Time Series Interpolation (TSI): Leveraging Google Trends and LSTM

Azizallah Farhadi*
Sara Ramazani†

August 11, 2024

Abstract

This study introduces the TSI-LSTM hybrid model, leveraging Long Short-Term Memory (LSTM) networks for temporal interpolation. The model effectively captures long-term dependencies in daily data sequences, using Google Trends as a proxy for official economic indicators. Through a meticulous hyperparameter tuning process , the architecture of the TSI-LSTM model is optimized to enhance forecasting accuracy.

The TSI-LSTM model demonstrates a high predictive accuracy, achieving an R^2 score of 97%. This approach successfully converts low-frequency data into high-frequency predictions, enabling more precise short-term nowcasting compared to traditional regression methods. The findings underscore the potential of integrating Google Trends data with advanced machine learning techniques for real-time economic analysis. The study's implications extend to policymakers and economists, providing a robust framework for timely and accurate economic forecasting. Future research could explore other machine learning models like GRU or Transformer networks, and the integration of multiple high-frequency data sources to further improve the robustness of economic nowcasting.
Keywords: LSTM,interpolation,disaggregation,now-casting,MIDAS,machine learning,google trends

1 Introduction

The dissemination of economic information, particularly during emergencies, often faces challenges in accuracy and timeliness. The ”information gap in accuracy” refers to the disparity between needed and available information, while the ”information gap in timeliness” highlights the delay in receiving necessary information Sax and Steiner (2013). Statistics are crucial for decision-making and reflecting economic policies and behaviors. However, delays between data collection and dissemination can hinder effective policy decisions, obscuring crucial information and preventing timely actions. Therefore, pursuing real-time statistics is critical to enhance decision-making accuracy despite the persistent lag between data collection and delivery.

Nowcasting has emerged as a promising approach to address this issue. This study introduces an innovative nowcasting method utilizing Google Trends (GT) as a real-time data source. Traditional nowcasting methods base predictions on the frequency of the target variable. In contrast, this study proposes basing predictions on the frequency of feature variables, allowing for daily forecasts even if the target variable is typically reported monthly Itkonen and Juvonen (2017). By leveraging the more frequent data from feature variables, this approach aims to provide more timely and accurate predictions.

When dealing with mixed frequency data, predicting daily outcomes poses a significant challenge. Existing libraries like MIDAS typically predict based on the frequency of the target variable, creating a gap when attempting to forecast at the higher frequency of the feature variables. Aligning low-frequency target variables with higher-frequency feature data results in missing values.

*PhD in Econometrics;E-Mail:azizfarhadi1974@gmail.com
†Faculty member of Gisma University of Applied Sciences

Current algorithms, such as LSTM networks, do not inherently handle missing values effectively, requiring pre-processing steps to manage these missing values before inputting the data into the network. Common methods for estimating missing target values include interpolation techniques and advanced imputation methods like Multiple Imputation by Chained Equations (MICE) and Kalman Filtering. In the literature on nowcasting, which has gained momentum in recent years, the focus is on short-term forecasting, with MIDAS models being among the best approaches. However, predictions with MIDAS are constrained by the frequency of the target variable. If we can use modern methods to increase the frequency of the target variable, it would allow us to make forecasts that are even shorter-term than traditional nowcasting. For this reason, we have termed this much shorter version of nowcasting as ”short-run nowcasting.”

This research proposes using Google Trends data as a proxy to address missing target values, leveraging the correlation between search trends and the target variable. This method provides an effective solution for imputation and improves the accuracy of higher frequency forecasts when dealing with mixed frequency data.

From a theoretical perspective, this research enhances the existing body of knowledge in time series analysis and data imputation. By introducing the use of Google Trends data, it presents an innovative technique that integrates external data sources with TSI and LSTM networks. This fusion broadens the scope of imputation techniques and showcases the potential of combining machine learning with econometric models.

In practice, this research offers substantial benefits for industries relying on accurate and timely data analysis. High-frequency data improves decision-making processes, enables precise forecasting, and enhances risk management strategies. The use of Google Trends data provides a cost-effective and accessible alternative for data augmentation, applicable in scenarios ranging from market research to public health monitoring.

In time series forecasting, aligning frequencies between target and feature variables is crucial for model performance. Traditional methods like MIDAS handle mixed frequency data but maintain the original frequency of the target variable. This research addresses the need for a method that increases the frequency of the target variable to match higher-frequency feature variables before forecasting. The proposed approach involves using TSI to increase the target variable’s frequency and applying an LSTM model for forecasting. This methodology aims to improve the timeliness and accuracy of predictions compared to traditional models.

This paper aims to address the challenge of augmenting data frequency, focusing on leveraging Google Trends as an alternative source of real-time economic indicators and adopting LSTM models over traditional regression models. The objectives are to:

1. Investigate the potential of Google Trends data as a real-time indicator of economic behaviors.
2. Develop LSTM models to analyze high-frequency data from Google Trends for interpolation purposes.
3. Compare the effectiveness of this approach against traditional statistical methods.
4. Assess the impact of integrating real-time data into economic decision-making processes.

Based on existing theories, the following hypotheses are formulated:

- **Null Hypothesis (H0):** Using TSI to align the frequency of the target variable with the feature variables does not significantly impact prediction accuracy.
- **Alternative Hypothesis (H1):** Using TSI to align the frequency of the target variable with the feature variables significantly improves prediction accuracy.

1.1 Methodology Overview

This study utilizes Long Short-Term Memory (LSTM) networks for temporal interpolation, capitalizing on their capacity to capture long-term dependencies in sequential data. The LSTM model in this study processes input sequences of daily data, which have been previously interpolated using appropriate methods (Figure 1). Each input sequence spans a length of 30 days or more.

To optimize the LSTM model's architecture and hyperparameters, hyperparameter tuning is conducted using the `kerastuner.engine.hyperparameters` module. This tuning process involves selecting the optimal number of layers and adjusting various hyperparameters to improve the model's performance.

After training and optimizing the LSTM model, it is evaluated using the Python programming language. The evaluation focuses on the model's ability to accurately predict the target variable based on Google Trends data.

Overall, the methodology presented in this study demonstrates how LSTM networks can be employed for temporal interpolation, offering a robust framework for forecasting economic variables with alternative data sources such as Google Trends.

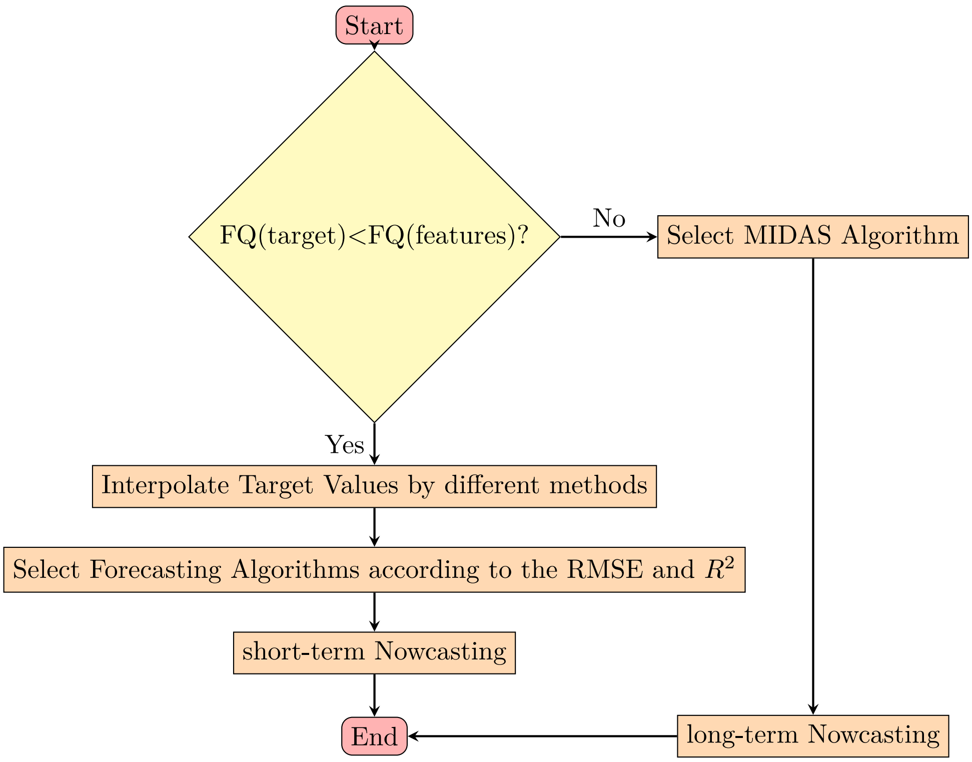


Figure 1: from increasing frequency(FQ) to Selecting Forecasting Algorithm(Source:Resercher)

If we want to understand Flowchart 1 more precisely, it is better to take a look at Tables 1 and 2. In Table 1, as you can see, the target data is placed in MIDAS model without any change, and the number for the next period (here the next month) is predicted. However, in Table 2, it can be seen that the data of the target variable has been changed in such a way that we are able to use the LSTM model to predict shorter periods. Since the term *nowcasting* is a general concept for near-term forecasting, to distinguish between monthly and daily forecasting, in this dissertation, I will introduce new terminology. We will refer to it as *short-term nowcasting* for daily forecasting and *long-term nowcasting* for monthly forecasting.

2 Literature Review

In time series analysis, algorithms predict the target variable based on its inherent frequency. For instance, if the target variable is recorded monthly, the forecast will also be generated on a monthly basis. However, to increase the prediction frequency, temporal interpolation is used. While Ghysels, Santa-Clara, and Valkanov (Ghysels et al. (2004)), Ball (Ball and Gallo (2018)), and Guay (Guay and Maurin (2015)) MIDAS[1] regression models to analyze time series data sampled at different

[1]MIDAS (Mixed Data Sampling) is a regression technique designed to handle and incorporate data with different frequencies (e.g., daily, weekly, monthly) into a single model. It is particularly useful in economic and financial

Table 1: Example of mixed frequency dataset and MIDAS prediction

Date	**Feature1(Daily)**	**Feature2(Daily)**	**Feature3(Daily)**	**Target(Monthly)**
2023-01-01	0.5	1.2	3.4	500
2023-01-02	0.6	1.3	3.5	
2023-01-03	0.4	1.1	3.2	
⋮	⋮	⋮	⋮	
2023-01-31	0.7	1.4	3.6	
2023-02-01	0.5	1.2	3.4	100
2023-02-02	0.6	1.3	3.5	
2023-02-03	0.4	1.1	3.2	
⋮	⋮	⋮	⋮	
2023-02-28	0.7	1.4	3.6	
2023-03-01	0.5	1.2	3.4	MIDAS prediction*
2023-03-02	0.6	1.3	3.5	
2023-03-03	0.4	1.1	3.2	
⋮	⋮	⋮	⋮	
2023-02-31	0.7	1.4	3.6	

* MIDAS monthly prediction

Table 2: Example of mixed frequency dataset TSI-LSTM prediction.

Date	**Feature1(Daily)**	**Feature2(Daily)**	**Feature3(Daily)**	**Target***
2023-01-01	0.5	1.2	3.4	10
2023-01-02	0.6	1.3	3.5	12
2023-01-03	0.4	1.1	3.2	15
⋮	⋮	⋮	⋮	⋮
2023-01-31	0.7	1.4	3.6	26
2023-02-01	0.5	1.2	3.4	10
2023-02-02	0.6	1.3	3.5	12
2023-02-03	0.4	1.1	3.2	11.5
⋮	⋮	⋮	⋮	⋮
2023-02-28	0.7	1.4	3.6	16
2023-03-01	0.5	1.2	3.4	TSI-LSTM prediction†
2023-03-02	0.6	1.3	3.5	TSI-LSTM prediction†
2023-03-03	0.4	1.1	3.2	TSI-LSTM prediction†
⋮	⋮	⋮	⋮	⋮
2023-03-31	0.7	1.4	3.6	TSI-LSTM prediction†

* Monthly variable that is converted to daily by TSI method
† TSI-LSTM daily prediction(short-term Nowcasting)

frequencies, these approaches are not optimal for predicting time series with feature frequencies higher than the target variable, suitable for short-run prediction. Hence, Time Series Interpolation (TSI) has gained attention, especially for understanding the short-term economic behavior.

analyses, where it's necessary to combine data from various time intervals. MIDAS allows you to leverage high-frequency data to improve predictions of lower-frequency variables, such as inflation rates or GDP.

2.1 Literature of Temporal Interpolation

The loss of information due to temporal data aggregation is significant. Monthly and quarterly data exhibit more complex time-series processes than annual data, which tends to show simpler processes with minimal cyclical variation. Aggregated data often display long-term persistence, losing cyclical variations seen in the original data Rossana and Seater (1995).

A variety of methods for TSI have been developed. Mathematical models treat unknown sub-annual series as deterministic, adhering strictly to annual constraints, while statistical models treat them as stochastic, allowing more flexibility. The choice depends on available information and preferences Denton (1971); Causey and Trager (1981); Boot et al. (1970).

This research evaluates methods for interpolation and temporal distribution in routine sub-annual estimations within national accounts. We focus on the Denton adjustment method Denton (1971), the Causey-Trager growth preservation method Causey and Trager (1981), and the smoothing technique by Boot, Feibes, and Lisman Boot et al. (1970). Additionally, we assess five extensions of the Chow-Lin regression approach Chow and Lin (1971), including AR(1) models, the Fernandez random walk model Fernández (1981), and the Litterman random walk-Markov model Litterman (1998).

2.1.1 Mathematical Methods for TSI

Denton Adjustment Method The Denton method is based on movement preservation, minimizing the deviation between sub-annual estimates x_t and indicator values z_t under temporal aggregation constraints: $\min_x (z-x)'A(z-x)$, subject to $y = B'x$, where y represents annual values, and B maps sub-annual estimates to annual constraints. Variants of this method include additive and proportional first and second differences, which aim to preserve different aspects of the sub-annual series' movements relative to the indicator series.

Causey-Trager Method The Causey-Trager model aims to preserve the proportional period-to-period growth rate $g_t = x_t/x_{t-1}$. The objective is to minimize the deviation of x_t from z_t under the condition $x_t/x_{t-1} = z_t/z_{t-1}$: $\min \sum_{t=1}^{T}(x_t - z_t)^2$, subject to $\frac{x_t}{x_{t-1}} = \frac{z_t}{z_{t-1}}$.

Boot-Feibes-Lisman Smoothing Method The Boot-Feibes-Lisman method fits a smooth curve to annual data to generate sub-annual estimates, assuming the annual series can be interpolated smoothly: $\sum_{t=1}^{T} x_t^2 \to \min$, subject to $\sum_{t=1}^{T} x_t = y$.

Chow-Lin Regression Method Extensions The Chow-Lin regression method is a widely used technique for temporal disaggregation, especially suited for stationary or cointegrated series. This method performs a Generalized Least Squares (GLS) regression on the low-frequency series using high-frequency indicators. Extensions of the Chow-Lin method include various models that account for different assumptions about the relationship between the high-frequency series and the low-frequency series, including the AR(1) model, Fernandez random walk model, and Litterman random walk-Markov model.

Fernandez Random Walk Model The Fernandez model is designed to handle non-cointegrated series by assuming that the sub-annual series follows a random walk: $x_t = x_{t-1} + \epsilon_t$, where $\epsilon_t \sim N(0, \sigma^2)$. This model is particularly useful when the underlying data-generating process of the series suggests non-stationarity, reflecting situations where the level of the series changes unpredictably over time.

Litterman Random Walk-Markov Model The Litterman model extends the Fernandez approach by incorporating a Markov switching mechanism, where the sub-annual errors follow an autoregressive process of order 1 (AR(1)): $u_t = \rho u_{t-1} + \epsilon_t$, with $\epsilon_t \sim N(0, \sigma^2)$ and $|\rho| < 1$. Additionally, the Litterman model assumes that the series can switch between different regimes, governed by a Markov process: $x_t = \mu_{s_t} + \phi x_{t-1} + \epsilon_t$, where s_t is a state variable with a regime-dependent mean μ_{s_t}. This approach allows the model to capture structural changes or regime shifts in the data, making it suitable for series that exhibit different behaviors over time. The AR(1) process is typically estimated using Maximum Likelihood (ML) or Generalized Least Squares (GLS) methods.

2.2 Comparing the power of Prediction Models

In the previous parts, various methods of TSI were discussed and investigated. But these methods are not able to predict, so we need to use special prediction algorithms. This part of the paper is dedicated to the topic of which is the best algorithm for forecasting for a very short period of time. As mentioned before, one of the famous algorithms in this regard is LSTM[2]. As its name suggests, it retains short-term memory and these characteristics have made many researchers use this algorithm in forecasting.

In study byElsaraiti and Merabet (2021), actual wind speed data were compared using the traditional ARIMA model and the deep learning-based LSTM model. The results demonstrated that the LSTM model is highly effective, with a lower error rate, making it preferable for forecasting compared to other models. LSTM's capability to recognize patterns over long periods enhances its prediction efficiency when implemented with deep learning. While previous academic studies indicated that the ARIMA model performed better with smaller datasets, this study shows that with larger datasets, deep learning algorithms like LSTM outperform traditional algorithms such as ARIMA. The study recommends further research with more real data and comparative analysis with other studies to validate these findings . Hopp (Hopp (2021)) found that LSTM outperformed the dynamic factor model (DFM) in nowcasting global merchandise export values and volumes, as well as global services exports.
Hopp(Hopp (2022)) further compared LSTM and DFM performance during the COVID-19 pandemic and found that LSTM had better performance in terms of mean absolute error and root mean square error for these variables. Additionally, Hopp introduced a methodology to introduce interpretability to LSTMs. These findings indicate that LSTM can be a valuable tool for economic nowcasting. A comparison of nowcasting methodologies(figure 2) for US quarterly GDP growth found that long short-term memory artificial neural networks (LSTM) and Bayesian vector autoregression (BVAR) performed best Hopp (2023). The sparse-group LASSO estimator, particularly in the form of structured machine learning regressions, also showed promise in nowcasting, outperforming unstructured LASSO and other alternatives (Babii, 2019; Babii, 2020). Additionally, the use of statistical learning on the social web, particularly Twitter, has been successful in inferring events and phenomena, such as rainfall levels and regional Influenza-like Illness rates (Lampos, 2012)Lampos and Cristianini (2012). The papers collectively suggest that LSTM artificial neural networks (LSTM) are effective in economic nowcasting.

In this section we provided a comprehensive review of several methods used for TSI of time series data. TSI refers to the process of converting low-frequency data, such as annual data, into higher-frequency data, such as quarterly or monthly data, while maintaining consistency with the original data. The methods discussed in the article include the Chow-Lin method (1971), Fernandez method (1981), Litterman method (1983), Denton method (1971), and the Denton-Cholette method (2006). Each method has its unique approach and assumptions, catering to different types of data and desired outcomes. For instance, the Chow-Lin method employs generalized least squares estimation, assuming an autoregressive process for quarterly residuals, while the Denton method focuses on minimizing deviations from a differenced indicator series. Moreover,we highlighted the evolution and advancements in TSI techniques over the years. The Denton method, originally introduced in 1971, was later modified by Cholette in 2006 to address the issue of spurious transient movements at the beginning of the resulting series. The Fernandez and Litterman methods, introduced in the 1980s, both assume non-stationary processes for quarterly residuals, with the latter incorporating an autoregressive component. These methods provide valuable tools for economists and statisticians, enabling more accurate and detailed analysis of economic indicators and other time series data. The article emphasizes the importance of selecting the appropriate method based on the specific characteristics of the data and the analytical goals(table 3).

[2]Long Short-Term Memory

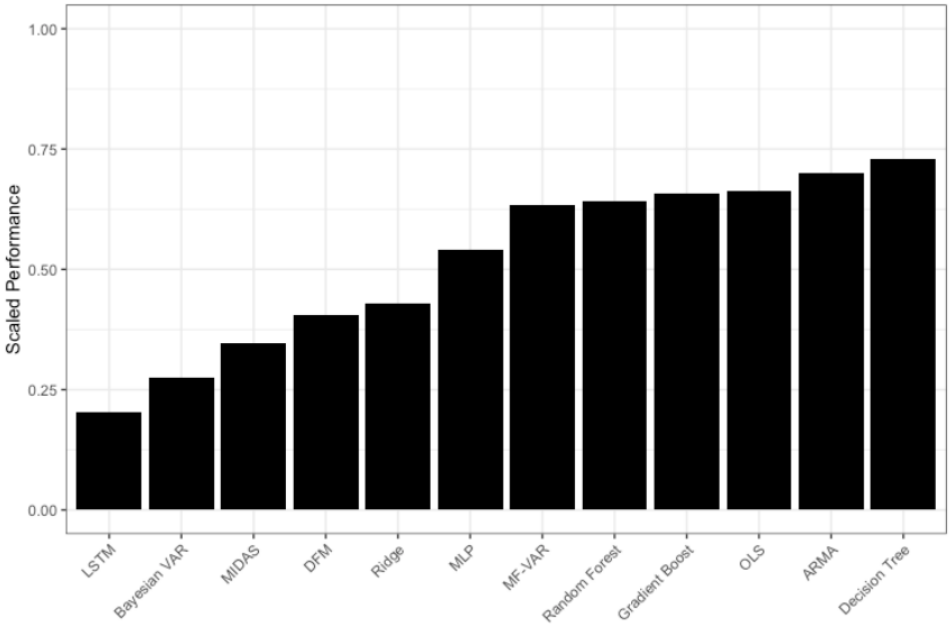


Figure 2: compare of different approaches for TSI(source:Hopp(2023))

Beyond the challenge of TSI, we also explored the literature on time series forecasting and discovered that the LSTM model excels at making predictions over very short time intervals.
Extensive research on temporal disaggregation has been conducted, with studies continuing up to 2014. This body of work has resulted in the introduction of nearly 18 scientific methods. However, in recent years, due to the emergence of COVID-19 and regional and global crises, the focus of major publications has shifted more towards the applications of these methods. One possible reason for this shift is the increased production of data and the need to enhance data frequency in recent years, utilizing these methods to increase the frequency of other datasets.
For instance, Gross Domestic Product (GDP) data, typically produced on a quarterly basis by official statistical agencies, has seen efforts to be calculated monthly. However, for many countries, the financial burden of monthly calculations proved challenging. As a result, the application of these techniques to increase data frequency has been widely adopted. Therefore, while the methods themselves have not changed significantly, their applications have seen a substantial increase.

3 Methodology

This study adopts a positivist research philosophy, emphasizing empirical testing of hypotheses derived from established theories. It combines time series interpolation (TSI) and Long Short-Term Memory (LSTM) models to address their individual limitations. TSI's difficulty in predicting out-of-sample data is mitigated by the predictive strength of LSTM, while TSI improves the handling of missing values in LSTM. This integration is empirically validated through experiments on various datasets.

A deductive approach is used, where hypotheses based on existing literature are tested with statistical and machine learning methods. For TSI, nine statistical approaches are evaluated, selecting the best method based on Mean Squared Error (MSE) and R^2. The LSTM model, known for short-term time series prediction, forecasts target data, with performance assessed using MSE and R^2. Hyperparameter tuning is conducted to optimize the model.

A mono-method quantitative approach is employed, relying on numerical data and statistical analysis. This study uses a non-sampling method, analyzing existing open data sources such as Google Trends and German foreign trade statistics. This secondary data analysis leverages publicly available datasets for insights.

Data analysis combines statistical techniques and machine learning algorithms. Statistical

Table 3: Summary of TSI Methods

Method Name	Year	Description
Denton Adjustment Method	1971	Based on the principle of movement preservation, ensures the sub-annual series preserves the movement in the indicator series.
Causey-Trager Method	1981	Minimizes the sum of squares of deviations under the condition that sub-annual and indicator series grow at the same rate.
Boot-Feibes-Lisman Smoothing Method	1970	Fits a smooth curve to annual data for sub-annual estimates. Assumes annual series can be interpolated by a smooth curve.
Chow-Lin Regression Method	1971	Five extensions including AR(1) model, Fernandez random walk model, and Litterman random walk-Markov model.
Multivariate Structural Time Series Model	2005	Uses a seemingly unrelated time series equations model and the Kalman filter for TSI.
AR(1) Model	2006	Sub-annual series follows an autoregressive process. Estimated using ML or GLS methods.
Fernandez Random Walk Model	1981	Sub-annual series follows a random walk process. Estimated using ML or GLS methods.
Litterman Random Walk-Markov Model	1998	Combines a random walk process with a Markov switching mechanism. Estimated using ML or GLS methods.
ANN Approach	2014	Uses artificial neural networks to disaggregate annual US GDP data into quarterly increments.
ARIMA and SUR Approach	1990	Uses ARIMA-based approach for TSI, developing an optimal estimator for disaggregated series.

methods explore patterns and relationships, while data science techniques enhance the depth of analysis. This comprehensive approach aims to increase data frequency for predictive purposes.

The research strategy integrates archival research with experimental design. Archival research analyzes existing datasets, such as Google Trends, for patterns. Experimental design involves developing and testing an LSTM model to predict economic indicators, focusing on increasing data frequency. The study uses time series data from Google Trends and German foreign trade data from 2006 to November 2023. The analysis involves data collection, preprocessing, frequency enhancement methods, forecasting techniques, and performance evaluation using RMSE and R^2.

This cross-sectional study analyzes data collected from 2006 to November 2023. It includes daily Google Trends data and monthly German foreign trade data, employing a multivariate time series analysis.

The primary technique is the TSI-LSTM method, a combination of time series interpolation and Long Short-Term Memory networks, a type of recurrent neural network suitable for time series forecasting. The process involves data collection, preprocessing, frequency enhancement, and forecasting, with model performance evaluated based on RMSE and R^2 metrics.

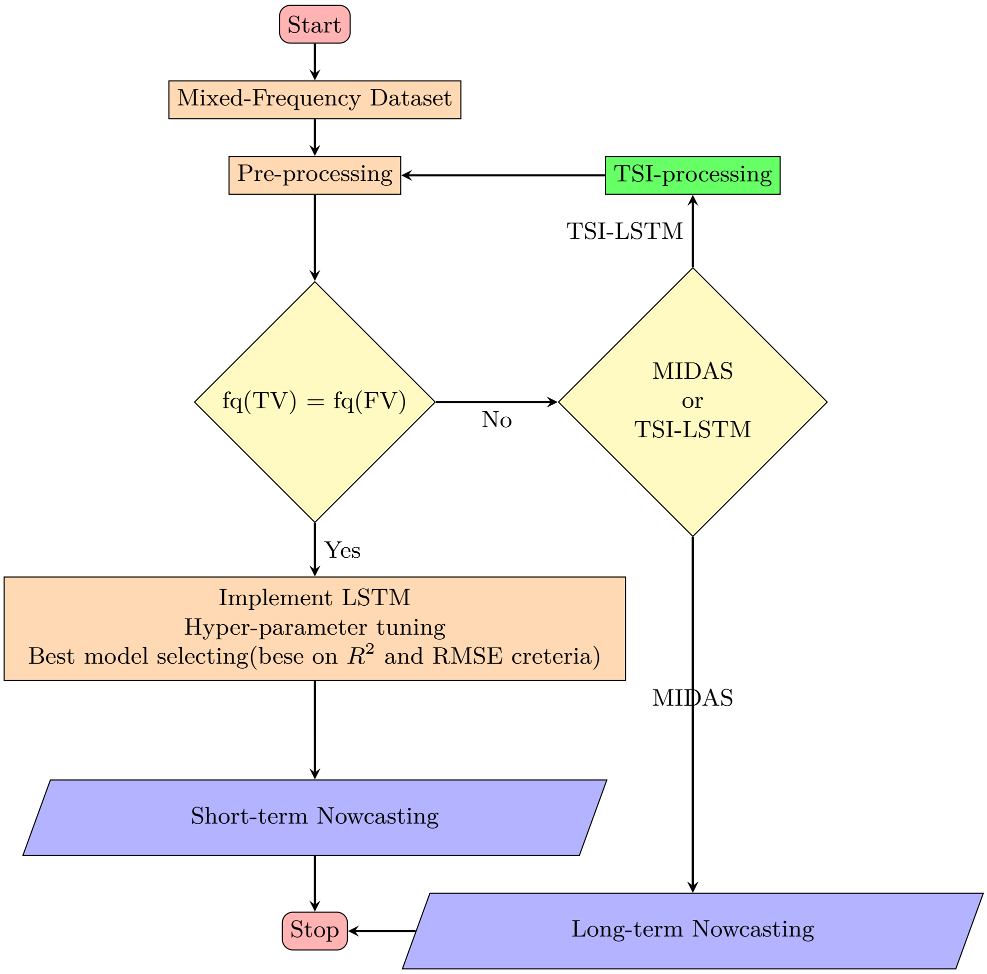


Figure 3: TSI-LSTM Implementation Pipeline

3.0.1 The LSTM architecture

LSTM,which were introduced by Sepp Hochreiter and Jürgen Schmidhuber in 1997(Hochreiter and Schmidhuber (1997)) , is a type of recurrent neural network (RNN) architecture designed to address the vanishing gradient problem in traditional RNNs.

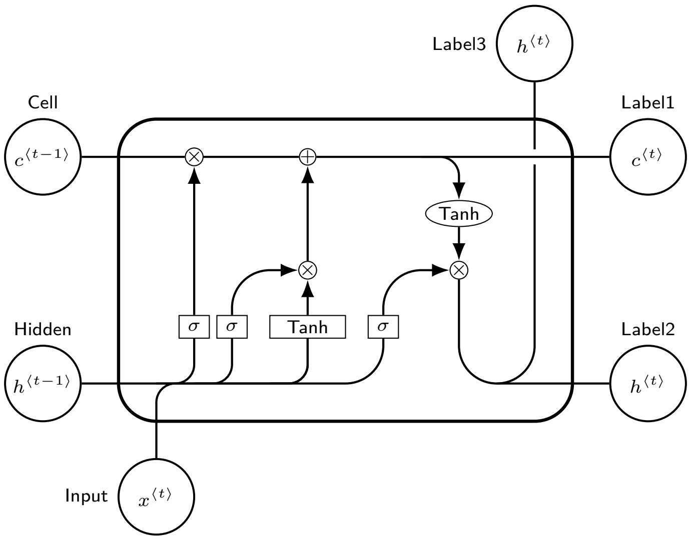


Figure 4: LSTM architecture source:Researcher

As is shown in Figure 4,the main cell box is represented by a large rectangle with rounded corners. This is the main structure of the LSTM cell where various operations and gates are defined. Gates and functions include the input gate (σ), forget gate (σ), output gate (σ), and the Tanh activation function. The input gates control how much of the new information to add to the cell state. The forget gate decides what portion of the cell state to keep from the previous timestep. The output gate controls how much of the cell state should be output. The Tanh activation function is used to squish values to be between -1 and 1, contributing to the cell state and output. Operators in the architecture include multiplication ($\times$) and addition ($+$). Multiplication is used for element-wise multiplication in the gating mechanism, while addition combines information from different sources to update the cell state.
External inputs and outputs involve the previous cell state c_{t-1}, previous hidden state h_{t-1}, and input x_t. The previous cell state represents the cell state from the previous timestep. The previous hidden state represents the hidden state from the previous timestep. The input is the input at the current timestep.
Internal components include the internal cell state c_t, internal hidden state h_t, and output h_t. The internal cell state represents the updated cell state for the current timestep. The internal hidden state is the hidden state output for the current timestep. The output represents the final output from the LSTM cell at the current timestep.
Connections and arrows represent the flow of data between different parts of the LSTM cell. External inputs to gates include the previous hidden state h_{t-1} which is fed into all gates, and the input x_t which is also fed into the Tanh activation function and other gates. Internal computations involve the forget gate and previous cell state being multiplied and added to the result of the input gate and candidate cell state. The updated cell state is then processed through a Tanh activation function and further multiplied by the output gate to produce the hidden state h_t.
In the examination of economic literature pertaining to time interpolation, recent research leans towards the utilization of LSTM models. This study adopts a similar methodology for temporal interpolation.
The utilization of the LSTM model(Figure 4) aimed to capture extended dependencies within sequential data. Input features comprised sequences of daily data, each with a sequence length of 30 days or more. The LSTM layer consisted of 50 nodes (49 features and bias), succeeded by a

Dense layer with 1 unit to forecast the daily target value.

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, X_t] + b_f) \\
i_t &= \sigma(W_i \cdot [h_{t-1}, X_t] + b_i) \\
o_t &= \sigma(W_o \cdot [h_{t-1}, X_t] + b_o) \\
\tilde{c}_t &= \tanh(W_c \cdot [h_{t-1}, X_t] + b_c) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$

where, X_t is the input sequence at time step t, h_t is the hidden state at time step t, c_t is the cell state at time step t, f_t is the forget gate at time step t, i_t is the input gate at time step t, o_t is the output gate at time step t, and $\tilde{c}_t$ is the candidate cell state at time step t. Also, (W_f, W_i, W_o, W_c) are weight matrices for the forget gate, input gate, output gate, and candidate cell state computations, respectively. (b_f, b_i,b_o, b_c) are bias vectors for the forget gate, input gate, output gate, and candidate cell state computations, respectively.(σ) represents the sigmoid activation function. (tanh) represents the hyperbolic tangent activation function. ($\odot$) denotes element-wise multiplication.

The output of the LSTM layer, h_i, is then passed through a Dense layer with 1 unit (output layer y which is computed by equation 1), Where W_d is the weight matrix and b_d is the bias vector for the Dense layer.

$$[h]y = W_\mathrm{d} \cdot h_i + b_\mathrm{d} \tag{1}$$

If we wish to examine this network from a different angle, it would be more advantageous to refer to the accompanying diagram(Figure 5). The key inquiry now is how to compute the value of node $a_1^{(1)}$. Here, we focus solely on the connection between the preceding layer, referred to as the input layer, and node $a_1^{(1)}$. As illustrated, the mathematical relationships between the input layer and node $a_1^{(1)}$ are expressed in the form(Figure 6). If we wish to examine this network from a different angle, it would be more advantageous to refer to the accompanying diagram (Figure 5). The key inquiry now is how to compute the value of nodes $a_1^{(1)} \ldots a_m^{(1)}$. Here, we focus solely on the connection between the preceding layer, referred to as the input layer $a_1^{(0)} \ldots a_m^{(0)}$, and the second layer $a_1^{(1)} \ldots a_m^{(1)}$. As illustrated in Figure 6, the mathematical relationships between the input layer and the second layer are expressed.

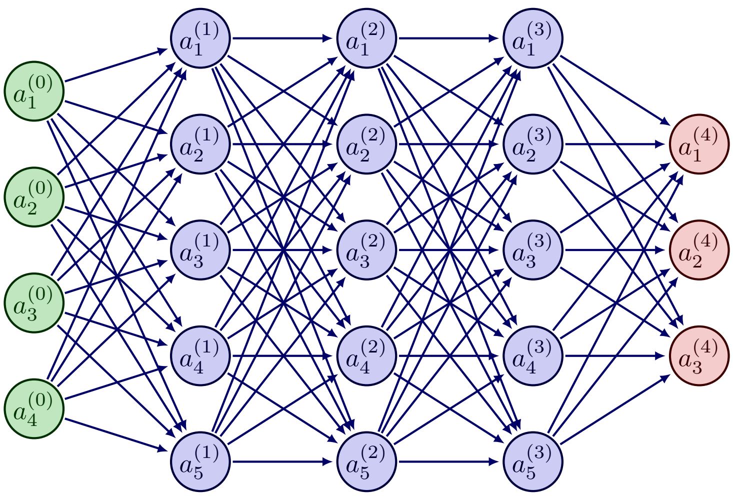
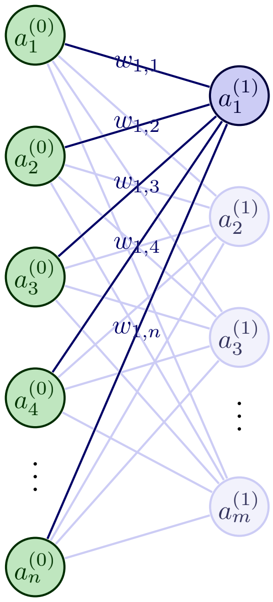


Figure 5: sample of neural network with three hidden layers



$$
\begin{pmatrix} a_1^{(1)} \\ a_2^{(1)} \\ \vdots \\ a_m^{(1)} \end{pmatrix} = \sigma \left[\begin{pmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,1} & w_{m,2} & \cdots & w_{m,n} \end{pmatrix} \begin{pmatrix} a_1^{(0)} \\ a_2^{(0)} \\ \vdots \\ a_n^{(0)} \end{pmatrix} + \begin{pmatrix} b_1^{(0)} \\ b_2^{(0)} \\ \vdots \\ b_m^{(0)} \end{pmatrix} \right]
$$

$$
\mathbf{a}^{(1)} = \sigma \left(\mathbf{W}^{(0)} \mathbf{a}^{(0)} + \mathbf{b}^{(0)} \right)
$$

Figure 6: mathematical explanation

3.0.2 Implementation of TSI-LSTM algorithm

The latest advancements in time series analysis and forecasting include nowcasting, which aims to predict the target variable using the maximum available data, even if it comes in mixed frequencies. However, a significant challenge arises because these predictions align with the variable frequency of the target. While the feature-related data is nearly real-time, predicting the target variable based on its own frequency is less appealing. In the implementation of the LSTM model, we intend to adjust the target variable using scientific methods (as shown in Figure 3), which is essentially a part of the preprocessing stage. This approach ensures that our predictions are in real-time, provided that our features are also in real-time.

3.0.3 Data collection, pre-processing and data analyses

features and target variables

Google Trends
Examining economic behavior on shorter time scales, such as daily, weekly, and monthly intervals, necessitates the production of economic indicators within these specific periods. The timeliness of producing these statistics is crucial, and their prompt publication is equally significant to capture sudden and immediate economic developments, requiring access to high-frequency data. Regrettably, within the realm of economics, with the exception of financial markets, there is a scarcity of statistics generated on a daily and weekly basis, and fewer still are those that are continuously accessible or promptly published. However, this limitation presents an opportune avenue for investigating economic behaviors through the utilization of Google Trends as a Big Data. This data source offers real-time accessibility, publishing information with minimal time delays, thus circumventing the challenges associated with the lag in traditional economic indicators. This article leverages Google Trends data along with analogous monthly export statistics provided by the Federal Statistical Office of Germany to explore economic dynamics at shorter time intervals.

In light of the scarcity of high-frequency economic indicators that can be both timely and strongly correlated with the target variable (typically of low frequencies), numerous studies underscore the pressing need for a comprehensive solutionDauphin et al. (2022). In this context, GT has emerged as a widely adopted proxy for high-frequency data in numerous studies over the past decadeChoi and Varian (2012) . Google Trends (GT) is conveniently available on Google's website, providing users with the ability to gather statistics for a wide range of search terms across multiple platforms. These platforms include web page search, image search, news search, product search, and YouTube search. Moreover, users can tailor their searches to specific geographical regions, such as countries, provinces, and cities. As an illustration, the website offers search results for the term **restaurant** at both the country and city levels.

The GT database offers users the flexibility to define the temporal and spatial scope of search terms. The data is available from 2004 to the present. It is noteworthy that the aforementioned database does not disclose search data in an absolute numerical form. Instead, it presents the data as a searchable ratio, utilizing the following formula for this purpose.

$$SR_{ig} = \frac{S_{ig}}{\sum_{i=1}^{N} \sum_{g=1}^{G} S_{ig}} \times 100 \tag{2}$$

The Search Ratio (SR) signifies the proportion of searches for term i in a geographic region g, divided by the cumulative searches for all terms within that specific region. This quotient is then multiplied by 100 to yield the Search Ratio. In simpler terms, the Search Ratio ranges from 0 to 100, with 100 indicating the popularity of the given term. This standardization facilitates seamless comparisons across different search terms due to variations in search volumes across countries and regions. GT is categorized into historical and real-time segments. Monthly data dating back to 2004 is accessible, while high-frequency real-time data from the past week is updated on an hourly basis. GT focuses on providing data for frequently searched terms. Google, however, asserts that the data excludes ”repeated searches from the same person over a short period of time,” reducing the likelihood of deliberate manipulation of search term popularity. Furthermore, GT data only encompasses search terms without special characters. Since August 2008, Google has categorized distinct search terms into various classes, accounting for the potential ambiguity of terms like ”apple” referring to either a fruit or a computer company. Google employs probabilities to assign these terms to specific categories; for instance, ”apple” is placed within the food and drink category. GT encompasses 27 expansive categories encompassing subjects such as news, shopping, and job-related searches. These broad categories further branch into over 1,400 subcategories, ranging from precise programming languages to niche subcultures like Gothic. Here it is necessary to mention the advantages and disadvantages of this database:

Utilizing Google Trends for economic analysis offers several advantages. Firstly, it provides businesses and policymakers with a valuable early window into consumer behavior. By analyzing shifts in search volumes related to products, services, and shopping, early insights can be gained into changing consumer preferences and demands. For instance, an increase in searches for ”online shopping deals” may indicate heightened consumer interest, potentially signaling economic growth. Additionally, Google Trends serves as an early indicator of economic changes, offering a preview before official statistics are released. Tracking search trends for terms like ”job opportunities” or ”unemployment benefits” provides valuable insights into potential shifts in the labor market, aiding timely decision-making.

Despite its advantages, Google Trends has limitations that need to be considered. The data is based on a sample of search queries, introducing potential biases that may not accurately represent the entire population’s behavior. Demographic biases can skew results, impacting the reliability of the insights. Moreover, Google Trends does not reveal individual searches, making it challenging to understand the exact context behind the queries, which limits the precision of the data. The accuracy of Google Trends also faces challenges due to undisclosed sampling methods, coverage biases, and incomplete representation of certain populations and regions. While it demonstrates stability in presenting the evolution of search popularity, the presence of ”0” values indicating searches below a popularity threshold may introduce some incompleteness, although this generally does not signify a significant issue.

Keyword Selection for google trends

The table 11 in the appendix displays the categorization of keywords available on Google Trends. In this table, categories and subcategories are categorized into various topics such as Arts & Entertainment, Autos & Vehicles, Beauty & Fitness, Books & Literature, Business & Industrial, Computers & Electronics, Finance, Food & Drink, Games, Health & Fitness, Hobbies & Leisure, Internet & Telecom, Jobs & Education, and Travel & Tourism. Categories are listed in the first and third columns of this table, and subcategories are specified in the second and fourth columns. Overall, this table covers a wide range of topics that people are interested in and search for related topics on the internet. The table 4 summarizes the categorization of Google Trends keywords into 23 categories and 342 subcategories.

Table 4: Google Trends keywords categories and number of subcategories

Categories	Subcategories	Categories	Subcategories
Arts & Entertainment	17	Law & Government	12
Autos & Vehicles	27	News & Media	21
Beauty & Fitness	20	People & Society	27
Books & Literature	18	Real Estate	8
Business & Industrial	54	Reference	12
Computers & Electronics	17	Science	16
Finance	10	Shopping	21
Food & Drink	11	Sports	44
Games	9	Travel	19
Health	25	Home & Garden	20
Hobbies & Leisure	24	Internet & Telecom	12
Jobs & Education	24		

Choosing the right keywords is a significant challenge when using Google Trends data. Different methods can be used to deal with this challenge. The initial approach involves relying on the findings of existing reports. Alternatively, one can select a set of keywords relevant to the research topic and subsequently refine the selection using lasso regression to remove irrelevant terms. In this paper, we used keywords that have been studied previously. However, since we use Google Trends for prediction in the LSTM model, this model assigns zero weight to irrelevant indicators and automatically excludes irrelevant keywords from the model.

The data related to Google Trends has been extracted daily for Germany using the *gtrendsR* from R libraryMassicotte et al. (2016).The keywords used in this paper are given in the table below. These keywords have been extracted based on the studies that have been done so far.

Table 5: list of keywords used in this paper

Stock trading, stock prices, stock market, stock market today, retirement planning, investment advice, unemployment, exchange year, GDP, taxation, income tax, Federal Reserve, financial report, financial crisis, financial markets, forex trading, monetary policy, trade deficit, budget deficit, mortgage rates, real estate market, inflation, investment, investment banking, capital investment, economic situation, credit score, credit advice, cryptocurrency, market analysis, market research, recession, savings account, tax deductions, tax reform, consumer, exchange rates, world trade, economy, economic situation, business news, economic policy, business companies, economic growth, central bank, interest rates, foreclosure, economic development, economic

target variable

Our target variable is Germany's foreign trade, which is published monthly. The required data was obtained from the website of the Federal Statistics Office [3]. These statistics have been available monthly since 1950. Since the time period of the target variable must align with that of the feature variables (which are based on Google Trends data available since 2006), for this paper , the target data has been used from 2006 onwards. Certainly, this is an example for a target variable, and any other variable could be used. The key is to introduce and test this approach. For instance, you could convert the monthly production output of a factory into daily values using the method proposed in this paper , provided that RMSE and R-squared metrics are considered.

[3] `https://www-genesis.destatis.de/`

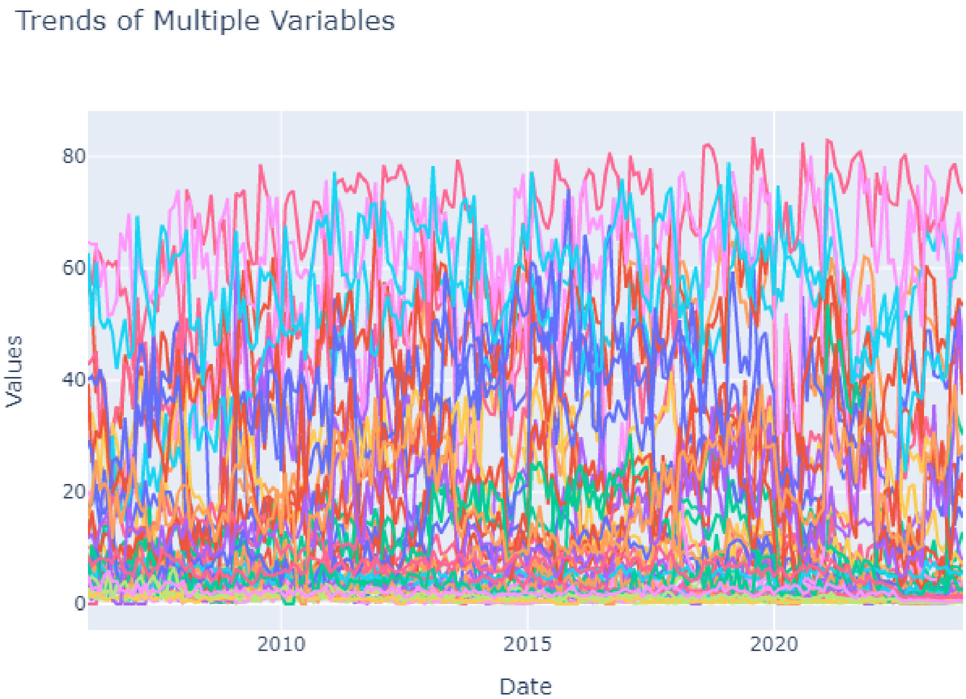


Figure 7: Features time series

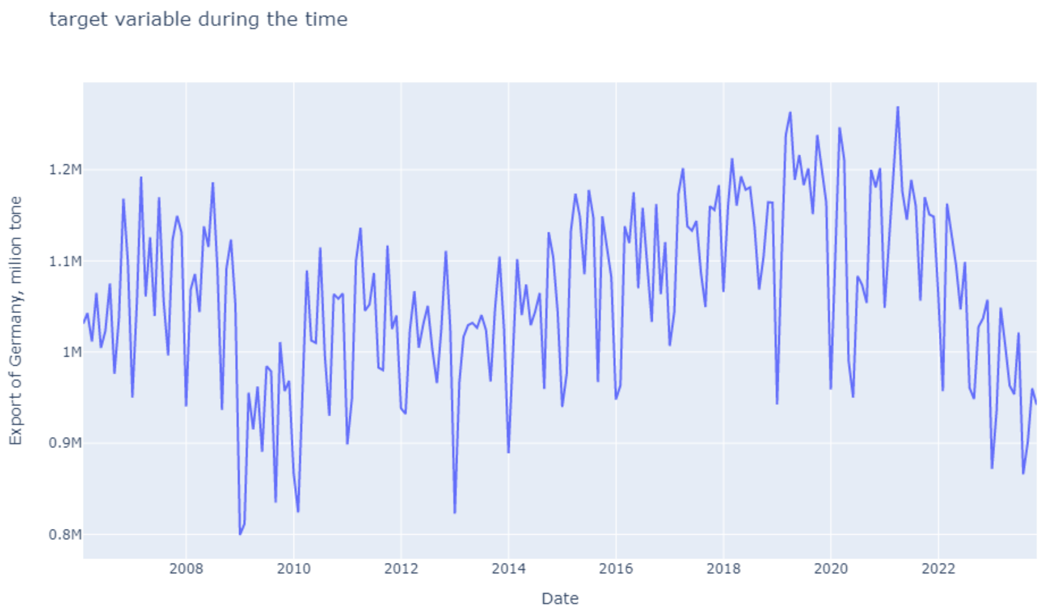


Figure 8: Target time series

3.0.4 Implementation of TSI and pre-processing

As previously mentioned (see Figure 3), in this paper , we intend to apply a process to the target variable if there is a frequency mismatch between the target variable and the feature variables. Temporal disaggregation methods are used to break down or interpolate a low-frequency time series into a higher-frequency series, ensuring that either the sum, average, first, or last value of the resulting high-frequency series matches the original low-frequency series. Temporal disaggregation can be performed with or without the use of one or more high-frequency indicator series. Methods include Chow-Lin, Santos-Silva-Cardoso, Fernandez, Litterman, Denton, and Denton-Cholette. These methods support most R,Python time series classes.
In the pre-processing section, a process emphasized in this paper is the use of the *timedisag* package to increase the frequency of the target variable to match the frequency of the feature variable. In this paper , the feature variable is daily, but the target variable is monthly. Therefore, to make daily predictions, there is no choice but to increase the frequency of the target variable. For this purpose, the *timedisag* package, which is developed in Python, has been used in this paper [4].As it can be seen, among the 9 temporal disaggregation methods, there are a few that exhibit higher R-squared values and lower RSS values. Here, we first considered the R-squared criterion and then selected the one with the lowest RSS among them(Table 6).

4 Results

As mentioned in the previous section s, in this paper , we used a combination of two scientific methods for frequency augmentation and forecasting. In the frequency augmentation method, referred to as TSI, there are several algorithms, of which we used nine. As observed, the best method for temporal disaggregation is the *chow-lin-maxlog* method, which is considered the preferred approach for increasing data frequency.

Table 6: Results for different methods with RMSE and R_squared values

No.	Methods	RMSE	R^2
1	chow-lin-maxlog	2189.264	0.993
2	chow-lin-minrss-ecotrim	1065.365	0.989
3	chow-lin-minrss-quilis	621.408	0.485
4	dynamic-maxlog	3898.123	0.994
5	dynamic-minrss	835.543	0.308
6	litterman-maxlog	624.291	0.473
7	litterman-minrss	149.362	0.246
8	dynamic-maxlog	3898.123	0.994
9	dynamic-minrss	835.543	0.308

4.0.1 Data Analyses

An LSTM model with three hidden layers is planned for implementation, based on prior experience. The first, second, and third layers will contain 96, 64, and 32 neurons respectively. Additionally, the dropout rate will be set to 0.5. With these assumptions, the structure of the model will be as follows:

4.0.2 Model tuning, Training and Evaluation

The LSTM model is trained using TensorFlow and Keras. The model's performance is evaluated using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to ensure its accuracy in predicting economic indicators. The next step involves compiling the model. Several parameters and hyperparameters need to be considered. For instance, the learning rate is set at 0.001, and *Adam* is chosen as the optimizer function. These parameters can be adjusted with

[4] `https://github.com/jstephenj14/timedisagg`

Table 7: Primary Network Layers and Parameters

Layer (type)	Output Shape	Param
lstm_111 (LSTM)	(None, 10, 96)	56,064
lstm_112 (LSTM)	(None, 10, 32)	16,512
lstm_113 (LSTM)	(None, 32)	8,320
dropout_37 (Dropout)	(None, 32)	0
dense_37 (Dense)	(None, 1)	33

the aim of achieving a R^2 and a low *RMSE*. Considering the multiple layers in an LSTM model and the numerous hyperparameters it includes, it is crucial to determine the ideal number of layers and the best hyperparameters. Consequently, this study concentrates on fine-tuning the neuron count and dropout rate to achieve optimal performance. The provided code defines a pipeline class that

Parameter/Hyperparameter	Primary Value
Number of layers	3
Number of neurons in each layer	[16, 64, 16]
Dropout rate	0.01
Learning rate	0.001
Number of epochs	100
Batch size	32
Activation function	tanh

Table 8: Parameters and Hyperparameters of the LSTM Model

automates the machine learning process, specifically for training a neural network model known as LSTM. The objective of this model is to predict future values based on past data. The pipeline encompasses several key steps(Figure 9):

1. **Data Import and Preprocessing**: The data is imported from a specified website and undergoes preprocessing to prepare it for analysis.
2. **Data Splitting**: The preprocessed data is divided into training, validation, and testing sets to ensure robust model evaluation.
3. **Data Standardization**: All features are standardized to ensure they share the same scale, which is crucial for model performance.
4. **Sequence Organization**: The data is organized into sequences suitable for the LSTM model, allowing it to learn temporal patterns effectively.
5. **Model Architecture Definition**: The LSTM model architecture is defined with multiple LSTM layers, followed by a dropout layer to mitigate overfitting.
6. **Model Training and Evaluation**: The model is trained across various combinations of LSTM units, and its performance is evaluated using the R^2 score, which indicates the model's predictive accuracy.
7. **Results Comparison**: The results are analyzed to identify the optimal combination of LSTM units and dropout rates for the model.

Hyper-parameter tuning

The LSTM model used in this paper includes several parameters and hyperparameters. Therefore, it is crucial to ensure that the optimal hyperparameters are selected before applying the model. This process is known as hyperparameter tuning. During this process, various hyperparameter combinations are tested by running the model, and their results are compared. The model that yields the highest R-squared value or the lowest RMSE is considered the best model. In this step, the focus is on automatically extracting the entire pipeline while tuning two specific hyperparameters: the number of neurons and the dropout rate in the three hypothesized hidden layers of the

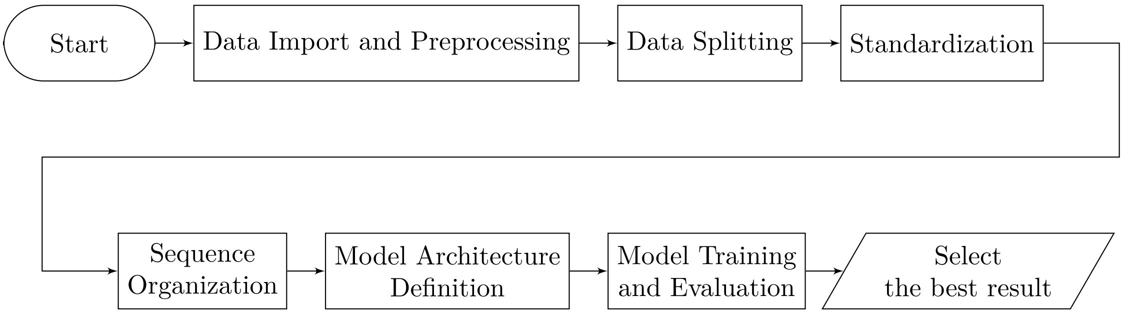


Figure 9: full pipeline

Table 9: LSTM Number of Nodes and Dropout Values in Different Scenarios

Scenario	Number of Nodes and Dropout Values			
	L_1	L_2	L_3	Dropout Values
1	16	64	32	[0.01, 0.015, 0.02]
2	16	64	16	[0.01, 0.015, 0.02]
3	16	64	8	[0.01, 0.015, 0.02]
4	16	32	32	[0.01, 0.015, 0.02]
5	16	32	16	[0.01, 0.015, 0.02]
6	16	32	8	[0.01, 0.015, 0.02]

LSTM. The results can be seen in Table 10, where the R^2 score is slightly more than 98%. As shown, with the number of neurons in the first to third layers and the following dropout, a higher number can be achieved. With LSTM units: **[16, 64, 16]** and Dropout: **0.015**, the R^2 score is 97%. In Figure 11, as observed, the validation loss and training loss stabilize around epoch 35. This indicates that the model reaches its optimal performance at this point, making 35 epochs the best choice for this model.

Table 10: Performance of various pipelines with different LSTM units and dropout rates.

Pipeline	LSTM_units	Dropout	R2_score
Pipe_Line_2_Dropout_0.01	[16, 64, 16]	**0.010**	**0.969108**
Pipe_Line_1_Dropout_0.015	[16, 64, 32]	0.015	0.954294
Pipe_Line_2_Dropout_0.015	[16, 64, 16]	0.015	0.964002
Pipe_Line_1_Dropout_0.02	[16, 64, 32]	0.020	0.937698
Pipe_Line_4_Dropout_0.015	[16, 32, 32]	0.015	0.926979
Pipe_Line_4_Dropout_0.02	[16, 32, 32]	0.020	0.931688
Pipe_Line_5_Dropout_0.015	[16, 32, 8]	0.015	0.938163
Pipe_Line_5_Dropout_0.01	[16, 32, 8]	0.010	0.894652
Pipe_Line_1_Dropout_0.01	[16, 64, 32]	0.010	0.916466
Pipe_Line_4_Dropout_0.01	[16, 32, 32]	0.010	0.906207
Pipe_Line_2_Dropout_0.02	[16, 64, 16]	0.020	0.961831
Pipe_Line_3_Dropout_0.02	[16, 64, 8]	0.020	-0.341800
Pipe_Line_3_Dropout_0.015	[16, 64, 8]	0.015	-0.367244
Pipe_Line_3_Dropout_0.01	[16, 64, 8]	0.010	-0.309727
Pipe_Line_5_Dropout_0.02	[16, 32, 8]	0.020	0.887368

Model evaluation

With the best parameters and hyperparameters selected, it is now time to predict the target variable. The following chart(Figure 12) visualizes the comparison between the actual and predicted values of the target variable. As previously mentioned, to validate the results, it is essential to consider the metrics such as RSS and R-squared in addition to the plotted chart. The scientific significance of using R^2 and Root Mean Squared Error (RMSE) lies in their ability to evaluate the performance of a predictive model.

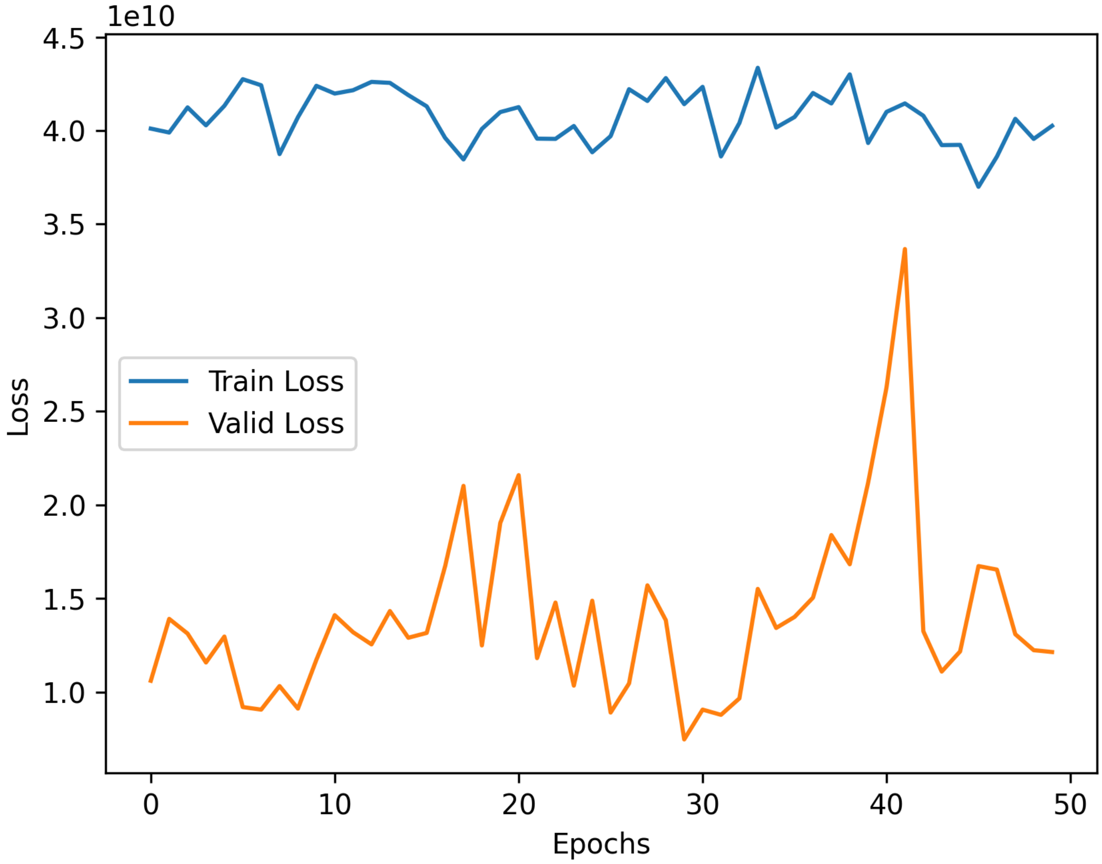


Figure 10: Train Loss and Validation loss before hyperparameter tuning

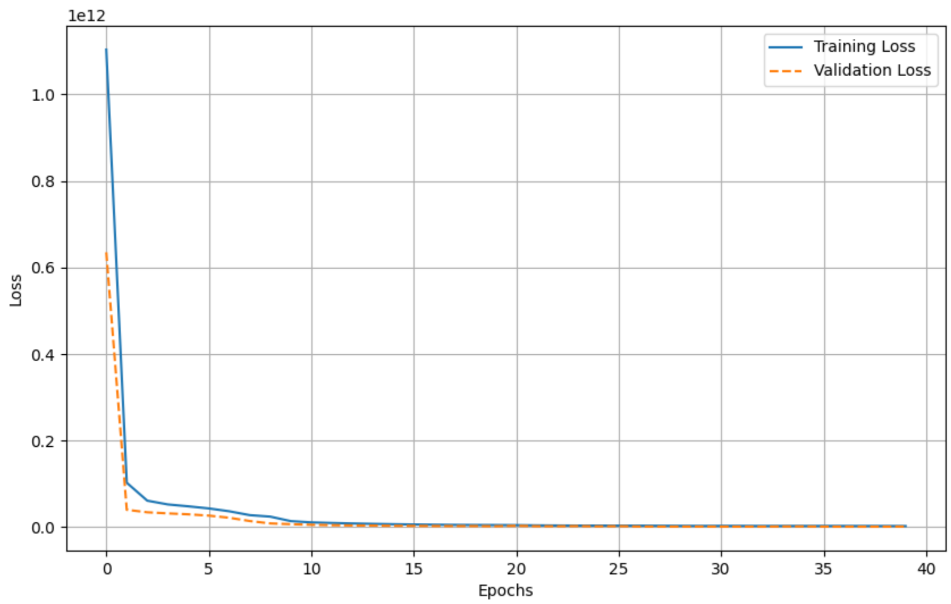


Figure 11: Train Loss and Validation loss after hyperparameter tuning

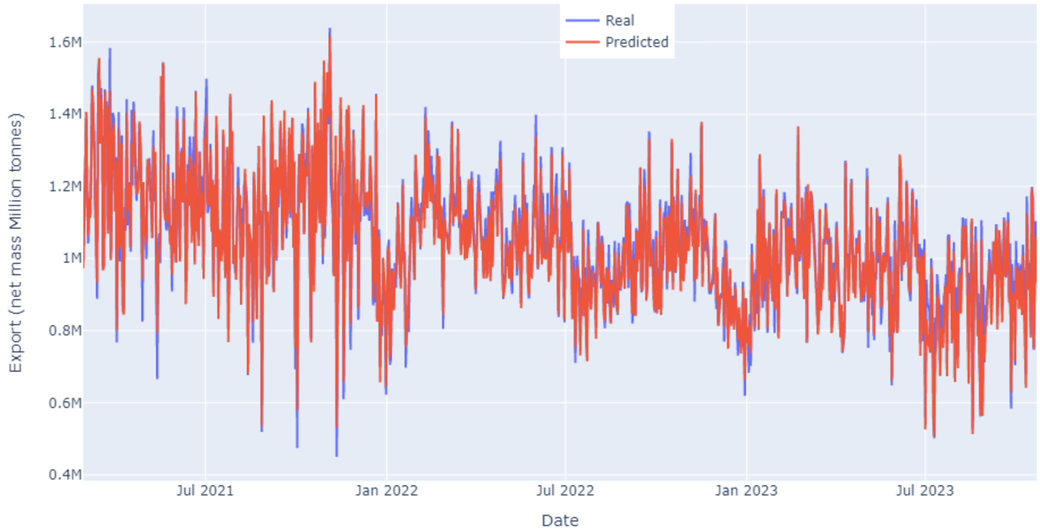


Figure 12: comparing real and predicted target variable after selecting best model

- **R-squared (R^2):** This metric measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It indicates how well the model's predictions match the actual data. A higher R^2 value suggests a better fit, meaning the model explains a greater portion of the data's variability.
- **Root Mean Squared Error (RMSE):** RMSE measures the average magnitude of the prediction errors, representing how close the observed data points are to the model's predicted values. A lower RMSE indicates better model accuracy, as it shows that the model's predictions are closer to the actual values.

Together, these metrics provide a comprehensive assessment of a model's predictive accuracy and its ability to generalize well to new data.

As shown in Figure 13, the prediction error falls within the 95% confidence interval, which means that in 95% of cases, the prediction aligns with the actual value. This implies that the model's predictions are generally reliable within a specified range of uncertainty. If most prediction errors fall within this interval, it means the model is well-calibrated and that there is a high level of confidence that the predictions are accurate. This helps in validating the model's performance and ensuring that it is a trustworthy tool for making predictions.

5 Summary, Implications, & Conclusions

The proposed TSI-LSTM hybrid model, featuring three LSTM layers, effectively increases the frequency of target variables and predicts near-term values, achieving high accuracy with an R^2 score of 97%.Key findings include:

- Successful conversion of low-frequency data to high-frequency data using LSTM models.
- Google Trends as a reliable proxy for official statistical indicators.
- Enhanced short-term fluctuation prediction with LSTM models compared to traditional regression methods.

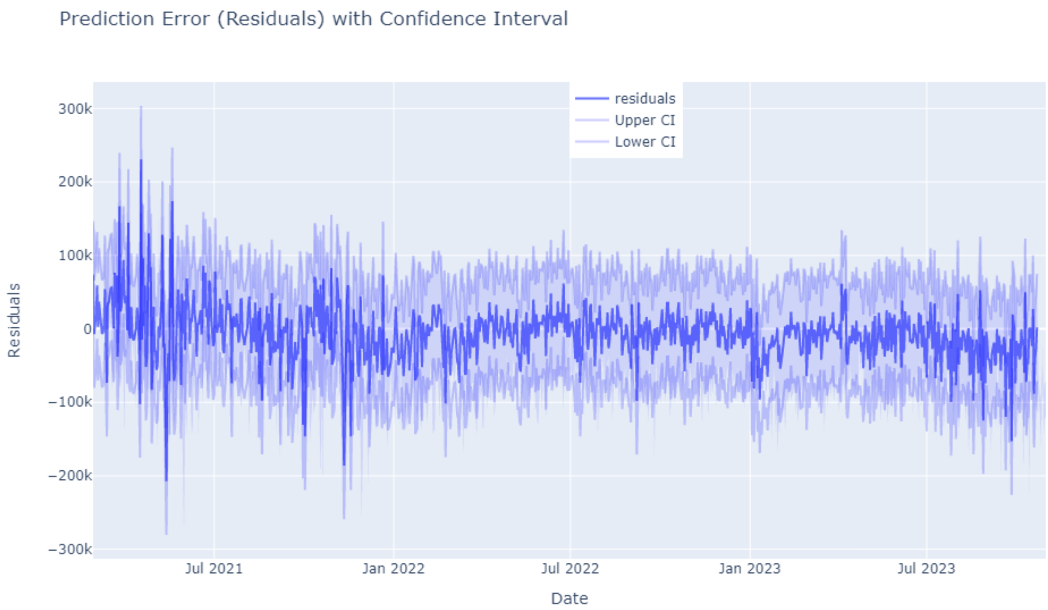


Figure 13: prediction error with confidence interval 95%

In Conclusions, Google Trends data can effectively nowcast economic indicators with high accuracy, confirming the hypothesis that real-time online search data offers immediate economic insights. Also, LSTM models outperform traditional regression methods in transforming low-frequency data into high-frequency data, capturing short-term dependencies and non-linear relationships within the data.

This study validates the use of Google Trends and LSTM models for accurate and timely economic data dissemination, offering significant theoretical and practical advancements in nowcasting and temporal interpolation.

The theoretical implications introduce a novel approach to nowcasting and validate Google Trends as a valuable data source for economic forecasting. The practical implications provide policymakers with real-time economic insights, enhancing decision-making. This approach is applicable in various economic and financial contexts for responsive and accurate forecasting. However, there are limitations to this approach, as it is limited to Google Trends data and specific LSTM models. Future research could explore additional data sources and alternative machine learning models.

Suggestions for future research include investigating the use of other machine learning models like GRU or Transformer models for temporal interpolation. Another suggestion is to integrate multiple high-frequency data sources to improve forecasting robustness. Additionally, conducting case studies in different economic contexts can help generalize findings and assess the method's versatility.

This research contributes a robust framework for increasing data prediction frequency, with significant implications for economic and statistical applications. Also, the prediction errors revealed that these errors fall within the 95% confidence interval, indicating the accuracy and reliability of the model's predictions. This analysis provides confidence that the model is generally trustworthy and suitable for accurate forecasting.

References

Ball, R. T. and Gallo, L. A. (2018), 'A mixed data sampling approach to accounting research', *Available at SSRN 3250445* .

Boot, J., Feibes, W. and Lisman, D. J. (1970), 'The smoothing of time series by moving averages', *Journal of the American Statistical Association* **65**(331), 520–536.

Causey, B. D. and Trager, R. L. (1981), 'An improved disaggregation procedure for non-symmetric time series models', *Journal of Business & Economic Statistics* **15**(2), 272–279.

Choi, H. and Varian, H. (2012), 'Predicting the present with google trends', *Economic record* **88**, 2–9.

Chow, G. C. and Lin, A.-L. (1971), 'Best linear unbiased interpolation, distribution, and extrapolation of time series by related series', *The Review of Economics and Statistics* **53**(4), 372–375.

Dauphin, M. J.-F., Dybczak, M. K., Maneely, M., Sanjani, M. T., Suphaphiphat, M. N., Wang, Y. and Zhang, H. (2022), *Nowcasting gdp-a scalable approach using dfm, machine learning and novel data, applied to european economies*, International Monetary Fund.

Denton, F. T. (1971), 'Adjustment of monthly or quarterly series to annual totals: an approach based on quadratic minimization', *Journal of the American Statistical Association* **66**(333), 99–102.

Elsaraiti, M. and Merabet, A. (2021), 'A comparative analysis of the arima and lstm predictive models and their effectiveness for predicting wind speed. energies 14 (20): 6782'.

Fernández, C. (1981), 'A methodological note on the use of a random walk for estimating cyclical phenomena', *Economics Letters* **8**(1), 65–68.

Ghysels, E., Santa-Clara, P. and Valkanov, R. (2004), 'The midas touch: Mixed data sampling regression models'.

Guay, A. and Maurin, A. (2015), 'Disaggregation methods based on midas regression', *Economic Modelling* **50**, 123–129.

Hochreiter, S. and Schmidhuber, J. (1997), 'Long short-term memory', *Neural computation* **9**(8), 1735–1780.

Hopp, D. (2021), 'Performance of lstm neural networks in nowcasting during the covid-19 crisis', *UNCTAD Research Paper* (74).

Hopp, D. (2022), 'Benchmarking econometric and machine learning methodologies in nowcasting', *arXiv preprint arXiv:2205.03318* .

Hopp, D. (2023), 'Benchmarking econometric and machine learning methodologies in nowcasting gdp', *Empirical Economics* pp. 1–57.

Itkonen, J. and Juvonen, P. (2017), Nowcasting the finnish economy with a large bayesian vector autoregressive model, Technical report, BoF Economics Review.

Lampos, V. and Cristianini, N. (2012), 'Nowcasting events from the social web with statistical learning', *ACM Transactions on Intelligent Systems and Technology (TIST)* **3**(4), 1–22.

Litterman, R. B. (1998), 'Forecasting with bayesian vector autoregressions - five years of experience', *Journal of Business & Economic Statistics* **16**(3), 239–246.

Massicotte, P., Eddelbuettel, D. and Massicotte, M. P. (2016), 'Package 'gtrendsr'', *R package* .

Rossana, R. J. and Seater, J. J. (1995), 'Temporal aggregation and economic time series', *Journal of Business & Economic Statistics* **13**(4), 441–451.

Sax, C. and Steiner, P. (2013), 'Temporal disaggregation of time series'.

6 Appendix

Table 11: Google trends keywords categories and subcategories

Categories	Subcategories	Categories	Subcategories
Arts & Entertainment	Celebrities & Entertainment News	Law & Government	Courts & Judiciary
Arts & Entertainment	Comics & Animation	Law & Government	Government
Arts & Entertainment	Entertainment Industry	Law & Government	Government Resources
Arts & Entertainment	Events & Listings	Law & Government	Immigration
Arts & Entertainment	Fun & Trivia	Law & Government	Legal Education
Arts & Entertainment	Humor	Law & Government	Legal Services
Arts & Entertainment	Movies	Law & Government	Military
Arts & Entertainment	Music & Audio	Law & Government	National & Public Holidays
Arts & Entertainment	Offbeat	Law & Government	Politics
Arts & Entertainment	Online Media	Law & Government	Public Safety
Arts & Entertainment	Performing Arts	Law & Government	Social Services
Arts & Entertainment	TV & Video	Law & Government	Taxes
Arts & Entertainment	Visual Art & Design	News & Media	Broadcast & Network News
Autos & Vehicles	Automotive Industry	News & Media	Business News
Autos & Vehicles	Bicycles & Accessories	News & Media	Celebrity Gossip
Autos & Vehicles	Boats & Watercraft	News & Media	College Media
Autos & Vehicles	Campers & RVs	News & Media	Entertainment News
Autos & Vehicles	Classic Vehicles	News & Media	International News
Autos & Vehicles	Commercial Vehicles	News & Media	Journalism
Autos & Vehicles	Custom & Performance Vehicles	News & Media	Local News
Autos & Vehicles	Hybrid & Alternative Vehicles	News & Media	Newspapers
Autos & Vehicles	Microcars & City Cars	News & Media	News Agencies
Autos & Vehicles	Motorcycles	News & Media	Sports News
Autos & Vehicles	Off-Road Vehicles	News & Media	Technology News
Autos & Vehicles	Personal Aircraft	News & Media	Weather
Autos & Vehicles	Scooters & Mopeds	Online Communities	Chat
Autos & Vehicles	Trucks & SUVs	Online Communities	Dating
Autos & Vehicles	Vehicle Brands	Online Communities	Forums & Message Boards
Autos & Vehicles	Vehicle Codes & Driving Laws	Online Communities	Image Sharing
Autos & Vehicles	Vehicle Licensing & Registration	Online Communities	Instant Messaging
Autos & Vehicles	Vehicle Maintenance	Online Communities	Online Goodies
Autos & Vehicles	Vehicle Parts & Accessories	Online Communities	Online Journals & Personal Sites
Autos & Vehicles	Vehicle Shopping	Online Communities	Online Portfolios
Autos & Vehicles	Vehicle Shows	Online Communities	Social Networks
Beauty & Fitness	Beauty Pageants	People & Society	Dating
Beauty & Fitness	Body Art	People & Society	Disabled & Special Needs
Beauty & Fitness	Cosmetic Procedures	People & Society	Family & Relationships
Beauty & Fitness	Cosmetology & Beauty Professionals	People & Society	Holidays & Special Occasions
Beauty & Fitness	Face & Body Care	People & Society	Military
Beauty & Fitness	Fashion & Style	People & Society	Seniors
Beauty & Fitness	Fitness	People & Society	Social Issues & Advocacy
Beauty & Fitness	Hair Care	People & Society	Social Sciences
Beauty & Fitness	Spas & Beauty Services	People & Society	Social Work
Beauty & Fitness	Weight Loss	People & Society	Teens
Books & Literature	Biographies & Quotations	People & Society	Work & Labor Issues
Books & Literature	Book Retailers	Real Estate	Property Development
Books & Literature	Children's Literature	Real Estate	Property Investment
Books & Literature	E-Books	Real Estate	Property Management
Books & Literature	Fan Fiction	Real Estate	Property Search
Books & Literature	Literary Classics	Real Estate	Real Estate Listings
Books & Literature	Magazines	Real Estate	Real Estate Services
Books & Literature	Poetry	Reference	Almanacs
Books & Literature	Writers Resources	Reference	Calculators
Business & Industrial	Advertising & Marketing	Reference	Dictionaries & Encyclopedias
Business & Industrial	Aerospace & Defense	Reference	Educational Resources
Business & Industrial	Agriculture & Forestry	Reference	General Reference
Business & Industrial	Automotive Industry	Reference	How-To & Expert Content
Business & Industrial	Business Education	Reference	Libraries & Museums
Business & Industrial	Business Finance	Reference	Time
Business & Industrial	Business News	Reference	Trivia
Business & Industrial	Business Operations	Reference	White Pages & Directories
Business & Industrial	Business Services	Science	Astronomy
Business & Industrial	Chemicals Industry	Science	Biology
Business & Industrial	Construction & Maintenance	Science	Chemistry
Business & Industrial	Energy & Utilities	Science	Computer Science
Business & Industrial	Enterprise Technology	Science	Earth Sciences
Business & Industrial	Entertainment Industry	Science	Ecology
Business & Industrial	Hospitality Industry	Science	Engineering & Technology
Business & Industrial	Industrial Materials & Equipment	Science	Environmental Science
Business & Industrial	Manufacturing	Science	Mathematics
Business & Industrial	Metals & Mining	Science	Physics
Business & Industrial	Pharmaceuticals & Biotech	Science	Social Sciences
Business & Industrial	Printing & Publishing	Shopping	Antiques & Collectibles
Business & Industrial	Professional & Trade Associations	Shopping	Apparel
Business & Industrial	Retail Trade	Shopping	Children's Products
Business & Industrial	Small Business	Shopping	Classifieds
Business & Industrial	Textiles & Nonwovens	Shopping	Consumer Advocacy
Business & Industrial	Transportation & Logistics	Shopping	Consumer Electronics
Computers & Electronics	CAD & CAM	Shopping	Coupons & Discount Offers
Computers & Electronics	Computer Hardware	Shopping	General Merchandise
Computers & Electronics	Computer Security	Shopping	Gifts & Special Event Items
Computers & Electronics	Consumer Electronics	Shopping	Luxury Goods
Computers & Electronics	Electronics & Electrical	Shopping	Mass Merchants & Department Stores
Computers & Electronics	Enterprise Technology	Shopping	Photo & Video Services
Computers & Electronics	Networking	Shopping	Shopping Portals
Computers & Electronics	Programming	Shopping	Swap Meets & Outlets
Computers & Electronics	Software	Shopping	Tobacco Products
Computers & Electronics	Technical Support	Shopping	Toys
Computers & Electronics	Technology News	Sports	Adventure Sports
Finance	Accounting & Auditing	Sports	Baseball
Finance	Banking	Sports	Basketball
Finance	Corporate Finance	Sports	College Sports
Finance	Financial News	Sports	Combat Sports
Finance	Financial Planning	Sports	Cycling
Finance	Grants & Financial Assistance	Sports	Equestrian

Table 11: Google trends keywords categories and subcategories

Categories	Subcategories	Categories	Subcategories
Finance	Insurance	Sports	Extreme Sports
Finance	Investing	Sports	Fantasy Sports
Finance	Retirement & Pension	Sports	Fishing
Finance	Taxation	Sports	Football
Food & Drink	Beverages	Sports	Golf
Food & Drink	Cooking	Sports	Gymnastics
Food & Drink	Food	Sports	Hockey
Food & Drink	Food Service	Sports	Hunting & Shooting
Food & Drink	Groceries	Sports	Individual Sports
Food & Drink	Restaurants	Sports	Motor Sports
Games	Board Games	Sports	Rodeo
Games	Card Games	Sports	Rugby
Games	Computer & Video Games	Sports	Running
Games	Puzzles	Sports	Soccer
Games	Roleplaying Games	Sports	Softball
Games	Table Games	Sports	Swimming
Games	Tile Games	Sports	Tennis
Games	Video Games	Sports	Track & Field
Health	Alternative & Natural Medicine	Sports	Volleyball
Health	Child Health	Sports	Walking
Health	Dentistry	Sports	Water Sports
Health	Health Conditions	Sports	Winter Sports
Health	Health Education	Travel	Adventure Travel
Health	Medical Devices & Equipment	Travel	Bus & Rail
Health	Men's Health	Travel	Car Rental & Taxi Services
Health	Nutrition	Travel	Cruises
Health	Occupational Health & Safety	Travel	Destinations
Health	Pharmacy	Travel	Ecotourism
Health	Public Health	Travel	Lodging
Health	Reproductive Health	Travel	Specialty Travel
Health	Senior Health	Travel	Tourist Boards & Visitor Centers
Health	Substance Abuse	Travel	Travel Agencies & Services
Health	Vision Care	Travel	Travel Guides & Travelogues
Health & Fitness	Weight Loss	Home & Garden	Bed & Bath
Hobbies & Leisure	Animal Shows	Home & Garden	Furniture
Hobbies & Leisure	Crafts	Home & Garden	Gardening
Hobbies & Leisure	Drawing & Coloring	Home & Garden	Home Appliances
Hobbies & Leisure	Events & Listings	Home & Garden	Home Furnishings
Hobbies & Leisure	Hobbies & Collecting	Home & Garden	Home Improvement
Hobbies & Leisure	Outdoor Recreation	Home & Garden	Home Storage & Shelving
Hobbies & Leisure	Radio Control & Modeling	Home & Garden	Housekeeping
Hobbies & Leisure	Sci-Fi & Fantasy	Home & Garden	Interior Decorating
Hobbies & Leisure	Smoking	Home & Garden	Kitchens
Hobbies & Leisure	Toys	Home & Garden	Laundry
Hobbies & Leisure	Water Activities	Home & Garden	Pest Control
Internet & Telecom	Communications Equipment	Home & Garden	Pools & Spas
Internet & Telecom	Email	Home & Garden	Yard & Patio
Internet & Telecom	ISPs	Internet & Telecom	Communications Equipment
Internet & Telecom	Mobile & Wireless	Internet & Telecom	Email
Internet & Telecom	Web Apps & Online Tools	Internet & Telecom	ISPs
Internet & Telecom	Web Hosting & Domain Registration	Internet & Telecom	Mobile & Wireless
Internet & Telecom	Web Services	Internet & Telecom	Web Apps & Online Tools
Jobs & Education	Adult Education	Internet & Telecom	Web Hosting & Domain Registration
Jobs & Education	Business Education	Internet & Telecom	Web Services
Jobs & Education	College & University	Jobs & Education	Adult Education
Jobs & Education	Distance Learning	Jobs & Education	Business Education
Jobs & Education	Education Resources	Jobs & Education	College & University
Jobs & Education	E-Learning	Jobs & Education	Distance Learning
Jobs & Education	Job Listings	Jobs & Education	Education Resources
Jobs & Education	Job Search	Jobs & Education	E-Learning
Jobs & Education	K-12	Jobs & Education	Job Listings
Jobs & Education	Language Learning	Jobs & Education	Job Search
Jobs & Education	Primary & Secondary Schooling	Jobs & Education	K-12
Jobs & Education	Special Education	Jobs & Education	Language Learning
Jobs & Education	Training & Certification	Jobs & Education	Primary & Secondary Schooling
Jobs & Education	Tutoring	Jobs & Education	Special Education
Jobs & Education	Vocational & Continuing Education	Jobs & Education	Training & Certification
Jobs & Education		Jobs & Education	Tutoring
		Jobs & Education	Vocational & Continuing Education