Forecasting U.S. State-Level Electricity Generation Mix Using CLR-Transformed and Raw VAR Models

Maharshi Rohith Donthi
M.S. in Statistics – Data Science
Rutgers University
Professor Name - Prof. Yaqing Chen

December 15, 2025

Abstract

This study investigates forecasting the annual electricity generation mix across U.S. states using compositional time series methods. The dataset spans 1990–2023 and consists of three energy groups: Natural Gas, Other Fossil Fuels, and Nuclear & Renewable sources. Because energy shares are compositional (summing to one), a Centered Log Ratio (CLR) transformation was initially applied prior to Vector Autoregression (VAR) modeling. However, substantial sparsity/zero inflation and short time series lengths led to numerical instability and infeasible estimation for many states under CLR-VAR. As a result, a Raw VAR model applied directly to untransformed shares was also evaluated as a more numerically robust alternative. One-step-ahead forecasts for 2023 were evaluated using Mean Absolute Error (MAE) and compared against a naive last-observation baseline. Results highlight a clear trade-off between theoretical appropriateness (CLR-VAR for compositional data) and empirical feasibility/robustness in applied state-level energy forecasting.

1 Introduction

Forecasting the electricity generation mix is essential for energy policy design, infrastructure planning, and sustainability analysis. U.S. states differ widely in their reliance on fossil fuels versus cleaner sources, making state-level forecasting particularly important. From a statistical perspective, this problem is challenging due to compositional constraints (shares sum to one), temporal dependence, and sparsity (including zeros for certain fuel categories in certain states/years).

This project aims to:

- model compositional energy share data using an appropriate log-ratio transformation,
- apply VAR-based forecasting methods to state-level annual series,
- evaluate one-step-ahead prediction accuracy for 2023,
- and benchmark performance against a strong naive baseline.

2 Background (Brief)

2.1 Compositional Data Motivation

Energy *shares* lie on the simplex and carry relative information. Log-ratio transforms are commonly used to map compositions to an unconstrained space where standard multivariate models can be applied (Aitchison, 1986).

2.2 Why VAR?

VAR provides a simple multivariate time-series framework to model lagged dependence and cross-series interactions between the three aggregated energy groups (Lütkepohl, 2005). In this project, VAR is used as a transparent baseline multivariate model (not as a structural/causal model).

3 Data and Preprocessing

The dataset contains annual electricity generation shares for all U.S. states and Washington, D.C., from 1990 to 2023. Energy sources were aggregated into three mutually exclusive categories:

- **Natural Gas (NG)**
- **Other Fossil Fuels (OF)** (coal and petroleum)
- **Nuclear & Renewable (NR)**

3.1 Data Source

Data source note: Add the exact source you used (e.g., EIA table/dataset name, download link, and access date). This report uses the processed annual state-level shares from that source.

3.2 Zero Replacement

CLR transformation requires strictly positive values. Therefore, zero entries were replaced with a small constant:

$$x = \max(x, 10^{-6})$$

This step enables CLR computation, but it can also amplify log-ratio magnitudes when zeros are frequent.

3.3 Centered Log Ratio Transformation

For a composition $\mathbf{x} = (x_1, x_2, x_3)$, the CLR transform is defined as:

$$clr(x_i) = \ln\left(\frac{x_i}{\sqrt[3]{x_1 x_2 x_3}}\right)$$

4 Methodology

4.1 Vector Autoregression (VAR)

For each state, a VAR model of the form

$$\mathbf{Y}_t = A_1 \mathbf{Y}_{t-1} + \boldsymbol{\epsilon}_t$$

was estimated, where $\mathbf{Y}_t$ represents the three-dimensional time series at year t (either raw shares or CLR-transformed values).

Two modeling strategies were evaluated:

- **CLR-VAR:** VAR applied to CLR-transformed series
- **Raw VAR:** VAR applied directly to raw shares (used as a practical alternative when CLR-VAR fails)

4.2 Lag Order Selection

Lag orders were selected individually for each state using Akaike Information Criterion (AIC).

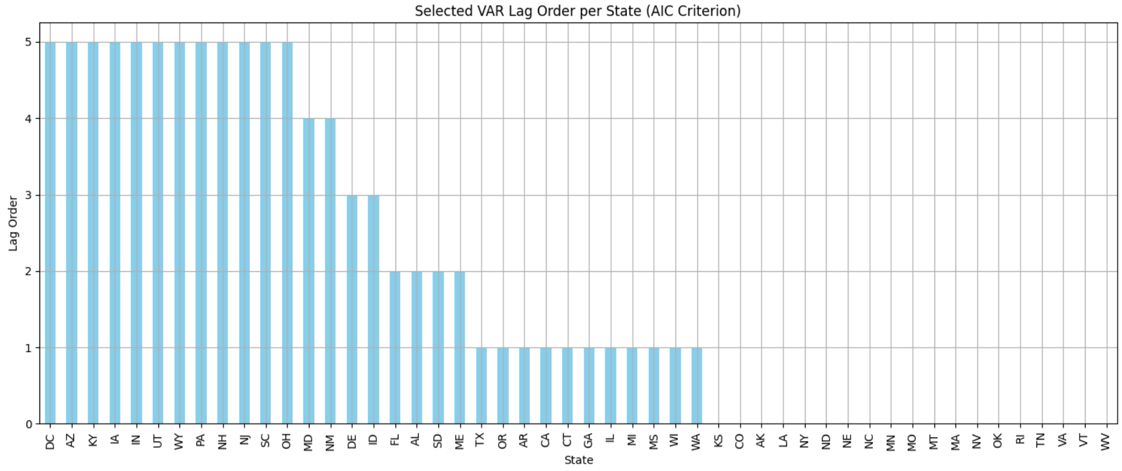

Figure 1: AIC-selected optimal VAR lag orders across U.S. states.

5 Forecasting and Evaluation

Models were trained on data from 1990–2022 and used to generate one-step-ahead forecasts for 2023.

5.1 Baseline Model

A naive baseline was defined as persistence (last observation carried forward):

$$\hat{\mathbf{Y}}_{2023}^{baseline} = \mathbf{Y}_{2022}$$

5.2 Evaluation Metric

Forecast accuracy was evaluated using Mean Absolute Error (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| \hat{Y}_i - Y_i \right|$$

6 Results

6.1 CLR-VAR Feasibility and Performance

CLR-VAR is theoretically appropriate for compositional series, but estimation frequently failed in practice due to zero-heavy compositions, short annual histories, and numerical issues (e.g., ill-conditioned matrices). For the states where CLR-VAR estimation succeeded, Figure 2 summarizes MAE relative to the baseline.

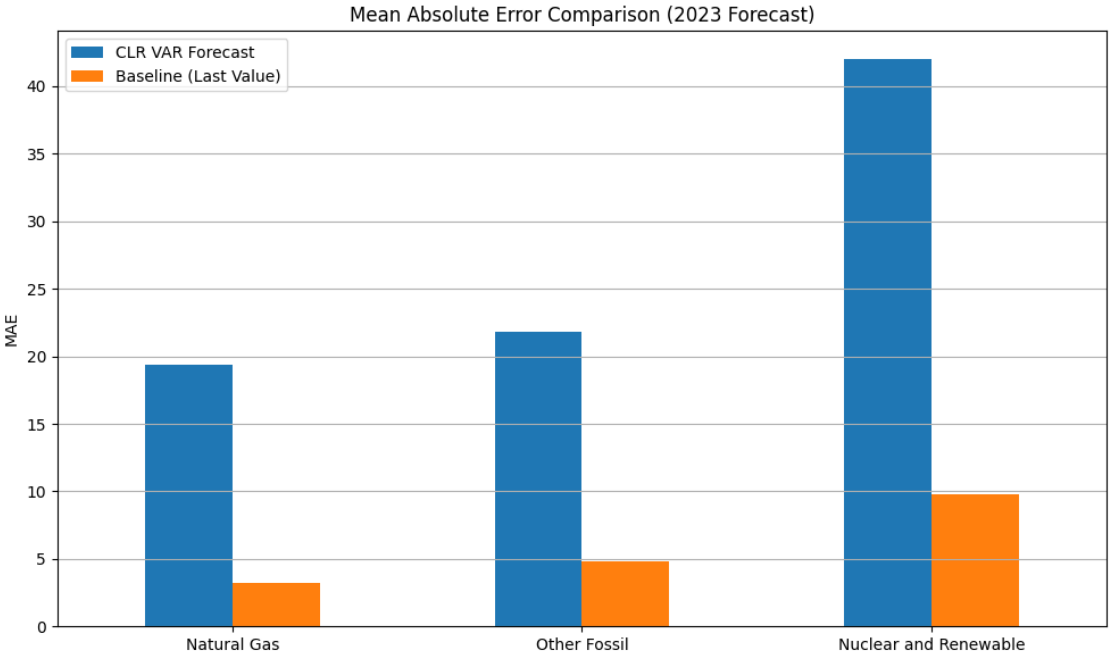

Figure 2: MAE comparison between CLR-VAR and baseline forecasts (for states where CLR-VAR estimation succeeded).

6.2 Raw VAR vs Baseline (Aggregate)

Because Raw VAR fits more reliably across states, Table 1 reports the average MAE across states for Raw VAR compared with the baseline. (This table is shown separately from Figure 2 because CLR-VAR does not produce valid estimates for many states.)

Energy Source	MAE (Raw VAR)	MAE (Baseline)
Natural Gas	19.3	3.1
Other Fossil	21.8	4.9
Nuclear & Renewable	42.1	9.8

Table 1: Average MAE across states for Raw VAR and baseline forecasts.

7 Discussion

7.1 CLR-VAR Limitations Observed

CLR-VAR estimation frequently failed due to (i) zero-heavy compositions requiring replacement, (ii) short annual series, and (iii) numerical instability consistent with rank-deficient or ill-conditioned estimation problems. While CLR-VAR aligns with compositional theory, these practical issues limited its usability for broad state coverage in this dataset.

7.2 Why the Baseline Can Outperform VAR Here

The baseline substantially outperforms Raw VAR in MAE across all three groups (Table 1). A likely explanation consistent with the observed results is that annual energy shares change slowly year-to-year for many states, making persistence a very strong one-step-ahead predictor. In addition, with relatively few annual observations per state, VAR can overfit noise and degrade out-of-sample accuracy.

7.3 Lag Order Figure Interpretation

Figure 1 summarizes the AIC-selected lag orders used for VAR estimation. This distribution helps contextualize model complexity choices across states and supports the observation that lag selection varies meaningfully across the panel.

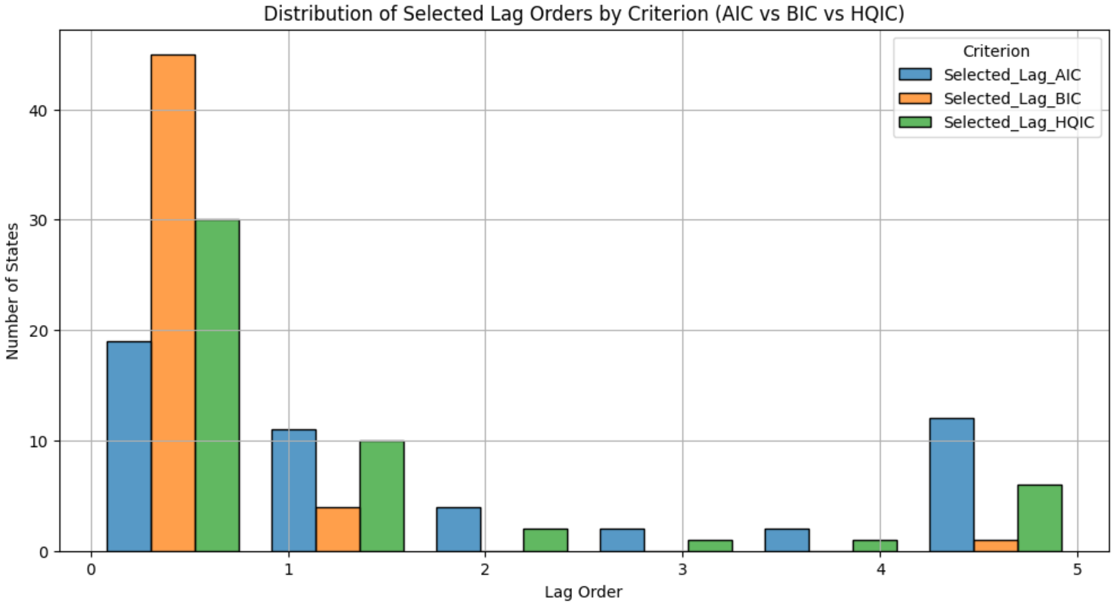

Figure 3: Distribution of selected lag orders under AIC, BIC, and HQIC.

8 Reproducibility

All analysis was performed in a Jupyter notebook (Python). The workflow includes data preprocessing (aggregation into three groups), zero replacement, CLR transformation, per-state VAR fitting with AIC lag selection, and one-step-ahead forecasting for 2023 with MAE evaluation against a persistence baseline.

9 Conclusion

This project demonstrates that while CLR-based methods are theoretically appealing for compositional data, practical forecasting performance and numerical feasibility can favor simpler approaches under real-world data constraints. CLR-VAR encountered frequent estimation failures, while Raw VAR was more robust. However, a naive persistence baseline outperformed VAR forecasts in MAE at a one-year horizon, suggesting that for annual state-level compositions, improvements over persistence may require additional structure (e.g., exogenous covariates) or alternative modeling strategies.

10 Future Work

Future extensions include incorporating exogenous variables, exploring alternative compositional transforms (ILR, ALR), and adopting Bayesian VAR frameworks.

Acknowledgments

I thank my instructor/advisor and course staff for guidance and feedback throughout this project.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman & Hall.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis.* Springer.