House Price Prediction Based on ARIMA Model and STARIMA Model

Introduction and Data Description

1. Experiment Summary

- Main task: Forecasting the house price of ward Dulwich Village in Southwark.
- Models / Methods: ARIMA and STARIMA
- Hypothesis: Considering spatial dependence can improve forecasting accuracy, i.e., STARIMA model performs better than ARIMA model.

2. Literature Review

ARIMA model has a strict theoretical statistical background, convenience, and ideal forecasting accuracy, making it a popular time-series forecasting model [1]. Peng et al., in their study, showcased that the AIRIMA model has the advantage of having lower error rates than basic time series models [2]. STARIMA performs better in spatial-temporal forecasting by leveraging spatial dependence. It is significantly used in timber price modelling, traffic flow modelling, electricity demand forecasting and analysis of regional unemployment [1].

3. Data Description

1) Temporal data – house price

- Seasonal median house price of wards in England and Wales from Dec 1995 to Jun 2022.Downloaded from the UK government website *Office for National Statistics* [3].
- 8320 rows×108 columns: Rows represent wards, and columns represent ward information (ward code, ward name, local authority code and local authority name) and time stamps. Data from Dec 1995, Mar 2022 and Jun 2022 were removed to generate an integrated seasonal time series from 1996 to 2021.

2) Spatial data – ward shapefile

- A shapefile of London wards information. Downloaded from the UK government website *The London Assembly* [4]
- 657 rows×5 columns: Rows represent wards, and columns represent ward information (same as the house price data).
- Finally, house price was aggregated to the ward data and written into a shapefile for spatial and temporal analysis.

4. Data Visualisation

Table.1 shows that the house price of Dulwich Village is greatly higher than the house price of Southwark. It can also be confirmed in the distribution histograms, where house prices in Dulwich Village distribute more concentratedly between 500,000 to 1,500,000, while house price in Southwark has a positively skewed distribution, with the majority of the data clustered on the left side between 0 to 500,000.

	Southwark	Dulwich Village
mean	310,989	659,013
median	263,250	593,750
min	41,000	140,250
max	1,815,000	1,500,000

Table.1 Statistical summary

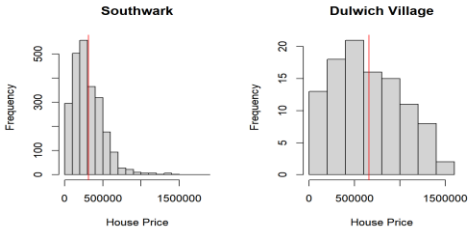
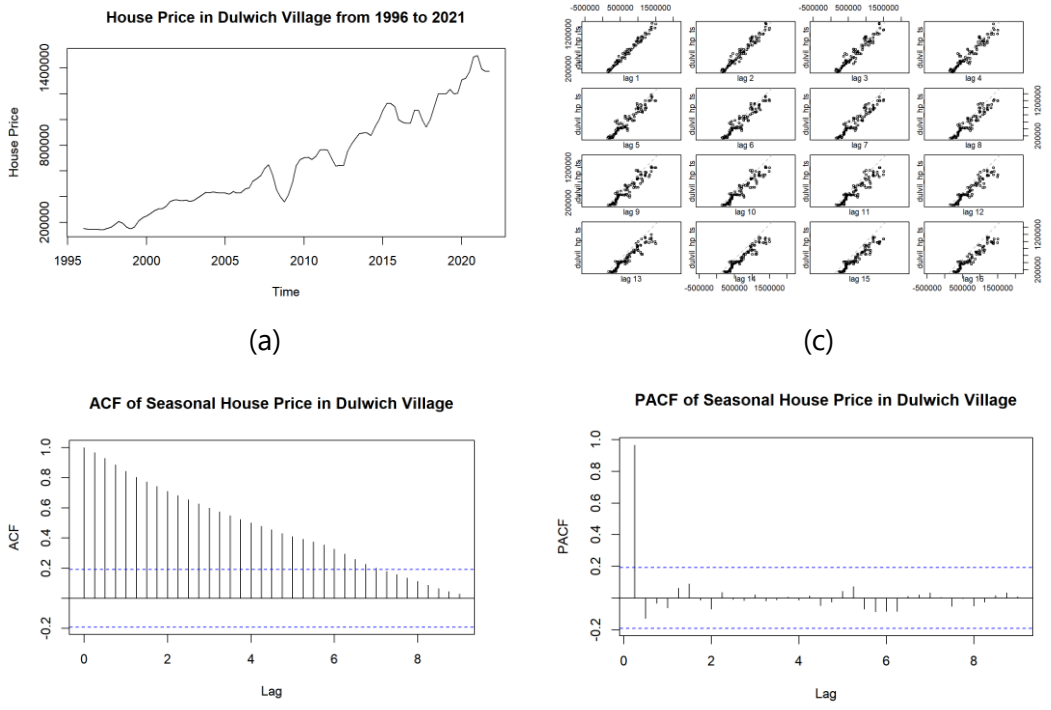


Figure.1 House price data distribution

Explanatory Spatio-Temporal Data Analysis

1. Temporal Characteristics

Figure.2(a) shows that with some fluctuations, house prices in Dulwich Village continued to grow from 1996 to 2021. From the scatter plot of temporal correlations (Fig.2(b)), strict autocorrelation can be seen at lag1; then, the autocorrelation weakens with lag values increasing. As ACF decays slowly to 0 and PACF only has lag1 being significant (due to default settings of Rstudio, lag values in the graph are the multiples of 4), it proves that significant temporal autocorrelation exists in the time series.



(a) (c)

(b) (d)

Figure.2 Temporal characteristics of house price in Dulwich Village

2. Spatial Autocorrelation

To have a general look at the spatial autocorrelation across wards in Southwark, seasonal data were aggregated into yearly data and put into the Moran'I test. As shown in Table 2, only about 1/3 years show significant spatial autocorrelation, which contributes to the result that the seasonal dataset had no significant spatial autocorrelation. Overall, from the spatial perspective, house prices in Southwark are positively autocorrelated. However, in some years, there is no sufficient evidence to reject the null hypothesis that spatial dependence doesn't exist between wards in Southwark.

Yearly	**Spatial Autocorrelation (Mean)**	**Significant Year Number**
	0.14	8
Seasonal	**Moran I**	**p-value**
	0.09	0.12

Table.2 Moran'I Test results

3. Spatio-temporal Autocorrelation

Figure.3 shows the STACF and STPACF of house prices in all the wards in Southwark. It indicates a similar pattern to the Dulwich Village ACF and PACF and also proves that significant temporal autocorrelation exists in the time series, which has included the factor of spatial dependence.

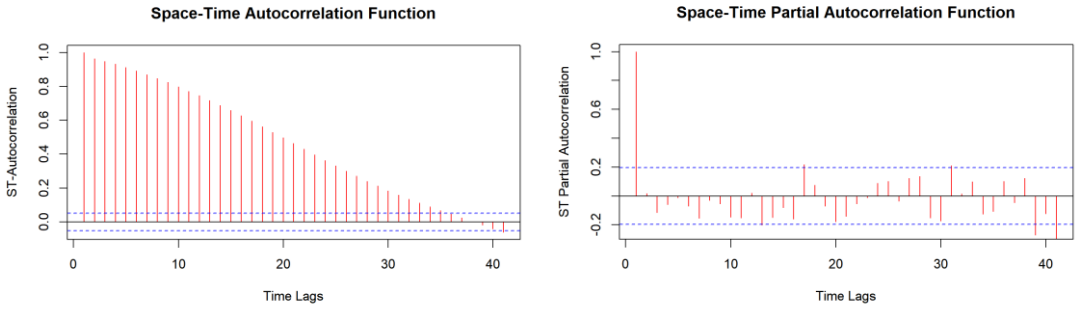


Figure.3 Spatio-Temporal characteristics of house price in Dulwich Village

Methodology and Results

1. Description of the method

ARIMA model consists of three components:

- AR (Auto regression): The autoregressive order of the model (*p*).
- I (Integrated): The integration order of the model (*d*), which is the number (and type) of differences required to make the series stationary.
- MA (Moving average): The moving average order of the model (*q*).

A model that has non-zero *p*, d and q is called an ARIMA (*p*, *d*, *q*) model. ARIMA models can also have a seasonal component, which is denoted (*P*, *D*, *Q*) s, where uppercase indicates that the components are seasonal, and s indicates the seasonal period. An ARIMA model with both seasonal and nonseasonal components is denoted as an ARIMA (*p*, *d, q*) (*P*, *D*, *Q*) S model. STARIMA model is the extension of the ARIMA model to space. It can capture space-time autocorrelation in series by means of a spatial weight matrix. The explanation of orders is the same as the ARIMA model.

2. Detailed Explanation of the Experimental Setup

- ARIMA

Figure.4 shows the time series is non-stationary with both trend and seasonal components. Therefore, it is necessary to do differencing and seasonal differencing before building the model.

After a few times of attempts, the trend was removed after one time of differencing. The seasonality was removed after one time of seasonal differencing with the period being 8 (Fig.5). Therefore, the differencing parameters are 1 differencing and 1 seasonal differencing with the period of 8.

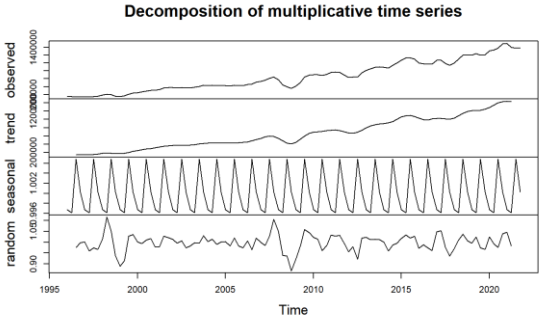
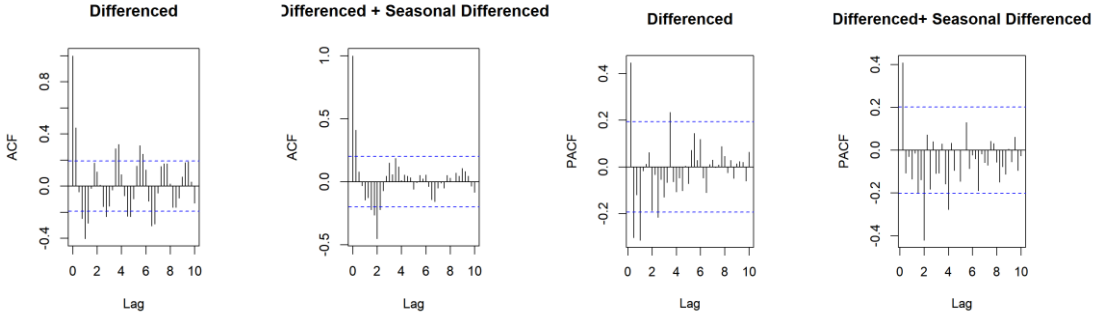


Figure.4 Decomposition components of the time series



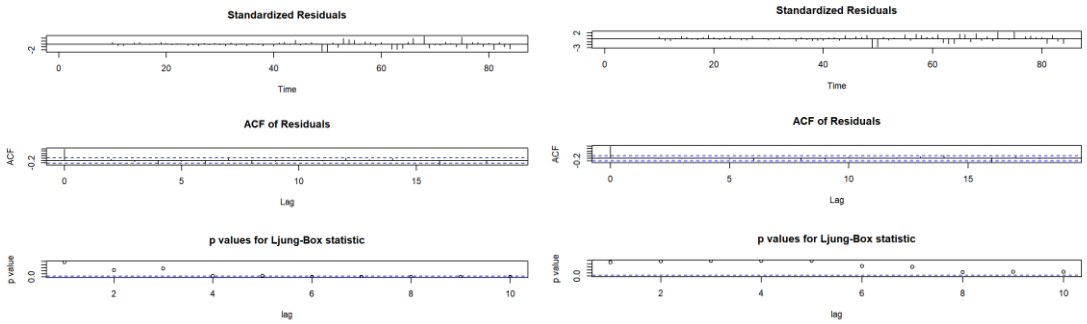
(a) ACF (b) PACF

Figure.5 ACF and PACF after differencing and seasonal differencing

According to the ACF and PACF after differencing and seasonal differencing (Fig.5), autocorrelation at lag 1 and lag 8 are strongly significant, and partial autocorrelation at lag 1 and lag 8 are also strongly significant, suggests an AR(1) model and a seasonal component with the order of (1,1,0)8. As there are no more strongly significant spikes in the ACF, an MA model may not be needed. So far, ARIMA(1,1,0)(1,1,0)8 can be a starting point for choosing the candidate model.

3) Diagnostic checking and prediction

This experiment chooses 80% of the dataset for training (first 21 years) and 20% for testing (last 5 years). The Box–Pierce (and Ljung–Box) test results are shown in Figure.6(a), which indicates the residuals are not random and further order-tuning is needed to find the model that can better explain the time series pattern. After attempting other potential models, ARIMA(4,1,4)(1,1,0)8 passed the Box-Pierce(and Ljung-Box) test(Fig.6(b)) and is considered the optimal model with a better balance between the order scale and the statistical test. Larger orders than this are likely to add more unnecessary errors to the data. The prediction effect and normalized root mean squared error (NRMSE) of fitting and testing are shown respectively in Figure.9(a) and Table.3.



(a) ARIMA(1,1,0)$(1,1,0)_8$ (b) ARIMA(4,1,4)$(1,1,0)_8$

Figure.6 The Box-Pierce(and Ljung-Box) test results

- STARIMA

After one time of differencing, there is no seasonality in the STACF and STPACF(Fig.7). In the meantime, they both have a strongly significant spike at lag 1, suggesting a STARIMA(1,1,0) model would be a proper starting point to fit the model, which will use the same training and testing dataset to the ARIMA model. Diagonisc checking (Fig.8) indicates residuals are random and normally distributed. This suggests STARIMA model has explained the time series pattern in a good way, and there are no remaining unexplained patterns in the dataset. The prediction effect and NRMSE of fitting and testing are shown respectively in Figure.9(b) and Table.3.

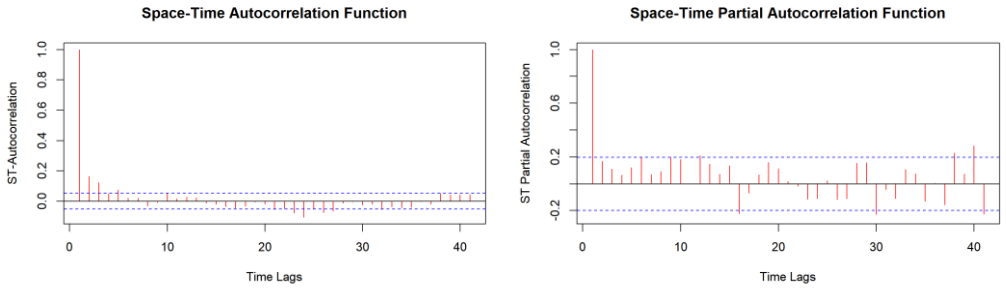


Figure.7 Differenced STACF and STPACF

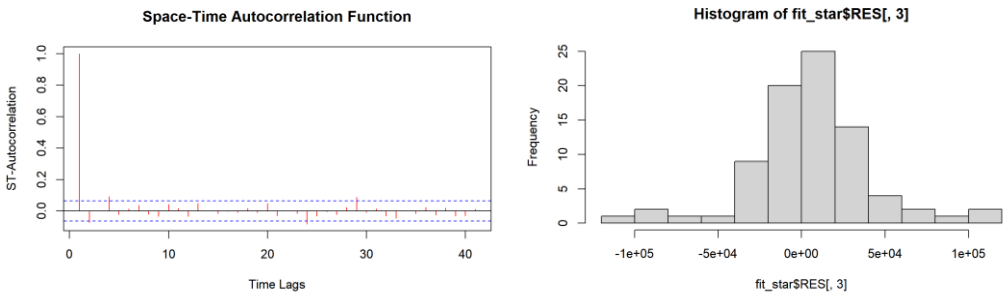
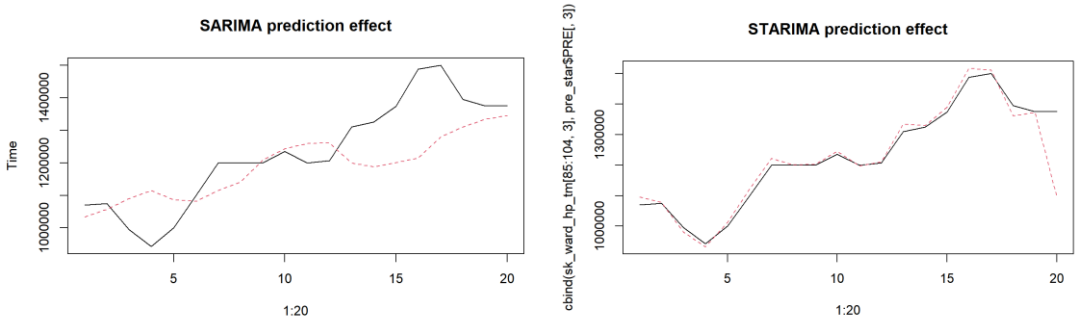


Figure.8 Diagonistic checking for STARIMA model



(a) ARIMA(4,1,4)(1,1,0)$_8$ (b) STARIMA(1,1,0)

Figure.9 Comparision of prediction effect

	Fitting NRMSE	**Testing NRMSE**
(S)ARIMA	0.10	2.03
STARIMA	0.06	1.27

Table.3 Comparision of NRMSE

Conclusion and Discussion

According to Figure.9 and Table.3, the STARIMA model has smaller fitting NRMSE and testing NRMSE, meaning that the STARIMA model can better explain the pattern of spatio-temporal series pattern and make more accurate predictions. This confirms the hypothesis of this experiment and is consistent with the previous study results [3]. The main reason why the STARIMA model can perform better than the ARIMA model is STARIMA model improves its interpretability by accounting for spatial dependencies of the data with spatial weights and spatial autocorrelation in the model [1]. This is particularly useful when dealing with spatial-temporal data such as temperature, air quality, or economic activity across different locations. In terms of ease of implementation, the R package *STARIMA* developed by the Spatio-temporal Data Mining Group from UCL includes all the essential functions for spatio-temporal analysis and makes the experiment easy to implement [5].

However, though the STARIMA model in this experiment makes great improvements to the ARIMA model, there are some limitations to consider in the context of house prices:

1. Spatial heterogeneity: Housing prices can be influenced by other spatial factors such as population, crime rates, and school quality. The effectiveness of STARIMA in capturing these spatial dependencies may be limited by data on these factors.

2. External factors: Housing prices can also be influenced by external factors such as economic conditions and government. While STARIMA can capture the spatial and temporal dependencies in the data, it may not be able to account for all of these external factors, which can limit the accuracy of the forecasts.

There are several ways to improve the performance of the STARIMA model, for example:

1. Combining STARIMA with other models, such as neural networks or decision trees, may improve forecasting performance by capturing non-linear relationships and interactions between the spatio-temporal features.

2. The effectiveness of STARIMA depends heavily on the quality of the spatial weights used in the model [3]. Improving the specification of these weights can improve the accuracy of the forecasts.

Overall, for this experiment, the STARIMA model has a significant advantage compared to the ARIMA model when forecasting the house price. While considering house price can be influenced by many factors other than space, more complex models can be compared or combined with the STARIMA model to improve forecasting accuracy.

Reference List

[1] Yang, Y., & Zhang, H. (2019). Spatial-temporal forecasting of tourism demand. *Annals of Tourism Research*, 75,106-1119. https://doi.org/10.1016/j.annals.2018.12.024.

[2] Peng, B., Song, H., & Crouch, G. I. (2014). A meta-analysis of international tourism demand forecasting and implications for practice. *Tourism Management*, 45, 181–193. https://doi.org/10.1016/j.tourman.2014.04.005

[3] Office for National Statistic. (2022). Median house prices by ward: HPSSA dataset 37.https://www.ons.gov.uk/peoplepopulationandcommunity/housing/datasets/medianpricepaidbywardhpssadataset37

[4] Great London Authority. (2018). Statistical GIS Boundary Files for London. https://data.london.gov.uk/dataset/statistical-gis-boundary-files-london

[5] Wang, F., Zou, Y., Zhang, H., & Shi, H. (2019). House price prediction approach based on deep learning and ARIMA model. In 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT) (pp. 303-307). IEEE. https://doi.org/10.1109/ICCSNT47585.2019.8962443