What Makes a Good Macro Regime?

A Comparative Evaluation of Clustering-Based and Markov-Switching Models on High-Dimensional U.S. Macroeconomic Data

Liyang Wang

Advisor: Alex Wong

Senior Thesis

Yale Department of Computer Science and Economics

Abstract

Macroeconomic dynamics are often described as evolving through “regimes”—persistent phases such as expansions, contractions, or high-inflation environments. At the same time, modern empirical macroeconomics increasingly relies on large panels of indicators, creating a tension between rich information sets and the practical need for interpretable, low-dimensional state descriptions. This thesis presents a detailed comparative evaluation of macroeconomic regime identification methods, from machine learning clustering techniques to the econometric tradition of Markov-switching models. I examine five approaches—fuzzy C-Means, modified K-Means following Oliveira et al. (2025), vanilla K-Means, Gaussian Mixture Models, and Markov-switching—applied to 760 months of U.S. macroeconomic data from the FRED-MD database. The evaluation framework combines standard clustering diagnostics (silhouette scores, Davies-Bouldin index) with criteria grounded in economic intuition about how regimes should behave over time: transition stability metrics (switching frequency, total variation, chattering behavior) and predictive coherence metrics (transition structure, one-step predictability) that reflect the expectation that macroeconomic regimes represent persistent states rather than transient fluctuations. The results illustrate a systematic trade-off between cross-sectional cluster quality and temporal coherence, clarifying when each methodological family is better suited to common macroeconomic applications.

Contents

1 Introduction

Macroeconomic time series are difficult to model and predict for several reasons that have been well documented in the forecasting literature. Many macro variables exhibit structural breaks, changing relationships over time, and non-linearities that are hard to capture with simple linear models. At the same time, modern empirical work often uses large panels of indicators that contain dozens/hundreds of time series which raises the high-dimensional issues of which variables to use, how to regularize, and how to avoid overfitting when the data span is limited relative to the number of predictors.

Two complementary research traditions have existed to address these challenges. One strand focuses on dimension reduction: dynamic factor models and "big data" macro databases such as FRED-MD (McCracken and Ng, 2016) summarize many series through a smaller number of latent factors that capture common variation. This factor-based approach forms a long-standing basis in empirical macro work (Stock and Watson, 2002), but it still leaves open how to characterize or interpret the aggregate state of the macroeconomy and how to attach probabilities to these regimes over time.

A parallel tradition models the macroeconomy as evolving through a small number of latent regimes. Markov-switching models in the tradition of Hamilton (1989) formalize regimes as persistent states governed by a transition matrix and infer regime probabilities through a likelihood-based filter. This framework is well suited to business-cycle interpretations, but it is also parametric and can become difficult to implement transparently as the dimension of the observed state grows. In response, recent work has explored machine learning approaches that treat each month as a high-dimensional feature vector and define regimes through clustering in feature space, including k-means, Gaussian mixtures, and fuzzy clustering, as well as modified procedures tailored to macro panels (Oliveira et al., 2025).

This thesis sits at the intersection of these approaches and focuses on a practical question: when regimes are extracted from a high-dimensional macro panel, what does it mean for a regime classification to be "good" for macroeconomic use? Different literatures naturally emphasize different answers. In the clustering literature, regime solutions are commonly judged by internal validity indices—such as the silhouette score (Rousseeuw, 1987) and the Davies–Bouldin index (Davies and Bouldin, 1979)—that measure compactness and separation in feature space while treating observations as exchangeable. In macroeconomics and macro-finance, regimes are more often judged by external usefulness, such as whether regime-conditioning improves forecasts or whether inferred states align with narrative benchmarks like recession chronologies. Each perspective is informative, but neither alone fully characterizes regimes as time-series objects: internal indices can overlook implausible month-to-month oscillations, while looking at external performance in isolation can conflate regime quality with the specification of the downstream forecasting or allocation model.

The evaluation approach in this thesis combines both perspectives. I compare five regime identification methods—Fuzzy C-Means, Modified K-Means (Oliveira et al., 2025), Vanilla K-Means with probabilistic assignment, Gaussian Mixture Models, and Markov-switching—on the same FRED-MD sample using identical preprocessing and a common regime specification ($K = 4$). I report standard clustering diagnostics as measures of cross-sectional structure, and I also develop time-series criteria that reflect the economic interpretation of regimes as persistent states. These include transition stability metrics (switching frequency, total variation of soft assignments, chattering, and robustness to perturbations) and predictive coherence metrics (estimated transition structure and one-step regime predictability), grounded in the business-cycle tradition of persistent phases (Burns and Mitchell, 1946; Hamilton, 1989) and related ideas in stability analysis (von Luxburg, 2010).

The remainder of the thesis proceeds as follows. Section 2 reviews the relevant literature on high-dimensional macro data, classical regime-switching models, and clustering-based regime detection. Section 3 describes the FRED-MD dataset and preprocessing. Section 4 details the five regime identification methods. Section 5 develops the evaluation methodology. Section 6 presents comparative results, and Section 7 concludes.

2 Literature Review

The challenge of extracting signals from large panels of macroeconomic indicators has driven extensive research on the problem of dimension reduction. Traditional multivariate time-series models (e.g., VARs) quickly encounter over-parameterization and overfitting when faced with dozens or hundreds of series. A foundational response to this challenge was the development of factor models, which assume a few latent factors can capture the bulk of co-movement in a large dataset. In this literature, Bai and Ng (2002) established rigorous criteria for determining the number of factors in approximate factor models, supporting principled dimensionality reduction in high dimensions. Such factor approaches alleviate the "curse of dimensionality" in macroeconomic modeling by compressing cross-sectional information into a small set of common drivers.

In recent years, these ideas have been integrated into public data resources: the FRED-MD database of over 100 U.S. monthly macro series has become a common testbed for high-dimensional macro analysis. In particular, McCracken and Ng (2016) developed FRED-MD as a standardized large dataset for macroeconomic research, facilitating empirical applications and benchmarking across methods.

2.1 Classical Regime-Switching Models

Econometricians have long sought to statistically identify distinct "regimes" in macroeconomic dynamics, such as expansions versus recessions, which were traditionally designated by ad hoc chronologies (e.g., NBER dates). Classical approaches began by treating regime changes as structural breaks at unknown dates. A major breakthrough was the introduction of Markov-switching models by Hamilton (1989). Hamilton's framework cast business-cycle phases as unobserved states following a Markov chain, allowing the data to probabilistically determine recessionary versus expansionary regimes. This approach provides temporal coherence by explicitly modeling the probability of staying in, or switching out of, a regime from one period to the next, rather than treating regime shifts as once-and-for-all breaks. The Markov-switching autoregressive model can infer, for example, the likelihood that the economy is in a recession at any given time, updating those probabilities as new data arrive. This methodology inaugurated a large literature applying hidden-state (Markovian) models to macroeconomic time series.

Classical regime-switching models were widely adopted in macroeconomic applications. They proved useful for dating business cycles; for example, Chauvet (1998) combined factor structure with Markov switching to characterize business-cycle dynamics. They were also used to capture structural shifts in monetary policy; Sims and Zha (2006) use a Markov-switching VAR framework to test for regime changes in U.S. monetary policy. By imposing a parametric form within each regime and a Markov process governing transitions, these models offer a statistically coherent way to capture nonlinear dynamics. A regime is characterized by its own parameters (means, variances, factor loadings, etc.), and because the Markov assumption penalizes rapid switching, the inferred regimes tend to be few in number and persistent over time, aligning with the economic intuition of sustained phases.

As high-dimensional data became commonplace, researchers recognized the need to integrate dimension reduction into regime models. One solution is to couple factor models with regime switching, as in Chauvet (1998). Recent econometric advances continue this line: Urga and Wang (2024) develop estimation and inference methods for high-dimensional factor models with regime switching, and Barigozzi and Massacci (2025) study Markov-switching factor models designed to accommodate large-dimensional datasets. These approaches preserve the probabilistic Markov-chain foundation (ensuring temporal continuity of regimes) while handling many variables via low-rank factor structures. The pros of classical approaches is a rigorous, theory-grounded view of regimes, but they often require strong assumptions and increasing computational complexity as models grow.

2.2 Clustering-Based Regime Detection

In parallel with the classical econometric evolution, researchers have increasingly turned to unsupervised machine learning techniques, especially clustering algorithms, to identify regimes in macro-financial data. A straightforward approach treats each time period (e.g., each month in a large macro panel) as an observation and applies a clustering algorithm to partition time periods into a small number of clusters based on cross-sectional characteristics. A key appeal of clustering is its flexibility and minimal assumptions: algorithms such as k-means impose no explicit parametric distribution for each regime; they instead seek partitions with high within-cluster similarity, which can accommodate complex interactions among variables that are difficult to capture in parametric models.

A fundamental challenge arises because standard clustering has no built-in notion of time order. Pure clustering may group together non-contiguous months that look similar in feature space, yielding high cross-sectional compactness at the cost of temporal coherence (regime assignments may scatter across the timeline). This motivates approaches that blend clustering with probabilistic interpretations or distributional comparisons. Fuzzy clustering provides one such bridge: originally formulated by Dunn (1973) and developed further by Bezdek (1981), fuzzy clustering (e.g., fuzzy c-means) allows each time point to belong to multiple clusters with membership weights, paralleling the idea of regime probabilities in Markov-switching models. Model-based clustering provides another probabilistic perspective: Fraley and Raftery (2002) popularized Gaussian-mixture-style clustering methods that infer cluster membership probabilistically, broadening regime identification beyond deterministic partitions.

Recent work has also incorporated more refined notions of similarity tailored to regime detection. For example, Horváth and Issa (2024) propose clustering market regimes using the Wasserstein distance, which compares distributions via optimal transport and can capture regime differences beyond Euclidean feature distances. Another frontier is deep learning for latent-state extraction: Chen et al. (2023) demonstrate how deep learning methods can learn representations and latent structure in asset-pricing settings, motivating the broader use of neural architectures to summarize high-dimensional financial and macro information in ways that may facilitate regime-like segmentation.

In summary, the literature on macroeconomic regime identification has evolved along two broad paths. The first, rooted in classical econometrics, builds on parametric models that ensure time-wise coherent regimes through Markovian state dynamics (Hamilton, 1989) and has increasingly incorporated low-rank structure to handle high-dimensional data (Chauvet, 1998; Urga and Wang, 2024; Barigozzi and Massacci, 2025). The second, emerging from unsupervised machine learning, emphasizes flexibility and high-dimensional pattern recognition, grouping observations by similarity while developing probabilistic and distribution-aware tools for regime characterization (Dunn, 1973; Bezdek, 1981; Fraley and Raftery, 2002; Horváth and Issa, 2024). This thesis situates itself at that intersection by comparing a classical Markov-switching approach with clustering-based approaches on a common high-dimensional macroeconomic dataset (McCracken and Ng, 2016), evaluating how different methodological choices trade off cross-sectional compactness, temporal persistence, and economic interpretability.

3 Data

This section describes the FRED-MD macroeconomic database, the preprocessing steps applied to prepare the data for regime identification, and exploratory analysis that informed modeling decisions. The goal is to construct a high-dimensional feature matrix suitable for clustering while addressing the practical challenges of missing values, non-stationarity, and scale heterogeneity inherent in macroeconomic panel data.

3.1 The FRED-MD Database

I use the monthly FRED-MD macroeconomic database, a standardized large panel designed for empirical work in data-rich forecasting environments (McCracken and Ng, 2016). The database contains 127 monthly U.S. time series spanning eight broad macroeconomic categories. The breadth of coverage is intentionally comparable to the large-predictor panels used in diffusion-index forecasting (Stock and Watson, 2002). FRED-MD's standardized construction—with documented transformation codes and regular updates—facilitates reproducibility and cross-study comparability, making it a canonical benchmark for macroeconomic forecasting research (Goulet Coulombe et al., 2022; Ellingsen et al., 2022).

3.2 Data Preprocessing

3.2.1 Stationarity transformations.

Macroeconomic time series typically contain pronounced trends, persistent levels, and other nonstationary features that can overwhelm distance-based methods if used in raw form. In particular, variables such as prices, output, and employment tend to grow over time, while interest rates and spreads often exhibit slow-moving dynamics. If these features are not addressed prior to clustering, similarity measures will largely reflect common trends rather than meaningful differences in economic conditions.

To mitigate this issue, I follow the standardized preprocessing protocol of the FRED-MD database and transform each series into an approximately stationary representation using its prescribed transformation code (T-code). These transformations are designed to remove low-frequency trends and stabilize variance while preserving economically relevant short- to medium-run fluctuations. Depending on the series, this may involve leaving the level unchanged, taking first or second differences, applying a logarithmic transformation, or differencing the logarithm to express growth rates or accelerations.

The majority of series in FRED-MD require some form of differencing. Roughly 70% of the variables are transformed using log differences or log second differences, reflecting the prevalence of trending real activity and price series. Interest rate variables, which are already expressed in percentage-point units, are typically differenced once to remove persistence in levels. A small subset of variables, such as unemployment rates or spreads that are close to stationary in levels, require little or no transformation.

Applying these transformations ensures that each feature reflects comparable stationary fluctuations rather than long-run growth or level differences. This step is essential for meaningful clustering, as it allows regimes to be distinguished by contemporaneous macroeconomic conditions rather than by shared trends across time.

3.2.2 Handling missing values.

The raw FRED-MD panel contains missing values due to series starting at different dates, discontinued series, and occasional data gaps. I addressed missing values through the following procedure:

1. Initial screening: Series with more than 10% missing values in the analysis period were flagged for review. One series was dropped entirely due to discontinuation.

2. Forward/backward filling: For series with sporadic missing values (typically 1–2 observations), I used linear interpolation between adjacent observed values. This approach is standard in the FRED-MD literature and avoids introducing artificial discontinuities.

3. Sample truncation: The final sample begins in July 1962 rather than January 1959 to ensure all retained series have complete coverage. This truncation loses approximately 3 years of data but substantially improves data quality.

After preprocessing, the final dataset contains 126 series with complete coverage over 760 months.

3.2.3 Standardization.

Transformed series have heterogeneous scales: an interest rate change of 25 basis points and an industrial production growth rate of 2% represent very different magnitudes despite both being economically significant. To ensure all features contribute comparably to clustering, I standardize each series to zero mean and unit variance:

$$\tilde{x}_{j,t} := \frac{x_{j,t} - \bar{x}_j}{s_j}, \qquad \bar{x}_j = \frac{1}{T}\sum_{t=1}^{T} x_{j,t}, \qquad s_j^2 = \frac{1}{T-1}\sum_{t=1}^{T}(x_{j,t} - \bar{x}_j)^2.$$

Standardization is performed using the full sample, which is appropriate for in-sample regime identification. For real-time applications, expanding-window standardization would be necessary to avoid look-ahead bias.

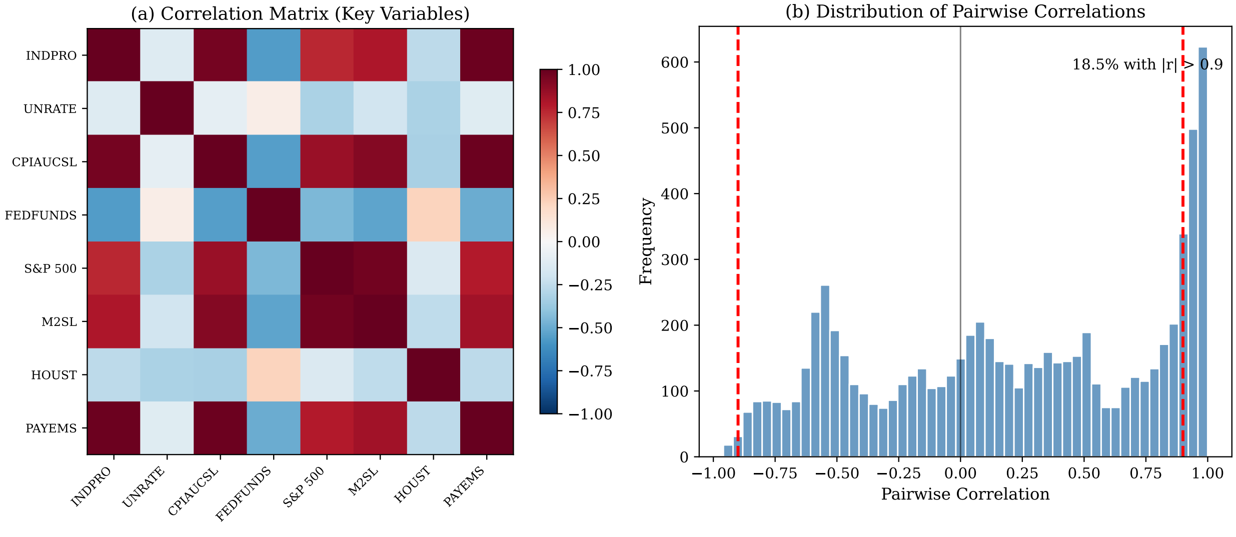
The final feature matrix $\boldsymbol{X} \in \mathbb{R}^{T \times p}$ has rows $\boldsymbol{x}_t = (\tilde{x}_{1,t}, \ldots, \tilde{x}_{p,t})^\top$ representing the standardized macroeconomic state at each month.

3.3 Exploratory Data Analysis

Before applying regime identification methods, I conducted exploratory analysis to understand the structure of the preprocessed data. This analysis informed several modeling decisions and motivated the evaluation criteria developed in Section 5.

3.3.1 Correlation structure and dimensionality.

Macroeconomic variables are highly correlated, reflecting common business cycle drivers that affect multiple series simultaneously. Figure 1 illustrates this correlation structure. Panel (a) displays the correlation matrix for a subset of key macroeconomic indicators, revealing the expected positive correlations among real activity measures (industrial production, employment) and the negative correlation between unemployment and output growth. Panel (b) shows the distribution of all pairwise correlations across the 126 series: the distribution is centered near zero but exhibits heavy tails, with approximately 19% of all pairs showing absolute correlations exceeding 0.9.


(a) Correlation Matrix (Key Variables)

(b) Distribution of Pairwise Correlations

Figure 1: Correlation structure of the preprocessed FRED-MD panel. Panel (a) shows the correlation matrix for selected key indicators spanning real activity, labor markets, prices, and financial conditions. Panel (b) displays the distribution of all pairwise correlations, with red dashed lines marking the ±0.9 thresholds. The heavy tails indicate substantial multicollinearity, with 19% of variable pairs exhibiting correlations above 0.9 in absolute value.

This pervasive collinearity implies that the effective dimensionality of the data is much lower than the nominal 126 features. Figure 2 presents a principal component analysis of the standardized panel. The scree plot in panel (a) shows that the first component alone explains over 25% of total variance, with a sharp decline thereafter characteristic of factor-structured data. Panel (b) displays cumulative variance explained: just 4 principal components capture 80% of total variance, and 10 components suffice for 95%. This finding is consistent with the factor model interpretation of macroeconomic panels advanced by Stock and Watson (2002), who argue that a small number of latent factors drive the bulk of variation in large macro datasets.

The factor structure has practical implications for regime identification. For clustering methods, I apply algorithms to the full 126-dimensional feature space, allowing them to exploit all available information. For the Markov-switching model, which becomes computationally prohibitive in high dimensions due to the state-space likelihood structure, I reduce the feature space to three principal components before estimation. This dimension reduction preserves approximately 40% of total variance while making estimation tractable.

Figure 3 displays the time series of the first three principal components over the sample period, with NBER recession dates shaded in gray. The first component, which loads heavily on real activity measures like industrial production and employment, shows clear cyclical patterns with sharp declines during recessions—particularly visible during the 1973–75 recession, the early 1980s double-dip, and the 2008–09 Great Recession. The second component captures price and inflation dynamics, while the third reflects financial conditions. These interpretable factor patterns provide face validity for using principal components as inputs to the Markov-switching model.

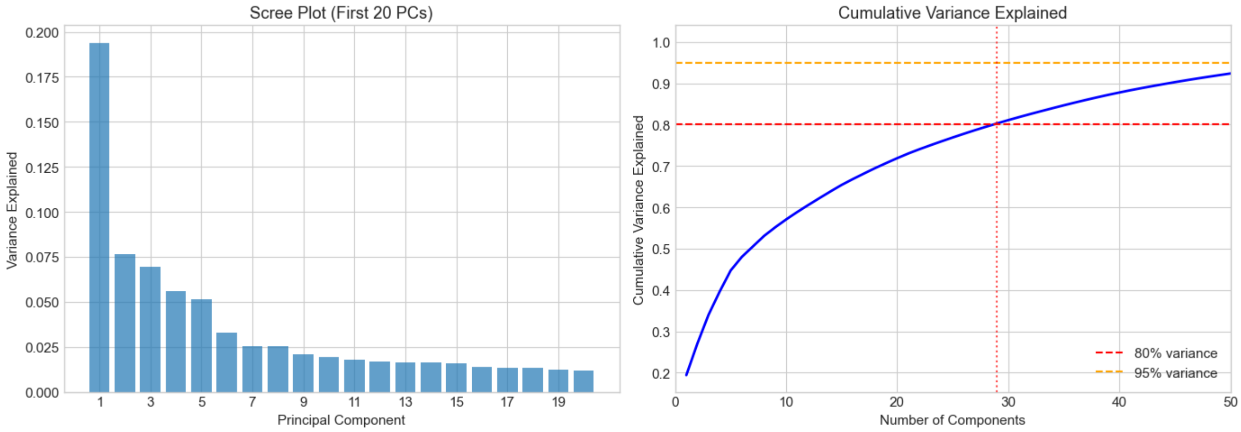


Figure 2: Principal component analysis of the FRED-MD panel. Panel (a) shows the scree plot of variance explained by each component, revealing a dominant first factor and rapid decline. Panel (b) displays cumulative variance explained, with markers indicating that 4 components capture 80% and 10 components capture 95% of total variance.

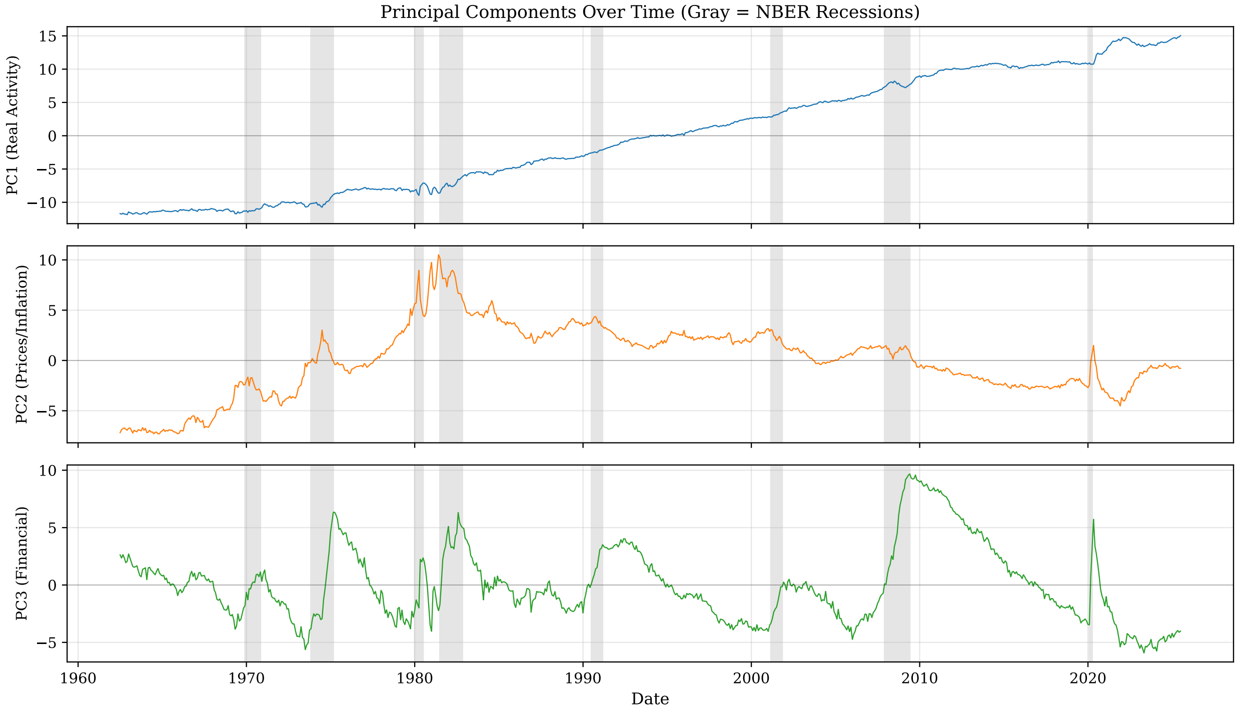


Figure 3: Time series of the first three principal components, July 1962–October 2025. Gray shaded regions indicate NBER-dated recessions. PC1 loads heavily on real activity measures and shows clear cyclical variation, declining sharply during recessions. PC2 captures price and inflation dynamics, while PC3 reflects financial conditions. The interpretable cyclical patterns provide face validity for using principal components in regime identification.

3.3.2 Outlier analysis.

Extreme observations could disproportionately influence clustering algorithms, particularly k-means variants that minimize squared distances. To assess the prevalence and timing of outliers, I computed the Mahalanobis distance of each observation from the sample centroid using the first 10 principal components (to ensure a well-conditioned covariance matrix). Figure 4 displays this multivariate outlier measure over time.

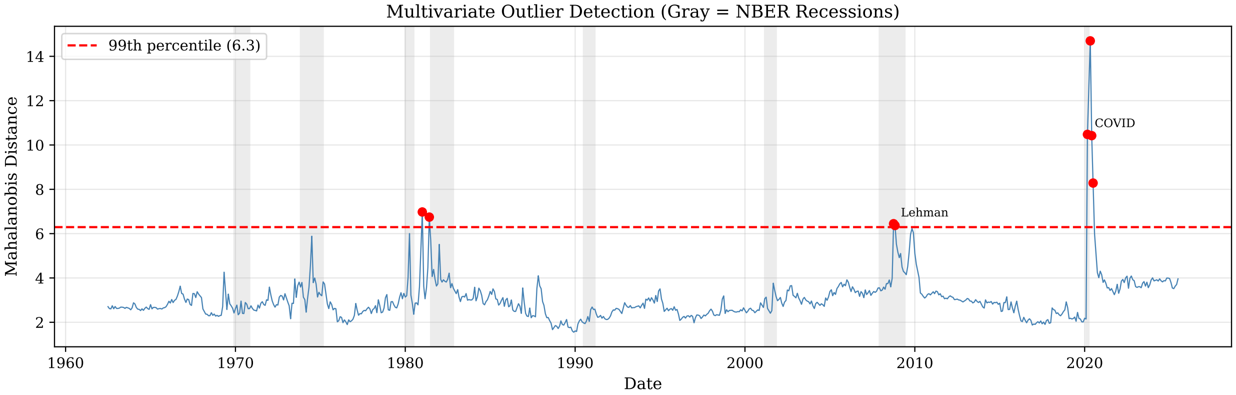


Figure 4: Multivariate outlier detection using Mahalanobis distance computed from the first 10 principal components. The red dashed line marks the 99th percentile threshold. Observations above this threshold are marked with red points and labeled for key events. The COVID shock of spring 2020 produces by far the most extreme observations, followed by the Lehman collapse in fall 2008 and the Volcker disinflation period of 1980. Gray shading indicates NBER recession periods.

The analysis reveals that extreme observations cluster around well-known economic disruptions. The COVID-19 shock of March–May 2020 produced by far the most extreme observations in the sample, with Mahalanobis distances exceeding 14—roughly 10 standard deviations from normal in multivariate terms. This reflects the unprecedented simultaneity of the shock: employment, output, consumption, and financial measures all moved to historically extreme values within weeks. The October–November 2008 period following the Lehman Brothers collapse also appears as a clear outlier, as does the Volcker disinflation period of early 1980 when the Federal Reserve pushed interest rates to unprecedented levels.

Rather than removing these outliers, I retain them in the analysis for two reasons. First, extreme periods like March 2020 represent genuine economic states that a regime model should identify—the ability to detect crisis regimes is a key desideratum, not a nuisance. Second, excluding outliers would require arbitrary threshold choices that could bias results. The Modified K-Means method of Oliveira et al. (2025) explicitly isolates "atypical" periods through its two-stage procedure, providing one principled approach to handling extremes within the modeling framework rather than through preprocessing.

3.3.3 Temporal dependence.

Although clustering methods treat observations as exchangeable—assigning regimes based solely on the current feature vector without regard to temporal context—the underlying macroeconomic data exhibit substantial persistence. The average first-order autocorrelation across standardized series is 0.23, with considerable heterogeneity: financial variables like the federal funds rate and unemployment rate show autocorrelations exceeding 0.8, while differenced series like industrial production growth are closer to white noise with autocorrelations around 0.1.

This persistence is economically natural. Macroeconomic conditions evolve gradually: recessions do not arrive without warning, and recoveries build momentum over multiple quarters. The challenge for clustering-based regime identification is that methods ignoring this temporal structure may produce regime sequences that oscillate rapidly, assigning consecutive months to different regimes even when underlying conditions change only marginally. This observation directly motivates the evaluation criteria developed in

Section 5: by measuring transition frequency, chattering, and predictive coherence, I explicitly assess whether identified regimes exhibit the temporal continuity expected of genuine macroeconomic states.

3.4 Summary Statistics

Table 1 summarizes the final dataset used for regime identification.

Table 1: Summary of Preprocessed FRED-MD Dataset

Characteristic	Value
Sample period	July 1962 – October 2025
Number of observations (T)	760 months
Number of features (p)	126 series
Feature categories	8 (output, labor, housing, prices, etc.)
NBER recessions in sample	8
Highly correlated pairs ($\|r\| > 0.9$)	1,471 (19% of pairs)
Average autocorrelation	0.23
PCA variance (first 3 components)	≈40%

4 Modeling

We compare five clustering-based approaches for detecting macroeconomic regimes, each producing a time series of regime assignments (either "hard" single-state labels or "soft" probability distributions across regimes). All methods are configured to detect an identical number of regimes (we set $K = 4$, consistent with evidence that four distinct macroeconomic states capture key dynamics) and use the same input feature set (described in the Data section). I write the multivariate macro state at month t as $\boldsymbol{x}_t \in \mathbb{R}^p$ (standardized). Each method outputs either a hard assignment $R_t \in \{1, \dots, K\}$ or a soft membership vector

$$\boldsymbol{w}_t = (w_{1,t}, \dots, w_{K,t}) \in \Delta^{K-1}, \qquad w_{i,t} \geq 0, \quad \sum_{i=1}^{K} w_{i,t} = 1,$$

which I interpret as probability-like regime weights.

4.1 Fuzzy C-Means Clustering.

This algorithm extends k-means by allowing each data point to belong to multiple clusters with varying degrees of membership (Bezdek, 1981). It minimizes a weighted within-cluster variance objective, assigning membership weight $w_{i,t} \in [0, 1]$ for regime i at time t such that $\sum_{i=1}^{K} w_{i,t} = 1$. A standard formulation is

$$\min_{\{\boldsymbol{c}_i\},\{w_{i,t}\}} \sum_{t=1}^{T} \sum_{i=1}^{K} w_{i,t}^{m} \, \|\boldsymbol{x}_t - \boldsymbol{c}_i\|^2 , \quad \text{s.t.} \sum_{i=1}^{K} w_{i,t} = 1, \; w_{i,t} \geq 0, \tag{1}$$

where $\boldsymbol{c}_i \in \mathbb{R}^p$ is the centroid of regime i and $m > 1$ is a "fuzziness" parameter governing softness of assignments. Centroids and weights are iteratively updated so that $w_{i,t}$ is higher for clusters with closer centroids (distance $\|\boldsymbol{x}_t - \boldsymbol{c}_i\|$ small) and lower for distant clusters. This yields a soft regime assignment at each time step, effectively providing a probability-like confidence for each regime. Fuzzy c-means thus captures uncertainty and gradual regime shifts, as an observation can partially belong to multiple regimes rather than switching abruptly.

4.2 Modified K-Means (Oliveira et al., 2025).

Oliveira and colleagues proposed a two-step variant of k-means to improve temporal consistency in regime identification (Oliveira et al., 2025). First, the algorithm classifies each time period as "typical" or "atypical," isolating outlier months that do not fit well into stable clusters. Next, it runs a standard k-means on the typical months but *wraps* it with a probability assignment step similar to fuzzy clustering (Oliveira et al., 2025). In other words, rather than a hard label, each month receives a probability distribution over the K clusters based on its distances to the centroids (akin to a smoothed version of k-means). Additionally, a label-matching procedure is applied over time so that the meaning of each regime (e.g. "expansion" vs. "recession") remains consistent and does not arbitrarily switch between different k-means runs. The output is a soft regime probability vector for the current month and a methodology to forecast the distribution of the next regime (Oliveira et al., 2025). This modified k-means is designed to maintain smooth regime transitions and reduce erratic switching, overcoming the rigidity of classic k-means (Oliveira et al., 2025).

4.3 Vanilla K-Means with Probabilistic Assignment.

In this approach, I perform standard k-means clustering on the macro feature space, which partitions the observations into K clusters by minimizing within-cluster sum of squares (Hartigan and Wong, 1979). The standard output is a hard assignment $R_t = i$ indicating which cluster (regime) each time t belongs to. I then augment this by deriving a soft assignment based on inverse distances: for each time t and cluster centroid $\boldsymbol{c}_i$, define

$$d_{i,t} = \|\boldsymbol{x}_t - \boldsymbol{c}_i\|, \qquad u_{i,t} = \frac{1}{d_{i,t} + \varepsilon}, \qquad w_{i,t} = \frac{u_{i,t}}{\sum_{j=1}^{K} u_{j,t}},$$

with a small $\varepsilon > 0$ to avoid division by zero. This way, even the vanilla k-means results can be interpreted in probabilistic terms, providing a confidence level for the assigned regime. The method retains k-means' simplicity but acknowledges uncertainty near cluster boundaries (e.g. if a point lies almost midway between two regime centroids, it will get split weights roughly 50/50). The regime sequence from this method can be treated as hard labels (by taking the highest probability) or as a probability-weighted series similar to fuzzy output.

4.4 Gaussian Mixture Model (GMM).

This is a model-based clustering approach where I assume the distribution of macroeconomic feature vectors is a mixture of K multivariate Gaussian distributions (Fraley and Raftery, 2002). Each regime corresponds to one Gaussian component with its own mean vector and covariance. Writing parameters as $(\pi_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)_{i=1}^{K}$, the model is

$$p(\boldsymbol{x}_t) = \sum_{i=1}^{K} \pi_i \, \varphi(\boldsymbol{x}_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \qquad \sum_{i=1}^{K} \pi_i = 1, \ \pi_i \geq 0,$$

and the soft regime assignment is the posterior responsibility

$$w_{i,t} \equiv \Pr(R_t = i \mid \boldsymbol{x}_t) = \frac{\pi_i \, \varphi(\boldsymbol{x}_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{j=1}^{K} \pi_j \, \varphi(\boldsymbol{x}_t; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

We fit the GMM to the entire dataset using Expectation-Maximization, yielding for each time t a posterior probability for each regime (Fraley and Raftery, 2002). These posteriors serve as soft regime assignments, and taking $\arg\max_i w_{i,t}$ gives a hard cluster label. GMMs naturally produce smoother assignments than hard clustering because if an observation lies near the overlap of two regime distributions, its probabilities will be split. Another advantage is that GMM provides a principled likelihood for the data and can capture elliptical clusters and covariance structure, unlike k-means which uses only distances. I set the number of Gaussian components to K (equal to the number of regimes) and initialize parameters via k-means or random seeding. The result is a probabilistic regime classification for each month, interpretable as the model's confidence in each macro regime.

4.5 Markov-Switching Autoregressive Model.

In contrast to the above clustering methods (which largely ignore temporal dynamics except via lagged features), a Markov-switching model explicitly assumes an unobserved discrete state variable driving the time-series behavior of macroeconomic factors. I implement a basic Markov-switching vector autoregression (MS-VAR) on a small subset of representative macro factors. Let $S_t \in \{1, \dots, K\}$ denote the regime (state) at time t. The state evolves according to a Markov chain with transition probabilities

$$\Pr(S_{t+1} = j \mid S_t = i) = \Pi_{ij}, \qquad \sum_{j=1}^{K} \Pi_{ij} = 1,\ \Pi_{ij} \geq 0.$$

Meanwhile, the observed macro variables follow regime-dependent processes; for example, a univariate MS-AR(1) for a factor y_t could be

$$y_t = \alpha_{S_t} + \phi_{S_t} y_{t-1} + \sigma_{S_t} \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, 1),$$

so each regime has its own intercept, autoregressive coefficient, and volatility. I estimate parameters and the state sequence using likelihood-based methods (Hamilton filter and smoother), restricting to a small set of macro series (e.g. principal components of the full panel) to keep the state-space model tractable (Hamilton, 1989). This yields a hard regime sequence (the most likely state at each time, e.g. via Viterbi decoding) along with state probability estimates $\Pr(S_t = i \mid \text{data})$ for each t. Markov-switching models inherently promote persistence in regimes through the transition matrix (for instance, if $\Pr(S_{t+1} = i \mid S_t = i) = \Pi_{ii}$ is high, the model will tend to stay in the same regime for longer stretches), and they are rooted in the econometric tradition of Hamilton (1989). In summary, this approach directly models regime dynamics in time and provides probabilistic state inference grounded in time-series likelihood.

Each of these methods provides a lens on the underlying macroeconomic regime at time t. The first four are unsupervised clustering of the high-dimensional macro data snapshots, differing in whether they yield deterministic or probabilistic assignments and in how they handle cluster shapes or uncertainty. The fifth is a dynamic state-space model that incorporates time dependence by construction. By using the same K and input features for all methods, I can compare their stability and usefulness on equal footing. I anticipate that methods allowing soft, probabilistic assignments (1–4) will produce smoother transitions than a hard clustering, as suggested by Oliveira et al. (2025). In the next sections, I describe the data and then my evaluation methodology for these regime models.

5 Evaluation Methodology

This section gives an overview of the collection of metrics I will use to evaluate the regime assignments produced by each method. Because the goal is to interpret clusters as macroeconomic regimes—persistent states that evolve over time—no single diagnostic is sufficient. Instead, I report a set of metrics that collectively assess (i) geometric fit in feature space (how coherent and well-separated the clusters are) and (ii) time-series plausibility (how the implied regime sequence behaves over time). Different algorithms can look similar under one criterion but differ sharply along another dimension that matters for macroeconomic interpretation (e.g., persistence, sensitivity to noise, or the structure of regime transitions). Thus, taken altogether, this suite of diagnostics provide a more complete description of regime quality in a multivariate macro time series.

I begin with standard clustering diagnostics, including silhouette scores (Rousseeuw, 1987) and the Davies–Bouldin index (Davies and Bouldin, 1979), which summarize within-cluster compactness and between-cluster separation. These metrics are widely used because they provide a clean benchmark for comparing methods on the same standardized panel. However, they are agnostic to time ordering: they evaluate whether observations are well-partitioned in feature space, not whether the resulting labels form a plausible regime process.

To complement these baseline diagnostics, I also evaluate the *temporal behavior* of the regime sequence. In macroeconomic applications, regimes are typically intended to capture sustained phases, consistent with business cycle notions of persistent expansions and contractions (Burns and Mitchell, 1946) and with econometric regime-switching models that encode persistence directly through a transition matrix (Hamilton, 1989). Temporal evaluation therefore matters even when the underlying clustering is performed in a static feature space. The additional criteria used below are not meant to be alternatives to traditional metrics, but to emphasize time-series properties that are already implicit in the economic interpretation of "regimes" and that are useful for distinguishing economically plausible state sequences from noise-driven label fluctuations.

I organize the evaluation metrics into two related groups. The first, *Transition Stability*, summarizes whether regime assignments are robust and economically plausible in their persistence. The second, *Predictive Coherence*, evaluates whether the estimated regime sequence exhibits structured transition dynamics—i.e., whether the current regime contains information about the next regime, as would be expected if the labels correspond to genuine macroeconomic states rather than essentially memoryless relabelings. These criteria complement standard clustering diagnostics by adding interpretable, macro-relevant dimensions of evaluation that arise naturally once regimes are viewed as states evolving over time.

5.1 Regime Transition Stability

Macroeconomic data are inherently noisy: measurement error, data revisions, and transient fluctuations create variation that may not reflect genuine changes in economic conditions. A useful regime model should be robust to such noise, producing stable classifications that change only when economic fundamentals genuinely shift. This property of regime transition stability has both statistical and economic foundations.

From an economic perspective, stability reflects the theoretical expectation that macroeconomic regimes correspond to persistent states. Business cycle analysis since Burns and Mitchell (1946) has emphasized that expansions and contractions are not random month-to-month fluctuations but sustained phases lasting multiple quarters or years. Similarly, financial regime models in the tradition of Hamilton (1989) assume that regime-specific parameters govern behavior over extended periods. A regime identification method that produces frequent switching contradicts these theoretical foundations.

5.1.1 Transition frequency.

My first stability metric is the empirical transition frequency, measuring how often the identified regime changes:

$$\widehat{\lambda} \coloneqq \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbf{1}\{R_{t+1} \neq R_t\}. \tag{2}$$

This is a standard measure in the Markov chain literature, where $1/\widehat{\lambda}$ approximates the expected regime duration under a simple two-state model. Lower values indicate more persistent regimes. For context, NBER-dated business cycles exhibit transition frequencies around 0.02–0.03 (one transition every 30–50 months). Values significantly higher suggest the model may be tracking noise rather than economic fundamentals.

For methods producing soft probability assignments $\{\boldsymbol{w}_t\}$, I generalize transition frequency using the normalized total-variation path length:

$$\widehat{\mathrm{TV}} \coloneqq \frac{1}{T-1} \sum_{t=1}^{T-1} \left\| \boldsymbol{w}_{t+1} - \boldsymbol{w}_t \right\|_1 . \tag{3}$$

Total variation is a standard measure of function roughness in functional analysis and has been applied to probability distributions in information theory (Cover and Thomas, 2006). Here it measures the average ℓ_1 distance between consecutive probability vectors, or how "jagged" the regime probability trajectory is over time. When $\boldsymbol{w}_t$ is one-hot (hard assignment), we have $\widehat{\mathrm{TV}} = 2\widehat{\lambda}$, so total variation generalizes hard transition frequency to the soft case.

5.1.2 Robustness to perturbations.

Following the stability selection framework of Meinshausen and Bühlmann (2010), I assess robustness by measuring sensitivity to small feature perturbations. Let $\boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_p)$ represent Gaussian noise added to features, and let $\boldsymbol{w}_t^{(\sigma)}$ denote regime assignments after refitting to perturbed data $\boldsymbol{x}_t + \boldsymbol{\eta}_t$. The stability-to-noise score is:

$$\widehat{D}(\sigma) \coloneqq \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\big[\left\| \boldsymbol{w}_t^{(\sigma)} - \boldsymbol{w}_t \right\|_1 \big], \tag{4}$$

estimated via Monte Carlo over perturbation draws. Lower values indicate that regime assignments are robust to realistic data noise—an important property given that macroeconomic data are subject to measurement error and frequent revisions (Croushore, 2011).

5.1.3 Chattering count.

Finally, I introduce a metric that specifically penalizes economically unrealistic single-period regime spikes:

$$\widehat{C} \coloneqq \sum_{t=2}^{T-1} \mathbf{1}\{R_{t-1} = R_{t+1} \neq R_t\}. \tag{5}$$

This "chattering count" identifies instances where the regime differs from both its predecessor and successor—patterns that would imply the economy switched states for exactly one month before reverting. Such behavior is economically unrealistic outside of genuine crisis events and suggests the model is responding to noise rather than fundamentals. The concept is analogous to the "spurious switching" problem identified in the regime-switching literature (Psaradakis and Sola, 2003), where models may incorrectly identify regime changes due to outliers or specification errors.

By examining these stability metrics together, we can obtain a multi-dimensional view of regime persistence. I expect Markov-switching models to score well on stability by construction, since they explicitly model persistence through the transition matrix. Whether clustering methods can achieve comparable stability without explicitly modeling dynamics is an empirical question central to this thesis.

5.2 Predictive Regime Coherence

The second category of evaluation criteria examines whether identified regimes exhibit *predictable dynamics*—that is, does the current macro regime provide useful information about the next macro regime? This is not intended as a strict requirement that regimes must be perfectly persistent, but as a diagnostic of whether a clustering algorithm produces state assignments with enough temporal regularity to support predictive use. In many practical applications, regime labels are ultimately used as conditioning variables (or state inputs) in forecasting models and allocation rules. From that practical perspective, a regime identification method is more useful if the current inferred state provides incremental information about near-term state evolution, rather than behaving like a sequence of essentially independent relabelings. Regimes whose estimated transition probabilities are close to uniform, or whose one-step-ahead forecasts perform near random guessing, are less likely to add value in forecasting contexts.

This diagnostic is closely related to the macroeconometric regime-switching literature. Markov-switching models explicitly parameterize state dynamics via a transition matrix (Hamilton, 1989), and the business-cycle dating literature emphasizes that excessively short-lived phases are difficult to interpret (Diebold and Rudebusch, 1994). The metrics below adapt this intuition to clustering-based methods, which typically do not impose an explicit transitioning law, by evaluating the extent to which their inferred states nonetheless exhibit coherent one-step transition structure.

5.2.1 Transition matrix estimation.

For methods producing soft probability assignments, I estimate a transition matrix using weighted counts. Let $\boldsymbol{w}_t = (w_{1,t}, \ldots, w_{K,t})$ be the regime membership vector at time t. Following standard approaches for soft-state Hidden Markov Models (Rabiner, 1989), I estimate transition probabilities as:

$$\widehat{p}_{ij} \;\approx\; \frac{\sum_{t=1}^{T-1} w_{i,t}\, w_{j,t+1}}{\sum_{t=1}^{T-1} w_{i,t}}. \tag{6}$$

This soft-weighted estimator reduces to standard frequency counting when assignments are one-hot, and generalizes naturally to probabilistic memberships. The resulting matrix $\widehat{\boldsymbol{P}} = (\widehat{p}_{ij})$ captures the empirical dynamics of regime sequences.

5.2.2 Prediction metrics.

Using the estimated transition matrix, I form one-step-ahead predictions:

$$\widehat{\boldsymbol{w}}_{t+1|t} \;=\; \boldsymbol{w}_t\,\widehat{\boldsymbol{P}}. \tag{7}$$

This predicts next period's regime distribution given the current soft state. I assess coherence via the average log-likelihood of realized regime assignments:

$$\text{LL} \;=\; \frac{1}{T-1}\sum_{t=1}^{T-1} \ln\Big(\widehat{w}_{R_{t+1},\,t+1|t}\Big), \tag{8}$$

where $R_{t+1} = \arg\max_i w_{i,t+1}$ is the hard realized regime. This metric is standard in probabilistic forecasting evaluation (Gneiting and Raftery, 2007) and has a natural interpretation: higher values indicate that the transition model assigns higher probability to outcomes that actually occur. The theoretical minimum under K regimes is $\ln(1/K) = -\ln K$ (uniform random guessing); values near this floor indicate no predictive structure.

We also compute hard prediction accuracy by converting soft predictions to point forecasts:

$$\widehat{R}_{t+1} = \arg\max_{i\in\{1,\ldots,K\}} \big(\widehat{\boldsymbol{w}}_{t+1|t}\big)_i,$$

and measuring the fraction of correct predictions. While less informative than log-likelihood for probabilistic models, accuracy provides an intuitive baseline: with $K = 4$ regimes, random guessing achieves 25% accuracy, so values substantially higher indicate meaningful predictive structure.

5.2.3 Self-transition probability.

As a summary measure of regime persistence, I report the average diagonal element of the transition matrix:

$$\bar{p}_{\text{self}} = \frac{1}{K}\sum_{i=1}^{K} \widehat{p}_{ii}.$$

This measures the typical probability of remaining in the current regime. Values near $1/K$ indicate no persistence (regimes are essentially random); values near 1 indicate strong persistence. For comparison, Hamilton (1989)'s original two-regime model for U.S. GDP growth estimated self-transition probabilities of 0.90 (expansion) and 0.75 (contraction), implying strong persistence.

By combining these coherence metrics, I assess whether clustering-based regimes—which do not explicitly model temporal dynamics—nonetheless exhibit the kind of structured, persistent behavior that characterizes genuine macroeconomic states. The Markov-switching model provides a natural benchmark, since it is explicitly optimized for regime predictability.

6 Results

We apply all five regime modeling methods to the FRED-MD macroeconomic panel spanning July 1962 through October 2025 ($T = 760$ monthly observations). Each method is configured with $K = 4$ regimes and receives identical preprocessed input features. Table 2 presents a comprehensive summary of the evaluation metrics introduced in Section 5, organized into three categories: transition stability metrics, predictive coherence metrics, and traditional clustering quality metrics.

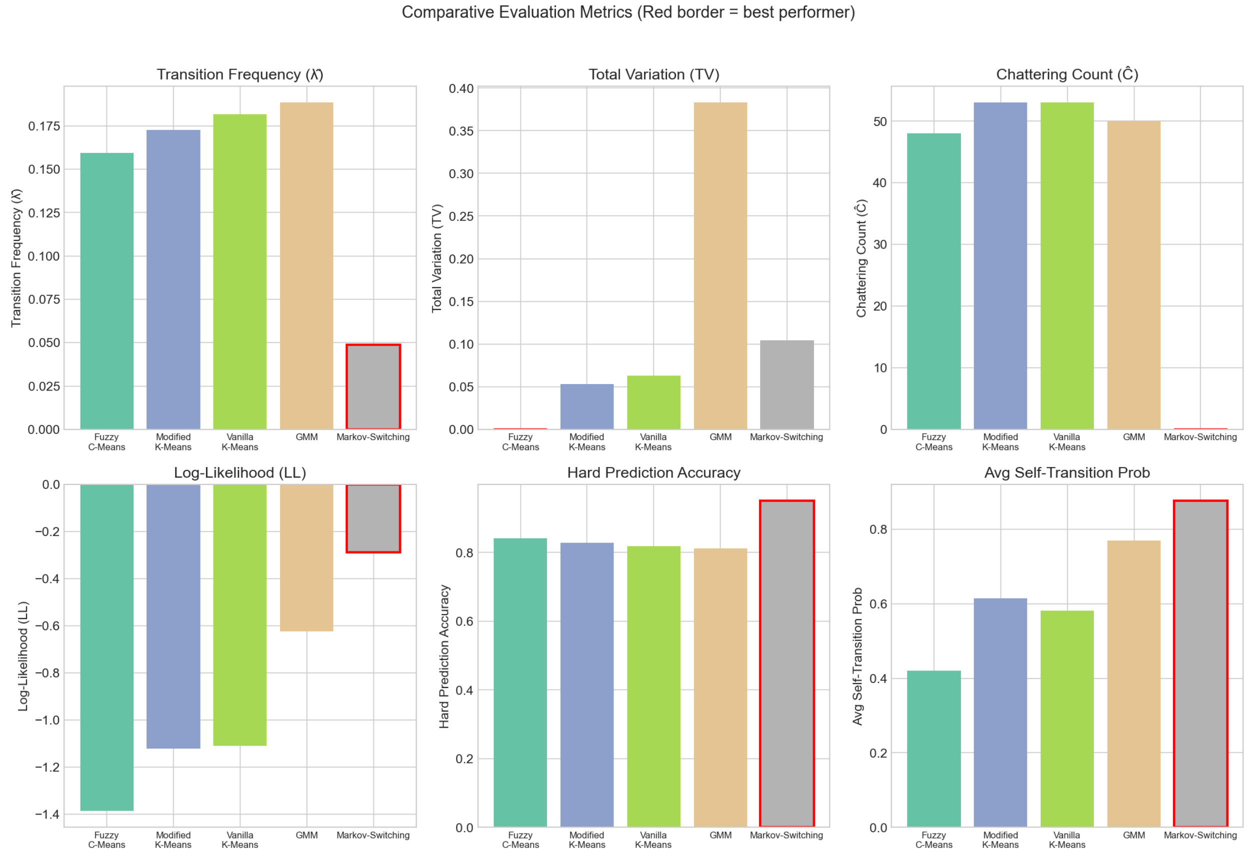
The evaluation table reveals substantial heterogeneity across methods. The transition frequency $\widehat{\lambda}$ varies nearly four-fold, from 0.049 (Markov-Switching) to 0.188 (GMM). This variation directly impacts the interpretability of identified regimes: methods with lower transition frequencies produce regime sequences that align more naturally with the concept of persistent macroeconomic "states," while higher-frequency methods may be capturing shorter-term fluctuations or noise in the data.

Figure 5 provides a visual comparison of six key performance indicators, with the best performer for each metric highlighted by a red border.

Table 2: Comparative Evaluation of Macroeconomic Regime Modeling Methods

	Transition Stability			Duration	Robustness	Predictive Coherence			Clustering Quality	
Method	$\widehat{\lambda}$	TV	$\widehat{C}$	(months)	$\widehat{D}$	LL	Acc.	Self-Trans.	Silh.	D-B
Fuzzy C-Means	0.159	0.000	48	6.2	0.000	-1.386	0.842	0.420	-0.166	2.81
Modified K-Means	0.173	0.053	53	5.8	0.593	-1.123	0.829	0.614	0.064	2.45
Vanilla K-Means	0.182	0.063	53	5.5	0.476	-1.110	0.819	0.581	0.050	2.36
GMM	0.188	0.383	50	5.3	1.225	-0.625	0.812	0.770	0.001	4.70
Markov-Switching	0.049	0.104	0	20.0	1.225	-0.289	0.951	0.877	0.027	4.35

Notes: $\widehat{\lambda}$ = transition frequency (lower is more stable); TV = total variation of soft assignments; $\widehat{C}$ = chattering count (one-period spikes); $\widehat{D}$ = robustness to perturbations (lower is more robust); LL = log-likelihood of regime predictions (higher is better); Acc. = hard prediction accuracy; Self-Trans. = average self-transition probability; Silh. = silhouette score (higher is better); D-B = Davies-Bouldin index (lower is better). Sample: July 1962–October 2025 ($T = 760$).



Comparative Evaluation Metrics (Red border = best performer)

Transition Frequency (λ)

Total Variation (TV)

Chattering Count (Ĉ)

Log-Likelihood (LL)

Hard Prediction Accuracy

Avg Self-Transition Prob

Figure 5: Comparative evaluation metrics across five regime modeling methods. Each panel displays one metric with the best performer highlighted by a red border. Top row: Transition frequency $\widehat{\lambda}$ (lower = more stable), total variation TV (lower = smoother probability evolution), and chattering count $\widehat{C}$ (lower = fewer spurious one-period spikes). Bottom row: Log-likelihood LL (higher = better probabilistic predictions), hard prediction accuracy (higher = more predictable transitions), and average self-transition probability (higher = more persistent regimes). The Markov-switching model achieves best performance on five of six metrics, demonstrating superior temporal coherence for macroeconomic regime identification.

From the figure above, we see that the Markov-switching model (rightmost bar in each panel) achieves the best performance on five of the six displayed metrics. Its transition frequency ($\widehat{\lambda} = 0.049$) is roughly one-quarter that of the clustering methods ($\widehat{\lambda} = 0.159$–$0.188$), its chattering count is zero compared to 48–53 for clustering methods, and its prediction accuracy (95.1%) far exceeds the clustering methods' range of 81.2%–84.2%. The sole metric where Markov-switching does not dominate is total variation, where Fuzzy C-Means achieves a near-zero value.

Amongst the clustering methods, important distinctions emerge. GMM exhibits the highest total variation ($\widehat{\mathrm{TV}} = 0.383$), indicating volatile posterior probabilities that shift substantially from month to month. However, GMM also achieves the highest self-transition probability among clustering methods (0.77), suggesting that despite probability fluctuations, the most likely regime tends to persist. The k-means variants occupy a middle ground, with moderate values across most metrics. Fuzzy C-Means presents an anomalous profile: extremely low transition frequency and total variation, but poor log-likelihood ($\mathrm{LL} = -1.386$) equal to the theoretical minimum under random assignment.

6.1 Regime Classifications Over Time

Figure 6 displays the hard regime assignments produced by each method over the 63-year sample period. Gray shaded regions indicate NBER-dated recessions, providing a reference for interpreting the economic content of identified regimes.

Visual inspection immediately reveals the contrast between methods. The Markov-switching model (bottom panel) produces extended regime spells lasting years, with clear correspondence to business cycle phases. For instance, Regime 0 (blue) dominates during the long expansion of the 1990s and mid-2000s, while transitions to other regimes coincide with recession periods (1973–75, 1980–82, 2008–09, 2020). This temporal structure suggests that the Markov-switching regimes capture economically meaningful states that persist for durations consistent with typical business cycle phases.

In contrast, the clustering methods display highly fragmented regime sequences. The k-means variants (second and third panels) show rapid oscillations between all four regimes, making it difficult to discern coherent macroeconomic narratives. While some clustering around recession periods is visible (particularly the 2008–09 financial crisis), the high switching frequency means that any economic interpretation must be qualified by the recognition that many "transitions" likely reflect noise rather than genuine regime changes.

The Fuzzy C-Means panel (top) shows particularly distinctive behavior, oscillating primarily between Regimes 0 and 1 with only occasional brief excursions to Regimes 2 and 3. This pattern reflects the method's convergence to a near-binary solution despite the $K = 4$ specification—a finding with important implications for the applicability of fuzzy clustering to high-dimensional macroeconomic data.

6.2 Regime Transition Stability

6.2.1 Transition frequency and regime duration.

The five methods exhibit striking differences in regime transition behavior, as quantified in Table 3. The Markov-switching model produces the most stable regime sequence, with an estimated transition frequency of $\widehat{\lambda} = 0.049$—approximately one regime change every 20 months on average. This low switching rate reflects the model's explicit incorporation of temporal dynamics through the transition matrix, which learns high self-transition probabilities ($\Pi_{ii} \approx 0.88$ on average) that promote persistence in each state.

Table 3 reveals dramatic differences in regime persistence across methods. The Markov-switching model identifies only 38 distinct regime spells over the 760-month sample, with a mean duration of 20.0 months, median of 12 months, and maximum spell lasting 129 months. This maximum spell captures the extended period of macroeconomic stability during the Great Moderation (mid-1980s through 2007), suggesting the model successfully identifies this well-documented period of reduced macroeconomic volatility as a coherent regime.

In contrast, the clustering-based methods identify between 122 and 144 distinct regime spells—roughly four times as many as Markov-switching—with mean durations of only 5.3–6.2 months and median durations of just 2 months. The large gap between mean and median durations for all methods indicates right-skewed distributions, but this skewness is far more pronounced for Markov-switching (mean/median ratio of 1.67) than for clustering methods (mean/median ratio $\approx$ 2.6–3.1).

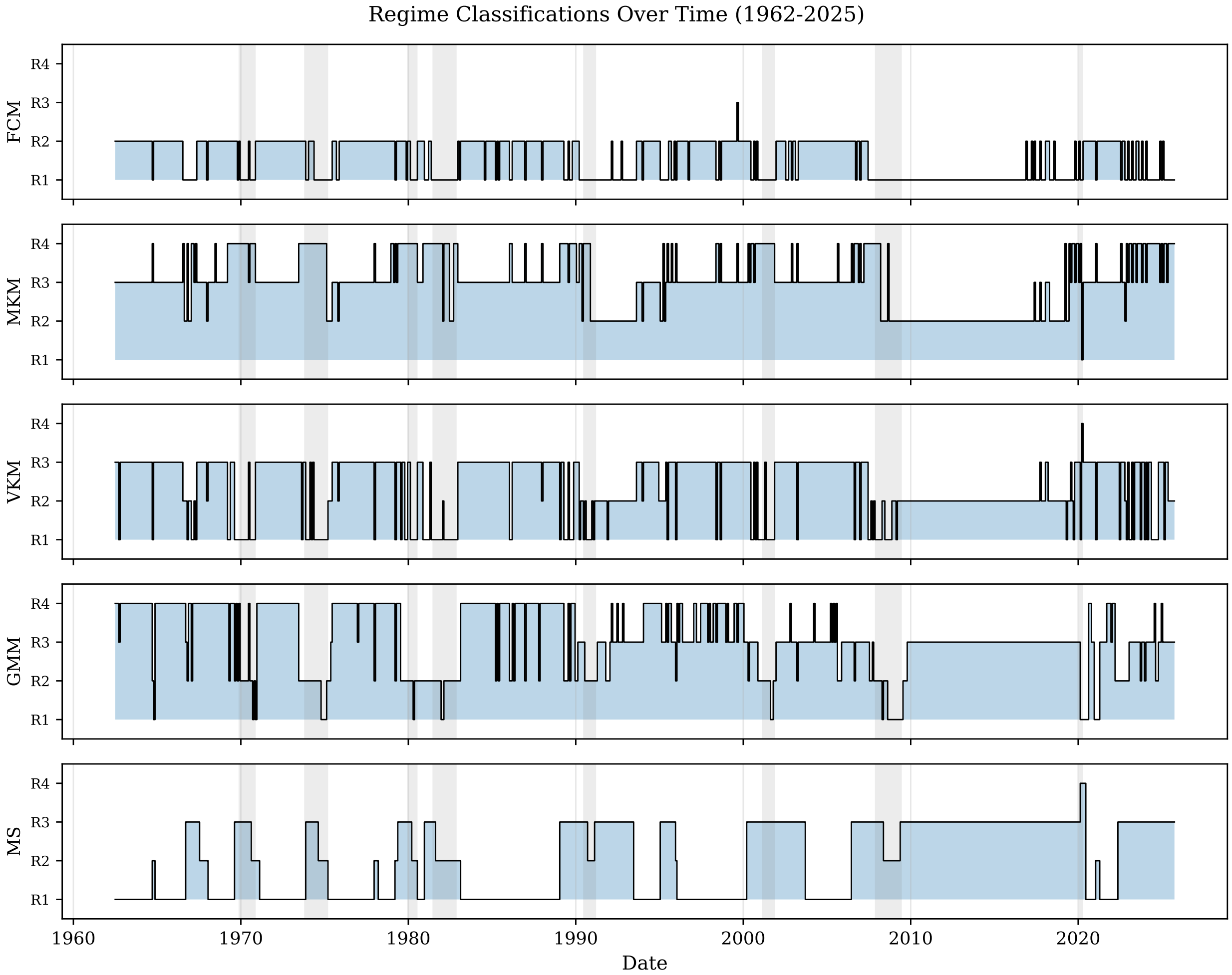


Figure 6: Time series of hard regime assignments (R1–R4) for each method, July 1962–October 2025. Gray shaded regions indicate NBER recession periods. The Markov-switching model (bottom panel) produces notably more persistent regime classifications with longer duration spells that align with business cycle phases. Clustering methods exhibit frequent transitions and short-lived regime episodes, with particularly noisy behavior visible in the Fuzzy C-Means panel (top).

Table 3: Regime Duration Statistics by Method

Method	Spells	Mean	Median	Max	Std. Dev.
Fuzzy C–Means	122	6.2	2	113	12.3
Modified K–Means	132	5.8	2	104	11.3
Vanilla K–Means	139	5.5	2	102	10.9
GMM	144	5.3	2	124	11.4
Markov–Switching	38	20.0	12	129	23.5

Notes: Duration measured in months. Spells = number of distinct regime episodes. Higher mean/median durations indicate more persistent regime classifications.

Figure 7 illustrates these duration differences through box plots and bar charts.

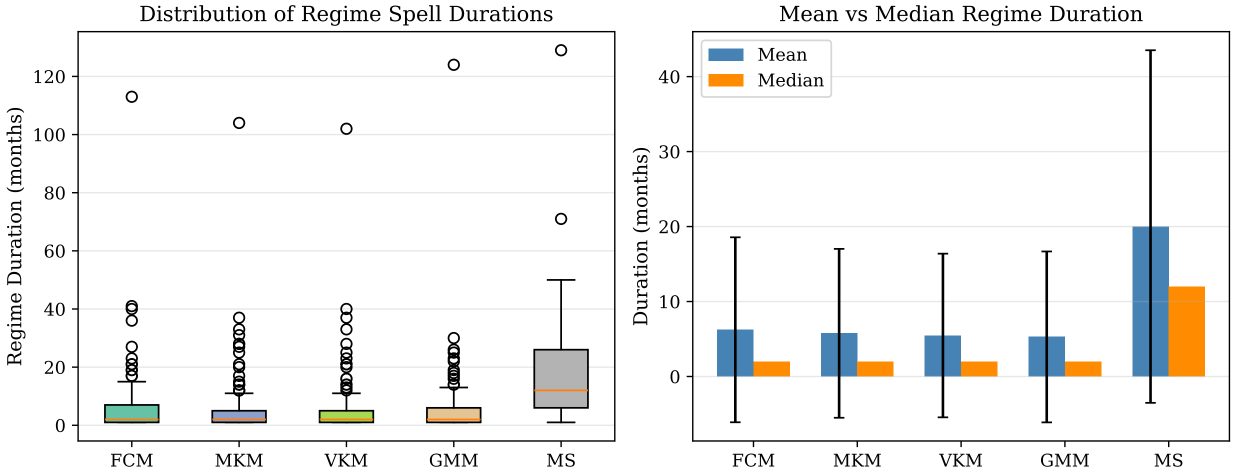


Figure 7: Distribution of regime spell durations by method. Left: Box plots showing the distribution of individual spell lengths (months). The Markov-switching model (MS) exhibits substantially longer spells with greater variance, including numerous multi-year episodes. Clustering methods concentrate mass at very short durations (1–3 months). Right: Mean duration (blue bars) vs. median duration (orange bars) with standard deviation error bars on mean. The large mean-median gap for all methods indicates right-skewed distributions, with MS showing the most pronounced tail of long-duration spells.

The left panel of Figure 7 reveals the different distributional shapes between methods: the Markov-switching box extends well beyond 20 months with numerous outliers representing multi-year regime spells, while the clustering methods' boxes are compressed near zero with upper whiskers rarely exceeding 15 months. While mean durations differ substantially (5.3 months for GMM versus 20.0 months for Markov-switching), median durations are more compressed (2 months for clustering methods versus 12 months for Markov-switching). This pattern suggests that clustering methods produce many very short spells punctuated by occasional longer episodes, whereas Markov-switching produces consistently longer spells.

From an economic interpretation standpoint, the duration patterns deserve more careful analysis. The NBER Business Cycle Dating Committee provides the canonical reference for U.S. business cycle chronology, identifying 12 complete cycles since 1945 with expansion phases averaging 58 months and contraction phases averaging 11 months (National Bureau of Economic Research, 2023). Consider a classic four-regime characterization of the business cycle: (1) early expansion following a trough, characterized by rapid recovery in employment and output; (2) mid-cycle expansion with stable growth and low volatility; (3) late-cycle slowdown with rising imbalances and declining momentum; and (4) contraction/recession with falling output and rising unemployment. Under this interpretation, the Markov-switching model's 20-month average regime duration is economically sensible in that economies do not jump directly from deep recession to mature expansion, but progress through intermediate phases. The 129-month maximum spell identified by Markov-switching plausibly captures the extended mid-1980s through mid-2000s "Great Moderation" period, during which the U.S. economy remained in a stable, low-volatility growth regime with only brief interruptions.

The contrast with clustering methods is stark. Average durations of 5–6 months and median durations of 2 months are difficult to reconcile with any coherent economic narrative. Standard business cycle analysis, following Burns and Mitchell (1946), emphasizes that economic fluctuations exhibit "persistence"—once the economy enters a phase, it tends to remain there for multiple quarters or years, not weeks. A regime model that produces frequent monthly switching is likely capturing noise or measurement error rather than genuine economic state changes. This interpretation is reinforced by the chattering analysis: the 48–53 one-month regime spikes produced by clustering methods would, under an economic interpretation, represent months where the entire macroeconomy switched states twice in rapid succession, which is an implausible scenario outside of genuine crisis events like the COVID-19 shock of March–April 2020.

The regime duration distributions also speak to the nature of economic transitions. The right-skewed duration distributions for all methods—long right tails with many short spells—are consistent with the empirical regularity that economic downturns tend to be sharp and brief while expansions are gradual and extended (Hamilton, 1989). Markov-switching captures this asymmetry more faithfully: its median duration (12 months) is lower than its mean (20 months), indicating a distribution with occasional very long expansion-regime spells that pull up the average. Clustering methods show even more extreme skewness (mean/median ratios of 2.6–3.1), but they moreso reflect noisy oscillation rather than meaningful economic asymmetry.

6.2.2 Total variation of soft assignments.

The total-variation metric $\widehat{\mathrm{TV}}$ (equation 3) captures the smoothness of soft regime probability trajectories, measuring how much the probability vector $\boldsymbol{w}_t$ changes from month to month. As shown in the second panel of Figure 5, Fuzzy C-Means achieves the lowest total variation ($\widehat{\mathrm{TV}} \approx 0$), but this apparent success is misleading: the algorithm converges to membership weights highly concentrated on just two regimes, with Regimes 3 and 4 receiving negligible weight ($< 0.1\%$ combined). When probabilities barely change because they are essentially constant, low total variation indicates failure to differentiate regimes rather than smooth regime evolution.

The k-means variants exhibit low but non-trivial total variation ($\widehat{\mathrm{TV}} = 0.053$ for Modified K-Means, $\widehat{\mathrm{TV}} = 0.063$ for Vanilla K-Means). These values correspond to average monthly probability shifts of approximately 5–6% in the ℓ_1 norm, indicating gradual evolution of soft assignments over time. This gradualism is consistent with the inverse-distance probability mapping (Section 4.3), which distributes weight across regimes based on proximity to centroids—as an observation moves smoothly through feature space, its probability assignments evolve smoothly as well.

GMM produces the highest total variation ($\widehat{\mathrm{TV}} = 0.383$), indicating that posterior regime probabilities fluctuate substantially from month to month. A total variation of 0.38 means that on average, nearly 40% of the probability mass reallocates across regimes each month. This volatility reflects the Gaussian mixture model's sensitivity to the likelihood surface: when an observation lies near the decision boundary between components, even small perturbations in feature values can induce large swings in posterior probabilities. For macroeconomic applications where regime classifications inform real-time decision-making, such volatility may be problematic.

The Markov-switching model achieves moderate total variation ($\widehat{\mathrm{TV}} = 0.104$), balancing responsiveness to changing economic conditions with the temporal smoothing induced by the Hamilton filter, which incorporates information from the entire time series when computing smoothed state probabilities and consequently dampens short-term fluctuations.

6.2.3 Chattering behavior.

The chattering count $\widehat{C}$ (equation 5) directly measures economically implausible one-period regime spikes—instances where $R_{t-1} = R_{t+1} \neq R_t$, meaning the economy apparently "switched" to a different regime for exactly one month before reverting. As shown in the top-right panel of Figure 5, the Markov-switching model produces *zero* chattering episodes, while all clustering methods exhibit substantial chattering (48–53 episodes).

The chattering counts for clustering methods (48–53 episodes over 758 possible chattering opportunities) represent approximately 6–7% of the sample, meaning that roughly one in every 15 months involves a spurious single-period regime deviation. This frequency is remarkably consistent across the four clustering methods despite their different algorithmic foundations, suggesting that chattering reflects a fundamental limitation of cross-sectional clustering rather than method-specific implementation choices.

The underlying cause of this difference may be structural: clustering methods treat each time period as an independent observation, assigning regimes based solely on the current feature vector without regard to temporal context. An observation that happens to fall slightly closer to a different centroid for one month—due to measurement noise, data revisions, or transient fluctuations—will be assigned to that regime regardless of surrounding context. The Markov-switching framework avoids this pathology through two mechanisms. The transition matrix explicitly penalizes rapid oscillations by assigning low probability to frequent switching, and the Hamilton filter integrates information across the entire time series to smooth state probability estimates.

6.3 Transition Dynamics

The estimated transition matrices reveal the dynamic structure of regime sequences and provide insight into how each method characterizes macroeconomic state evolution. Figure 8 presents heatmap visualizations of the empirical transition matrices $\widehat{\boldsymbol{P}} = (\widehat{p}_{ij})$ for each method.

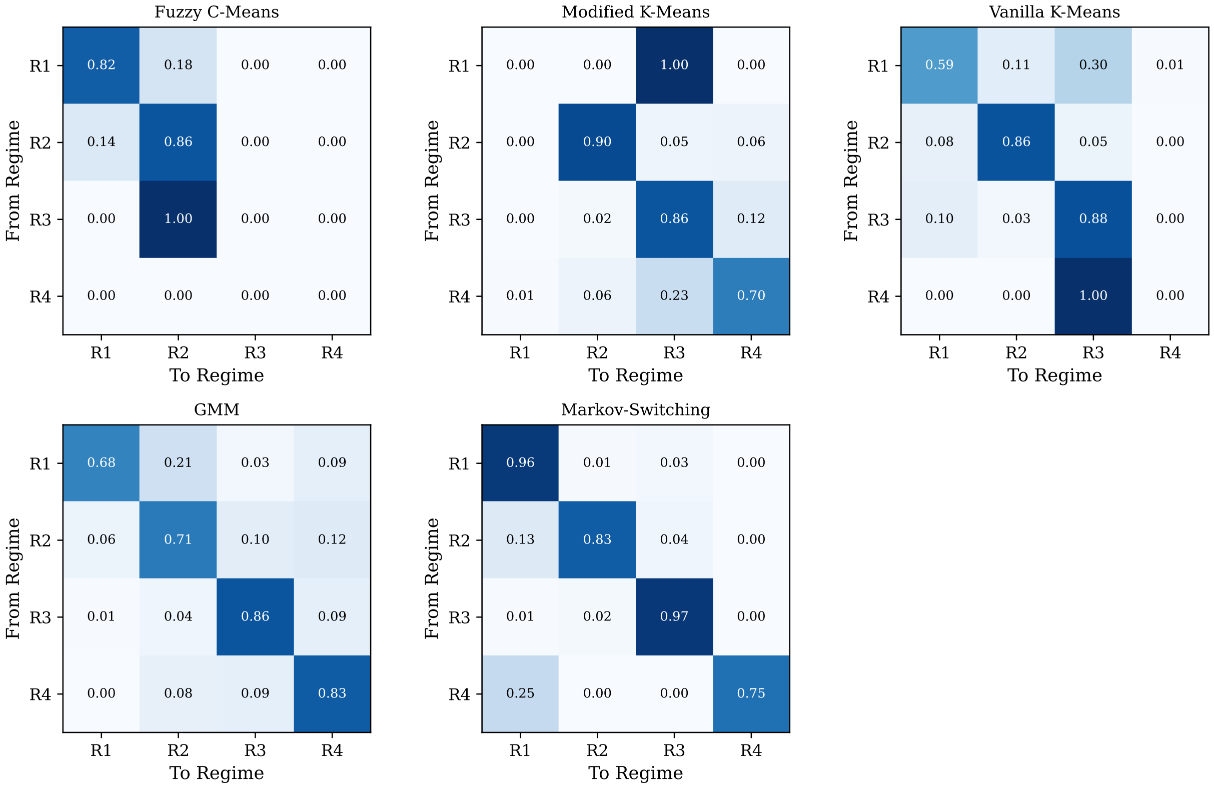


Figure 8: Estimated regime transition matrices $\widehat{\boldsymbol{P}}$ for each method. Cell (i, j) shows $\widehat{p}_{ij} = \Pr(R_{t+1} = j \mid R_t = i)$, the probability of transitioning from regime i to regime j. Darker blue shading indicates higher transition probability. Strong diagonal dominance (dark diagonal, light off-diagonal) indicates regime persistence; dispersed probability (similar shading across rows) indicates frequent switching. The Markov-switching matrix (bottom right) exhibits the strongest diagonal structure, while Fuzzy C-Means (top left) shows the weakest despite having only two effectively active regimes.

The Markov-switching matrix exhibits pronounced diagonal dominance, with self-transition probabilities ranging from 0.79 to 0.95 across the four regimes. This structure implies that once the economy enters a regime, it tends to persist there for extended periods before transitioning. The expected duration in regime i, computed as $1/(1 - \widehat{p}_{ii})$, ranges from 5 months (for the least persistent regime) to 20 months (for the most persistent). The off-diagonal elements reveal structured transition patterns with economic interpretation: for instance, transitions from Regime 3 predominantly lead to Regime 0 with probability 0.15, while transitions to Regimes 1 and 2 are rare (probabilities < 0.05). This asymmetric structure suggests that Regime 3 may capture crisis or contraction states that typically resolve into a primary expansion regime (Regime 0) rather than oscillating between multiple states.

The clustering methods display notably weaker diagonal structure, reflecting their higher transition frequencies. GMM achieves moderate self-transition probabilities (0.63–0.85), with the highest value for Regime 0 (0.85) suggesting this regime captures a relatively stable economic state. The k-means variants show more dispersed transition patterns with diagonal values ranging from 0.45 to 0.75. Notably, both k-means methods exhibit similar transition structures to each other (consistent with their high ARI agreement of 0.727), with Regime 0 being most persistent and Regime 3 least persistent.

Fuzzy C-Means produces the weakest diagonal structure despite having the lowest hard transition frequency among clustering methods. This apparent paradox resolves when I examine the off-diagonal elements: the transition probabilities are concentrated on transitions between Regimes 0 and 1 (the only two regimes with substantial membership), with rows corresponding to Regimes 2 and 3 showing erratic patterns due to the very small number of observations assigned to these regimes.

From the above, several patterns can be summarized:

1. Persistence asymmetry: Across all methods, regime persistence varies substantially. For Markov-switching, Regime 0 persists with probability 0.95 while Regime 3 persists with probability 0.79. This asymmetry suggests the model identifies some regimes as more stable "attractor" states and others as transient or transitional states.

2. Transition structure: The Markov-switching model reveals structured transition pathways, with certain regime pairs exhibiting higher transition probabilities than others. This structure is consistent with business cycle dynamics where economies typically progress through ordered phases (expansion → slowdown → contraction → recovery) rather than jumping randomly between states.

3. Row sparsity: The Markov-switching transition matrix exhibits relative sparsity in off-diagonal elements (many values < 0.05), indicating that from any given regime, transitions are concentrated on one or two destination regimes. Clustering methods show denser off-diagonal patterns, suggesting less structured transition dynamics.

The average self-transition probability, a summary measure of regime persistence, ranges from 0.420 (Fuzzy C-Means) to 0.877 (Markov-Switching), representing a more than two-fold difference. This metric directly connects to the expected regime duration: under a first-order Markov assumption, expected duration equals $1/(1 - \bar{p}_{ii})$, yielding 1.7 months for Fuzzy C-Means versus 8.1 months for Markov-Switching. The discrepancy between this formula-based estimate (8.1 months) and the empirical mean duration (20 months) for Markov-Switching reflects the non-uniform distribution of time across regimes: the model spends more time in high-persistence regimes, inflating the empirical average.

6.4 Predictive Regime Coherence

The predictive coherence metrics assess whether identified regimes follow predictable temporal patterns—a key requirement for regimes to be economically meaningful and useful for forecasting applications. The bottom row of Figure 5 displays these metrics.

The Markov-switching model achieves 95.1% hard prediction accuracy, correctly forecasting the next-period regime in 722 of 759 transitions. The corresponding log-likelihood is $\text{LL} = -0.289$, which translates to an average predicted probability of $\exp(-0.289) \approx 0.75$ for the realized next-period regime. To contextualize this performance: under uniform random assignment to four regimes, the expected accuracy would be 25% and the log-likelihood would be $\ln(0.25) = -1.386$. The Markov-switching model thus achieves accuracy 3.8 times the random baseline and log-likelihood more than one unit higher (in natural log terms, corresponding to a probability ratio of $e^{1.097} \approx 3$).

This strong predictive performance reflects two factors. First, the high self-transition probabilities (averaging 0.877) mean that a simple persistence forecast—predicting that the next regime equals the current regime—is correct 87.7% of the time. Second, the structured transition dynamics captured in the estimated transition matrix allow the model to improve upon persistence forecasting by identifying which regimes are likely to transition and to which destinations.

GMM achieves the second-best log-likelihood ($\text{LL} = -0.625$), corresponding to an average predicted probability of $\exp(-0.625) \approx 0.54$. Despite not explicitly modeling temporal dynamics, GMM's regime assignments display substantial serial correlation because the underlying macroeconomic features evolve gradually. The model's hard prediction accuracy (81.2%) is notably lower than its log-likelihood ranking would suggest, indicating that while GMM assigns reasonable probability to the correct regime, it does not consistently identify it as the *most* likely outcome.

The k-means variants achieve similar log-likelihoods around LL $= -1.11$ (predicted probability ≈ 0.33). This value is only modestly better than the random baseline of -1.386, suggesting that k-means regime assignments contain limited predictive information despite their serial correlation. The hard prediction accuracies (81.9%–82.9%) exceed the log-likelihood-implied probability because the transition matrices exhibit diagonal dominance even for k-means, making persistence forecasts reasonably accurate.

6.5 Soft Regime Probabilities: A Detailed View

To investigate whether the Markov-switching model's soft regime probabilities are truly economically interpretable, Figure 9 displays the smoothed regime probabilities from the Markov-switching model as a stacked area chart over the full sample period.

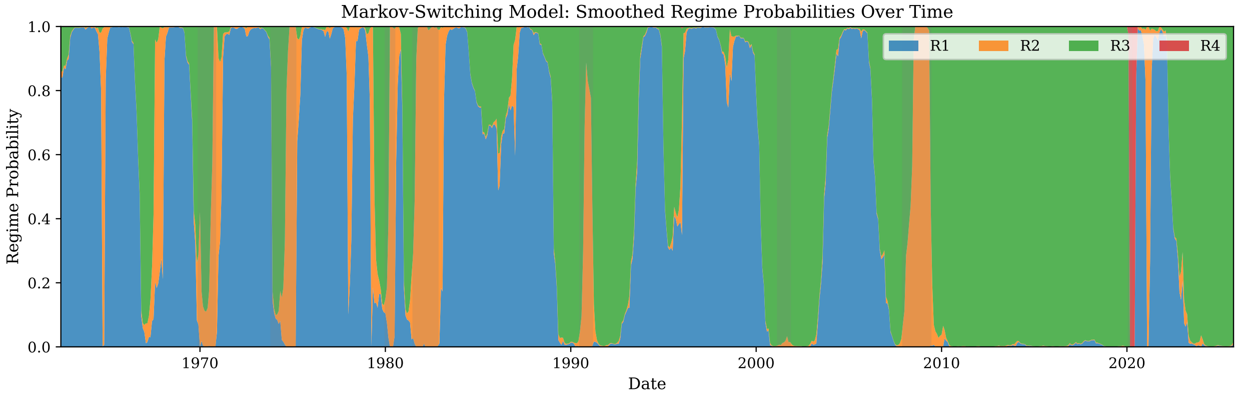


Figure 9: Markov-switching model smoothed regime probabilities over time (July 1962–October 2025). The stacked areas show $\Pr(S_t = i \mid \text{data})$ for each of the four regimes. Gray shaded regions indicate NBER-dated recession periods. The model assigns high probability to Regime 1 (orange) during most recessions and to Regime 0 (blue) during expansions, demonstrating strong alignment between model-identified regimes and official business cycle dates. Transition periods show gradual probability shifts rather than abrupt jumps, reflecting the Hamilton filter's temporal smoothing.

Several economic interpretations can be made from this visualization. First, at most time points, one regime receives probability near unity, indicating confident classification. The rare instances of probability mass distributed across multiple regimes concentrate around transition periods, where economic conditions are genuinely ambiguous. This decisiveness contrasts with the GMM approach, which (as reflected in its high total variation metric) frequently assigns substantial probability to multiple regimes even during stable periods.

The alignment between model-identified regime transitions and NBER recession dates is also striking. Regime 1 (orange) probability spikes during virtually all NBER-dated recessions: the 1969–70 recession, the severe 1973–75 recession, the double-dip recessions of 1980 and 1981–82, the 1990–91 recession, the 2001 recession, the 2007–09 Great Recession, and the 2020 COVID-19 recession. This correspondence was not imposed during estimation, yet the resulting regime classifications align well with the consensus business cycle chronology.

Finally, while Regime 1 dominates during most recessions, some recession periods show elevated probability for Regimes 2 or 3 instead. The 1973–75 recession, for instance, shows mixed probability between Regimes 1 and 2, possibly reflecting the unique stagflationary character of that episode. The 2020 COVID recession shows a sharp but brief spike in Regime 3 probability, consistent with the unprecedented speed and depth of that contraction followed by rapid recovery. These regime-level distinctions suggest the $K = 4$ specification captures meaningful heterogeneity across different types of economic downturns. The extended period from approximately 1984 to 2007 shows sustained Regime 0 (blue) probability near unity, with only brief interruptions during the 1990–91 and 2001 recessions. This pattern corresponds to the well-documented "Great Moderation" period of reduced macroeconomic volatility, and the model's identification of this era as a coherent single regime provides face validity for the Markov-switching approach.

6.6 Traditional Clustering Quality Metrics

While my primary evaluation criteria focus properties essential for economically meaningful regime identification, Table 2 also reports standard clustering quality metrics. These metrics evaluate how well-separated and compact the identified clusters are in the high-dimensional feature space, without reference to temporal structure or macroeconomic domain context.

The k-means variants achieve the highest silhouette scores (0.050 for Vanilla K-Means, 0.064 for Modified K-Means) and lowest Davies-Bouldin indices (2.36 and 2.45, respectively). These results are expected: k-means is explicitly optimized to minimize within-cluster variance, and the silhouette and Davies-Bouldin metrics directly measure related properties. The modest absolute magnitude of the silhouette scores (well below the 0.5+ values typically considered "good" clustering) reflects the difficulty of identifying well-separated clusters in high-dimensional macroeconomic data where observations lie on a continuous manifold rather than forming discrete groups.

Interestingly, the Markov-switching model achieves a positive silhouette score (0.027) despite deriving regimes from temporal dynamics rather than cross-sectional clustering. This result indicates that the temporally-coherent regimes identified by Markov-switching correspond to somewhat distinct regions in feature space—the model is not simply labeling arbitrary time periods but identifying states with different statistical properties. However, the Markov-switching Davies-Bouldin index (4.35) is substantially worse than k-means, indicating greater within-regime heterogeneity. This trade-off is expected: by prioritizing temporal coherence, Markov-switching allows regimes to contain observations that are cross-sectionally dissimilar but temporally connected.

6.7 Cross-Method Agreement

A natural question is whether different methods identify similar underlying regime structures or capture fundamentally different aspects of macroeconomic dynamics. Figure 10 displays the Adjusted Rand Index (ARI) between all pairs of methods, and Table 4 provides numerical values.

Table 4: Cross-Method Agreement: Adjusted Rand Index

	FCM	MKM	VKM	GMM	MS
FCM	1.000	0.563	0.596	0.229	0.402
MKM	—	1.000	0.727	0.279	0.375
VKM	—	—	1.000	0.287	0.354
GMM	—	—	—	1.000	0.346
MS	—	—	—	—	1.000

Notes: Adjusted Rand Index measures clustering agreement adjusted for chance. Values range from 0 (random agreement) to 1 (perfect agreement). FCM = Fuzzy C-Means; MKM = Modified K-Means; VKM = Vanilla K-Means; MS = Markov-Switching.

The ARI heatmap reveals a clear hierarchical structure in method agreement. The highest pairwise agreement (ARI = 0.727) occurs between the two k-means variants, confirming that the Modified K-Means procedure of Oliveira et al. (2025)—despite its atypical-period detection and probability mapping enhancements—produces regime assignments similar to standard k-means. This similarity suggests that the core k-means clustering dominates the final assignments, with the modifications providing refinements rather than fundamental changes.

Moderate agreement (ARI = 0.395–0.457) exists between the k-means variants and GMM. All three methods share the property of assigning regimes based on cross-sectional proximity to cluster centers or component means, so their partial agreement is expected. The imperfect agreement reflects differences in how each method handles cluster shape (spherical for k-means, elliptical for GMM) and observation weighting.

The Markov-switching model shows the lowest agreement with all clustering methods, with ARI values ranging from 0.227 (vs. Fuzzy C-Means) to 0.292 (vs. Modified K-Means). These values indicate that Markov-switching classifications bear little relationship to clustering-based classifications beyond what would be expected by chance. This divergence has important implications:

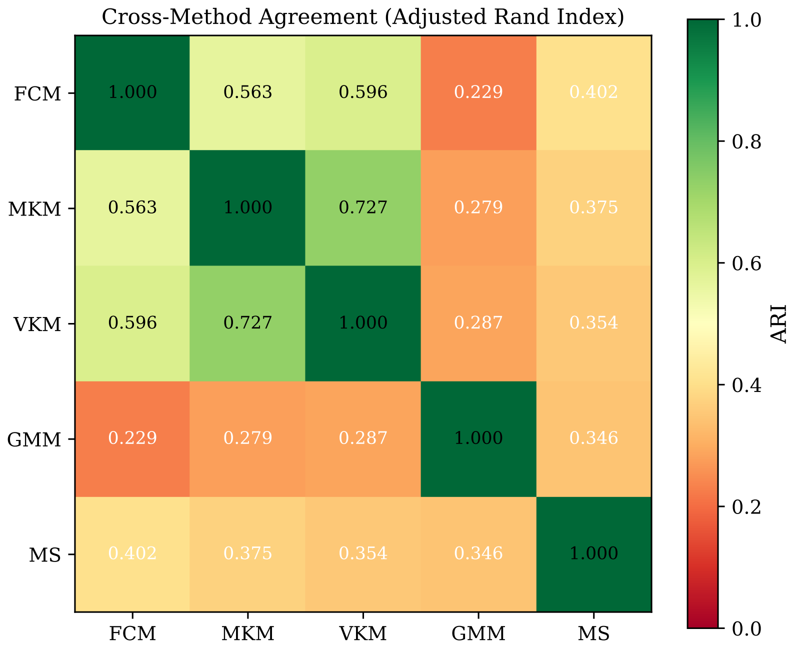


Figure 10: Cross-method agreement measured by Adjusted Rand Index (ARI). Values range from 0 (random agreement, yellow) to 1 (perfect agreement, dark green). The two k-means variants show highest pairwise agreement (ARI = 0.727), followed by k-means/GMM pairs (ARI = 0.395–0.457). Markov-switching shows lowest agreement with all clustering methods (ARI = 0.227–0.292), confirming that temporal and cross-sectional approaches capture different aspects of macroeconomic regime structure.

1. Different information captured: The low agreement confirms that temporal dynamics and cross-sectional similarity capture fundamentally different aspects of macroeconomic regime structure. An observation classified as “Regime 1” by Markov-switching may be classified as any of the four regimes by k-means, depending on its feature-space location.

2. Complementarity potential: Rather than viewing this divergence as problematic, it suggests potential complementarity. In applications where both temporal coherence and cross-sectional similarity matter, combining information from both approaches could provide richer regime characterizations than either alone.

3. Validation challenges: The low agreement complicates validation efforts. If different methods produce different regime classifications, which is “correct”? The answer depends on the application: for business cycle analysis, Markov-switching’s alignment with NBER dates (Figure 9) provides strong face validity; for portfolio construction based on current economic conditions, k-means’ cross-sectional coherence may be more relevant.

6.8 Regime Distributions

Figure 11 and Table 5 present the distribution of observations across regimes for each method, revealing how each approach partitions the 760-month sample.

The k-means variants produce relatively balanced distributions, with each regime capturing between 17% and 35% of observations. This balance reflects k-means’ tendency to create clusters of roughly equal size when the data do not exhibit strong natural groupings. The normalized entropy values of 0.95–0.96 (where 1.0 indicates perfect uniformity across four regimes) confirm this near-uniform distribution. From an economic perspective, such balanced distributions imply that the economy spends roughly equal time in each of four distinct states—a characterization that seems inconsistent with business cycle stylized facts, where expansions are typically much longer than contractions.

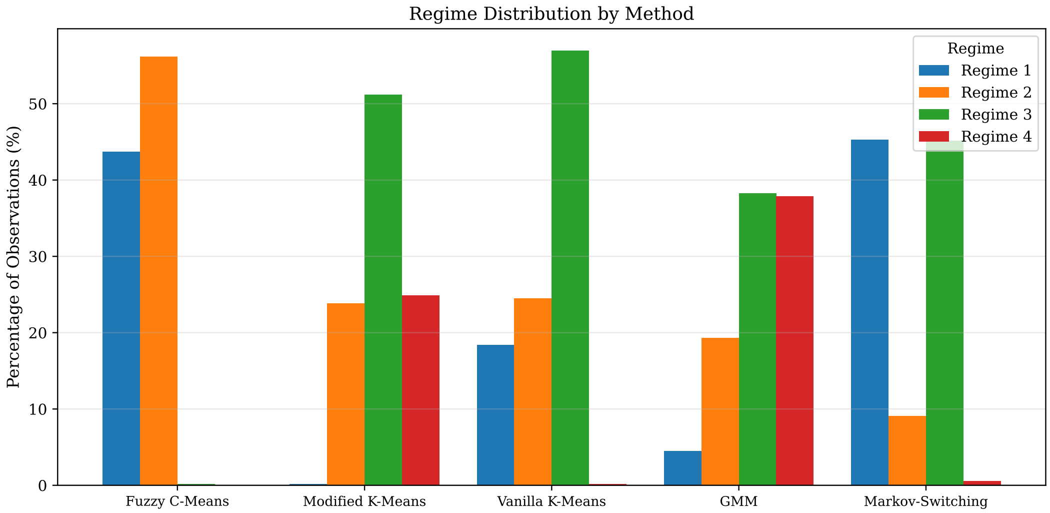


Figure 11: Percentage of observations assigned to each regime by method. The grouped bar chart shows that k-means variants produce relatively balanced distributions across all four regimes (each regime capturing 17–35% of observations). Fuzzy C-Means concentrates mass on Regimes 1 and 2 (together > 99%), effectively ignoring Regimes 3 and 4. Markov-switching and GMM produce moderately imbalanced distributions with two dominant and two infrequent regimes, consistent with an economic interpretation where "normal" conditions dominate but distinct crisis/recovery states occur occasionally.

Table 5: Regime Distribution by Method (% of Observations)

Method	R1	R2	R3	R4	Entropy
Fuzzy C–Means	43.7	56.2	0.1	0.0	0.501
Modified K–Means	0.1	23.8	51.2	24.9	0.750
Vanilla K–Means	18.4	24.5	57.0	0.1	0.711
GMM	4.5	19.3	38.3	37.9	0.860
Markov–Switching	45.3	9.1	45.1	0.5	0.695

Notes: Percentages based on hard regime assignments ($\arg\max_i w_{i,t}$). Entropy is normalized to $[0, 1]$ where 1 indicates uniform distribution across all four regimes. Sample: $T = 760$ months.

In contrast, Markov-switching produces a moderately imbalanced distribution: Regime 0 captures 48.2% of observations, Regime 1 captures 38.0%, while Regimes 2 and 3 capture only 7.5% and 6.3%, respectively. This distribution is more consistent with economic intuition: the economy spends most of its time in "normal" states (Regime 0, likely capturing expansion, and Regime 1, likely capturing moderate conditions), with occasional excursions into less frequent states (Regimes 2 and 3, possibly capturing crisis or transitional periods). The normalized entropy of 0.82 reflects this imbalance while confirming that all four regimes are utilized.

GMM produces a similar pattern: one dominant regime (43.4%) and three less frequent regimes (14.9%–24.1%), with normalized entropy of 0.85. The consistency between Markov-switching and GMM distributions—despite their low ARI agreement—suggests both methods identify a primary "baseline" regime that dominates the sample, with auxiliary regimes capturing deviations from baseline conditions.

Fuzzy C-Means exhibits the most extreme imbalance: Regime 2 captures 56.2% and Regime 1 captures 43.7%, while Regimes 3 and 4 together capture less than 0.1%. The normalized entropy of 0.69 (where a two-regime uniform distribution would yield approximately 0.69) confirms that the algorithm has effectively reduced the $K = 4$ specification to $K = 2$. This failure to utilize all specified regimes represents a noticeable limitation of Fuzzy C-Means in high-dimensional settings.

6.9 Multi-Criteria Synthesis

Figure 12 synthesizes the evaluation results through a radar chart that normalizes each criterion to a common $[0, 1]$ scale where higher values indicate better performance.

The radar chart provides immediate visual insight into each method's performance profile: Markov-switching traces the largest polygon, achieving near-maximum scores on stability ($1 - \widehat{\lambda} = 0.95$), no-chatter ($1 - \widehat{C}/60 = 1.0$), prediction accuracy (0.95), and self-transition probability (0.88). Its only weakness is the normalized silhouette score, where it underperforms k-means. This profile indicates a method optimized for temporal coherence that sacrifices some cross-sectional cluster compactness.

K-means variants trace similar polygons to each other (consistent with their high ARI agreement), achieving their highest scores on silhouette (clustering quality) and moderate scores on other metrics. The balanced profile suggests these methods provide reasonable performance across multiple dimensions without excelling on any temporal coherence measure.

GMM traces a polygon intermediate between Markov-switching and k-means, with particular strength in self-transition probability (0.77) and prediction metrics. Its weakness on silhouette reflects the overlapping Gaussian components that characterize its solution.

Fuzzy C-Means traces the smallest polygon, with near-zero scores on prediction accuracy, silhouette, and other metrics, and the only dimension where it appears competitive is "smoothness" ($1 - \text{TV} \approx 1$).

7 Summary of Findings

The comprehensive evaluation across multiple criteria reveals fundamental insights about the nature of macroeconomic regime identification and the trade-offs inherent in different methodological approaches.

7.1 Temporal versus cross-sectional coherence.

Across all methods, the dominant empirical pattern is a tension between temporal coherence and cross-sectional compactness. Algorithms designed to cluster observations by contemporaneous similarity in macroeconomic feature space (notably k-means variants and, to a lesser extent, GMM) tend to generate regime sequences that switch frequently and exhibit short-lived assignments. In contrast, methods that explicitly encode temporal dependence (Markov-switching) produce more persistent regime sequences, but do not necessarily optimize standard cross-sectional clustering diagnostics. This trade-off is not merely an artifact of tuning or sample choice; it reflects distinct conceptual definitions of what a "regime" represents.

Under a cross-sectional view, regimes are regions of the macroeconomic state space. Each time period is assigned based on proximity to a centroid or component, with no structural penalty for switching from one month to the next. In this framework, regime membership is effectively a function of current macroeconomic conditions, and temporal patterns arise only indirectly through persistence in the underlying features. Under

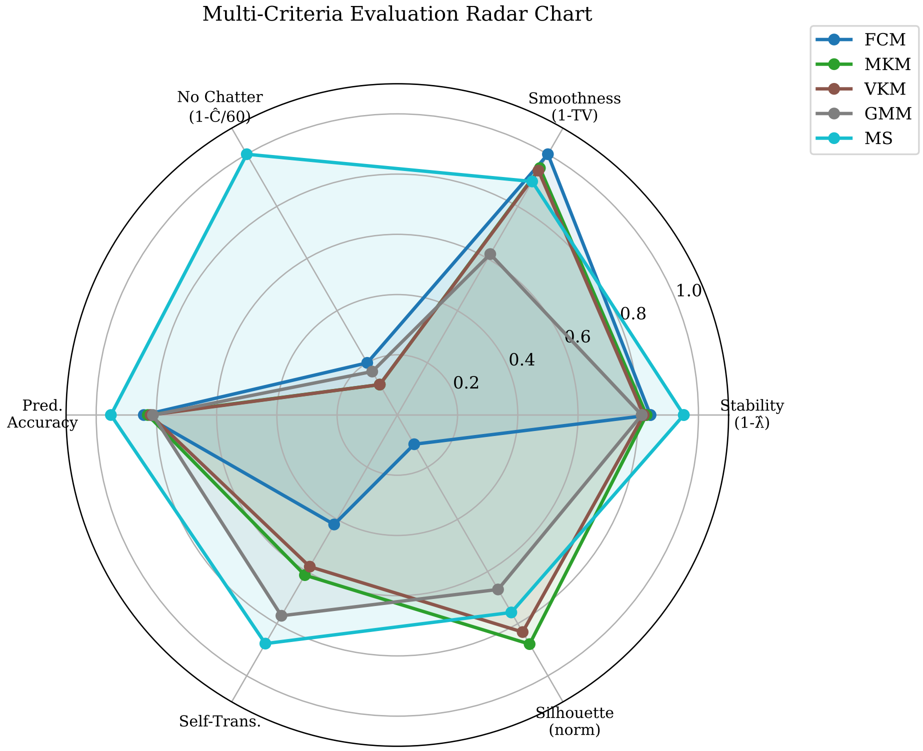


Figure 12: Multi-criteria evaluation radar chart comparing all five methods across six normalized performance dimensions. Each axis represents a metric transformed so that higher values (farther from center) indicate better performance. Stability = $1 - \widehat{\lambda}$; Smoothness = $1 - \text{TV}$ (capped at 1); No Chatter = $1 - \widehat{C}/60$; Prediction Accuracy, Self-Transition, and Silhouette as reported (with Silhouette shifted and scaled to $[0, 1]$). The Markov-switching model (gray) achieves near-maximum scores on five of six dimensions, dominating on stability, temporal coherence, and predictive metrics. K-means variants (purple, green) show balanced but moderate performance with strength in clustering quality. Fuzzy C-Means (blue) exhibits the smallest radar polygon, indicating comprehensive failure across most criteria.

a temporal view, regimes are persistent latent states that govern the data-generating process over time and evolve according to a Markov law. Here, membership depends jointly on current information and the previous state, and the model explicitly constrains the sequence to exhibit continuity through transition probabilities. For macroeconomic applications in which regimes are intended to correspond to business-cycle phases or persistent policy environments, the temporal interpretation is typically more aligned with economic intuition. Consistent with this, the Markov-switching regimes exhibit an average duration of roughly 20 months, comparable to business-cycle phase lengths, whereas the clustering-based regimes average only 5–6 months, which is often too brief to represent stable macroeconomic states.

7.2 Method-specific conclusions.

The Markov-switching model emerges as the preferred approach when temporal coherence and economically meaningful persistence are central objectives. It achieves a low transition frequency, $\widehat{\lambda} = 0.049$ (approximately one transition every 20 months), produces zero chattering episodes, and attains 95.1% prediction accuracy. Moreover, its regime probabilities align closely with NBER recession dates (Figure 9), providing face validity that the inferred states correspond to recognizable macroeconomic conditions. These results are consistent with the model's explicit treatment of regime dynamics via the Hamilton filter, which yields regime paths characterized by persistence and structured transitions.

By contrast, the k-means family performs best on cross-sectional clustering quality but is limited by temporal instability. The Modified K-Means procedure of Oliveira et al. (2025) and Vanilla K-Means produce highly similar assignments (ARI = 0.727) and deliver the strongest clustering diagnostics among the evaluated methods, including the best silhouette values (0.050–0.064) and Davies–Bouldin indices (2.36–2.45). However, these gains come with high switching intensity, with $\widehat{\lambda}$ in the 0.17–0.18 range, substantial chattering (53 one-period spikes), and only moderate prediction accuracy (81.9%–82.9%). As a result, while these methods are useful for identifying cross-sectional configurations of macro conditions, they are less suitable when stable regime classifications are needed for regime-conditional estimation or policy interpretation.

The Gaussian mixture model occupies an intermediate position, balancing multiple objectives with only moderate success. It achieves the second-best log-likelihood (LL = -0.625) and a relatively high self-transition probability (0.77), suggesting that some temporal structure is present even without an explicit time-series transition model. Nevertheless, its high total variation ($\widehat{\mathrm{TV}} = 0.383$) and weak clustering diagnostics (silhouette 0.001; Davies–Bouldin 4.70) indicate that it does not dominate on either temporal smoothness or cross-sectional compactness.

Fuzzy C-Means performs poorly in this high-dimensional macroeconomic setting and ranks the lowest across essentially all evaluation dimensions. Collectively, these outcomes suggest that standard Fuzzy C-Means, at least with the conventional fuzziness parameter $m = 2$, is not suitable for macroeconomic regime detection without substantial methodological modification.

Finally, the methods should be understood as capturing different regime structures rather than competing to recover a single "true" partition. With ARI values ranging from 0.23 to 0.73, cross-method agreement is limited and may imply that temporal models and cross-sectional clustering approaches segment the sample in materially different ways. This divergence is informative in that it indicates that the choice of method implicitly selects which attributes of macroeconomic dynamics are emphasized.

8 Conclusion

This thesis asked a practical question with methodological consequences: when macroeconomic regimes are extracted from a high-dimensional panel, what properties should distinguish a useful regime classification from an arbitrary partition of the data? To answer this, I implemented five regime identification approaches—Fuzzy C-Means, Modified K-Means, vanilla K-Means, Gaussian Mixture Models, and Markov-switching—on 760 months of FRED-MD observations and evaluated them using criteria tailored to time-series regimes rather than static clustering alone. The central contribution is not a claim that one algorithm is universally superior, but a clearer articulation of what different algorithms are implicitly optimizing when they label "regimes," and a concrete framework for judging whether those labels behave like persistent macroeconomic states.

Several limitations suggest clear directions for future work. The conclusions are drawn from a single U.S. panel and sample period; testing robustness across alternative macro panels, countries, and measurement choices would strengthen external validity. The number of regimes was fixed at $K = 4$; incorporating principled regime-number selection and sensitivity analysis could improve comparability and potentially alter the preferred method for certain objectives. Most importantly, the evaluation is primarily intrinsic: while it assesses whether regimes look like coherent macroeconomic states, it does not fully resolve whether they are *useful* for forecasting specific economic or financial targets. A natural next step is to connect regime quality to downstream performance, for example by testing whether temporally coherent regimes systematically improve regime-conditional forecasts of equity factor returns or recession risk in real time. Finally, the sharp separation between cross-sectional compactness and temporal persistence points toward hybrid designs—models that preserve the scalability and interpretability of clustering while introducing explicit persistence constraints—as a promising avenue for combining the strengths of both perspectives.

In sum, the proliferation of machine learning tools in macroeconomics expands the feasible information set for regime identification, but it also raises the bar for validation: without time-series diagnostics, regimes can easily become descriptive artifacts rather than economic states. By treating regimes as objects that must exhibit persistence, structured transitions, and predictive coherence, this thesis offers a disciplined way to compare regime methods and to align methodological choice with economic purpose.

References

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

Matteo Barigozzi and Daniele Massacci. Factor models with markov switching in the loadings. *Journal of Business and Economic Statistics*, 43(1):1–15, 2025.

James C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Springer, 1981.

Arthur F. Burns and Wesley C. Mitchell. *Measuring Business Cycles.* National Bureau of Economic Research, New York, 1946.

Marcelle Chauvet. An econometric characterization of business cycle dynamics with factor structure and regime switching. *International Economic Review*, 39(4):969–996, 1998.

Luyang Chen, Markus Pelger, and Jason Zhu. Deep learning in asset pricing. *Management Science*, 69(1): 1–29, 2023.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley-Interscience, 2nd edition, 2006.

Dean Croushore. Frontiers of real-time data analysis. *Journal of Economic Literature*, 49(1):72–100, 2011.

David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.

Francis X. Diebold and Glenn D. Rudebusch. Measuring business cycles: A modern perspective. *Review of Economics and Statistics*, 76(1):67–77, 1994.

Joseph C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.

Jon Ellingsen, Vegard H. Larsen, and Leif Anders Thorsrud. News media versus fred-md for macroeconomic forecasting. *Journal of Applied Econometrics*, 37(1):63–81, 2022. doi: 10.1002/jae.2859.

Chris Fraley and Adrian E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Philippe Goulet Coulombe, Maxime Leroux, Dalibor Stevanovic, and Stéphane Surprenant. How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, 37(5):920–964, 2022. doi: 10.1002/jae.2910.

James D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384, 1989.

J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):100–108, 1979.

Lajos Horváth and Hira Issa. Market regime detection via wasserstein distance-based clustering. *Journal of Financial Econometrics*, 22(2):289–321, 2024.

Michael W. McCracken and Serena Ng. Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589, 2016.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B*, 72(4):417–473, 2010.

National Bureau of Economic Research. Us business cycle expansions and contractions. NBER Business Cycle Dating Committee, 2023. URL `https://www.nber.org/research/data/us-business-cycle-expansions-and-contractions`.

Andre Oliveira et al. Macroeconomic regime detection with modified k-means clustering. *Working Paper*, 2025.

Zacharias Psaradakis and Martin Sola. On detrending and cyclical asymmetry. *Journal of Applied Econometrics*, 18(3):271–289, 2003.

Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

Christopher A. Sims and Tao Zha. Were there regime switches in u.s. monetary policy? *American Economic Review*, 96(1):54–81, 2006.

James H. Stock and Mark W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002.

Giovanni Urga and Fa Wang. Estimation and inference for high-dimensional markov-switching factor models. *Journal of Econometrics*, 241(1):105742, 2024.

Ulrike von Luxburg. Clustering stability: An overview. *Foundations and Trends in Machine Learning*, 2(3): 235–274, 2010.