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 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 data 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.

2.1 Exploratory Analysis

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.

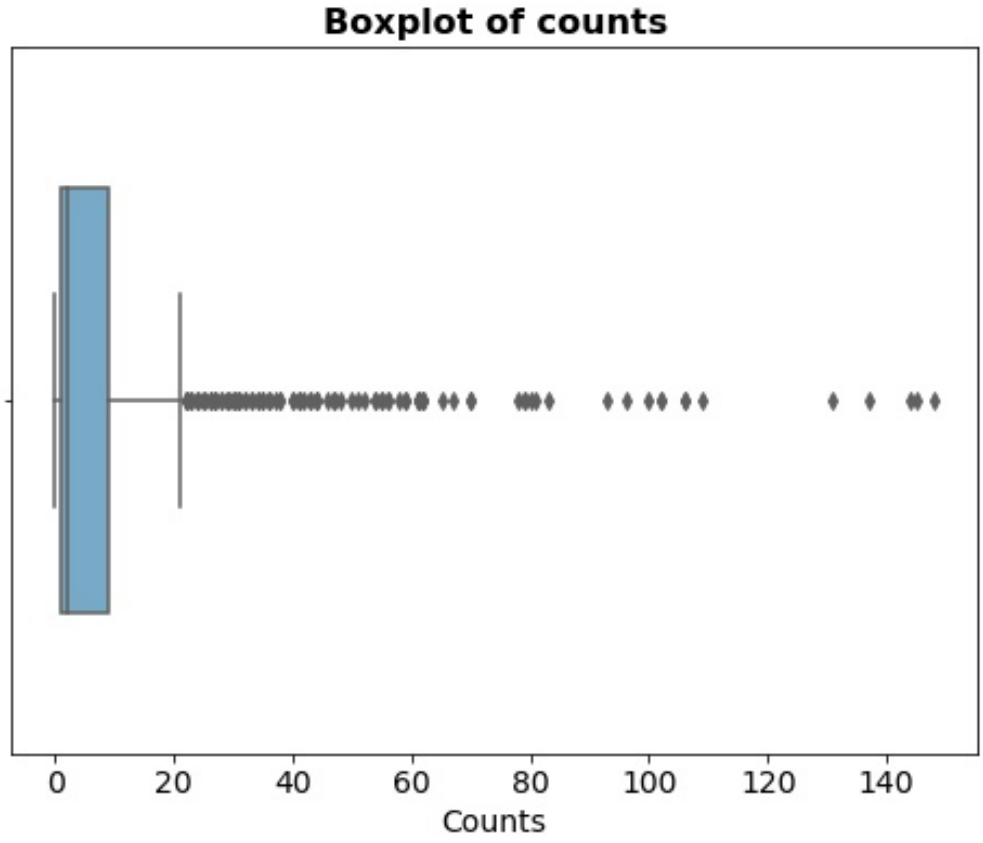

Figure 1: Household Attributes

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

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

3.1 Train-Test Split

Since there are many locations that have no demand or orders for the majority of the time period, the data is very sparse. 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 b used initially. The dataset was 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.

3.2 ARIMA models

Autoregressive Integrated Moving Average (ARIMA) models.....

3.3 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.

4 VAR Model

Vector Autoregressive (VAR) models are.... To validate if the multi-variate time series is stationary, the Johansen's test for cointegrating time series would be performed.

Assumption 1. *The first assumption of a VAR model is....*

Assumption 2. *It is generally true...*

4.1 VARX Model

VAR models can also be extended to include exogenous variables.

4.2 Model Checking

To validate our model,

4.3 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 46.641 on the 42 locations.

5 GLM Model

Assumption 3. *Generalised Linear Models (GLM) are.....*

Assumption 4. *An assumption is that the data follows a poisson process or a non-homogenous poisson process.*

5.1 Model Checking

To validate our model,

5.2 Insights and Implementation

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

6 Conclusion

Conclude your efforts and main findings.

7 Appendix

Append extra plots, graphs, analysis, etc. Walliamson Zengyi (2001)

References

Walliamson Zengyi, P, H. (2001). A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small area microdata.