Expert advice from experts

Marie Curie
Nobel Prize, PhD

Report for
Acme Corporation

3 May 2024

MONASH
BUSINESS
SCHOOL

Department of Econometrics & Business Statistics

☎ (03) 9905 2478
✉ mcurie.notreal@gmail.com

ABN: 12 377 614 012

Contents

1 Abstract

- Overview: Briefly summarize the purpose, methodology, key findings, and implications of the forecast.
- Key Recommendations: High-level actionable recommendations based on the forecast results.

2 Introduction

2.1 Background:

Overview of LTD and its economic significance.

2.2 Objective:

Purpose of forecasting LTD for DTF. Outline the specific goals of the project, such as predicting revenue from LTD or analyzing trends.

3 Rationale for the Forecast

3.1 Need for Forecasting:

Discuss why forecasting LTD is critical for budgetary planning and economic analysis at DTF.

3.2 Benefits:

How the forecasting results will aid in policy making, financial planning, etc.

4 Data Description

4.1 Data Source:

Detail the sources of the LTD data, including collection methods, frequency, and historical range.

4.2 Variables Description:

Describe each variable used in the analysis, including dependent and independent variables.

5 Initial Data Analysis (IDA)

5.1 Data Cleansing:

Steps taken to clean and preprocess the data.

5.2 Descriptive Statistics:

Basic statistics like mean, median, etc., and initial observations.

5.3 Data Integrity Checks:

Ensuring data completeness, accuracy, and consistency.

6 Exploratory Data Analysis (EDA)

6.1 Visualization:

Use plots (time series plots, histograms, etc.) to visualize trends, cycles, and outliers in LTD data.

6.2 Correlation Analysis:

Identify relationships between LTD and potential predictors.

6.3 Preliminary Findings:

Summarize insights gained about the data and any implications for modeling.

7 Methodology

7.1 Hierarchical time series

Upon analyzing the data characteristics and the disaggregation of LTD, it becomes evident that a three-level hierarchy structure can be established. There is utility in generating forecasts at various levels of aggregation, driven by diverse reasons and objectives. For instance, forecasting solely at the total or top level may lead to inaccuracies due to the limited number of time series encompassed. Additionally, each level of aggregation may exhibit distinct characteristics; for instance, transactions involving residential properties might vary from those involving non-residential properties due to differences in market dynamics or market size for each property type.

In an ideal scenario, forecasts from different levels of aggregation could seamlessly sum up to the top level. However, practical implementation often reveals incoherent among independently produced forecasts. Consequently, it becomes vital for forecasts to align and aggregate according to the hierarchical structure organizing the array of time series. As a solution to this challenge, reconciliation forecasting emerges as one of the most prevalent and effective methodologies employed today.

Within the domain of hierarchical time series forecasting, there are three traditional single level approaches for generating forecasts for hierarchical time series. The first, known as the bottom-up approach, initiates by producing forecasts for each series at the lowest level and subsequently aggregates these to generate forecasts for the upper levels of the hierarchy. Conversely, the top-down approach starts with a forecast at the highest level, which is then disaggregated to lower levels using predetermined proportions—typically based on historical data distributions. Lastly, the middle-out approach amalgamates elements of both the bottom-up and top-down methods.

7.2 Forecast reconciliation

Recall from the data structure and insights from EDA section, we can construct this hierarchical structure for LTD:

$$\begin{bmatrix} \text{Total}_t \\ \text{Non-residential}_t \\ \text{Residential}_t \\ \text{Commercial}_t \\ \text{Industrial}_t \\ \text{Other}_t \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \text{Residential}_t \\ \text{Commercial}_t \\ \text{Industrial}_t \\ \text{Other}_t \end{bmatrix}$$

or in a more compact notation:

$$\text{LTD}_t = \text{Sb}_t,$$

where S represents the summing matrix defining how bottom-level series are aggregated.

As indicated by Hyndman and Athanasopoulos (2021), reconciliation forecasting involves the introduction of a mapping matrix, denoted as **G**, to the base forecast, which is determined by the adopted methodology. This matrix, when multiplied by **SG**, yields a coherent set of forecasts.

However, the traditional single level approaches may have their limitations since only base forecast at one level is used. In response to this, Wickramasuriya et al. (2019) introduced the *MinT* (Minimum Trace) optimal reconciliation methodology, which devises a **G** matrix aimed at minimizing the total forecast variance within the coherent forecast set.

This leads to the need for estimating W_h, the forecast error variance of h-step-ahead base forecasts. There are four simplifying approximations in place that have been shown to work well:

- **OLS**: $\mathbf{W}_h = \text{k}_h\mathbf{I}$
- **Variance Scaling**: $\mathbf{W}_h = \text{k}_h\text{diag}(\hat{\mathbf{W}}_1)$
- **Structural Scaling**: $\mathbf{W}_h = \text{k}_h\mathbf{\Lambda}$
- **MinT Shrinkage**: $\mathbf{W}_h = \text{k}_h\mathbf{W}_1$

7.3 Model Selection:

As depicted in , land transfer duty (LTD) exhibits mutual correlations with variables such as sales and the home value index (HVI). It is important to note that these relationships are not unidirectional; factors like sales and HVI may influence LTD, but changes in LTD can reciprocally impact these variables. This dynamic interaction is also observed with economic indicators like inflation and interest rates.

The Vector Autoregression (VAR) model, which predicts future values of multiple time series based on their historical data, is well-suited for capturing these complex interactions. Furthermore, as illustrated in , the detection of cointegration patterns indicating a stable long-term equilibrium relationship among variables such as LTD, sales, and HVI, which necessitates the application of the Vector Error Correction Model (VECM). This highlights the pertinence of employing both VAR and VECM to adequately model these relationships.

One limitation of these models is their reliance on ample data to produce reliable parameters. Monthly data spanning over a decade for LTD has been found to be sufficient. Nevertheless, the subsequent section on cross-validation will elaborate on the precise sizing of the training dataset required to ensure compatibility with the VAR and VECM models.

In addition to VAR and VECM, the Autoregressive Integrated Moving Average (ARIMA) model has also been employed, primarily for purposes of comparison and validation of improvements. The efficacy of this comparison will be assessed through time series cross-validation, utilizing various accuracy metrics such as the Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the Mean Absolute Scaled Error (MASE).

8 Time Series Cross-Validation

8.1 Definition and rationale

Time series cross-validation represents a sophisticated adaptation of the conventional training/test set approach for model selection, as stated by Hyndman and Athanasopoulos (2021). This methodology is particularly well-suited to time series data because it exclusively includes observations from periods prior to those being forecasted, distinguishing it from traditional cross-validation techniques.

In the scope of this project, time series cross-validation is employed to rigorously evaluate whether the VAR/VECM or ARIMA models yield more accurate forecasts for this dataset across forecasting horizons ranging from 1 to 12 steps ahead. Additionally, this approach is used to gauge the extent of improvement introduced by the reconciliation process in comparison to the b base forecasts. Another critical application of time series cross-validation in this context is to generate rolling forecasts over a 12-month period for various time intervals. This is essential for the Department of Treasury and Finance (DTF) to analyze the effects of market dynamics or policy changes on land transfer duty (LTD).

8.1.1 VECM

As VECM fitting is not supported by 'fpp3' package, more complicated procedure is required. Firstly, 'ca.jo()' will be used to conduct Johansen test for cointegration between LTD, sales and hvi. This test checks if a set of time series variables have a long-term, stable relationship. Value of argument K is initially set to 2, indicating the lag order in the VAR model used within the Johansen procedure is 2. This value can also be put into a cross-validation for optimal value, generally between 1 and 3 is suitable for this data.

Afterwards, the test result at 5% level of significance identifies value for r, indicating the number of cointegrating relationships to include in the VECM model. The function for fitting VECM is 'vars::vec2var()', belongs to 'vars' package in R.

This procedure will be applied to each folds, using 'group_by()' and then producing forecast. Similarly, only forecast result for LTD will be used.

8.2 Forecasting

To create folds for the time series cross-validation procedure, we will use `stretch_tsibble()` function with user defined value for two arguments, `.init` for size of initial training set, and `.step` to define how many steps training sets will roll forward each time.

For example, `stretch_tsibble(.init = 10, .step = 1)` produces series of training sets, where first training set has size of 10 observations and will roll 1 step forward each time, i.e. the size of second set and third set will be 11 and 12, respectively. Each folds will be denoted by a unique `.id` value.

8.2.1 Generating forecast

To advancing the forecast accuracy, apart from cross-sectional hierarchical structure, forecast will also be generated across temporal structure, which later combines to cross-temporal reconciliation forecasting. The result generated by this approach will also be compared against cross-sectional and temporal reconciliation forecast individually to assess if there is an improvement.

In the need for temporal hierarchical structure, and since the data is in monthly frequency, other aggregation frequencies can be generated are: bi-monthly, quarterly, four-monthly, semi-annually and annually, which is the top level of the temporal structure.

Models will then be fitted across cross-sectional level at each temporal frequencies. By using a `for` loop, forecast and residuals will be produced and assign into a matrix. And then all matrices for each temporal frequencies: monthly (denoted as k1), bi-monthly (k2), quarterly (k3), 4-monthly (k4), semi-annually (k6), annually (k12) will then be nested inside a list.

By using this fitting and forecast producing procedure, it is easy to fit different model for different temporal frequency, as well as modify argument values if needed. On the other hand, if fitting different model required for different cross-sectional level, `if` condition will be used nested inside each temporal frequencies loop to achieve this. Forecast and residuals will then be assigned to `base` and `res` data.

8.2.1.1 ARIMA

8.2.1.2 VAR/VECM Procedures and models used to forecast LTD. - Apply models and produce forecast, as well as residuals across aggregation levels for one temporal frequency at a time. - Monthly data Temporal frequency considering will be: -

- -

8.3 Reconciliation Process

8.3.1 Approach:

- After combine all forecast and residuals output to ‘base’ and ‘res’ data, FoReco package will be used for reconciliation process, since it supports cross-sectional, temporal, optimal cross-temporal reconciliation all in one place.
- ‘htsrec()’ for cross-sectional reconciliation
- ‘thfrec()’ for temporal reconciliation
- ‘octrec()’ for optimal cross-temporal reconciliation
- Base forecast as well as reconciled forecast of different methods will be plotted against the test set or observation
- Accuracy metrics used: RMSE, MASE, MAPE

9 Results

9.1 Model Performance:

Presentation of model accuracy and comparison.

9.2 Forecast Results:

Detailed discussion of the forecasted values and their confidence intervals.

10 Discussion

10.1 Interpretation of Results:

Analysis of what this forecasts mean for DTF.

10.2 Limitations and Assumptions:

Any limitations encountered during the forecasting process.

11 Conclusion and Recommendations

11.1 Summary:

Recap the findings and their implications.

11.2 Future Work:

Suggestions for improving future forecasts.

12 Appendices

12.1 Additional Data:

Any supplementary data or detailed tables.

12.2 Code:

Include or reference the R scripts used for analysis.

13 References

13.1 Bibliography:

Cite all data sources, literature, and software used in the report the insights generated by your work.

Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3. Accessed on .