**Technical Report: Vector Autoregression (VAR) Model for Time Series Forecasting**

**Executive Summary**

This report provides a technical overview of the Vector Autoregression (VAR) algorithm implemented for forecasting transportation ridership data across multiple service categories. The VAR model was selected for its ability to capture interdependencies between multiple time series variables and generate simultaneous forecasts for all transportation modes.

**Algorithm Overview**

**Vector Autoregression (VAR) Model**

VAR is a multivariate extension of the autoregressive (AR) model that captures linear interdependencies among multiple time series. Unlike univariate models that forecast each series independently, VAR considers the dynamic relationships between all variables in the system.

**Model Specification**

**Endogenous Variables (k=5):**

* Local\_Route: Local bus service ridership
* Light\_Rail: Light rail system usage
* Peak\_Service: Peak hour service demand
* Rapid\_Route: Express bus service ridership
* School: School-related transportation usage

**Model Parameters and Configuration**

**1. Lag Order Selection (p)**

**Method:** Akaike Information Criterion (AIC)

* **Primary:** AIC-based automatic selection with fallback to BIC
* **Constraint:** Maximum lags ≤ min(12, T/5) where T is sample size
* **Fallback:** Fixed lag order of 2 if automatic selection fails

**Rationale:** AIC balances model fit against complexity, preventing overfitting while capturing sufficient temporal dynamics. The constraint ensures adequate degrees of freedom for parameter estimation.

**2. Information Criteria**

**AIC Formula:** AIC = -2ln(L) + 2k **BIC Formula:** BIC = -2ln(L) + k×ln(T)

Where L is the likelihood, k is the number of parameters, and T is the sample size. BIC applies a stronger penalty for model complexity than AIC.

**3. Estimation Method**

**Technique:** Ordinary Least Squares (OLS)

* Each equation in the VAR system is estimated separately using OLS
* Provides consistent and efficient estimates under standard assumptions
* Computationally efficient for moderate-sized systems

**4. Data Preprocessing Parameters**

**Date Parsing:**

* Multiple format attempts: DD/MM/YYYY, DD-MM-YYYY, YYYY-MM-DD, MM/DD/YYYY
* Error handling with coercion to NaN for unparseable dates
* Monotonic index enforcement through sorting and duplicate removal

**Numeric Conversion:**

* Coercion of non-numeric values to NaN using pd.to\_numeric()
* Removal of infinite values and missing observations
* Non-negativity constraint on forecasts (transportation ridership cannot be negative)

**5. Forecast Generation**

**Methodology:** Direct multi-step forecasting

* Uses the fitted VAR model to generate point forecasts
* Forecast horizon: 7 days ahead
* Input: Last p observations from the cleaned dataset

**Formula for h-step ahead forecast:**

Ŷ\_{T+h|T} = μ + Σᵢ₌₁ᵖ Âᵢ × Ŷ\_{T+h-i|T}

**Model Assumptions and Limitations**

**Assumptions**

1. **Stationarity:** All time series are stationary (mean and variance constant over time)
2. **Linearity:** Relationships between variables are linear
3. **Normality:** Error terms are normally distributed
4. **Homoscedasticity:** Constant error variance across time
5. **No autocorrelation:** Error terms are independently distributed

**Limitations**

1. **Curse of dimensionality:** Parameter count increases quadratically with variables and lags
2. **Stationarity requirement:** May require differencing non-stationary series
3. **Linear relationships only:** Cannot capture non-linear dependencies
4. **No structural breaks:** Assumes stable relationships over time
5. **Short-term forecasting:** Accuracy degrades rapidly beyond short horizons

**Implementation Details**

**Software Environment**

* **Primary Library:** statsmodels.tsa.api.VAR
* **Data Processing:** pandas for data manipulation and cleaning
* **Numerical Computing:** numpy for array operations

**Model Validation Approach**

1. **Data Quality Checks:** Missing values, infinite values, data type validation
2. **Minimum Sample Size:** Ensures at least 10 observations for model stability
3. **Convergence Monitoring:** Automatic fallback if primary estimation fails
4. **Post-estimation Diagnostics:** AIC, BIC, and observation count reporting

**Forecast Post-processing**

* **Non-negativity Constraint:** Applied using .clip(lower=0) to ensure realistic forecasts
* **Aggregation Reporting:** Daily totals and category averages provided
* **Date Alignment:** Proper date indexing for forecast periods

**Performance Considerations**

**Computational Complexity:** O(k²p²T) for estimation where k is variables, p is lags, T is observations **Memory Requirements:** Moderate - scales with dataset size and number of variables **Scalability:** Suitable for systems with 2-20 variables; performance degrades with larger systems

**Conclusion**

The VAR model provides a robust framework for multivariate time series forecasting in transportation applications. Its ability to capture cross-variable dependencies makes it particularly suitable for ridership forecasting where different transportation modes may exhibit correlated demand patterns. The implementation includes comprehensive error handling, data validation, and multiple fallback mechanisms to ensure reliable operation across diverse data conditions.

**Recommended Use Cases:**

* Short-term operational forecasting (1-7 days)
* Systems with moderate variable count (5-15 series)
* Applications requiring simultaneous forecasts across related time series
* Scenarios where cross-variable relationships are important

**Model Limitations to Consider:**

* Requires stationary input data
* Limited to linear relationships
* Forecast accuracy decreases with longer horizons
* Sensitive to outliers and structural breaks

**Sample Analysis:**

**Output:**

Loading data...

Raw data shape: (1919, 7)

First few rows:

date Local\_Route Light\_Rail Peak\_Service Rapid\_Route School \

0 Date Local Route Light Rail Peak Service Rapid Route School

1 30/08/2024 16436 10705 225 19026 3925

2 15/09/2023 15499 10671 267 18421 4519

3 28/12/2021 1756 2352 0 3775 0

4 11/01/2023 10536 8347 223 14072 0

Other

0 Other

1 59

2 61

3 13

4 48

Parsing dates...

Trying automatic date parsing...

Warning: 1 dates could not be parsed

Preparing numeric data...

Data types after conversion:

Local\_Route int64

Light\_Rail int64

Peak\_Service int64

Rapid\_Route int64

School int64

dtype: object

Missing values:

Local\_Route 0

Light\_Rail 0

Peak\_Service 0

Rapid\_Route 0

School 0

dtype: int64

Cleaning data...

Shape before cleaning: (1918, 5)

Shape after cleaning: (1918, 5)

Removed 0 rows

Final dataset summary:

Date range: 2019-07-01 to 2024-09-29

Total observations: 1918

Sample data:

Local\_Route Light\_Rail Peak\_Service Rapid\_Route School

date

2019-07-01 15987 9962 407 21223 3715

2019-07-02 16895 10656 409 21715 3993

2019-07-03 16613 10658 427 22025 3638

2019-07-04 16604 10445 437 21868 3576

2019-07-05 16040 10532 400 20697 2856

Descriptive statistics:

Local\_Route Light\_Rail Peak\_Service Rapid\_Route School

count 1918.00 1918.00 1918.00 1918.00 1918.00

mean 9891.40 7195.45 179.58 12597.21 2352.69

std 6120.72 3345.62 156.53 6720.49 2494.77

min 1.00 0.00 0.00 0.00 0.00

25% 3044.50 4463.50 0.00 6383.00 0.00

50% 11417.00 7507.00 193.00 13106.50 567.50

75% 15517.50 10008.25 313.75 17924.75 4914.00

max 21070.00 15154.00 1029.00 28678.00 7255.00

==================================================

FITTING VAR MODEL

==================================================

Maximum lags to consider: 12Loading data...

Raw data shape: (1919, 7)

First few rows:

date Local\_Route Light\_Rail Peak\_Service Rapid\_Route School \

0 Date Local Route Light Rail Peak Service Rapid Route School

1 30/08/2024 16436 10705 225 19026 3925

2 15/09/2023 15499 10671 267 18421 4519

3 28/12/2021 1756 2352 0 3775 0

4 11/01/2023 10536 8347 223 14072 0

Other

0 Other

1 59

2 61

3 13

4 48

Parsing dates...

Trying automatic date parsing...

Warning: 1 dates could not be parsed

Preparing numeric data...

Data types after conversion:

Local\_Route int64

Light\_Rail int64

Peak\_Service int64

Rapid\_Route int64

School int64

dtype: object

Missing values:

Local\_Route 0

Light\_Rail 0

Peak\_Service 0

Rapid\_Route 0

School 0

dtype: int64

Cleaning data...

Shape before cleaning: (1918, 5)

Shape after cleaning: (1918, 5)

Removed 0 rows

Final dataset summary:

Date range: 2019-07-01 to 2024-09-29

Total observations: 1918

Sample data:

Local\_Route Light\_Rail Peak\_Service Rapid\_Route School

date

2019-07-01 15987 9962 407 21223 3715

2019-07-02 16895 10656 409 21715 3993

2019-07-03 16613 10658 427 22025 3638

2019-07-04 16604 10445 437 21868 3576

2019-07-05 16040 10532 400 20697 2856

Descriptive statistics:

Local\_Route Light\_Rail Peak\_Service Rapid\_Route School

count 1918.00 1918.00 1918.00 1918.00 1918.00

mean 9891.40 7195.45 179.58 12597.21 2352.69

std 6120.72 3345.62 156.53 6720.49 2494.77

min 1.00 0.00 0.00 0.00 0.00

25% 3044.50 4463.50 0.00 6383.00 0.00

50% 11417.00 7507.00 193.00 13106.50 567.50

75% 15517.50 10008.25 313.75 17924.75 4914.00

max 21070.00 15154.00 1029.00 28678.00 7255.00

==================================================

FITTING VAR MODEL

==================================================

Maximum lags to consider: 12

✓ Model fitted successfully!

Selection method: AIC

Selected lags: 12

Observations used: 1906

AIC: 60.75

BIC: 61.64

==================================================

GENERATING FORECASTS

==================================================

Using last 12 observations for forecasting

============================================================

FORECAST RESULTS - NEXT 7 DAYS

============================================================

Local\_Route Light\_Rail Peak\_Service Rapid\_Route School

2024-09-30 517.0 450.6 0.0 60.6 39.7

2024-10-01 0.0 0.0 0.0 0.0 52.0

2024-10-02 1873.2 768.6 41.0 1721.9 486.5

2024-10-03 1366.1 702.1 26.1 1527.6 0.0

2024-10-04 1433.8 748.8 25.7 1585.8 149.9

2024-10-05 1673.5 985.6 32.1 1917.7 104.8

2024-10-06 1826.4 1056.0 29.5 2030.7 151.9

==================================================

FORECAST SUMMARY

==================================================

Daily totals:

2024-09-30: 1067.9

2024-10-01: 52.0

2024-10-02: 4891.3

2024-10-03: 3621.9

2024-10-04: 3944.0

2024-10-05: 4713.6

2024-10-06: 5094.6

Average daily forecast by category:

Local\_Route: 1241.4

Light\_Rail: 673.1

Peak\_Service: 22.1

Rapid\_Route: 1263.5

School: 140.7

Total 7-day forecast: 23385.3

==================================================

MODEL DIAGNOSTICS

==================================================

Model: VAR(12)

Variables: 5

Sample period: 2019-07-01 to 2024-09-29

Training observations: 1906

Forecast horizon: 7 days

==================================================

ANALYSIS COMPLETE

==================================================