NUS Business School Honors Dissertation

Peng Seng Ang

AY 2019/2020 Semester 2

Abstract

This paper studies how we can use various spatial-temporal time series model to better predict demand across different locations.

1 Introduction

Having an accurate forecast of delivery demand for food service providers would help them more effectively and efficiently assign orders to drivers to improve the overall delivery time. Currently, most Autoregressive (AR) or Autoregressive Integrated Moving Average (ARIMA) models only consider temporal features when predicting demand. However, we believe including spatial features between the data points might improve forecast accuracy. This paper would focus on and explore models that include both spatial and temporal features to improve forecast accuracy.

2 Literature Review

To be written

3 Data

The data source used was an operational dataset from a food delivery service provider from Shanghai that includes delivery information for a 2-month period from 10 August 2015 to 30 September 2015 (excluding Saturdays) in 2015. The provider only provides delivery service for 90 minutes during lunchtime and the dataset has split the data into 15-minute time periods, and as such, each day would only consists of demand data for 6 time periods. Hence, our dataset has 839 locations with demand count data, in integer, for 204 time periods in total.

To include other exogenous variables, data from https://www.worldweatheronline.com/shanghai-weather-history/shanghai/cn.aspx was used to include weather and rainfall data as well as encoding of the weekadys for all the respective days.

3.1 Exploratory Analysis

We would first do some exploratory analysis and check if there are any obvious relationships between the variables.

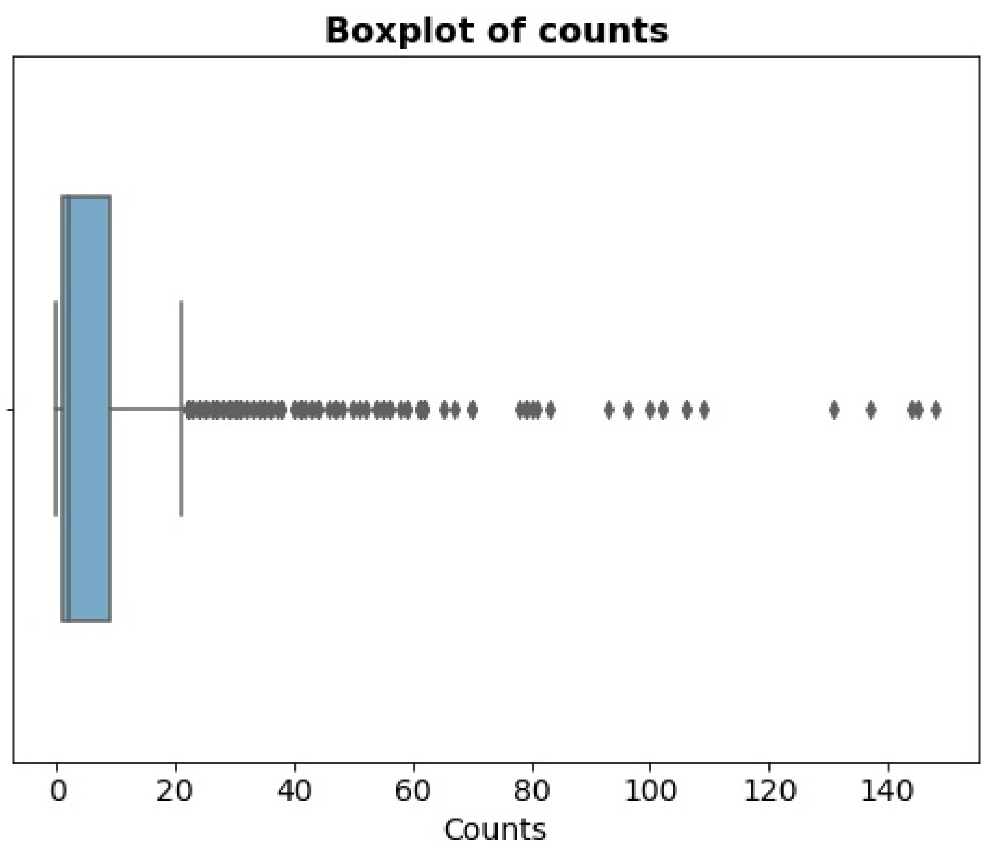

Figure 1: Boxplot of counts

We can see from the boxplot that most of the locations have extremely low number of non-zero orders, with about 335 locations having just a maximum of one non-zero order throughout the 204 time periods.

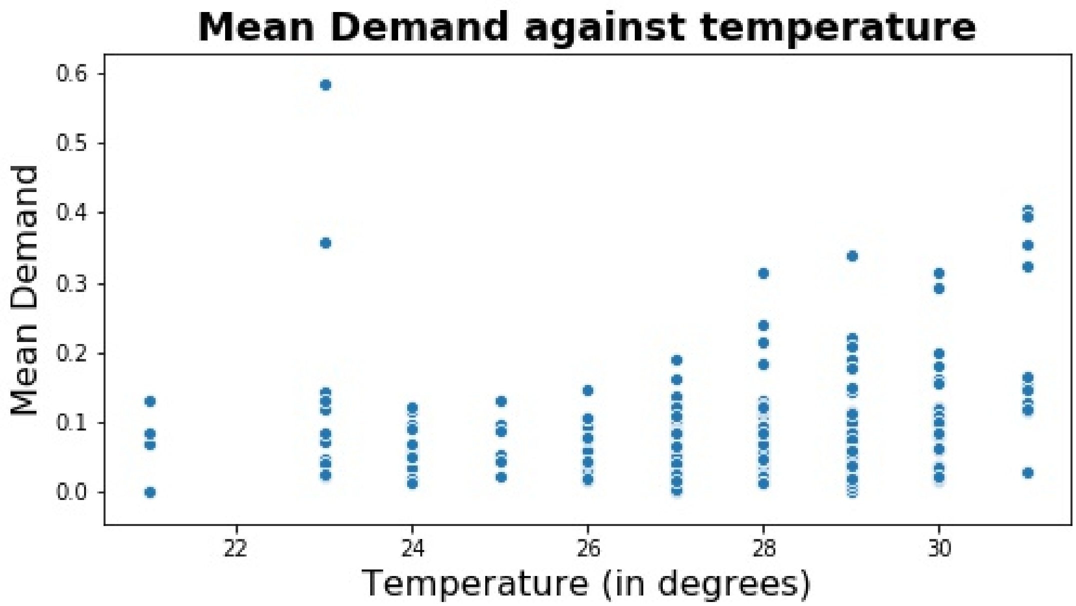

Figure 2: Scatter plot of mean counts against temperature

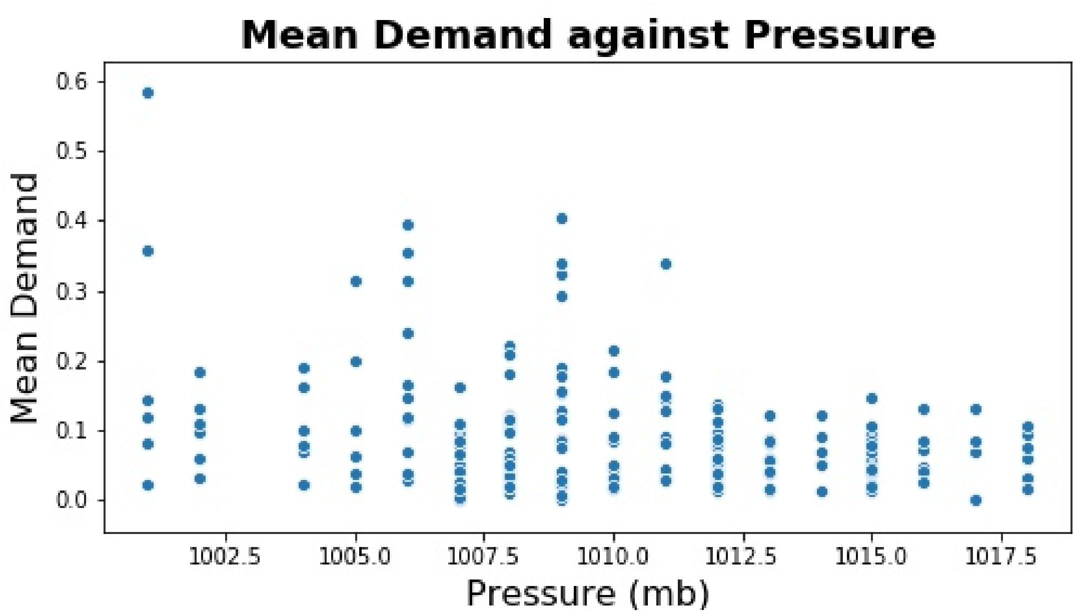

Figure 3: Scatter plot of mean counts against pressure

The scatter plot in Figure 2 visually display a slight positive relationship between temperature and mean demand across all locations whereas Figure 3 visually display a slight negative relationship between pressure and mean demand across all locations.

The distribution plots for the rest of the exogenous variables can be found in the appendix.

4 Baseline Model

In this section, we would build a simple baseline model. Following which, we would try other different spatial temporal time series models and compare the results to the baseline model.

4.1 Metric Used

The main metric that would be used for comparison would be Mean Squared Forecast Error (MSFE), which is calculated by:

$$MSFE = \frac{1}{n}\sum_{t=1}^{n} \|\hat{y}_t - y_t\|_2^2$$

where n is the number of data points, $\hat{y_t}$ is the predicted demand at time t and y_t is the actual demand at time t.

4.2 Train-Test Split

From Figure 1 in Section 3.1, the data is very sparse as there are many locations that have no demand counts for the majority of the time period. Hence, to get a better idea of how our models would work, only locations with at least 50 non-zero counts across the time period would be used initially, leaving us with 42 locations that meet this criteria. The dataset was then split into training and test set by considering the first 27 days as the training set and the next 7 days as the test set. Our training set would then have 162 demand data for each location and test set would have 42 demand data for each location.

4.3 ARIMA models

Autoregressive Integrated Moving Average (ARIMA) models are one of the most commonly used models for time series (Z. Asha Farhath (2016)). ARIMA models are made up of 3 processes, mainly the Autoregressive

(AR) process, the Integrated (I) process and the Moving Average (MA) process (Jamal Fattah (2018)). The AR process assumes that each observation can be expressed as a linear combination of its past values. An AR(x) process would mean using x lagged values. The MA process assumes that each observation can be expressed as a linear combination of its current error term as well as its past error terms. The Integrated Process states that the time series can undergo differencing to ensure that the series is stationary. A MA(x) process would mean using x number of past observations. Hence, an ARIMA model is usually represented by ARIMA(p,d,q), where p represents the number of autoregressive terms, d represents the number of differences needed for stationarity, and q represents the number of lagged forecast errors.

4.4 Baseline ARIMA Result

As a baseline model, each of the locations was assessed individually and a suitable ARIMA model was built for each location. Auto-arima function from Python was used to implement this. The out-of-sample MSFE for this baseline model on the 42 locations is **58.80**. A sample forecast plot is shown below:

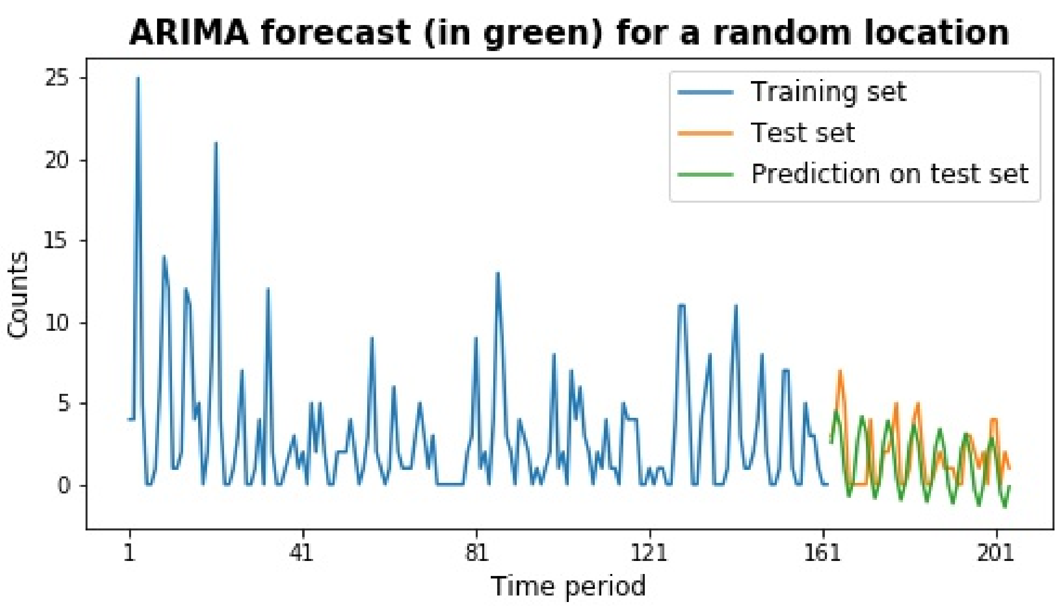

Figure 4: ARIMA forecast on a random location

5 VAR Model

Vector Autoregressive (VAR) models are the most commonly used model for multivariate time series, particularly in economics and financial time series as shown in Bjrnland (2000). VAR models are very similar to multivariate linear regression models and methods used to perform inferencing on linear regression models can also be applied to VAR models. VAR(p) represents a VAR model of order p if the time series can be written as:

$$y_t = v + \sum_{i=1}^{p} \phi_i y_{t-i} + \alpha_t$$

where p is the number of lagged endogenous variables used, y_t is the value at time t, v is a constant vector, ϕ_i are coefficient matrices for $i > 0$ and α_t are independent and identically distributed random vectors.

5.1 Stationarity Condition

For a univariate time series, it is important for the time series to be transformed into a stationary series and Augmented Dickey-Fuller (ADF) test can be used to perform unit root test for stationarity, as shown in Zhijie X. (1998) and Mushtaq (2011). For a multi-variate time series, if the series are unit-root nonstationary, applying the VAR model could lead to spurious regression, as shown in Baumhl (2009). While it is possible to perform differencing on every series, it might cause over-differencing, as mentioned in Tsay (2014). In this paper, we would check for cointegration between the series instead.

5.1.1 Cointegration

Box (1977) shows that it is possible to linearly combine various unit-root nonstationary time series to form a stationary series. The term Cointegration, first mentioned in Granger (1983), states that although some or all the time series might be unit-root nonstationary individually, these time series can be said to be cointegrated if there exists a possible linear combination of them that would form a stationary series. Intuitively, 2 series are cointegrated if they move together and the distance between them remain stable over time.

5.1.2 Johansen Test for Cointegration

While Cointegrated Augmented Dickey Fuller Test, commonly used for Pairs Trading, can be used, it is only able to be applied on 2 separate series. In our dataset, we have 839 locations at least, hence we would apply the popular approach to cointegrating tests for VAR model, called the Johansen's Cointegration Test. However, one limitation is that it can only be used to check for cointegration between a maximum of 12 variables. For further elaboration on the Johansen's Cointegration Test, please refer to Johansen (1991).

5.2 VARX Model

VAR models can also be extended to include exogenous variables. A VARX(p,s) (with exogenous variables) model can be expressed as:

$$y_t = v + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{s} \beta_j x_{t-j} + \alpha_t$$

where p is the number of lagged endogenous variables used, s is the number of lagged exogenous variables used, y_t is the value at time t, v is a constant vector, ϕ_i are coefficient matrices for endogenous coefficient matrix for $i > 0$, β_i are coefficient matrices for exogenous coefficient matrix for $i > 0$ and α_t are independent and identically distributed random vectors.

5.3 Model Checking

To validate and verify if our fitted model is adequate, model checking would be performed by performing the following residual analysis:

5.3.1 Whiteness of Residuals

To ensure our fitted model is adequate, the resiudals should behave like a white noise series.

The plots below shows the distribution of our residuals for the VAR model and the VARX model.

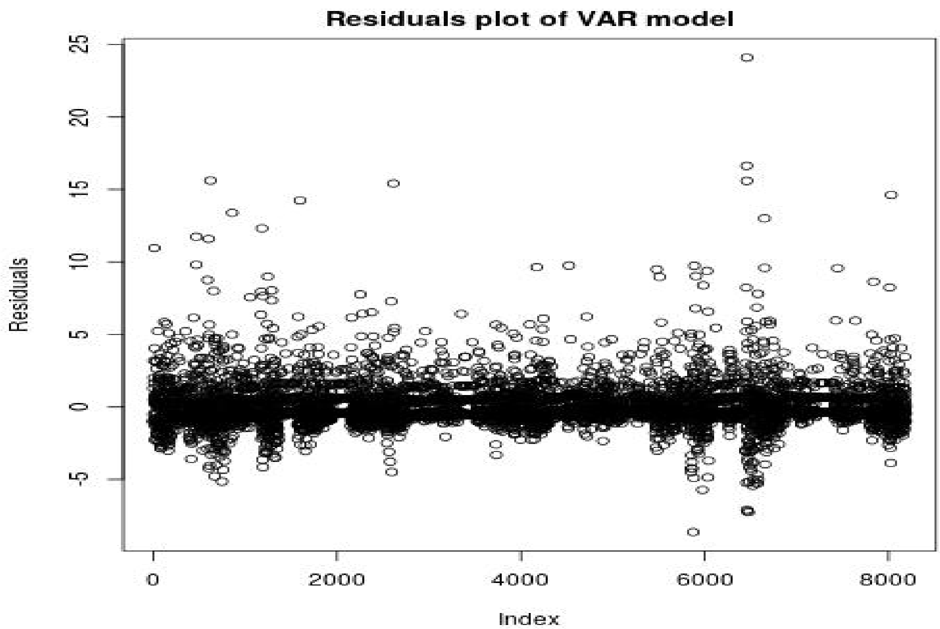

Figure 5: Residuals for VAR Model

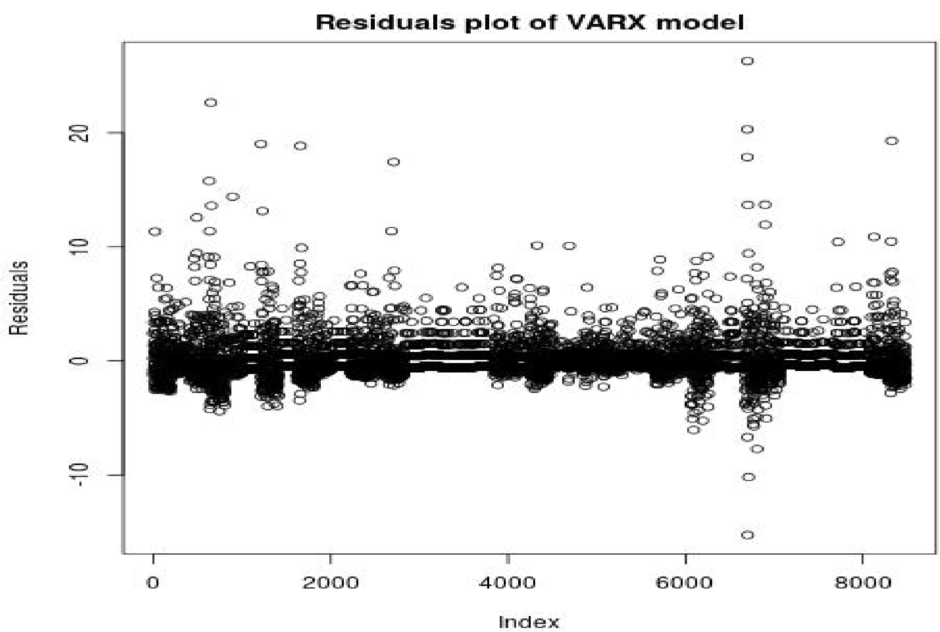

Figure 6: Residuals for VARX Model

5.3.2 Correlation of Residuals

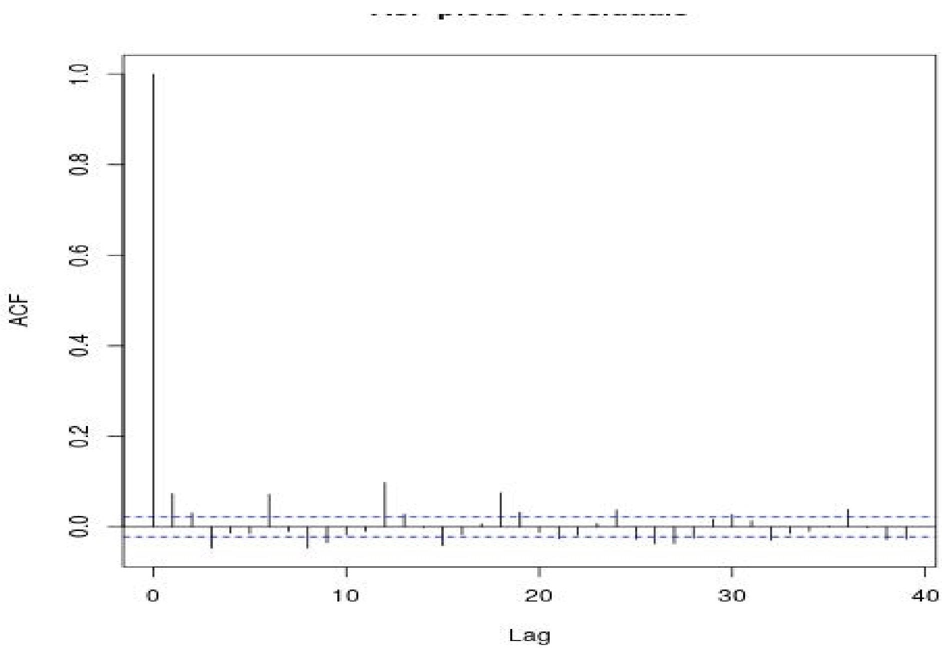

Figure 7: ACF of Residuals for VAR Model

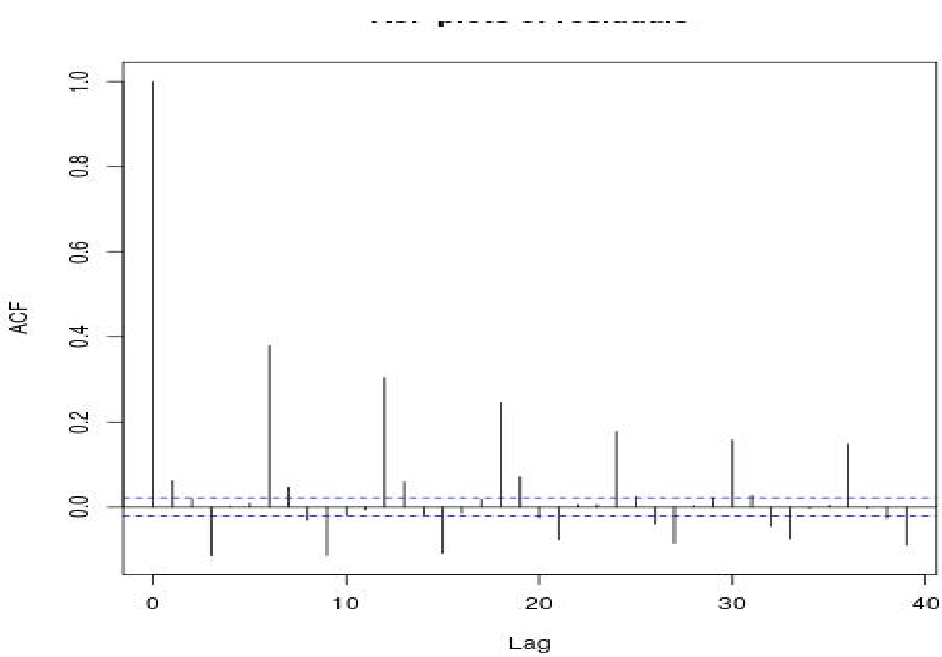

Figure 8: ACF of Residuals for VARX Model

5.4 Results

BigVAR library in R was used to implement the VAR models. The results from the VAR model without exogenous variables gives an out-of-sample MSFE of XXX on the 42 locations.

6 GLM Model

The dataset that we are using follows a count time series, which means the observations are non-negative integers. A flexible and commonly used model for count time series is the Generalized Linear Model (GLM) Nelder JA (1972). GLM normally take the form of:

$$g(\lambda_t) = \eta^T X_t$$

Using the R package from Tobias Liboschik (2017), the GLM used would be an extension of the above equation and can be expressed in the form of:

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \tilde{g}(Y_{t-i_k}) + \sum_{l=1}^{q} \alpha_l g(\lambda_{t-j_l}) + \eta^T X_t$$

where g represents a link function and $\tilde{g}$ represents a transformation function. η represents a parameter vector that corresponds to the covariates.

Since there are many locations which have values that are all 0 throughout all the time period, the GLM model would run into an error if applied on those. Hence, only locations with at least 1 non-zero value would be considered. Similar to before, each of the locations was assessed individually and a suitable GLM model was fitted for each location. The out-of-sample MSFE for this baseline model is **XXX**.

6.1 Model Checking

To validate our model, we shall analyse the residuals distribution. Only residuals from a randomly selected number of locations would be shown for conciseness.

6.1.1 Whiteness of Residuals

6.1.2 Correlation of Residuals

6.2 Results

7 VGLM Model

VGLM

7.1 Model Checking

To validate our model,

7.2 Results

7.3 Insights and Implementation

Any findings from the results? If there are any benefits or issues in implementing the proposed model...

8 Limitations

9 Conclusion

Conclude your efforts and main findings.

10 Appendix

Append extra plots, graphs, analysis, etc.

References

Baumhl, Eduard, L. . (2009). Stationarity of time series and the problem of spurious regression.

Bjrnland, H. C. (2000). Var models in macroeconomic research.

Box, G. E. P., T. G. C. (1977). A canonical analysis of multiple time series. biometrika64(2):355365.

Granger, C. (1983). Co-integrated variables and error-correcting models, ucsd discussion paper 83-13.

Jamal Fattah, Latifa Ezzine, Z. A. H. E. M. A. L. (2018). Forecasting of demand using arima model.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models.

Mushtaq, R. (2011). Augmented dickey fuller test.

Nelder JA, W. R. (1972). generalized linear models. journal of the royal statistical society a, 135(3), 370384.

Tobias Liboschik, Konstantinos Fokianos, R. F. (2017). tscount: An r package for analysis of count time series following generalized linear models.

Tsay, R. S. (2014). Multivariate time series analysis with r and financial applications.

Z. Asha Farhath, B. Arputhamary, L. A. (2016). A survey on arima forecasting using time series model.

Zhijie X., P. C. P. (1998). An adf coefficient test for a unit root in arma models of unknown order with empirical applications to the us economy.