.39%	4.12%	4.34%	4.31%	**4.01%**
One quarter	8.23%	8.37%	9.01%	8.47%	8.21%	8.01%	**7.64%**
Half year	11.23%	10.89%	10.18%	10.23%	10.01%	11.21%	**9.87%**
One year	13.78%	13.58%	13.15%	13.45%	12.69%	12.27%	**11.78%**

Hybrid time-series/neural network models				
Time dimension	ATES + CNN	ATES + DNN	ATES + GRU	ATES + GRU–CNN–LSTM
RMSE				
One month	0.1241	0.1356	0.1230	**0.0851**
One quarter	0.1723	0.1733	0.2145	**0.1409**
Half year	0.2562	0.2312	0.2876	**0.2133**
One year	0.3578	0.3289	0.3512	**0.2809**
MAPE				
One month	16.12%	15.32%	14.93%	**12.43%**
One quarter	18.21%	18.81%	18.99%	**16.22%**
Half year	24.34%	26.22%	26.44%	**22.65%**
One year	30.18%	31.45%	30.70%	**28.11%**
PFE				
One month	4.32%	4.12%	4.22%	**2.43%**
One quarter	7.88%	7.78%	7.65%	**5.38%**
Half year	10.83%	10.67%	10.27%	**8.88%**
One year	13.32%	12.51%	13.01%	**11.32%**

References

Ali, Ö. G., & Gürlek, R. (2020). Automatic interpretable retail forecasting with promotional scenarios. *International Journal of Forecasting, 36*(4), 1389–1406.

Ali, O. G., & Pinar, E. (2016). Multi-period-ahead forecasting with residual extrapolation and information sharing-utilizing a multitude of retail series. *International Journal of Forecasting, 32*(2), 502–517.

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.

Berry, L. R., Helman, P., & West, M. (2020). Probabilistic forecasting of heterogeneous consumer transaction-sales time series. *International Journal of Forecasting, 36*(2), 552–569.

Bohanec, M., Borštnar, M. K., & Robnik-Šikonja, M. (2017). Explaining machine learning models in sales predictions. *Expert Systems with Applications, 71*, 416–428.
Boone, T., Ganeshan, R., Jain, A., et al. (2019). Forecasting sales in the supply chain: Consumer analytics in the big data era. *International Journal of Forecasting, 35*(1), 170–180.
Bose, R. (2009). Advanced analytics: Opportunities and challenges. *Industrial Management & Data Systems, 109*(2), 155–172.
Box, G. E., Jenkins, G. M., Reinsel, G. C., et al. (2015). *Time series analysis: Forecasting and control*. Wiley.
Caiado, J., Crato, N., & Poncela, P. (2020). A fragmented-periodogram approach for clustering big data time series. *Advances in Data Analysis and Classification, 14*(1), 117–146.
Chen, F., & Ou, T. (2011). Sales forecasting system based on gray extreme learning machine with Taguchi method in retail industry. *Expert Systems with Applications, 38*(3), 1336–1345.
Chen, I. F., & Lu, C. J. (2017). Sales forecasting by combining clustering and machine-learning techniques for computer retailing. *Neural Computing and Applications, 28*(9), 2633–2647.
Choi, T. M., Hui, C. L., Ng, S. F., et al. (2011). Color trend forecasting of fashionable products with very few historical data. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42*(6), 1003–1010.
Chu, C. W., & Zhang, G. P. (2003). A comparative study of linear and nonlinear models for aggregate retail sales forecasting. *International Journal of Production Economics, 86*(3), 217–231.
Chu, T. H., Nguyen, Q. U., & Cao, V. L. (2018) Semantics based substituting technique for reducing code bloat in genetic programming. In *Proceedings of the ninth international symposium on information and communication technology* (pp. 77–83).
Ding, J., Chen, Z., Xiaolong, L., & Lai, B. (2020) Sales forecasting based on catboost. In *2020 2nd international conference on information technology and computer application (ITCA)* (pp. 636–639). IEEE.
Disney, S. M., Ponte, B., & Wang, X. (2021). Exploring the nonlinear dynamics of the lost-sales order-up-to policy. *International Journal of Production Research, 59*(19), 5809–5830.
do Nascimento Camelo, H., Lucio, P. S., Junior, J. B. V. L., et al. (2018). Innovative hybrid models for forecasting time series applied in wind generation based on the combination of time series models with artificial neural networks. *Energy, 151*, 347–357.
Eachempati, P., Srivastava, P. R., Kumar, A., et al. (2022). Can customer sentiment impact firm value? An integrated text mining approach. *Technological Forecasting and Social Change, 174*(121), 265.
Efat, M. I. A., Bashar, R., Uddin, K. I., & Bhuiyan, T. (2018) Trend estimation of stock market: An intelligent decision system. In *International conference on cyber security and computer science (ICONCS'18)* (pp. 44–49).
Ferreira, S. L., Caires, A. O., Borges, Td. S., et al. (2017). Robustness evaluation in analytical methods optimized using experimental designs. *Microchemical Journal, 131*, 163–169.
Flores, B. E. (1989). The utilization of the Wilcoxon test to compare forecasting methods: A note. *International Journal of Forecasting, 5*(4), 529–535.
Gahirwal, M, (2013) Inter time series sales forecasting. arXiv preprint arXiv:1303.0117
Ganesan, V. A., Divi, S., Moudhgalya, N. B., Sriharsha, U., & Vijayaraghavan, V. (2019) Forecasting food sales in a multiplex using dynamic artificial neural networks. In *Science and information conference* (pp. 69–80). Springer.
Gelper, S., Fried, R., & Croux, C. (2010). Robust forecasting with exponential and Holt–Winters smoothing. *Journal of Forecasting, 29*(3), 285–300.
Harsoor, A. S., & Patil, A. (2015). Forecast of sales of Walmart store using big data applications. *International Journal of Research in Engineering and Technology, 4*(6), 51–59.
Huang, T., Fildes, R., & Soopramanien, D. (2019). Forecasting retailer product sales in the presence of structural change. *European Journal of Operational Research, 279*(2), 459–470.
Iwok, I. A. (2016). Seasonal modelling of Fourier series with linear trend. *International Journal of Statistics and Probability, 5*(6), 65–72.
Jha, A., Ray, S., Seaman, B., & Dhillon, I. S. (2015) Clustering to forecast sparse time-series data. In *2015 IEEE 31st international conference on data engineering* (pp. 1388–1399). IEEE.
Ji, S., Wang, X., Zhao, W., & Guo, D. (2019). An application of a three-stage xgboost-based model to sales forecasting of a cross-border e-commerce enterprise. *Mathematical Problems in Engineering*. https://doi.org/10.1155/2019/8503252
Jiménez, F., Sánchez, G., García, J. M., et al. (2017). Multi-objective evolutionary feature selection for online sales forecasting. *Neurocomputing, 234*, 75–92.
Kaggle (2018) Corporación favorita grocery sales forecasting. Retrieved February 3, 2020, from https://www.kaggle.com/c/favorita-grocery-sales-forecasting
Kechyn, G., Yu, L., Zang, Y., & Kechyn, S. (2018) Sales forecasting using WaveNet within the framework of the Kaggle competition. arXiv preprint arXiv:1803.04037

Kharfan, M., Chan, V. W. K., & Firdolas Efendigil, T. (2021). A data-driven forecasting approach for newly launched seasonal products by leveraging machine-learning approaches. *Annals of Operations Research, 303*(1), 159–174.

Kim, T. Y., & Cho, S. B. (2019). Predicting residential energy consumption using CNN–LSTM neural networks. *Energy, 182*, 72–81.

Klimberg, R., & Ratick, S. (2000) A new measure of relative forecast error. In *INFORMS fall meeting*

Kolassa, S. (2016). Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting, 32*(3), 788–803.

Kraus, M., Feuerriegel, S., & Oztekin, A. (2020). Deep learning in business analytics and operations research: Models, applications and managerial implications. *European Journal of Operational Research, 281*(3), 628–641.

Kuleshov, V., Fenner, N., & Ermon, S. (2018) Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning, PMLR* (pp. 2796–2804).

Kumar, A., Shankar, R., & Aljohani, N. R. (2020). A big data driven framework for demand-driven forecasting with effects of marketing-mix variables. *Industrial Marketing Management, 90*, 493–507.

Li, C., Cheang, B., Luo, Z., et al. (2021). An exponential factorization machine with percentage error minimization to retail sales forecasting. *ACM Transactions on Knowledge Discovery from Data (TKDD), 15*(2), 1–32.

Li, C., & Lim, A. (2018). A greedy aggregation-decomposition method for intermittent demand forecasting in fashion retailing. *European Journal of Operational Research, 269*(3), 860–869.

Li, M., Ji, S., & Liu, G. (2018). Forecasting of Chinese e-commerce sales: an empirical comparison of Arima, nonlinear autoregressive neural network, and a combined ARIMA–NARNN model. *Mathematical Problems in Engineering, 2018*, 1–12.

Liang, Y., Wu, J., Wang, W., Cao, Y., Zhong, B., Chen, Z., & Li, Z. (2019) Product marketing prediction based on xgboost and lightGBM algorithm. In *Proceedings of the 2nd international conference on artificial intelligence and pattern recognition* (pp. 150–153).

Lim, B., Arık, S. Ö., Loeff, N., et al. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting, 37*(4), 1748–1764.

Liu, N., Ren, S., Choi, T. M., et al. (2013). Sales forecasting for fashion retailing service industry: A review. *Mathematical Problems in Engineering, 738*, 675.

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.

Lu, C. J., Lee, T. S., & Lian, C. M. (2012). Sales forecasting for computer wholesalers: A comparison of multivariate adaptive regression splines and artificial neural networks. *Decision Support Systems, 54*(1), 584–596.

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

Ma, S., Fildes, R., & Huang, T. (2016). Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra-and inter-category promotional information. *European Journal of Operational Research, 249*(1), 245–257.

Misiorek, A., Trueck, S., & Weron, R. (2006). Point and interval forecasting of spot electricity prices: Linear vs non-linear time series models. *Studies in Nonlinear Dynamics & Econometrics*. https://doi.org/10.2202/1558-3708.1362

Navratil, M., & Kolkova, A. (2019). Decomposition and forecasting time series in business economy using prophet forecasting model. *Central European Business Review, 8*(4), 26–39.

Noh, J., Park, H. J., Kim, J. S., et al. (2020). Gated recurrent unit with genetic algorithm for product demand forecasting in supply chain management. *Mathematics, 8*(4), 565.

Pan, H., Zhou, H., et al. (2020). Study on convolutional neural network and its application in data mining and sales forecasting for e-commerce. *Electronic Commerce Research, 20*(2), 297–320.

Paria, B., Sen, R., Ahmed, A., & Das, A. (2022) Hierarchically regularized deep forecasting. arXiv preprint arXiv:2106.07630

Pavlyshenko, B. (2018) Using stacking approaches for machine learning models. In *2018 IEEE second international conference on data stream mining & processing (DSMP)* (pp. 255–258). IEEE.

Pavlyshenko, B. M. (2016) Linear, machine learning and probabilistic approaches for time series analysis. In *2016 IEEE first international conference on data stream mining & processing (DSMP)* (pp. 377–381). IEEE.

Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. *Data, 4*(1), 15.

Proietti, T., & Lütkepohl, H. (2013). Does the Box–Cox transformation help in forecasting macroeconomic time series? *International Journal of Forecasting, 29*(1), 88–99.

Ramos, P., Santos, N., & Rebelo, R. (2015). Performance of state space and Arima models for consumer retail sales forecasting. *Robotics and Computer-Integrated Manufacturing, 34*, 151–163.
Ren, S., Chan, H. L., & Siqin, T. (2020). Demand forecasting in retail operations for fashionable products: Methods, practices, and real case study. *Annals of Operations Research, 291*(1), 761–777.
Sagaert, Y. R., Aghezzaf, E. H., Kourentzes, N., & Desmet, B. (2018a). Tactical sales forecasting using a very large set of macroeconomic indicators. *European Journal of Operational Research, 264*(2), 558–569.
Sagaert, Y. R., Aghezzaf, E. H., Kourentzes, N., & Desmet, B. (2018b). Temporal big data for tire industry tactical sales forecasting. *Interfaces, 48*(2), 121–129.
Škare, M., & Porada-Rochoń, M. (2020). Forecasting financial cycles: Can big data help? *Technological and Economic Development of Economy, 26*(5), 974–988.
Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting, 36*(1), 75–85.
Sprangers, O., Schelter, S., & de Rijke, M. (2022). Parameter-efficient deep probabilistic forecasting. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2021.11.011
Sun, Z. L., Choi, T. M., Au, K. F., et al. (2008). Sales forecasting using extreme learning machine with applications in fashion retailing. *Decision Support Systems, 46*(1), 411–419.
Taylor, J. W. (2010). Exponentially weighted methods for forecasting intraday time series with multiple seasonal cycles. *International Journal of Forecasting, 26*(4), 627–646.
Taylor, J. W. (2011). Multi-item sales forecasting with total and split exponential smoothing. *Journal of the Operational Research Society, 62*(3), 555–563.
Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician, 72*(1), 37–45.
Tehrani, A. F., & Ahrens, D. (2016) Improved forecasting and purchasing of fashion products based on the use of big data techniques. In *Supply management research* (pp. 293–312). Springer.
Teunter, R. H., Syntetos, A. A., & Babai, M. Z. (2011). Intermittent demand: Linking forecasting to inventory obsolescence. *European Journal of Operational Research, 214*(3), 606–615.
Thomassey, S. (2010). Sales forecasts in clothing industry: The key success factor of the supply chain management. *International Journal of Production Economics, 128*(2), 470–483.
Ulrich, M., Jahnke, H., Langrock, R., et al. (2021). Distributional regression for demand forecasting in e-grocery. *European Journal of Operational Research, 294*(3), 831–842.
Vairagade, N., Logofatu, D., Leon, F., & Muharemi, F. (2019) Demand forecasting using random forest and artificial neural network for supply chain management. In *International conference on computational collective intelligence* (pp. 328–339). Springer.
von Sachs, R. (2020). Nonparametric spectral analysis of multivariate time series. *Annual Review of Statistics and Its Application, 7*, 361–386.
Wang, X., Smith, K., & Hyndman, R. (2006). Characteristic-based clustering for time series data. *Data Mining and Knowledge Discovery, 13*(3), 335–364.
Wang, X., Smith-Miles, K., & Hyndman, R. (2009). Rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series. *Neurocomputing, 72*(10–12), 2581–2594.
Weng, T., Liu, W., & Xiao, J. (2019). Supply chain sales forecasting based on lightGBM and LSTM combination model. *Industrial Management & Data Systems, 120*(2), 265–279.
Wong, W., & Guo, Z. (2010). A hybrid intelligent model for medium-term sales forecasting in fashion retail supply chains using extreme learning machine and harmony search algorithm. *International Journal of Production Economics, 128*(2), 614–624.
Wu, L., Kong, C., Hao, X., et al. (2020). A short-term load forecasting method based on GRU–CNN hybrid neural network model. *Mathematical Problems in Engineering, 1428*, 104.
Xia, M., & Wong, W. K. (2014). A seasonal discrete grey forecasting model for fashion retailing. *Knowledge-Based Systems, 57*, 119–126.
Zhang, G. P. (2003). Time series forecasting using a hybrid Arima and neural network model. *Neurocomputing, 50*, 159–175.
Zhang, G. P., & Qi, M. (2005). Neural network forecasting for seasonal and trend time series. *European Journal of Operational Research, 160*(2), 501–514.
Zhang, Q., Yang, L. T., Chen, Z., et al. (2018). A survey on deep learning for big data. *Information Fusion, 42*, 146–157.
Zhao, K., & Wang, C. (2017) Sales forecast in e-commerce using convolutional neural network. arXiv preprint arXiv:1708.07946