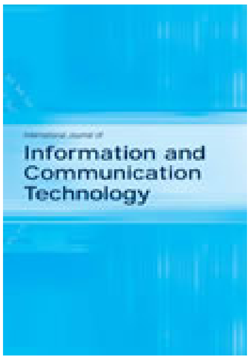
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of
Information and Communication Technology

International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Regional economic forecasting based on structural equation modelling and time series

Yuxuan Wu

Article History:

Received:	31 December 2024
Last revised:	13 January 2025
Accepted:	14 January 2025
Published online:	25 February 2025

Regional economic forecasting based on structural equation modelling and time series

Yuxuan Wu

School of Economics and Management,
Weifang University,
WeiFang 261061, China
Email: ywxk2024@wfu.edu.cn

Abstract: Policymaking, economic planning, and corporate decision-making all benefit from regional economic forecasting. Traditional forecasting techniques, nevertheless, can struggle to handle time dependency and complicated causality and this study thus suggests a regional economic forecasting model based on the combination of structural equation modelling (SEM) and time series analysis (TSA), SEMTSA-Region. The SEMTSA-Region model shows better forecasting performance in several assessment criteria when the model is employed in the experiment to anticipate the economic data of a region in the previous ten years and compared with many conventional models. Furthermore, this work performed parameter optimisation trials on the model to support its great stability and adaptability even more. This paper offers a fresh theoretical framework for regional economic forecasting as well as a useful base for actual implementations.

Keywords: structural equation modelling; SEM; time series analysis; TSA; regional economic forecasting.

Reference to this paper should be made as follows: Wu, Y. (2025) ‘Regional economic forecasting based on structural equation modelling and time series’, *Int. J. Information and Communication Technology*, Vol. 26, No. 4, pp.89–105.

Biographical notes: Yuxuan Wu received his PhD in Economics from the Nankai University in 2020. He is currently working as a Lecturer at the School of Economics and Management of Weifang University. His research areas and directions include digital economy and time series analysis.

1 Introduction

Regional economic forecasting has grown in relevance for policy formation, economic planning, and business strategic decision-making given the fast growth of the global economy and the growing connection among national economies (Acs and Szerb, 2007; Yang et al., 2008). More and more scholars have started to investigate modelling approaches relevant to complicated economic data in order to raise the accuracy and dependability of forecasts. Mostly depending on classical statistical models like time series analysis (TSA) and regression analysis, traditional regional economic forecasting techniques with the complexity of the economic environment, single statistical approaches progressively disclose their shortcomings in handling nonlinear relationships, dynamic fluctuations and multivariate interactions, even if these approaches can somewhat reflect the trend of economic data.

TSA techniques dominated most of the early work on regional economic forecasting (Hannaford et al., 2001). Widely applied in trend forecasting and cyclical fluctuation analysis of economic data, autoregressive integral sliding average model (ARIMA) is among the most often used time series forecasting models (Schaffer et al., 2021; Yu et al., 2014). By use of historical data analysis, the ARIMA model estimates the future values while capturing the time-series properties of the data. The model demonstrates some restrictions when confronted with non-smooth and unexpected events; nevertheless it mostly depends on the smoothness assumption of past data. Researchers have suggested extended models including generalised autoregressive conditional heteroskeasticity model (GARCH) and vector autoregressive model (VAR) to offset this deficiency in order to capture more volatility and multivariate relationships (Amado and Teräsvirta, 2014; Sato and Matsuda, 2021), which have been extensively applied especially in financial markets and macroeconomic forecasting.

These conventional time series models still suffer from the difficulty to handle complicated causal interactions and lengthy time dependencies even if they can somewhat increase forecasting accuracy. Consequently, more and more research has started to present increasingly sophisticated modelling techniques including machine learning and deep learning models. Support vector machines (SVM), random forests (RF), and neural networks (ANN) have progressively found use in regional economic forecasting (Mutale et al., 2024; Lei et al., 2019); they can automatically learn complex patterns in data, so capturing nonlinear relationships; and they have produced good forecasting results in many different fields. In terms of prediction accuracy, these techniques are far superior than conventional statistical models; moreover, they particularly exhibit great performance considering large-scale and multi-dimensional data.

Though they greatly increase accuracy, machine learning and deep learning approaches sometimes lack a thorough knowledge of the intrinsic structure and causal relationships between data, particularly in the field of economics where causality analyses are fundamental to the formulation of sensible economic policies. Consequently, structural equation modelling (SEM) has progressively become a useful instrument extensively applied in the disciplines of economics, sociology (Westland, 2015), psychology, etc. Because it may more naturally portray the interaction and internal mechanism of variables by means of causation between variables, SEM is well suited for simulating multivariate complex systems. In order to raise the interpretability and prediction accuracy of the models, some researchers have lately attempted to combine SEM with other prediction models (Karimi and Meyer, 2014). For market volatility prediction, some studies have, for instance, merged SEM with SVM and produced improved results.

Nevertheless, especially in handling long-term time-dependent and non-stationary data, SEM still finds significant difficulties modelling and forecasting time series data. More and more academics have begun to investigate the prospect of merging TSA techniques with SEM in order to solve these problems (Lou et al., 2020). While SEM can expose possible causal linkages between variables, TSA is good in handling time-dependent characteristics in data. Combining the two not only efficiently captures the time-series properties of economic data but also extensively analyses the causal pathways between variables, thereby enhancing the forecasting power and stability of the model.

Generally speaking, even with the several successes of the current regional economic forecasting systems, the conventional models still have some restrictions and cannot fully and successfully address the difficult economic forecasting issues. Therefore, in order to cope with the multilevel, multidimensional and nonlinear characteristics in the economic system, how to mix the advantages of several approaches? This has become a major topic in present study. This paper suggests a regional economic forecasting model (SEMTSA-Region) based on the mix of SEM and TSA to tackle this difficulty. This paper intends to increase the accuracy of regional economic forecasting and offer a new theoretical framework and practical basis for economic forecasting by including these two approaches.

This paper's innovations consist as follows:

1 Combination of SEM and TSA. We present a regional economic forecasting model (SEMTSA-Region) combining SEM and TSA. The model achieves a complete multi-level and multi-dimensional analysis by fully playing SEM's advantages in causality modelling and TSA's capacity to record time-dependent aspects.

2 Multifactor fusion modelling. This work addresses the restriction that conventional economic forecasting models cannot fully reflect the complicated interaction effects in economic data by merging the causality analysis of economic variables with the dynamic evolution of time series data.

3 Optimising model adaptivity. By optimising the settings of the TSA approach, the model increases the flexibility of the forecasting model to many economic cycles and event shocks, thereby aiming at the unpredictability and uncertainty of regional economic data.

4 Experimental validation and comparative analysis. The SEMTSA-Region model is confirmed for its benefits in forecasting accuracy and stability by means of comparison with numerous conventional economic forecasting models, therefore offering an experimental basis for the future enhancement of regional economic forecasting techniques.

2 Relevant technologies

2.1 SEM

Widely applied in economics and social sciences, SEM is a statistical technique for analysing the link between latent and observable variables (Muthén, 2002). Though they are indirectly measured through a collection of observed variables, latent variables are not immediately observable. Particularly appropriate for the study of complicated multivariate systems, SEM has strong theoretical modelling capacities and flexibility and can cope with several causal paths and correlations between variables at the same time (Lowry and Gaskin, 2014). Two components make up SEM: latent causal links between variables and the measurement model and structural model that explain the link between observed variables and latent variables.

The measurement model first addresses the link between the latent variables and the observable ones. Expression of the measurement model is Y as the observed variable and ξ as the latent variable:

$$Y = \left(y_1, y_2, \ldots, y_m\right)^T \tag{1}$$

$$\xi = \left(\xi_1, \xi_2, \ldots, \xi_k\right)^T \tag{2}$$

$$Y = \Lambda_y \xi + \epsilon \tag{3}$$

where Λ_y is the factor loading matrix linking the latent to observed variables; ξ is the latent variable; ϵ is the error term for the observed variable. Every observable variable includes some measurement error and is a linear mix of the latent variables. The equation shows how random errors of the latent variables affect the observable variables.

Conversely, the structural model explains the causal link among the latent variables (Fan et al., 2016). Let the latent variables have a causal link; the structural model may thus be expressed in the following form:

$$\xi = B\xi + \Gamma Z + \zeta \tag{4}$$

where Γ is the matrix representing the effect of the exogenous variable Z on the latent variables; ζ is the error term of the latent variables; B is the matrix of path coefficients between the latent variables, therefore reflecting the causal relationship between the latent variables. Covering the roles of endogenous and exogenous factors, this equation shows the interactions among the latent variables.

Maximum likelihood estimation (MLE) is the most often applied approach in SEM model parameter estimation (Asosega et al., 2022). Under a model parameter θ, let the observed dataset be Y and its probability density function be $f(Y|\ \theta)$. By optimising the likelihood function with the goal function, MLE identifies the model parameters:

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} f\left(y_i \,|\theta\right) \tag{5}$$

Often used to streamline the computations, the log-likelihood function is stated as:

$$L(\theta) = \sum_{i=1}^{n} \log f\left(y_i|\theta\right) \tag{6}$$

Optimising the log-likelihood function produces the best model parameter estimations. Commonly used optimisation strategies such the Newton-Raphson technique and the proposed Newton method update the parameters in the iterative process until convergence to the optimal solution, therefore improving the estimation efficiency.

Considering the measurement error, SEM may estimate the path coefficients of every latent variable in addition to exposing the causal linkages between them. SEM clearly benefits the study of complicated economic systems since it can simultaneously manage several causal routes and correlations among variables unlike in conventional regression analysis. In regional economic forecasting, SEM can be applied to expose the effects of possible factors such economic growth potential and market activity on observed variables such GDP and employment rate, so enabling researchers to thoroughly grasp the several causal relationships in economic systems.

Semantic analysis can construct the following model, for instance, if the observed variables include GDP (y_1) and unemployment rate (y_2) while the capacity for economic growth (ξ_1) and market activity (ξ_2) in the regional economic system are potential variables:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \lambda_{y_1\xi_1} & \lambda_{y_1\xi_2} \\ \lambda_{y_2\xi_1} & \lambda_{y_2\xi_2} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix} \tag{7}$$

And logical links between possible variables:

$$\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} + \begin{bmatrix} r_1 z_1 \\ r_2 z_2 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix} \tag{8}$$

One can determine the interrelationships between economic elements and their influence on regional economic development by means of the preceding equations. This modelling method offers a theoretical framework for policy analysis and forecasting as well as a means of clearly exposing the influence of possible elements on economic results.

Still, SEM has significant difficulties as well. First, SEM depends on high data quality and sample size; inadequate sample size could result in erratic estimations. Second, the accuracy of the theoretical framework determines the validity of the model; so, improper setup of the model could produce biassed estimation results. Consequently, while applying SEM, model validation and goodness-of-fit evaluation are very crucial.

Finally, being a potent statistical analysis tool, SEM finds great use in regional economic forecasting. By exposing the causal links among possible variables, it can offer more accurate theoretical basis for economic forecasting and simultaneously offer important policy references for decision makers.

2.2 TSA

Particularly for economic variables, TSA is absolutely important in regional economic forecasting (Ahlert, 2008). Like GDP, unemployment rate, inflation rate, economic data often show certain time series and trends. Appropriate time series models are absolutely essential for effective prediction of future economic developments.

Particularly suited for handling non-stationarity in economic time series data, autoregressive integral sliding average (ARIMA) model is a classic TSA technique (Dorais, 2024). Three components form the ARIMA model: sliding average (MA), autoregressive (AR), and differencing (I). The autoregressive part shows the link between the time series values and their historical values; the differencing part converts a non-stationary time series into a stationary one; and the sliding average part explains the effect of random perturbations in the time series on the current values.

Defined as the lagged relationship of the time series, the lag operator L in the ARIMA model reflects:

$$L^k y_t = y_{t-k} \tag{9}$$

where L^k refers to lagging the time series y_t by k time units. The ARIMA model has as its standard form:

$$\epsilon_t = \left(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p\right)\left(1 - \theta_1 L - \theta_2 L^2 - \cdots - \theta_q L^q\right) y_t \tag{10}$$

where ϕ_1, ϕ_2, …, ϕ_p is the autoregressive coefficient; θ_1, θ_2, …, θ_q is the sliding average coefficient; y_t is the observation at time point t; L is the lag operator; ϵ_t is the error term. Thus, the ARIMA model can effectively capture both long-term patterns and transient variations in the economic time series.

Usually non-stationary, that is, their mean and variance change with time, are economic time series data. Data has to be smoothed if it is to be acceptable for an ARIMA model. Smoothing usually is accomplished by differencing (Diaz et al., 2016). By computing the difference between the current value and the value at the prior moment, differencing eliminates the trend component. The difference operation defined for first order differencing is:

$$y_t' = y_t - y_{t-1} \tag{11}$$

Should non-stationarity persist following the first-order differencing, a second-order differencing procedure can help to further remove the trend component of the data. Following the differencing process, the time series can attain a smooth condition enabling the application of ARIMA model for forecasting and modelling.

Usually, partial autocorrelation function (PACF) of the data and autocorrelation function (ACF) define the orders p and q of the ARIMA model (Petrusevich, 2019). Whereas the PACF shows the correlation following the removal of the mediating lag components, the ACF explains the relationship between the time series and its lagged values. One may find the order of the MA and AR parts by use of analysis of the ACF and PACF graphs. The following two ideas guide the model order selection assuming smooth time series data after difference processing:

The order of the AR portion should be that order if the PACF plot collapses significantly following a given lag order.

The order of the MA portion should be that if the ACF plot falls significantly following a given lag sequence.

Usually depending on the MLE approach for parameter estimation, ARIMA models undergo training. Maximum likelihood estimate maximises the likelihood function of the observed data to guarantee the best fit, hence, guiding the model parameters. Estimating the autoregressive coefficient ϕ and the sliding average coefficient θ helps one to minimise the residuals – that is, the variation between the anticipated and actual values of the model.

Verifying the efficiency of the ARIMA model fit requires first residual analysis. The residuals should ideally be white noise, that is, have a zero mean, constant variance, and not be connected with temporal lags. Should the residuals show notable trends or relationships, the model fit is incomplete and either re-selection of the model parameters or adjustment of the model is necessary.

Forecasting future economic indicators can be done using the ARIMA model if it is trained and passes residuals diagnostics. ARIMA models have a forecasting formula:

$$\hat{y}_{t+h} = \mu + \sum_{i=1}^{p} \phi_i y_{t+h-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t+h-j} \tag{12}$$

where $\hat{y}_{t+h}$ is the anticipated value for the following h time steps; μ is the mean of the time series; ϵ_{t+h-j} is the historical error term applied in the forecast.

In particular in regional economic forecasting, the ARIMA model can assist in capturing the long-term trends and short-term variations of economic variables (such as the GDP). For instance, firstly we must do smoothness test and difference processing on the GDP data, then use ACF and PACF plots to choose the suitable model order, and lastly fit the ARIMA model by MLE and use the model to project the future GDP.

The Algorithm 1 pseudo-code reflects ARIMA model training and forecasting:

Algorithm 1 ARIMA model training and prediction

```
Input: Historical GDP time series data, chosen order parameters ($(p, d, q)$), forecast horizon $N$
Output: Forecasted GDP values for future time steps
begin
    Import GDP time series data;
    Check for missing values and handle them (e.g., imputation);
    Perform augmented Dickey-Fuller (ADF) test to check if the data is stationary;
    **If** p-value > 0.05 then
        Apply differencing: $y'_t = y_t - y_{t-1}$
        Recheck stationarity of the differenced series;
    **end if**
    **If** the series is still non-stationary, apply second-order differencing or further transformations;
        Plot the ACF and PACF to identify appropriate values for AR ($p$) and MA ($q$);
        Determine $p$ based on PACF plot (look for cut-off after lag $p$);
        Determine $q$ based on ACF plot (look for cut-off after lag $q$);
        Choose d based on the number of differencing applied;
        Initialise ARIMA model with parameters ($p$, $d$, $q$);
        Fit the ARIMA model to the differenced data;
        Estimate the coefficients $\phi$ (AR coefficients) and $\theta$ (MA coefficients);
        Calculate residuals: $e_t = y_t - \overline{y}_t$ (predicted values from the model);
        Plot residuals to check for white noise (zero mean, constant variance, no autocorrelation);
        **If** residuals are not white noise **then**
            Adjust $p$, $q$, or apply different transformations to the model;
    **end if**
    **For** $h = 1$ to $N$ (forecast horizon) **do**
    Forecast future GDP: $y_{t+h} = \mu + \sum_{i=1}^{p} \phi_i y_{t+h-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t+h-j}$;
    **end for**
    Visualise the forecasted values against the actual historical data;
    Calculate prediction intervals (e.g., 95%) to quantify uncertainty in the forecasts;
    Re-train the model periodically as new data becomes available;
    Output forecasted GDP values for future time steps;
end
```

By means of these processes, the ARIMA model can offer efficient support for regional economic forecasting, therefore enabling us to grasp economic trends and generate reliable forecasts. To guarantee the dependability of the model, the training process of the model calls not only appropriate smoothing and parameter selection but also residual analysis. ARIMA model is clearly a quite useful and efficient tool for economic data with robust time series.

3 Regional economic forecasting model based on structural equation modelling and time series: SEMTSA-Region

By means of multilevel modelling, this chapter suggests a regional economic forecasting model combining SEM and TSA, SEMTSA-Region, which is able to simultaneously consider the potential causal relationships in economic data and the dynamic characteristics of time series, so improving the accuracy of future trend of the regional economy. This model particularly captures the long-term trend of economic indicators by modelling the possible causal factors in the structural equation model using TSA techniques, therefore preserving sensitivity to short-term volatility.

3.1 Data pre-processing and differencing module

First, data preparation in economics. Usually, economic time series data shows trend or seasonal fluctuations; so, we must apply a differencing operation on the data to bring it into a smooth condition. Should the data show a non-smooth series, differencing removes the trend component to satisfy the smoothness assumption – a necessary condition for next modelling.

With an original economic time series y_t, the first-order differencing formula is as follows:

$$\Delta y_t = y_t - y_{t-1} \tag{13}$$

Second or higher order differencing can be carried on until the data attain a smooth condition if their stillness is still lacking.

Subsequent modelling and analysis using the data Δy_t following the differencing process guarantees that the data is fit for the time series model.

3.2 Module for modelling time series dynamics

This module's major goal is to use TSA techniques – especially AR and MA models to model the economic data and guarantee that the fundamental patterns and fluctuations in the time series are caught – capturing the dynamic aspects in the regional economic data. To get more accurate economic projections, this modelling approach will combine the dynamics of latent variables in SEM with the fundamental theory of time series.

We approach the time series data as a combination form of AR and MA models. ARMA models specifically allow one to depict the time series of the economic variable y_t:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} + \epsilon_t \tag{14}$$

where ϵ_t is the residual of the model and ε_t is the white noise error term.

This model helps one to understand how the past residuals affect the present value of an economic variable in addition to its historical value. Since many latent elements in the economic system often interact and are influenced by the external environment, in economic forecasting we often mix this time series modelling approach with latent variable modelling in SEM. Introducing the time series differential data Dt and the lag term of the autoregressive model helps one to consider the dynamic aspects of the time series in line with the causality in the structural equations.

This fusion method lets us evaluate the causal link between the latent variables in addition to catching the dynamic trend of the time series using historical data. In the end, for regional economic forecasting the multidimensional framework integrating SEM and TSA offers a more accurate modelling tool.

3.3 Model integration and optimisation module

This module aggregates the outputs of the two models using weighted integration since SEM and TSA respectively describe causality and time series aspects in economic systems. We characterise the weighted average of the SEM and TSA models as the forecast output $\hat{y}_t$ of the integrated model:

$$\hat{y}_t = w_1 \hat{y}_{SEM,t} + w_2 \hat{y}_{TSA,t} \tag{15}$$

where $\hat{y}_{SEM,t}$ is the structural equation model's prediction result; $\hat{y}_{TSA,t}$ is the prediction result of the TSA; w_1 and w_2 are the weight coefficients to be optimised; the sum of the weight coefficients is one:

$$w_1 + w_2 = 1 \tag{16}$$

The appropriate weighting factors are determined by optimising the objective function:

$$L(w_1, w_2) = \sum_{t=1}^{N} (y_t - \hat{y}_t)^2 \tag{17}$$

3.4 Prediction effectiveness evaluation module

Commonly used assessment criteria include mean square error (MSE), root mean square error (RMSE), and mean absolute error (MAE), which this module evaluates the prediction efficacy of the integrated model using. These measures give us a platform for more model modifications and enable us to measure the forecast accuracy of the model.

- MSE:

$$\text{MSE} = \frac{1}{N} \sum_{t=1}^{N} (y_t - \hat{y}_t)^2 \tag{18}$$

- RMSE:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{t=1}^{N} (y_t - \hat{y}_t)^2} \tag{19}$$

- MAE:

$$\text{MAE} = \frac{1}{N}\sum_{t=1}^{N}\left|y_t - \hat{y}_t\right| \tag{20}$$

These evaluation criteria enable the quantification and adaptation to actual needs of the predicted effectiveness of the model.

Algorithm 2 can depict SEMTSA-Region with the following pseudo-code.

Algorithm 2 Pseudo-code for calculating SEMTSA-Region

```
begin
    Initialise time series data $y_t$, latent variables $x_t$, and external variables $b_t$
    Initialise SEM parameters $\lambda_y$, $B_1$, and $B_2$
    Initialise ARMA model parameters $\phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q$
    Set maximum iterations max_iter and convergence threshold epsilon
    Set iteration count iter = 0
    while iter < max_iter do
        # Step 1: Pre-processing – check if the time series is stationary
        if not stationary($y_t$) then
            Perform differencing on $y_t$,:
    end if
        # Step 2: SEM-based latent variable modelling
        for t = 1 to length($y_t$) do
            $y_t = \lambda_y x_t, + \epsilon_t$ # SEM equation for observed variables
            $x_t = B_1 x_{t-1} + B_2 b_t + \eta_t$ # Update latent variables
        end for
        # Step 3: Time series modelling (ARMA)
        for $t = p + 1$ to length($y_t$) do
            $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$ # ARMA model
        end for
        # Step 4: Model fusion (Combine SEM and ARMA predictions)
        for $t = 1$ to length($y_t$) do
            $\gamma_{t\text{-pred}} = \lambda_y\ (B_1 x_{t-1} + B_2 b_t)$ # Combined SEM and ARMA prediction
        end for
        # Step 5: Check convergence
        if abs($\gamma_{t\text{-pred}}$) < epsilon then
            break # Convergence reached
        end if
        iter = iter + 1
    end while
  return $\gamma_{t\text{-pred}}$, $x_t$ # Return the predicted values and latent variables
  end
```

4 Experimental results and analyses

4.1 Experimental source

The China Statistical Yearbook of the National Bureau of Statistics provides the dataset used in this study. Covering a wide spectrum of economic variables, including gross domestic product (GDP), per capita income, investment, consumer expenditure, and inflation rates, the dataset comprises annual economic data for all Chinese provinces, municipalities, and autonomous areas. Highly representative and quite valuable, the data span 2011 to 2020.

Table 1 shows particular details on the dataset.

Table 1 Data overview

Indicator	*Data source*	*Time range*	*Number of records*	*Key content*
Gross domestic product (GDP)	National Bureau of Statistics	2011–2020	Ten years × 31 provinces	Annual GDP values for each province and city
Per capita income	National Bureau of Statistics	2011–2020	Ten years × 31 provinces	Per capita disposable income for each province and city
Fixed asset investment	National Bureau of Statistics	2011–2020	Ten years × 31 provinces	Fixed asset investment values for each province and city
Retail sales of consumer goods	National Bureau of Statistics	2011–2020	Ten years × 31 provinces	Retail sales of consumer goods in each province and city
Unemployment rate	National Bureau of Statistics	2011–2020	Ten years × 31 provinces	Annual unemployment rate for each province and city
Consumer price index (CPI)	National Bureau of Statistics	2011–2020	Ten years × 31 provinces	Annual changes in the consumer price index for each province and city

All of the data were pre-processed in the following stages throughout data usage to guarantee the accuracy of the analyses:

- Linear interpolation: it was used to fill in missing values for some provinces with missing values in particular years therefore guaranteeing the continuity of the data.
- Standardisation: every economic indicator has a different scale; so, all the data were normalised to guarantee that various variables had equal impact on the model.

To further enhance data quality, the data were checked for outliers and values much outside the typical range were eliminated.

4.2 Comparison experiments

The aim of this experiment is to evaluate in regional economic forecasting the performance variation between the SEMTSA-Region model and with several other conventional and sophisticated models. Particularly in its capacity to identify possible causal links between economic variables, we aim to confirm, via a comparative experiment, if the SEMTSA-Region model can offer better forecasting accuracy. We decide to compare the following models:

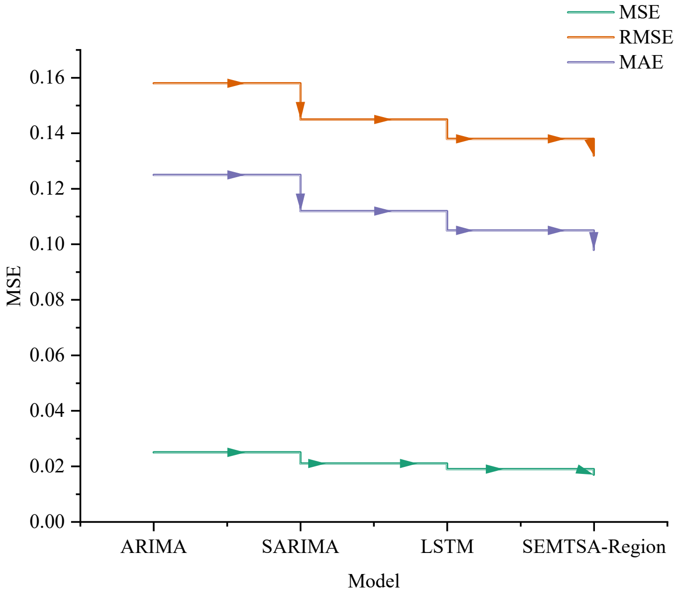
Using just historical data and forecasts for economic indicators, traditional time series model (ARIMA) generates this paradigm dismisses the possible causal links among economic factors.

Appropriate for economic data with seasonal fluctuations, a seasonal component is included to the ARIMA model (SARIMA).

Particularly fitting for nonlinear and very complex economic systems, long short-term memory network (LSTM) uses the LSTM model in deep learning to capture long-term dependencies in time-series data.

Combining SEM (SEM) to model causal linkages between economic variables, which are subsequently included into TSA to increase forecasting accuracy, SEMTSA-Region is a model.

Figure 1 Results of the comparison experiment (see online version for colours)



The experimental steps include data pre-processing, model training, prediction and evaluation. First, the raw data are processed with missing values and standardised to ensure the consistency of the input data of each model. Then, we use ARIMA, SARIMA, LSTM and SEMTSA-Region models to train the training set. Following all the models' training, we evaluated each one using MAE, RMSE, and MSE by means of the test set. Figure 1 exhibits the experimental outcomes.

The SEMTSA-Region model stands out from the experimental data as doing the best on all evaluation criteria. Particularly showing lower error rates than the other analysed models, the SEMTSA-Region model has an MSE of 0.017, an RMSE of 0.132, and an MAE of 0.098. This suggests that the model greatly helps to forecast economic trends and capture the possible relationship between economic factors.

Although they have some benefits in handling seasonality and time dependency, the conventional ARIMA and SARIMA models fall short in capturing the intricate causal linkages among economic variables, hence, reducing the forecast accuracy. Although it can capture complicated nonlinear patterns, the LSTM model does not fully use the causal relationships among economic variables and suffers from a certain degree of overfitting risk during the model training process; hence, its prediction performance is rather worse than that of SSTM. Its prediction ability is thus rather worse than that of the SEMTSA-Region model.

4.3 Parameter optimisation experiment

The objective of this experiment is to evaluate predicting accuracy by means of parameter optimisation of the SEMTSA-Region model. By changing the main model parameters, we investigate how better performance of regional economic forecasting results from. We confirm that the SEMTSA-Region model can attain the best prediction effect in actual applications by means of the optimisation experiment, so verifying the performance of the model under several parameter configurations.

This experiment mostly consists in tuning the essential parameters of the SEMTSA-Region model, which mostly relates with the following aspects:

SEM path choice: Forecasting effectiveness in SEM models depends in great part on causal path selection. We will evaluate the effect of several path topologies on model accuracy by trying several mixes of economic factors.

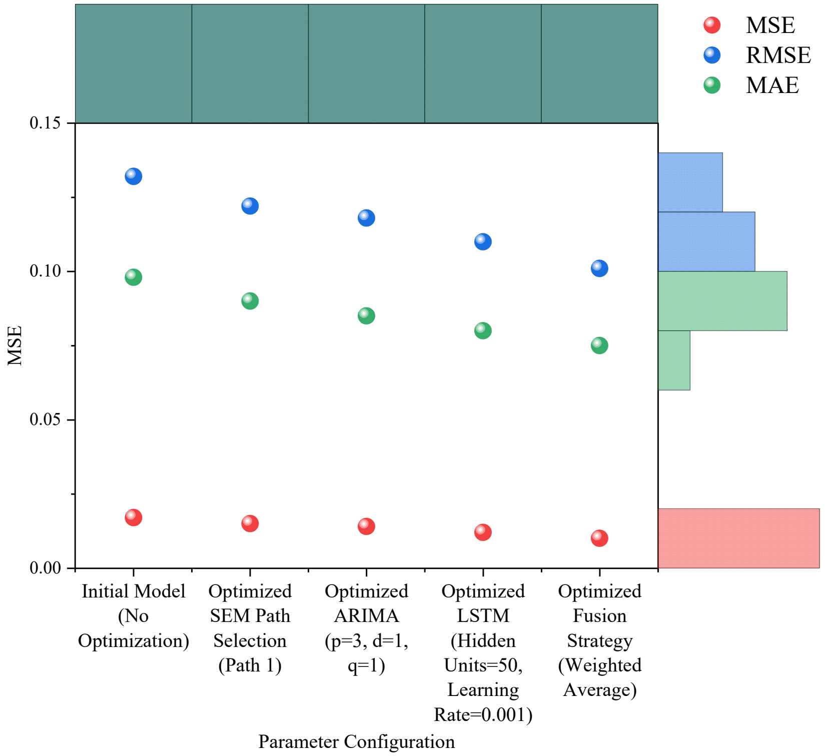
TSA parameter adjustment consists on component tuning of ARIMA and LSTM. We shall specifically change the ARIMA model's p, d, and q parameters as well as the hyperparameters learning rate in the LSTM network and number of hidden layer units respectively.

The SEMTSA-Region model combines SEM with TSA (e.g., ARIMA or LSTM), hence, optimising the fusion strategy. In this experiment, we will investigate several fusion techniques, e.g., by fusing the two portions of the predictions through a weighted averaging strategy or by changing the SEM outputs as time series inputs.

Following the models' training, we assess them under several parameter settings using metrics like MAE, RMSE, and MSE. Figure 2 displays the experimental outcomes.

In this work, the SEMTSA-Region model was parameterised to greatly raise the prediction accuracy. The first (unoptimised) model performs with MSE = 0.017, RMSE = 0.132, and MAE = 0.098; so, the baseline model and the outcomes show the fundamental prediction performance when untuned. Path optimisation of the SEM (SEM) helps to lower the MSE of the model to 0.015, RMSE to 0.122, and MAE to 0.090, thereby demonstrating the favourable effect of path optimisation in capturing the causal linkages between economic variables. After modifying its p, d, and q parameters, the MSE is dropped to 0.014, RMSE dropped to 0.118, and MAE dropped to 0.085. The optimisation of the ARIMA model also brings about a notable enhancement. Although ARIMA can fit linear time series data better, its fitting effect on nonlinear connection is still poor compared to LSTM.

Figure 2 Experimental results of parameter optimisation (see online version for colours)



By optimising the hyperparameters such the number of hidden layer units and the learning rate, the MSE is lowered to 0.012, the RMSE is lowered to 0.110, and the MAE is lowered to 0.080, so indicating the superiority of the LSTM in handling the complicated nonlinear relationships. Proving the efficient merger of SEM and TSA, the weighted average strategy's optimal outcomes in MSE (0.010), RMSE (0.101) and MAE (0.075) are ultimately best. The experimental results show that the prediction accuracy of SEMTSA-Region model in regional economic forecasting is much enhanced by precisely changing each parameter, so illustrating its advantages in handling economic data, particularly in combining causality and time dependence.

5 Conclusions

In order to address the causality and time-dependence issues in intricate economic data, we suggest in this work a regional economic forecasting model, SEMTSA-Region, based on SEM and TSA. First, we use SEM to capture the causal links between regional economic variables; then, we mix time series models (e.g., ARIMA and LSTM) to forecast economic variables in time series, and lastly we combine the benefits of both by means of the weighted average fusion strategy to raise the forecasting accuracy. In terms of prediction accuracy, especially when dealing with nonlinear relationships and complex time series data, which clearly shows benefits, the model verifies its effectiveness in the experimental part by means of comparison experiments and parameter optimisation trials, so surpassing the conventional single model.

There are still certain restrictions even although the SEMTSA-Region model suggested in this paper has shown notable achievements in regional economic forecasting. First of all, success of this model depends much on the quality and variety of the data. The timeliness, completeness, and correctness of the data may still influence the prediction of the model even if the dataset utilised in this study comprises of multiple economic factors. Second, the model implies that the economic data used in this study reflect future economic trends; so, they are mostly based on historical statistics. Nevertheless, erratic outside events (such as natural catastrophes, policy changes, unexpected epidemics, etc.) often influence the real economic environment, therefore producing historical data that might not be relevant for future projections. Furthermore, although the SEMTSA-Region model effectively combines SEM and TSA, the great model complexity and significant computing resource consumption – particularly in relation to large-scale data – may cause computational bottlenecks in pragmatic applications. This will influence the model's real-time prediction capacity particularly in situations involving economic decision-making that call for quick reaction.

Future research could be expanded in several areas:

1 Optimising the computational efficiency of the model. Although the SEMTSA-Region model has a good prediction accuracy, its computational cost is somewhat significant, particularly in cases of large-scale datasets where computational bottlenecks could arise. Thus, by using effective algorithm optimisation strategies like parallel computing and distributed computing to increase the feasibility of its use, particularly in the big data environment, the computational efficiency of the model can be increased in the future.

2 Comprehensive model considering the influence of multiple factors. Regional economic forecasting can be influenced by social, environmental, and other elements in addition to economic ones. Through multimodal learning approaches (e.g., multi-input networks, graph neural networks, etc.), future research can include more external aspects – such as climate change, social policies, international markets, etc. – and integrate these factors into the model.

3 Multi-level fusion models. Models' fusion technique is implemented using weighted averaging at present. Multilevel fusion techniques can be tried to mix several model outputs through more sophisticated combination methods (e.g., model integration, meta-learning, reinforcement learning, etc.) to further improve the prediction accuracy, especially the performance when different kinds of data sources are fused.

Acknowledgements

This work is supported by the Shandong Provincial Natural Science Foundation Project 'Research on Macro and Micro Mechanisms of PPPs' (No. ZR2021QG068).

Informed consent declaration

Written informed consent was obtained from all individual participants included in the study.

References

Acs, Z.J. and Szerb, L. (2007) 'Entrepreneurship, economic growth and public policy', *Small Business Economics*, Vol. 28, pp.109–122.

Ahlert, G. (2008) 'Estimating the economic impact of an increase in inbound tourism on the German economy using TSA results', *Journal of Travel Research*, Vol. 47, No. 2, pp.225–234.

Amado, C. and Teräsvirta, T. (2014) 'Conditional correlation models of autoregressive conditional heteroscedasticity with nonstationary GARCH equations', *Journal of Business & Economic Statistics*, Vol. 32, No. 1, pp.69–87.

Asosega, K.A., Iddrisu, W.A., Tawiah, K. et al. (2022) 'Comparing Bayesian and maximum likelihood methods in structural equation modelling of university student satisfaction: an empirical analysis', *Education Research International*, Vol. 2022, No. 1, p.3665669.

Diaz, J.M., Dormido, S. and Rivera, D.E. (2016) 'ITTSAE: a set of interactive software tools for time series analysis education [lecture notes]', *IEEE Control Systems Magazine*, Vol. 36, No. 3, pp.112–120.

Dorais, S. (2024) 'Time series analysis in preventive intervention research: a step-by-step guide', *Journal of Counseling & Development*, Vol. 102, No. 2, pp.239–250.

Fan, Y., Chen, J., Shirkey, G. et al. (2016) 'Applications of structural equation modeling (SEM) in ecological studies: an updated review', *Ecological Processes*, Vol. 5, pp.1–12.

Hannaford, J., Lloyd-Hughes, B., Keef, C. et al. (2011) 'Examining the large-scale spatial coherence of European drought using regional indicators of precipitation and streamflow deficit', *Hydrological Processes*, Vol. 25, No. 7, pp.1146–1162.

Karimi, L. and Meyer, D. (2014) 'Structural equation modeling in psychology: the history, development and current challenges', *International Journal of Psychological Studies*, Vol. 6, No. 4, pp.123–133.

Lei, C., Deng, J., Cao, K. et al. (2019) 'A comparison of random forest and support vector machine approaches to predict coal spontaneous combustion in gob', *Fuel*, Vol. 239, pp.297–311.

Lou, W., Zhang, D. and Bayless, R.C. (2020) 'Review of mineral recognition and its future', *Applied Geochemistry*, Vol. 122, p.104727.

Lowry, P.B. and Gaskin, J. (2014) 'Partial least squares (PLS) structural equation modeling (SEM) for building and testing behavioral causal theory: when to choose it and how to use it', *IEEE Transactions on Professional Communication*, Vol. 57, No. 2, pp.123–146.

Mutale, B., Withanage, N.C., Mishra, P.K. et al. (2024) 'A performance evaluation of random forest, artificial neural network, and support vector machine learning algorithms to predict spatio-temporal land use-land cover dynamics: a case from lusaka and colombo', *Frontiers in Environmental Science*, Vol. 12, p.1431645.

Muthén, B.O. (2002) 'Beyond SEM: general latent variable modeling', *Behaviormetrika*, Vol. 29, No. 1, pp.81–117.

Petrusevich, D. (2019) 'Time series forecasting using high order ARIMA functions', *International Multidisciplinary Scientific GeoConference: SGEM*, Vol. 19, No. 2.1, pp.673–679.

Sato, T. and Matsuda, Y. (2021) 'Spatial extension of generalized autoregressive conditional heteroskedasticity models', *Spatial Economic Analysis*, Vol. 16, No. 2, pp.148–160.

Schaffer, A.L., Dobbins, T.A. and Pearson, S-A. (2021) 'Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions', *BMC Medical Research Methodology*, Vol. 21, pp.1–12.

Westland, J.C. (2015) 'Structural equation models', *Stud. Syst. Decis. Control*, Vol. 22 No. 5, p.152.

Yang, L., Wall, G. and Smith, S.L. (2008) 'Ethnic tourism development: Chinese Government perspectives', *Annals of Tourism Research*, Vol. 35, No. 3, pp.751–771.

Yu, L., Zhou, L., Tan, L. et al. (2014) 'Application of a new hybrid model with seasonal auto-regressive integrated moving average (ARIMA) and nonlinear auto-regressive neural network (NARNN) in forecasting incidence cases of HFMD in Shenzhen, China', *PloS One*, Vol. 9, No. 6, p.e98241.