**Technical Report: SARIMAX Algorithm for Time Series Forecasting**

**Introduction**

Time series forecasting is an essential component in various fields, such as finance, supply chain management, and transportation. The Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors (SARIMAX) model is a popular method for time series forecasting, particularly when data exhibits seasonal trends. This report provides an overview of the SARIMAX algorithm, its application, and the significance of its parameters.

**Why Use SARIMAX?**

SARIMAX is an extension of the ARIMA (Autoregressive Integrated Moving Average) model that incorporates seasonality and exogenous variables. The following reasons justify its usage:

1. **Seasonality Handling**: Many real-world datasets exhibit seasonal patterns. SARIMAX captures these seasonal variations effectively, making it suitable for datasets with periodic fluctuations, such as public transport passenger journeys.
2. **Flexibility**: The model allows for integration of exogenous variables, which can improve forecasting accuracy by including additional information, such as holidays or special events.
3. **Robustness**: SARIMAX is robust to different types of time series data, providing accurate forecasts even with non-stationary data.

**Model Parameters**

SARIMAX is defined by several parameters that determine the model's structure. The general notation for SARIMAX is SARIMAX(p,d,q)(P,D,Q,s)SARIMAX(p, d, q)(P, D, Q, s)SARIMAX(p,d,q)(P,D,Q,s), where:

* **p**: The order of the autoregressive (AR) term. It indicates how many past observations are used to predict future values.
* **d**: The degree of differencing. It is the number of times the data needs to be differenced to achieve stationarity.
* **q**: The order of the moving average (MA) term. It indicates how many past forecast errors are considered in the model.

The seasonal part of the model is defined by:

* **P**: The seasonal autoregressive order. It represents the number of seasonal lagged observations included.
* **D**: The seasonal differencing order. It is the number of seasonal differences needed for stationarity.
* **Q**: The seasonal moving average order. It indicates how many past seasonal forecast errors are considered.
* **s**: The length of the seasonal cycle. For example, in daily data with a yearly seasonality, sss would be 365.

**Example Configuration**

In the provided code, the SARIMAX model is configured with the following parameters:

* **order=(1, 1, 1)**:
  + p=1p=1p=1: One lag of the dependent variable is used.
  + d=1d=1d=1: The series is differenced once to achieve stationarity.
  + q=1q=1q=1: One lag of the forecast error is included.
* **seasonal\_order=(1, 1, 1, 7)**:
  + P=1P=1P=1: One seasonal lag is used.
  + D=1D=1D=1: Seasonal differencing is applied once.
  + Q=1Q=1Q=1: One seasonal lag of the forecast error is included.
  + s=7s=7s=7: The model assumes a weekly seasonal pattern.

**Model Fitting and Evaluation**

Once the parameters are set, the SARIMAX model is fitted to the training dataset. The effectiveness of the model can be evaluated using metrics such as Root Mean Square Error (RMSE), which measures the average magnitude of the forecast errors. A lower RMSE indicates better model performance.

**Conclusion**

The SARIMAX algorithm is a powerful tool for time series forecasting, particularly for data with seasonal trends. Understanding its parameters allows practitioners to tailor the model to specific datasets, enhancing forecasting accuracy. The flexibility of including exogenous variables further expands its applicability across various domains. In the context of public transport passenger journeys, SARIMAX can provide valuable insights into passenger trends, enabling better resource allocation and operational planning.