

Article

Fuzzy and Explainable AI for CMB Polarization Segmentation: Regional Stability Under Controlled Perturbations

Gabriel Marín Díaz ^{1,2}¹ Faculty of Statistics, Complutense University, Puerta de Hierro, 28040 Madrid, Spain; gmarin03@ucm.es² Science and Aerospace Department, Universidad Europea de Madrid, Villaviciosa de Odón, 28670 Madrid, Spain

Abstract

The cosmic microwave background (CMB) contains key information about the early Universe, particularly through its polarization structure. This work proposes a Fuzzy and Explainable Artificial Intelligence framework (FAS-XAI) for the regional analysis of CMB polarization using Planck SMICA data. From the Stokes components Q and U , the polarization amplitude P and the scalar polarization modes E and B are derived. Regional features are then extracted over a HEALPix grid, considering only polarization-valid regions defined by the Planck polarization mask. Fuzzy C-Means identifies four interpretable polarization regimes: high-polarization structured regions, E -dominated medium-polarization regions, B -enhanced medium-polarization regions, and low-polarization regions. An XGBoost-SHAP layer is used to explain the resulting fuzzy memberships. XGBoost accurately reproduces the memberships, with $R^2 > 0.98$ for all clusters, while SHAP confirms the physical relevance of amplitude-related features and the $\log(B/E)$ balance. Finally, controlled perturbations in P and $\log(B/E)$ reveal a globally robust fuzzy structure with localized sensitivity. The proposed framework provides an interpretable methodology for studying regional CMB polarization patterns and their stability under controlled perturbations.

Keywords: cosmic microwave background; CMB polarization; explainable AI; Fuzzy C-Means; SHAP; HEALPix; perturbation analysis

MSC: 62H30; 68T20; 85A40

1. Introduction

The cosmic microwave background (CMB) is one of the most important observational probes of the early Universe. Its temperature anisotropies and polarization patterns preserve information about primordial fluctuations, the growth of cosmic structure, reionization, gravitational lensing, and possible tensor perturbations. While temperature anisotropies have been extensively studied, CMB polarization provides an additional and highly sensitive channel for investigating the physical conditions of the early Universe [1,2].

CMB polarization is commonly described through the Stokes parameters Q and U , which encode the local linear polarization field. From these quantities, one can derive the polarization amplitude P and decompose the signal into scalar modes E and B . The E -mode is mainly associated with gradient-like polarization patterns generated by scalar perturbations, whereas the B -mode is more delicate to interpret, since it can arise from

Academic Editor: Sanjar Abrarov

Received: 23 May 2026

Revised: 12 June 2026

Accepted: 23 June 2026

Published: 25 June 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

gravitational lensing, residual foregrounds, instrumental noise, leakage effects, or, in specific scenarios, primordial tensor modes. For this reason, the relationship between E and B is a key element in the physical interpretation of CMB polarization [3].

Most standard CMB analyses are based on angular power spectra, such as C_ℓ^{TT} , C_ℓ^{EE} , and C_ℓ^{BB} . These approaches are fundamental for cosmological parameter estimation and for testing theoretical models. However, they provide a mainly global and spectral description of the CMB. Complementary regional approaches can provide additional insight into how polarization structures are spatially organized across the celestial sphere, especially when combined with interpretable machine learning methods [4].

In this work, we propose a Fuzzy and Explainable Artificial Intelligence framework (FAS-XAI) for the regional analysis of CMB polarization [5]. The goal is not to replace standard cosmological analyses, but to complement them with a regional, fuzzy, and interpretable representation of polarization regimes. Instead of assigning each sky region to a single rigid class, the proposed framework models each region through membership degrees across physically interpretable regimes.

The analysis is performed using Planck SMICA data [1]. Starting from the Stokes components Q and U , we derive the polarization amplitude P and the scalar polarization modes E and B . The original maps are kept at $NSIDE = 2048$ to preserve high-resolution polarization information. However, clustering is not performed directly on the full pixel-level maps. Instead, regional features are extracted over a coarser HEALPix grid with $NSIDE_{\text{region}} = 64$. This strategy preserves information from the high-resolution maps while producing a computationally manageable and physically interpretable regional representation.

A relevant methodological aspect of this work is the use of the Planck polarization mask. Only pixels considered valid for polarization analysis are used, and regions with insufficient valid coverage are excluded. This prevents the clustering process from being driven by Galactic foregrounds, point sources, residual contamination, or mask-induced artifacts. Therefore, the identified regimes are based on regions where the polarization signal is considered sufficiently reliable.

For each valid region, we compute a set of physically meaningful features: the regional mean and standard deviation of P , the regional energy of the E -mode, and the logarithmic ratio $\log(B/E)$, which captures the relative balance between the B - and E -mode energies. These features are then used as input to Fuzzy C-Means clustering [6]. The resulting fuzzy memberships allow us to identify regional polarization regimes and, importantly, to quantify ambiguity and transitions between them.

To provide explainability, we introduce a supervised XAI layer based on XGBoost [7] and SHAP [8]. For each fuzzy cluster, an XGBoost regression model is trained to predict the degree of membership of each region from the selected physical features. SHAP values are then used to quantify how each feature contributes to increasing or decreasing membership in each regime. This allows us to interpret the fuzzy clusters not merely as numerical groups, but as physically meaningful polarization configurations.

Finally, we introduce a controlled perturbation analysis. Small counterfactual perturbations are applied to two key dimensions: the polarization amplitude P and the relative balance $\log(B/E)$. The objective is to evaluate the stability of the regional fuzzy structure under microvariations of physically relevant variables. This analysis allows us to distinguish robust regions from sensitive ones and to study how alternative regional configurations may emerge under controlled perturbations.

The main contributions of this work are as follows:

1. A regional FAS-XAI framework for CMB polarization analysis based on P , E , and B .
2. A fuzzy segmentation of valid CMB polarization regions into physically interpretable regimes.

3. An XGBoost-SHAP explainability layer for interpreting fuzzy memberships.
4. A perturbation-based stability analysis of regional CMB polarization regimes.
5. A controlled counterfactual framework for studying localized sensitivity in the CMB polarization structure.

The paper is organized as follows. Section 2 reviews the state of the art on CMB polarization, machine learning, explainable AI, and fuzzy clustering. Section 3 describes the proposed methodology, including data preprocessing, regional feature construction, fuzzy segmentation, XAI modeling, and counterfactual perturbation analysis. Section 4 presents empirical results. Section 5 discusses physical interpretation, limitations, and future research directions. Finally, Section 6 summarizes the main conclusions.

2. Related Work

2.1. Previous Studies on Cosmic Microwave Background Polarization

The Cosmic Microwave Background (CMB) is one of the most important observational probes of the early Universe [2]. Its temperature anisotropies and polarization patterns encode information about primordial fluctuations, acoustic oscillations, reionization, gravitational lensing, and the large-scale structure of the Universe. Over the last decades, CMB research has evolved from the detection and characterization of temperature anisotropies toward increasingly detailed studies of polarization, foreground removal, component separation, and the search for weak B -mode signals. To assess the evolution of this research area, a Web of Science topic search was conducted using the query: $TS = (\text{"cosmic microwave background" OR CMB}) \text{ AND } TS = (\text{polarization OR polarisation OR "E-mode" OR "B-mode"})$. The search returned 3404 publications and 247,230 citations, confirming that CMB polarization is a mature and highly active field. As shown in Figure 1, publication activity increased substantially with the development of dedicated CMB missions and experiments, particularly after the availability of high-precision data from WMAP and Planck.

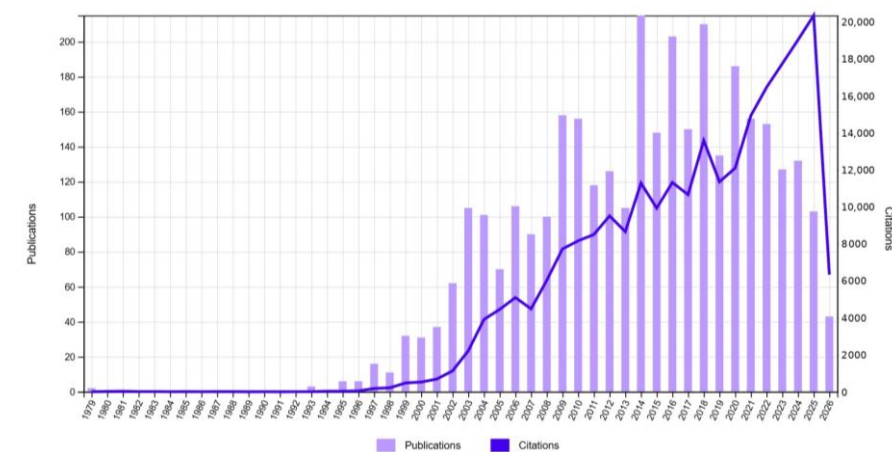


Figure 1. Publications (3404) and Citations (247,230). $TS = (\text{"cosmic microwave background" OR CMB}) \text{ AND } TS = (\text{polarization OR polarisation OR "E-mode" OR "B-mode"})$.

These missions enabled a more detailed characterization of the CMB temperature and polarization fields, including the analysis of the Stokes parameters Q and U , the decomposition into E - and B -modes, and the estimation of angular power spectra such as C_ℓ^{EE} and C_ℓ^{BB} . At the same time, they highlighted major challenges associated with polarization analysis, including foreground contamination, Galactic dust emission, synchrotron radiation, point-source masking, instrumental noise, and leakage between E - and B -modes. These issues are especially relevant for the interpretation of the B -mode

component, whose weak signal can be affected by gravitational lensing, residual foregrounds, or systematic effects.

Most existing CMB polarization studies are based on harmonic-space and spectral analyses. Angular power spectra remain the standard tool for cosmological inference and for the comparison between theoretical models and observational data. However, these approaches provide mainly global statistical descriptions of the CMB signal. Comparatively fewer studies focus on regional, spatially interpretable characterizations of polarization patterns over the celestial sphere.

2.2. Machine Learning and Deep Learning in CMB Analysis

The growing availability of high-resolution cosmological data products has encouraged the application of machine learning and deep learning techniques to CMB-related problems. These approaches have been used for tasks such as CMB signal recovery, foreground cleaning, lensing reconstruction, cosmological parameter inference, map simulation, and the analysis of secondary anisotropies. To assess the development of this research line, a Web of Science topic search was conducted using the query: $TS = ("cosmic\ microwave\ background")\ AND\ TS = ("machine\ learning"\ OR\ "deep\ learning"\ OR\ "neural\ network"\ OR\ "artificial\ intelligence")$.

The results show that the number of studies remains relatively limited compared with the broader literature on CMB physics and polarization. As shown in Figure 2, publication activity was almost negligible before the late 2000s, increased gradually between 2007 and 2018, and expanded more clearly from 2019 onward, coinciding with the broader adoption of machine learning in astrophysics and cosmology.

Despite this growth, machine learning in CMB research has mainly been applied to predictive, reconstruction, or inference-oriented tasks. Existing studies typically use machine learning to recover the CMB signal, remove foreground contamination, accelerate parameter estimation, emulate cosmological simulations, or reconstruct lensing-related information. Comparatively fewer works focus on the interpretability of learned representations or on the explanation of regional polarization structures using physically meaningful features.

A more focused inspection of the CMB-related subset shows that machine learning and deep learning applications are concentrated in four main subareas: CMB reconstruction and foreground cleaning, CMB polarization and B -mode-related problems, CMB anomalies and non-Gaussianity, and secondary anisotropies associated with clusters and small-scale foregrounds. Table 1 summarizes these subareas, including the number of works, accumulated citations, and representative highly cited contributions identified in the Web of Science sample.

Table 1. Main CMB-related subareas identified among the most cited works.

CMB-Related Subarea	No. of Works	Citations
CMB ML/lensing/foreground cleaning/reconstruction [9–15]	7	244
CMB polarization/ B -modes [16,17]	2	31
CMB anomalies/non-Gaussianity/cosmic strings [18–21]	4	77
SZ/clusters/small-scale CMB foregrounds [22–26]	5	109

This distribution suggests that machine learning in CMB research has so far been mainly developed as a predictive or reconstructive tool rather than as an explanatory framework. In particular, the literature is still comparatively sparse regarding interpretable map-level analyses of CMB polarization, especially those aimed at characterizing the spatial organization of E - and B -mode structures across sky regions.

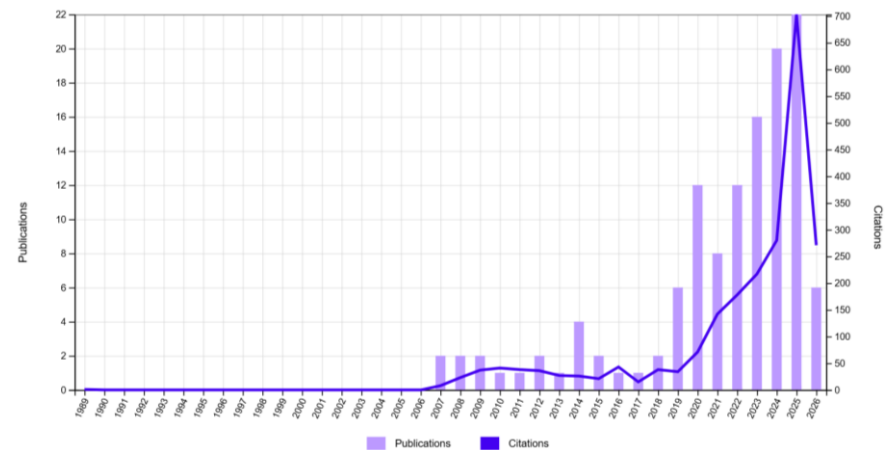


Figure 2. Publications (123) and Citations (2250). TS = (“cosmic microwave background”) AND TS = (“machine learning” OR “deep learning” OR “neural network” OR “artificial intelligence”).

2.3. Explainable Artificial Intelligence and Fuzzy Clustering in CMB Studies

To evaluate the specific adoption of explainable artificial intelligence in CMB studies, a restrictive Web of Science search was conducted using the query: TS = (“cosmic microwave background”) AND TS = (“explainable artificial intelligence” OR XAI OR SHAP). The search returned only one directly related publication [27]. This result suggests that the explicit integration of explainability methods into CMB analysis remains very limited. The identified study applies neural networks to cosmological model selection using CMB data and includes an analysis of feature importance. However, it does not address fuzzy regional segmentation, fuzzy membership explanation, SHAP-based interpretation of polarization regimes, or perturbation-based stability analysis.

In addition, a second Web of Science search was conducted to evaluate the presence of clustering-based approaches in CMB studies: TS = (“cosmic microwave background”) AND TS = (“fuzzy clustering” OR clustering). This query returned no results. This absence suggests that clustering-based approaches, and particularly fuzzy clustering methods, have not been explicitly explored in the CMB literature indexed by Web of Science under these search terms.

Therefore, although feature-importance methods have begun to appear in CMB-related machine learning studies, the use of XAI for explaining regional CMB polarization structures remains largely unexplored. Similarly, the absence of results for CMB-related fuzzy clustering searches highlights the lack of previous work on fuzzy regional segmentation of CMB polarization. This reinforces the novelty of the present work, which combines Fuzzy C-Means clustering, XGBoost-based membership modeling, SHAP explanations, and controlled perturbation analysis over physically interpretable CMB polarization features.

3. Materials and Methods

This section describes the methodological design adopted to analyze the regional structure of Cosmic Microwave Background (CMB) polarization using a fuzzy and explainable analytical framework. The objective is to establish a structured pipeline that integrates physical variable construction, regional feature extraction, fuzzy clustering, supervised membership modeling, explainable artificial intelligence, and controlled perturbation analysis. Rather than relying on predefined sky regions or rigid classifications, the proposed approach explores CMB polarization through an uncertainty-aware regional framework. By combining physically interpretable variables derived from the Stokes parameters with fuzzy clustering and explainable modeling techniques, the method enables

the identification of regional polarization regimes and the interpretation of the variables that characterize them.

The methodology is applied to the Planck PR3 SMICA CMB IQU map, COM_CMB_IQU-smica_2048_R3.00_full [1], which provides full-sky estimates of CMB intensity and polarization. The analysis uses the Stokes parameters I , Q , and U , together with the Planck polarization mask PMASK, to ensure that only regions considered reliable for polarization analysis are included. This step reduces the influence of Galactic foregrounds, point sources, residual contamination, and mask-induced artifacts on the clustering process.

The workflow is organized as a sequence of interconnected analytical phases, ranging from data preparation and physical variable construction to regional feature extraction, fuzzy segmentation, explainable modeling, and perturbation-based stability analysis. Figure 3 summarizes the overall methodological architecture.

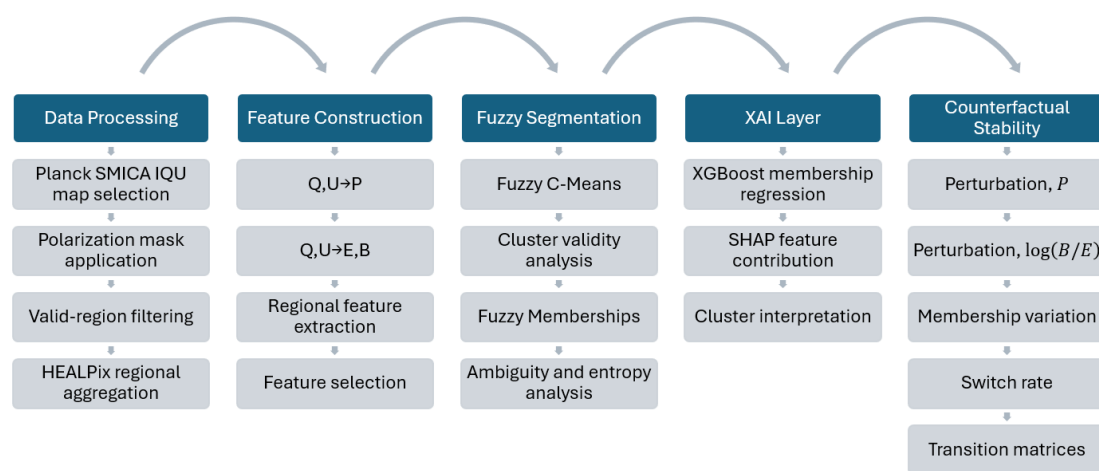


Figure 3. Framework methodological architecture.

The methodological workflow is summarized below, with each stage described in detail in Sections 3.1–3.10.

- **Data acquisition and preprocessing.** The analysis starts from the Planck PR3 SMICA CMB IQU map at its original HEALPix resolution. The original fields include the CMB intensity, the Stokes polarization components, and the corresponding temperature and polarization masks. Since the present work focuses on polarization, the Planck polarization mask is used to identify valid pixels for the analysis. Regions with insufficient valid coverage are excluded to prevent foreground-contaminated or unreliable areas from affecting the clustering results.
- **Construction of polarization variables.** From the Stokes parameters Q and U , the polarization amplitude P is computed. In addition, the E - and B -mode polarization maps are obtained through harmonic decomposition. These variables provide complementary physical information: P represents the total linear polarization amplitude, E captures the gradient-like component of the polarization field, and B represents the curl-like component.
- **Regional aggregation and feature extraction.** The original CMB maps are kept at high resolution to preserve polarization information. However, clustering is not performed directly at pixel level. Instead, a coarser HEALPix regional grid is used to summarize local information. For each valid region, several features are computed from P , E , and B , including regional averages, variability measures, energy-based descriptors, and the relative balance between B - and E -mode energies.

- **Masking and valid-region selection.** All regional features are computed only from polarization-valid pixels according to the Planck polarization mask. For each region, the fraction of valid pixels is calculated, and only regions with sufficient valid coverage are retained. This step ensures that the subsequent clustering and explainability analysis are not driven by foregrounds, Galactic contamination, or mask-induced artifacts.
- **Fuzzy clustering of regional polarization regimes.** Fuzzy C-Means clustering is applied to the normalized regional feature space. Unlike hard clustering, this method assigns each region a degree of membership to each cluster. This is especially suitable for CMB polarization, where spatial regimes may overlap and transition gradually rather than forming sharply separated classes.
- **Supervised membership modeling.** To assess the internal consistency and explainability of the fuzzy segmentation, XGBoost regression models are trained to predict the membership degree of each region to each fuzzy cluster. Instead of predicting only the dominant cluster, one model is trained for each membership function. This stage evaluates whether the fuzzy memberships can be accurately reconstructed from the selected physical features.
- **Explainable analytics.** SHAP values are computed from the XGBoost membership models to quantify the contribution of each feature to the predicted membership degrees. This stage provides an interpretable explanation of each fuzzy regime by identifying which physical variables increase or decrease the membership of a region to a given cluster.
- **Ambiguity and entropy analysis.** Since Fuzzy C-Means provides a full membership vector for each region, the degree of ambiguity can be quantified. The maximum membership value is used to identify clear, moderate, and ambiguous regions. In addition, fuzzy entropy is computed to measure how distributed the memberships are across clusters. This analysis allows the identification of transitional regions and helps characterize the continuous nature of the CMB polarization field.
- **Controlled perturbation and stability analysis.** Controlled perturbations are introduced into two key dimensions of the regional feature space: the polarization amplitude and the relative balance between B - and E -mode energies. Perturbed regional features are projected onto the original fuzzy cluster structure, and the resulting memberships are compared with the baseline memberships. Stability is evaluated through membership variation, dominant-cluster switching, entropy change, and transition matrices.
- **Physical and counterfactual interpretation.** The final stage integrates the fuzzy segmentation, SHAP explanations, and perturbation results to interpret the regional structure of CMB polarization. The analysis distinguishes globally stable patterns from locally sensitive regions and evaluates how small variations in physically meaningful variables may generate alternative regional configurations. These counterfactual configurations are interpreted as controlled scenarios for studying the informational stability of the observed CMB polarization structure.

3.1. Data Acquisition and Preprocessing

The analysis was conducted using the Planck PR3 SMICA CMB IQU map, COM_CMB_IQU-smica_2048_R3.00_full, distributed through the Planck Legacy Archive and the NASA/IPAC Infrared Science Archive [1]. This product provides estimates of the CMB intensity and linear polarization components obtained through the SMICA component-separation method.

The map is provided in HEALPix format at an original resolution of $NSIDE = 2048$, corresponding to 50,331,648 sky pixels and an approximate angular resolution of 1.72

arcmin. This high resolution was retained during the initial processing stages to preserve fine-scale polarization information before regional aggregation.

The FITS file contains ten fields: I_STOKES, Q_STOKES, U_STOKES, TMASK, PMASK, I_STOKES_INP, Q_STOKES_INP, U_STOKES_INP, TMASKINP, and PMASKINP. The present work focuses primarily on the component-separated Stokes polarization fields Q_STOKES and U_STOKES, together with the polarization mask PMASK. The intensity field I_STOKES was used for initial descriptive inspection and contextual analysis, but the regional fuzzy segmentation was based on polarization-derived variables.

The choice of the SMICA IQU product is motivated by its suitability for CMB polarization analysis and by the availability of both the Stokes components and the associated temperature and polarization masks within the same data product. Since the present study focuses on polarization, PMASK was used as the main quality-control mask. This step is essential because the subsequent clustering and explainability stages require reliable regional descriptors that are not dominated by foregrounds, Galactic contamination, point sources, or unreliable sky areas.

Therefore, the selected data product provides the necessary basis for constructing a regional, physically interpretable, and mask-controlled representation of CMB polarization. The selected fields serve as the starting point for the subsequent masking, polarization-variable construction, and regional feature extraction stages.

3.2. Construction of Polarization Variables

CMB polarization is described by the Stokes parameters Q and U , which encode the local linear polarization state of the radiation field. Unlike the temperature intensity I , the quantities Q and U depend on the choice of local coordinate system on the sphere. Nevertheless, they provide the fundamental observational basis from which physically meaningful polarization descriptors can be constructed.

In this work, the first derived polarization variable is the total polarization amplitude P , summarizes the total polarized intensity at each sky pixel, computed from the Stokes components as:

$$P = \sqrt{Q^2 + U^2} \quad (1)$$

In addition to P , the Stokes polarization field was decomposed into the scalar E - and B -mode components. This decomposition transforms the coordinate-dependent fields Q and U into two scalar polarization maps with different physical interpretations. The E -mode represents the gradient-like component of the polarization field, whereas the B -mode represents the curl-like component.

The E -mode is mainly associated with scalar density perturbations and constitutes the dominant cosmological polarization signal. The B -mode is more delicate to interpret, since it can arise from different sources, including gravitational lensing, foreground residuals, instrumental noise, leakage effects, and, in specific theoretical scenarios, primordial tensor perturbations. For this reason, the relative behavior of E and B is particularly relevant in the analysis of CMB polarization [28].

The variables P , E , and B provide complementary information. The polarization amplitude P captures the total strength of the linear polarization signal, while E and B describe its geometric structure. In the proposed framework, these variables constitute the physical basis for the subsequent regional feature extraction stage.

3.3. Regional Aggregation and Valid-Region Selection

The original CMB polarization maps were provided at high angular resolution, with $NSIDE = 2048$. Although this resolution preserves detailed pixel-level polarization information, clustering directly at this scale would be computationally inefficient and difficult

to interpret from a regional perspective. For this reason, the analysis was reformulated over a coarser HEALPix regional grid [29]:

$$NSIDE_{\text{region}} = 64 \quad (2)$$

This regionalization allows the sky to be partitioned into spatially coherent units while retaining the information contained in the original high-resolution polarization maps. Given the relation between both resolutions, each regional cell contains approximately

$$\left(\frac{2048}{64}\right)^2 = 1024 \quad (3)$$

high-resolution pixels. Let r denote a region in the coarser grid. The set of high-resolution pixels belonging to region r is defined as:

$$\mathcal{P}_r = \{p_1, p_2, \dots, p_{n_r}\} \quad (4)$$

where n_r is the total number of high-resolution pixels contained in that region.

Since the present study focuses on polarization, the Planck polarization mask PMASK was used to identify which high-resolution pixels could be considered reliable for analysis. Let M_p denote the value of the mask at pixel p . A pixel was considered valid when, $M_p > 0.5$.

For each region r , the number of valid pixels was computed as:

$$N_{\text{valid}}(r) = \sum_{p \in \mathcal{P}_r} \mathbb{I}(M_p > 0.5) \quad (5)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. The valid fraction of region r was then defined as:

$$f_{\text{valid}}(r) = \frac{N_{\text{valid}}(r)}{N_{\text{total}}(r)} \quad (6)$$

where $N_{\text{total}}(r) = n_r$ is the total number of high-resolution pixels contained in the region.

Only regions satisfying the condition $f_{\text{valid}}(r) \geq 0.8$ were retained for the subsequent feature extraction, clustering, and explainability stages.

This threshold guarantees that regional descriptors are computed from regions with sufficient polarization-valid sky coverage. Regions with insufficient valid fraction were excluded to avoid clustering structures driven by Galactic foregrounds, point sources, residual contamination, or mask-induced artifacts. This step is especially important because the quality of the fuzzy clustering depends directly on the reliability of the regional feature space. Therefore, the use of PMASK and the valid-region criterion provides a necessary quality-control stage before constructing the regional representation of CMB polarization.

3.4. Regional Feature Extraction

Once the valid regions were defined, regional features were extracted from the polarization-derived maps P , E , and B . The objective of this stage was to transform the high-resolution polarization maps into a structured regional feature matrix suitable for fuzzy clustering and explainable modeling [30].

For each region r , only the pixels satisfying the polarization mask criterion were used. Let $\mathcal{P}_r^{\text{valid}}$ denote the set of valid high-resolution pixels contained in region r , and let $N_{\text{valid}}(r)$ be the number of such pixels.

The regional mean polarization amplitude was defined as:

$$P_{\text{mean}}(r) = \frac{1}{N_{\text{valid}}(r)} \sum_{p \in \mathcal{P}_r^{\text{valid}}} P_p \quad (7)$$

This feature measures the average polarization amplitude within each region.

The regional standard deviation of polarization amplitude was defined as:

$$P_{\text{std}}(r) = \sqrt{\frac{1}{N_{\text{valid}}(r)} \sum_{p \in \mathcal{P}_r^{\text{valid}}} (P_p - P_{\text{mean}}(r))^2} \quad (8)$$

This variable captures the internal variability of the polarization amplitude within a region.

In addition to the mean and standard deviation, energy-based descriptors were computed from the P , E , and B maps. In general, the regional energy of a variable X was defined as the mean squared value over the valid pixels of the region:

$$X_{\text{energy}}(r) = \frac{1}{N_{\text{valid}}(r)} \sum_{p \in \mathcal{P}_r^{\text{valid}}} X_p^2 \quad (9)$$

where X may represent P , E , or B . In particular, the E -mode and B -mode regional energies were computed as:

$$E_{\text{energy}}(r) = \frac{1}{N_{\text{valid}}(r)} \sum_{p \in \mathcal{P}_r^{\text{valid}}} E_p^2 \quad (10)$$

$$B_{\text{energy}}(r) = \frac{1}{N_{\text{valid}}(r)} \sum_{p \in \mathcal{P}_r^{\text{valid}}} B_p^2 \quad (11)$$

These quantities summarize the regional strength of the gradient-like and curl-like polarization components, respectively.

To characterize the relative balance between B - and E -mode energy, the following ratio was computed:

$$\frac{B}{E}(r) = \frac{B_{\text{energy}}(r) + \varepsilon}{E_{\text{energy}}(r) + \varepsilon} \quad (12)$$

where ε is a small positive constant introduced to avoid numerical instability when the denominator is close to zero. A logarithmic transformation was then applied:

$$\log(B/E)(r) = \log\left(\frac{B_{\text{energy}}(r) + \varepsilon}{E_{\text{energy}}(r) + \varepsilon}\right) \quad (13)$$

This transformation provides a more stable representation of the relative B/E balance. Values close to zero indicate comparable B - and E -mode energies, negative values indicate relative dominance of E , and positive values indicate relative dominance of B .

The initial regional feature set included:

$$\{P_{\text{mean}}, P_{\text{std}}, P_{\text{energy}}, E_{\text{energy}}, B_{\text{energy}}, B/E, \log(B/E)\} \quad (14)$$

However, several of these variables were highly correlated, particularly the amplitude and energy-based descriptors. To reduce redundancy and preserve physical interpretability, the final feature space used for fuzzy clustering was defined as:

$$\{P_{\text{mean}}, P_{\text{std}}, E_{\text{energy}}, \log(B/E)\} \quad (15)$$

These four variables summarize complementary dimensions of the regional polarization structure: the average polarization amplitude, the internal variability of polarization, the regional E -mode energy, and the relative balance between B - and E -mode contributions.

3.5. Fuzzy Clustering of Regional Polarization Regimes

After constructing the regional feature matrix, Fuzzy C-Means (FCM) clustering was applied to identify regional polarization regimes in the CMB [31]. Unlike hard clustering methods, FCM assigns each region a degree of membership to every cluster. This is particularly suitable for CMB polarization analysis, since polarization structures may vary gradually across the sky and do not necessarily form sharply separated classes.

Before applying FCM, the selected features were standardized using z-score normalization:

$$z_{ik} = \frac{x_{ik} - \mu_k}{\sigma_k} \quad (16)$$

where x_{ik} is the value of feature k for region i , and μ_k and σ_k are the mean and standard deviation of feature k , respectively. This normalization ensures that all variables contribute comparably to the clustering process.

Let x_i denote the standardized feature vector of region i , with:

$$x_i = [P_{\text{mean}}, P_{\text{std}}, E_{\text{energy}}, \log(B/E)]_i \quad (17)$$

FCM seeks to minimize the objective function:

$$J_m = \sum_{i=1}^N \sum_{j=1}^c u_{ij}^m \|x_i - v_j\|^2 \quad (18)$$

where N is the number of regions, c is the number of clusters, u_{ij} is the membership degree of region i to cluster j , v_j is the centroid of cluster j , and m is the fuzziness parameter. In this study, the standard value $m = 2$ was used.

The membership degrees satisfy, $0 \leq u_{ij} \leq 1$, and

$$\sum_{j=1}^c u_{ij} = 1 \quad (19)$$

for each region i . Therefore, each region is represented not by a single rigid label, but by a membership vector describing its degree of association with each polarization regime.

Solutions with $c = 2, 3, 4, 5, 6$ were compared using the objective function, the Fuzzy Partition Coefficient, and the Partition Entropy. The Fuzzy Partition Coefficient measures the sharpness of the fuzzy partition, higher values of FPC indicate crisper memberships, while Partition Entropy measures the uncertainty or ambiguity of the memberships, higher values of PE indicate greater membership uncertainty.

The Fuzzy Partition Coefficient was computed as:

$$FPC = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^c u_{ij}^2 \quad (20)$$

The Partition Entropy was defined as:

$$PE = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^c u_{ij} \log(u_{ij}) \quad (21)$$

Based on the balance between quantitative metrics and physical interpretability, the solution with four clusters was selected. An additional stability analysis was also performed and is reported in Section 4.4, supporting ($c = 4$) as the most suitable solution among the evaluated configurations. This configuration allowed the identification of four regional polarization regimes: high-polarization structured regions, medium-polarization regions dominated by E , medium-polarization regions with relatively enhanced B , and low-polarization regions.

The resulting fuzzy memberships were subsequently used to generate dominant-cluster maps, membership maps, and ambiguity measures, and they served as the target variables for the supervised explainability stage.

3.6. Supervised Membership Modeling

After obtaining the fuzzy memberships from Fuzzy C-Means, a supervised learning layer was introduced to model and validate the relationship between the regional physical features and the membership degrees [7]. This stage was not designed to replace the unsupervised clustering process, but to assess whether the fuzzy memberships could be accurately reconstructed from the selected polarization features.

For each region i and cluster j , the predicted membership was defined as:

$$\hat{u}_{ij} = f_j(x_i) \quad (22)$$

where x_i is the feature vector of region i , and f_j is the XGBoost regression model trained for cluster j .

The input feature vector was the same as that used in the fuzzy clustering stage:

$$x_i = [P_{\text{mean}}, P_{\text{std}}, E_{\text{energy}}, \log(B/E)]_i \quad (23)$$

This formulation preserves the fuzzy nature of the segmentation, since the target variables are continuous membership degrees rather than hard class labels. As a result, the supervised layer evaluates the consistency of the fuzzy partition in terms of the original physical features.

The dataset was divided into training and test subsets. Model performance was evaluated using three regression metrics: the coefficient of determination, the mean absolute error, and the root mean squared error.

The coefficient of determination was computed as:

$$R^2 = 1 - \frac{\sum_i (u_{ij} - \hat{u}_{ij})^2}{\sum_i (u_{ij} - \bar{u}_j)^2} \quad (24)$$

where u_{ij} is the original Fuzzy C-Means membership, $\hat{u}_{ij}$ is the XGBoost-predicted membership, and $\bar{u}_j$ is the mean membership for cluster j .

The mean absolute error was defined as:

$$MAE = \frac{1}{N} \sum_i |u_{ij} - \hat{u}_{ij}| \quad (25)$$

and the root mean squared error was defined as:

$$RMSE = \sqrt{\frac{1}{N} \sum_i (u_{ij} - \hat{u}_{ij})^2} \quad (26)$$

High predictive performance in this stage indicates that the fuzzy memberships are not arbitrary but can be consistently explained from the selected physical descriptors. This provides an additional validation of the fuzzy segmentation and prepares the basis for the subsequent explainability analysis using SHAP values.

3.7. Explainable Analytics

After training the XGBoost membership models, SHAP values were used to interpret the contribution of each regional feature to the predicted fuzzy memberships. This stage provides the explainability layer of the proposed framework, allowing each fuzzy polarization regime to be characterized in terms of the physical variables that increase or decrease membership.

SHAP is based on Shapley values from cooperative game theory and decomposes a model prediction into additive feature contributions [8]. For a given model f and an input region x , the prediction can be expressed as:

$$f(x) = \phi_0 + \sum_{k=1}^K \phi_k \quad (27)$$

where ϕ_0 is the baseline prediction, K is the number of input features, and ϕ_k represents the SHAP contribution of feature k to the prediction.

In this study, SHAP values were computed separately for each XGBoost model associated with each fuzzy cluster. Therefore, for cluster j , the predicted membership can be interpreted as:

$$\hat{u}_j(x) = \phi_{0,j} + \sum_{k=1}^K \phi_{k,j} \quad (28)$$

where $\phi_{k,j}$ represents the contribution of feature k to the predicted membership of a region to cluster j .

The interpretation of SHAP values is direct. A positive SHAP value indicates that a feature increases the predicted membership to a given cluster, whereas a negative SHAP value indicates that the feature decreases that membership. The magnitude of the SHAP value reflects the strength of the contribution.

This analysis allows the fuzzy clusters to be interpreted in terms of physically meaningful variables. For example, SHAP can determine whether the balance between high values of P_{mean} , P_{std} , or E_{energy} increase membership to a high-polarization regime, or whether the relative balance $\log(B/E)$ drives membership toward E -dominated or B -enhanced regimes.

Global SHAP importance values were obtained by averaging the absolute SHAP values across regions:

$$I_k^{(j)} = \frac{1}{N} \sum_{i=1}^N |\phi_{k,j}^{(i)}| \quad (29)$$

where $I_k^{(j)}$ represents the global importance of feature k for cluster j , and $\phi_{k,j}^{(i)}$ is the SHAP contribution of feature k for region i in the membership model of cluster j .

In addition to global feature importance, SHAP beeswarm plots were used to analyze the direction and distribution of feature effects. This makes it possible to determine not only which variables are important, but also whether high or low values of those variables increase membership to each fuzzy polarization regime.

Thus, SHAP provides a physically interpretable explanation of the fuzzy segmentation by linking each cluster membership to specific regional polarization descriptors.

3.8. Ambiguity and Entropy Analysis

Since Fuzzy C-Means assigns each region a membership degree to every cluster, the resulting segmentation contains more information than a hard cluster label. This allows the degree of ambiguity of each region to be quantified and interpreted [32]. Regions with a dominant membership close to one can be considered clearly associated with a specific polarization regime, whereas regions with memberships distributed across several clusters can be interpreted as transitional or mixed regions.

For each region i , the maximum membership value was defined as:

$$u_{\max}(i) = \max_j u_{ij} \quad (30)$$

where u_{ij} is the membership degree of region i to cluster j . This value measures how strongly a region belongs to its dominant cluster.

Based on $u_{\max}(i)$, regions were classified into three ambiguity levels: clear, $u_{\max}(i) \geq 0.80$; moderate, $0.60 \leq u_{\max}(i) < 0.80$ and ambiguous, $u_{\max}(i) < 0.60$. This classification provides a simple way to distinguish regions with well-defined cluster membership from regions located in transitional areas of the fuzzy feature space.

In addition to maximum membership, fuzzy entropy was computed to quantify how the memberships are distributed across clusters. For each region i , entropy was defined as:

$$H_i = - \sum_{j=1}^c u_{ij} \log(u_{ij}) \quad (31)$$

where c is the number of clusters. Low entropy indicates that most of the membership is concentrated in one cluster, whereas high entropy indicates that the memberships are more evenly distributed across several clusters.

To make entropy values comparable across models with a fixed number of clusters, the entropy was normalized by its theoretical maximum:

$$H_i^{\text{norm}} = \frac{H_i}{\log(c)} \quad (32)$$

The normalized entropy satisfies, $0 \leq H_i^{\text{norm}} \leq 1$. Values close to zero indicate regions with clear membership to one dominant regime, whereas values close to one indicate highly ambiguous regions with memberships distributed across multiple regimes.

This ambiguity and entropy analysis is relevant for the interpretation of the CMB polarization field because it allows the identification of regions that do not belong sharply to a single polarization regime. Such regions may represent gradual transitions between high- and low-polarization areas, or between E -dominated and B -enhanced intermediate regimes. Therefore, ambiguity is not treated as a limitation of the model, but as an informative property of the fuzzy regional structure.

3.9. Controlled Perturbation and Stability Analysis

To evaluate the robustness of the fuzzy regional structure, controlled perturbations were introduced into two key dimensions of the regional feature space: the polarization amplitude and the relative balance between B - and E -mode energies. The objective of this stage was to analyze whether small variations in physically meaningful variables could modify the membership structure or induce transitions between polarization regimes.

Two types of perturbation were considered. First, multiplicative perturbation was applied to the regional polarization amplitude features. For each region r , the perturbed features were defined as:

$$P'_{\text{mean}}(r) = P_{\text{mean}}(r)(1 + \epsilon_P) \quad (33)$$

$$P'_{\text{std}}(r) = P_{\text{std}}(r)(1 + \epsilon_P) \quad (34)$$

where ϵ_P is a random perturbation term controlling the relative variation in polarization amplitude.

Second, an additive perturbation was applied to the relative B/E balance:

$$\log(B/E)'(r) = \log(B/E)(r) + \epsilon_{BE} \quad (35)$$

where ϵ_{BE} represents a controlled perturbation in the logarithmic ratio between B - and E -mode energies.

For each perturbed scenario, the modified regional feature vectors were standardized using the same scaling parameters obtained from the baseline data. The perturbed regions

were then projected onto the original Fuzzy C-Means cluster structure, keeping the original centroids fixed. This strategy ensures that all comparisons are made with respect to the same baseline polarization regimes.

Let U_i denote the original membership vector of region i , and let U'_i denote the membership vector obtained after perturbation. The total membership variation was measured using the L_1 distance:

$$\Delta U_i = \|U'_i - U_i\|_1 \quad (36)$$

Since the maximum possible L_1 distance between two membership vectors is 2, a normalized stability index was defined as:

$$S_i = 1 - \frac{\Delta U_i}{2} \quad (37)$$

Values of S_i close to 1 indicate high stability, whereas lower values indicate stronger membership variation under perturbation.

In addition to membership variation, dominant-cluster switching was evaluated. Let C_i be the dominant cluster of regions i in the baseline scenario, and C'_i the dominant cluster after perturbation. A cluster switch occurs when $C'_i \neq C_i$. The switch rate was then computed as the proportion of regions whose dominant cluster changed after perturbation.

The change in fuzzy entropy was also analyzed:

$$\Delta H_i = H'_i - H_i \quad (38)$$

where H_i and H'_i are the normalized fuzzy entropies before and after perturbation, respectively. Positive values of ΔH_i indicate that the region becomes more ambiguous under perturbation, whereas negative values indicate that the region becomes more clearly associated with a dominant cluster.

Finally, transition matrices were constructed to identify the direction of cluster changes. For baseline cluster a and perturbed cluster b , the transition matrix was defined as:

$$T_{ab} = \#\{i: C_i = a, C'_i = b\} \quad (39)$$

Each element T_{ab} counts the number of regions that move from cluster a in the baseline scenario to cluster b after perturbation. Row-normalized transition matrices were used to compare cluster-specific stability and to identify the most frequent transitions between polarization regimes.

This perturbation analysis provides a controlled counterfactual assessment of the regional CMB polarization structure. Rather than retraining the clustering model for each perturbation, the original fuzzy regimes are kept fixed and perturbed regions are evaluated against them. This allows the identification of robust regions, sensitive regions, and transitions between physically interpretable polarization regimes.

3.10. Physical and Counterfactual Interpretation

The final methodological stage integrates the fuzzy segmentation, the XAI explanations, and the perturbation-based stability analysis into a physical and counterfactual interpretation of the regional CMB polarization structure.

The fuzzy memberships obtained from Fuzzy C-Means allow each region to be interpreted as expressing different degrees of association with several polarization regimes. This is important because the CMB polarization field is spatially continuous, and transitions between regimes may occur gradually rather than through sharp boundaries.

Within this framework, regions with high stability under perturbation are interpreted as robust regions. These are areas whose membership profile and dominant cluster remain largely unchanged when small variations are introduced into the feature space.

Such regions can be understood as structurally stable components of the regional polarization pattern.

In contrast, regions that show large membership variations, increased entropy, or dominant-cluster switching are interpreted as sensitive regions. These areas are especially relevant because they may represent transitional zones between polarization regimes or regions where the balance between physical descriptors is more delicate.

The transition matrices provide an additional level of interpretation by identifying the direction of changes between clusters. This makes it possible to distinguish whether perturbations in polarization amplitude mainly induce transitions between low-, medium-, and high-polarization regimes, or whether perturbations in the B/E balance mainly affect the boundary between E -dominated and B -enhanced regimes.

The perturbation analysis is interpreted as a controlled counterfactual experiment. In this context, counterfactual configurations are treated as alternative regional configurations generated under small, controlled variations in physically meaningful CMB polarization variables.

This distinction is important. The proposed framework provides a data-driven and interpretable way to study the informational stability of the observed CMB polarization pattern. In other words, the method evaluates how much the regional fuzzy structure changes when specific polarization features are perturbed.

From this perspective, the combination of fuzzy memberships, SHAP explanations, and perturbation-based stability metrics allows the CMB polarization field to be analyzed as a globally structured but locally sensitive system. Robust regions indicate stable components of the observed pattern, while sensitive regions reveal where small variations may generate alternative regional configurations within the same fuzzy feature space.

Thus, the physical and counterfactual interpretation stages provide the bridge between the computational results and their cosmological meaning. It supports a cautious interpretation of the CMB regional structure as an explainable, fuzzy, and stability-aware representation of primordial polarization patterns.

4. Results

This section presents the results obtained from the proposed FAS-XAI framework for the regional analysis of CMB polarization. The analysis follows the methodological pipeline described above: initial inspection of the Planck PR3 SMICA IQU product, construction of polarization-derived variables, regional feature extraction, fuzzy segmentation, supervised membership modeling, SHAP-based interpretation, ambiguity analysis, and controlled perturbation assessment. Together, these stages allow the CMB polarization field to be analyzed as a regional, fuzzy, explainable, and stability-aware structure. The complete pipeline is provided as Supplementary Material.

4.1. Initial Inspection of the Planck SMICA IQU Map

The first stage consisted of inspecting the structure of the Planck PR3 SMICA CMB IQU map. The FITS file contains ten fields, including the CMB intensity, the Stokes polarization components, the corresponding masks, and their inpainted counterparts. The original map is provided at $NSIDE = 2048$, corresponding to 50,331,648 pixels and an approximate angular resolution of 1.72 arcmin. Table 2 summarizes the fields available in the data product.

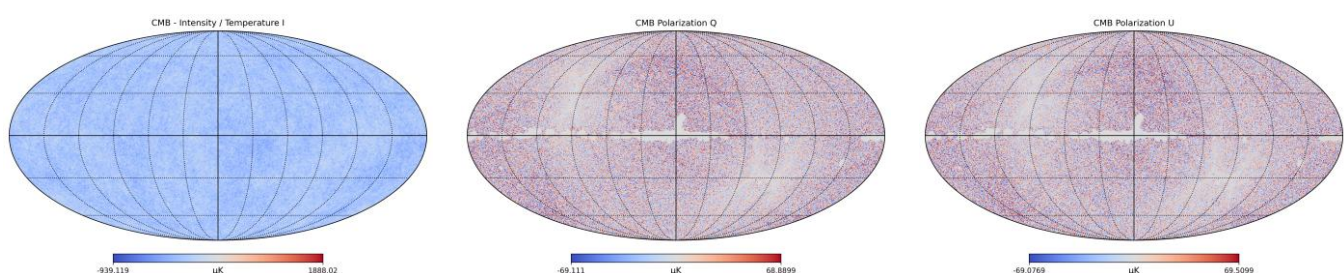
Table 2. Fields included in the Planck PR3 SMICA CMB IQU map.

Field	Name	Description
0	I_STOKES	CMB intensity/temperature field
1	Q_STOKES	Stokes (Q) polarization component
2	U_STOKES	Stokes (U) polarization component
3	TMASK	Temperature analysis mask
4	PMASK	Polarization analysis mask
5	I_STOKES_INP	Inpainted intensity field
6	Q_STOKES_INP	Inpainted (Q) polarization component
7	U_STOKES_INP	Inpainted (U) polarization component
8	TMASKINP	Inpainting temperature mask
9	PMASKINP	Inpainting polarization mask

The analysis focused on the non-inpainted Stokes components Q_STOKES and U_STOKES, together with the polarization mask PMASK. The temperature field I_STOKES was retained for initial descriptive inspection, but it was not included in the final fuzzy segmentation, which was based on polarization-derived variables.

The initial descriptive statistics showed that the temperature field has a substantially larger amplitude than the polarization components. In the original maps, the standard deviation of I_STOKES was approximately 1.08×10^{-4} K, while the standard deviations of Q_STOKES and U_STOKES were approximately 2.83×10^{-5} K and 2.84×10^{-5} K, respectively. This confirms that the polarization signal is weaker than the temperature anisotropy signal and motivates careful treatment of masks, resolution, and regional aggregation. The polarization mask PMASK had a mean value of approximately 0.929, indicating that about 92.9% of the high-resolution pixels were valid for polarization analysis before regional filtering.

Figure 4 shows the initial full-sky maps of the CMB intensity and the Stokes polarization components Q and U . A prominent central band associated with the Galactic plane is visible, reflecting foreground contamination from the Milky Way. This motivates the use of the polarization mask before deriving P , E , and B , and before applying regional fuzzy clustering.

**Figure 4.** Initial Planck SMICA CMB intensity and polarization maps at $NSIDE = 2048$.

The panels show the CMB intensity field I and the Stokes polarization components Q and U . The central Galactic band illustrates the need for mask-controlled analysis before constructing polarization-derived variables.

4.2. Polarization-Derived Maps and Resolution Assessment

The second stage of the analysis consisted of constructing physically meaningful polarization variables from the Stokes components Q and U . From these maps, the polarization amplitude P was computed, and the scalar E - and B -mode polarization maps were obtained through harmonic decomposition.

Figure 5 shows the resulting polarization-derived maps. The P map is strictly non-negative and represents the total amplitude of linear polarization across the sky. By contrast, the E - and B -mode maps retain signed fluctuations around zero and describe complementary geometric components of the polarization field: E captures the gradient-like structure, whereas B represents the curl-like component.

Their global statistics are consistent with these definitions. The P map has a positive mean value, while the E - and B -mode maps are approximately centered around zero. Although the global dispersions of E and B are of similar magnitude, their role in the proposed framework is not interpreted only through full-sky statistics, but mainly through their regional energy and relative balance.

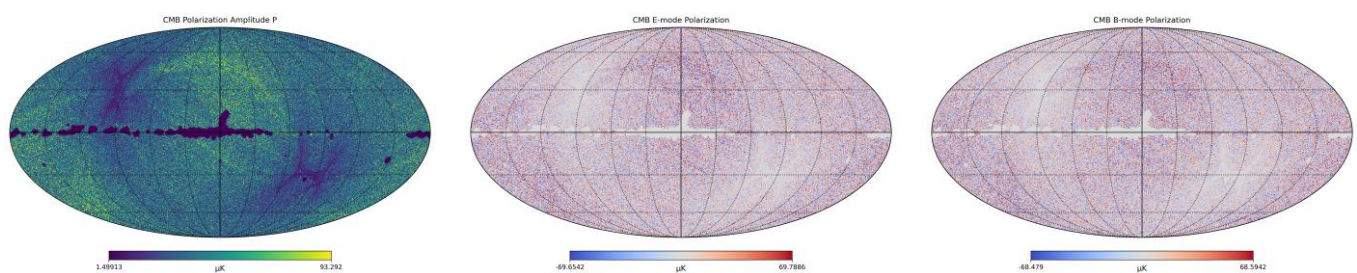


Figure 5. Derived CMB polarization maps.

The panels show the polarization amplitude P , the E -mode map, and the B -mode map derived from the Planck SMICA polarization fields.

A resolution stability analysis was also performed to assess how the dispersion of temperature and polarization fields changes under HEALPix degradation. Figure 6 summarizes the standard deviation of the temperature field and the Stokes polarization components across different resolutions. The temperature field retains most of its variability as resolution decreases, whereas the polarization components Q and U show a much stronger reduction in standard deviation. The two polarization curves are almost fully superposed, indicating that Q and U behave nearly identically across resolutions.

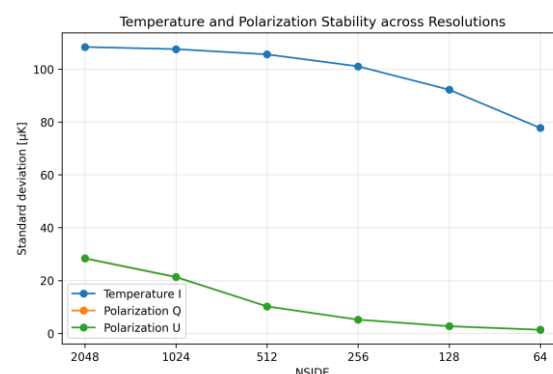


Figure 6. Temperature and polarization stability across HEALPix resolutions.

This result indicates that polarization contains fine-scale structure that is rapidly smoothed under degradation. For this reason, the original $NSIDE = 2048$ resolution was retained during the construction of P , E , and B . Regional aggregation was applied only afterwards, so that the analysis preserved high-resolution polarization information while still producing a manageable regional representation for clustering.

The plot shows the standard deviation of the temperature field and the Stokes polarization components Q and U across progressively degraded HEALPix resolutions. The Q and U curves are nearly superposed.

4.3. Regional Aggregation and Feature Extraction

After constructing the polarization-derived maps, the analysis was reformulated at the regional level using a HEALPix grid with $NSIDE_{\text{region}} = 64$. This grid produced 49,152 regional cells. After applying the polarization mask, regional descriptors could be computed for 47,082 regions. A second quality-control step retained only regions with a valid-pixel fraction greater than or equal to 0.8, resulting in a final clean dataset of 45,020 regions. All subsequent statistical summaries, correlation analyses, and fuzzy clustering were performed on this cleaned regional feature matrix.

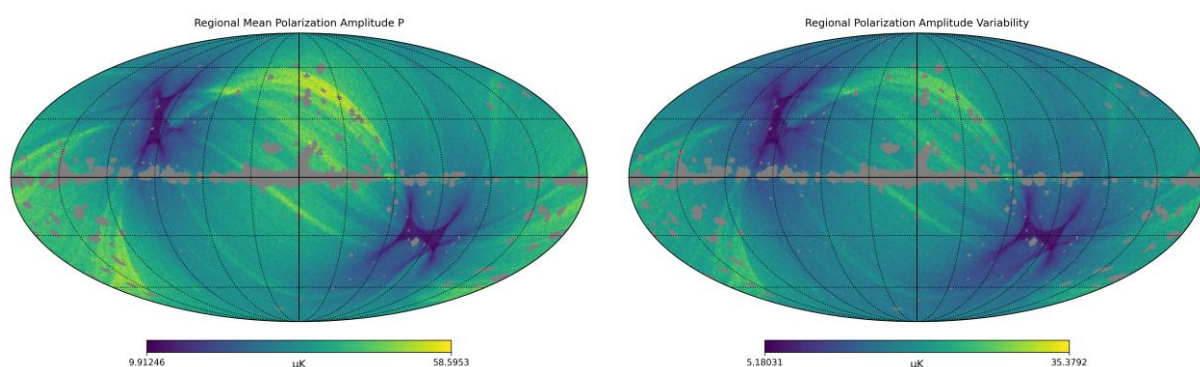
Regional descriptors were computed from the P , E , and B maps. The initial feature set included the valid-pixel fraction, the number of valid pixels, amplitude descriptors of P , energy-based descriptors for P , E , and B , and the relative balance between B - and E -mode energy. Table 3 summarizes the descriptive statistics of the cleaned regional descriptors.

Table 3. Descriptive statistics of regional polarization descriptors after valid-region filtering.

	P_{mean}	P_{std}	P_{energy}	E_{std}	E_{energy}	B_{std}	B_{energy}	$\log(B/E)_{\text{energy}}$
count	45,020	45,020	45,020	45,020	45,020	45,020	45,020	45,020
mean	35.11	18.37	1629.40	28.27	831.33	27.69	798.76	−0.048
std	6.81	3.58	589.62	5.43	299.42	5.54	297.89	0.114
min	9.91	5.18	127.58	8.41	74.50	6.75	46.13	−0.705
25%	31.50	16.46	1263.92	25.35	645.06	24.73	612.69	−0.123
50%	35.55	18.58	1608.78	28.58	819.08	28.03	786.88	−0.047
75%	39.37	20.62	1974.03	31.67	1005.90	31.18	973.87	0.029
max	58.60	35.38	4328.18	48.08	2314.07	47.64	2271.16	0.441

The retained regions showed high valid-pixel coverage, with a median valid fraction of 1.000 and a median of 1024 valid pixels per region. The selected regional descriptors also showed coherent behavior: P_{mean} and P_{std} summarized the regional amplitude and variability of the polarization field, while E_{energy} and B_{energy} captured the energy of the scalar polarization components. The mean $\log(B/E)$ value was slightly negative, indicating a modest overall dominance of E -mode energy over B -mode energy in the retained regional dataset.

Figure 7 shows the spatial distribution of the main regional descriptors; gray areas indicate masked or invalid regions excluded from the analysis. The maps reveal coherent large-scale structures in polarization amplitude and energy. Regions with higher P_{mean} and P_{std} correspond to stronger and more variable linear polarization, whereas E_{energy} captures the regional intensity of the gradient-like component. The B/E energy ratio provides complementary information about the local balance between curl-like and gradient-like contributions. This ratio was subsequently transformed into $\log(B/E)$, which was used as the final balance feature because it provides a more stable and symmetric representation of the relative B - to E -mode contribution.



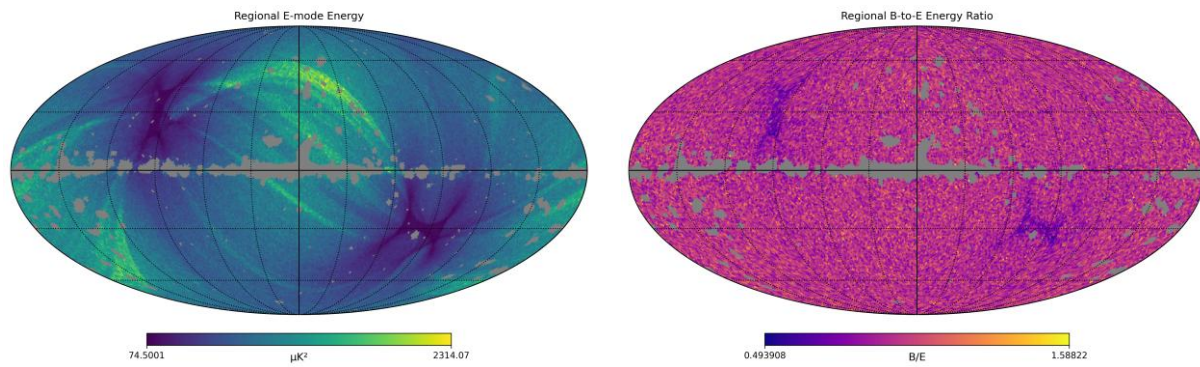


Figure 7. Spatial distribution of regional CMB polarization descriptors.

To reduce redundancy and preserve interpretability, a correlation analysis was performed on the candidate regional features. As shown in Figure 8, several amplitude- and energy-related descriptors were strongly correlated, indicating that including all candidate variables would introduce redundant information into the clustering process.

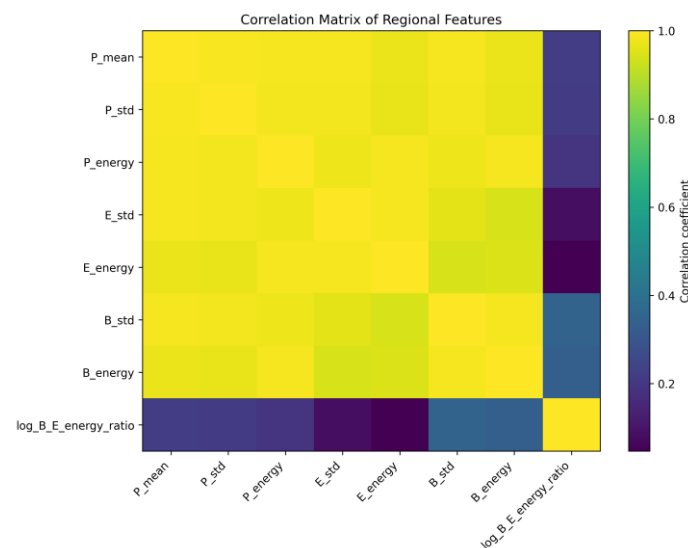


Figure 8. Correlation matrix of candidate regional polarization features.

Based on this analysis, the final feature space was restricted to four variables:

$$P_{\text{mean}}, P_{\text{std}}, E_{\text{energy}}, \log(B/E)$$

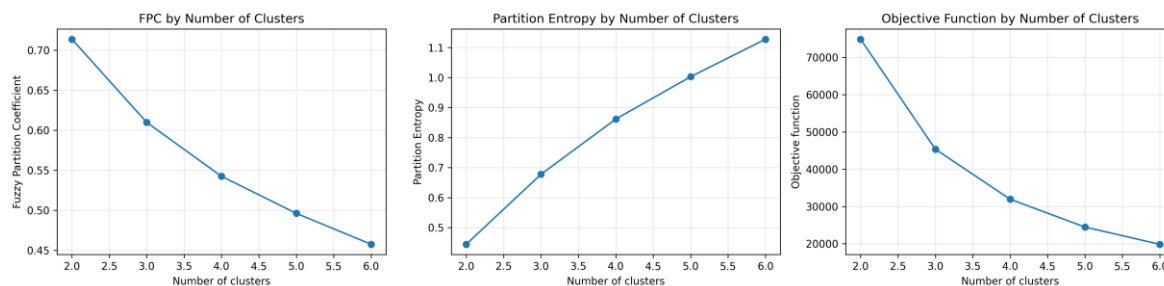
These variables were selected because they summarize complementary aspects of the regional polarization field: mean amplitude, internal variability, E -mode energy, and the relative B/E balance. This compact feature set was used as input for the subsequent Fuzzy C-Means segmentation.

4.4. Fuzzy C-Means Segmentation of Regional Polarization Regimes

Before applying Fuzzy C-Means, the selected regional features were standardized using z-score normalization, since P_{mean} , P_{std} , E_{energy} , and $\log(B/E)$ have different physical units and numerical ranges. These standardized variables defined the input space for the clustering process. Fuzzy C-Means was applied considering different numbers of clusters: $c = 2, 3, 4, 5, 6$. For each solution, the objective function, the Fuzzy Partition Coefficient (FPC), and the Partition Entropy (PE) were evaluated. The results are summarized in Table 4 and Figure 9.

Table 4. Fuzzy C-Means performance metrics for different numbers of clusters.

Number of Clusters	Objective	Iterations	FPC	PE
2	74,856.48	28	0.7136	0.4443
3	45,349.31	39	0.6098	0.6778
4	31,925.76	119	0.5425	0.8618
5	24,470.89	269	0.4963	1.0028
6	19,853.02	300	0.4576	1.1278

**Figure 9.** Fuzzy C-Means validation metrics for different numbers of clusters.

The panels show the Fuzzy Partition Coefficient, Partition Entropy, and objective function for solutions with $c = 2, \dots, 6$ clusters.

To further assess the robustness of the selected number of clusters, an additional stability analysis was performed using 30 random initializations for each candidate solution with $c = 3, 4, 5, 6$. The four-cluster solution showed very high stability, with mean $ARI = 0.9997$ and mean $NMI = 0.9992$ across random initializations. Although the three and five cluster solutions were also stable, the former merged the two intermediate E/B -balance regimes, whereas the latter fragmented existing regimes without adding a clearly distinct physical interpretation. The six-cluster solution showed a clear loss of stability, with mean $ARI = 0.6905$ and mean $NMI = 0.7771$. Therefore, $c = 4$ was retained as the most suitable compromise between stability, parsimony, and physical interpretability.

As expected, the objective function decreased as the number of clusters increased, while FPC decreased and PE increased. This indicates that solutions with more clusters improved the numerical fit but also produced more diffuse and less sharply defined memberships. The largest reduction in the objective function occurred between two and three clusters, whereas improvements beyond four clusters showed diminishing returns. The four-cluster solution was therefore selected as a compromise between goodness of fit, model complexity, and physical interpretability.

The centroids of the selected four-cluster solution were transformed back to the original feature scale to support physical interpretation. Table 5 and Figure 10 summarize the centroid profiles.

Table 5. Centroids of the four-cluster FCM solution and their physical interpretation.

Cluster	P_{mean}	P_{std}	E_{energy}	$\log(B/E)$	B/E	Interpretation
0	43.19	22.64	1217.50	−0.036	0.964	High-polarization structured regions
1	35.29	18.46	846.31	−0.117	0.889	E -dominated medium-polarization regions
2	35.02	18.31	770.44	0.041	1.042	B -enhanced medium-polarization regions
3	24.80	12.98	422.17	−0.098	0.907	Low-polarization regions

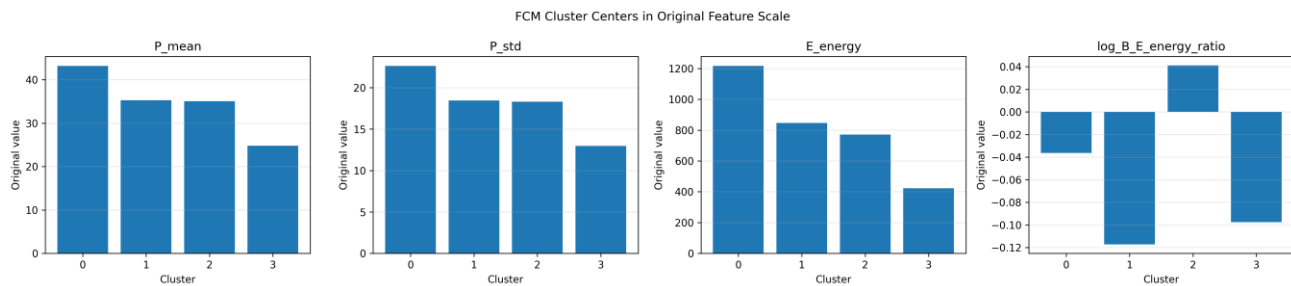


Figure 10. FCM cluster centers in the original feature scale.

The four centroids define distinct polarization regimes. Cluster 0 represents high-polarization structured regions, characterized by the highest values of P_{mean} , P_{std} , and E_{energy} . Cluster 3 represents low-polarization regions, with the lowest amplitude and energy values. Clusters 1 and 2 correspond to medium-polarization regimes with similar amplitude levels but different relative B/E balance. Cluster 1 is more clearly E -dominated, while Cluster 2 shows a relatively enhanced B -mode contribution. This latter interpretation must remain cautious: a higher relative B -mode contribution does not imply primordial B -mode detection, but may reflect lensing, residual foregrounds, noise, leakage effects, or complex polarization structure.

The four-cluster solution also produced a relatively balanced partition of the valid sky regions, as shown in Table 6. No cluster was residual or marginal, and all groups represented a relevant fraction of the retained regional dataset.

Table 6. Cluster size and dominant-membership summary for the four-cluster FCM solution.

Cluster	Number of Regions	Fraction	Mean Max. Membership	Median Max. Membership
0	10,293	0.2286	0.6828	0.6802
1	13,124	0.2915	0.6603	0.6496
2	13,465	0.2991	0.6646	0.6570
3	8138	0.1808	0.6859	0.6893

The mean maximum membership ranged from approximately 0.66 to 0.69, indicating that regions were not assigned rigidly to a single group. Instead, many regions exhibited partial membership to multiple regimes, which is consistent with the continuous nature of CMB polarization.

Figure 11 shows the dominant FCM cluster map, where each region is assigned to the cluster with the highest membership degree. The spatial distribution is not random, but displays extended areas dominated by specific polarization regimes, suggesting that the fuzzy segmentation captures coherent regional structure. Each valid region is assigned to the cluster with the highest membership degree.

Figure 12 shows the membership maps for the four clusters separately. These maps provide richer information than the dominant-cluster map because they show the degree to which each region belongs to each polarization regime. The resulting patterns reveal gradual transitions, mixed memberships, and ambiguous regions across the sky.

The panels show the membership degree of each region to the four FCM clusters: high-polarization structured regions, E -dominated medium-polarization regions, B -enhanced medium-polarization regions, and low-polarization regions.

Overall, the FCM segmentation provides an interpretable fuzzy partition of the CMB polarization field. The four regimes are physically distinguishable in terms of polarization amplitude, internal variability, E -mode energy, and relative B/E balance, while the membership maps support the use of fuzzy rather than hard segmentation.

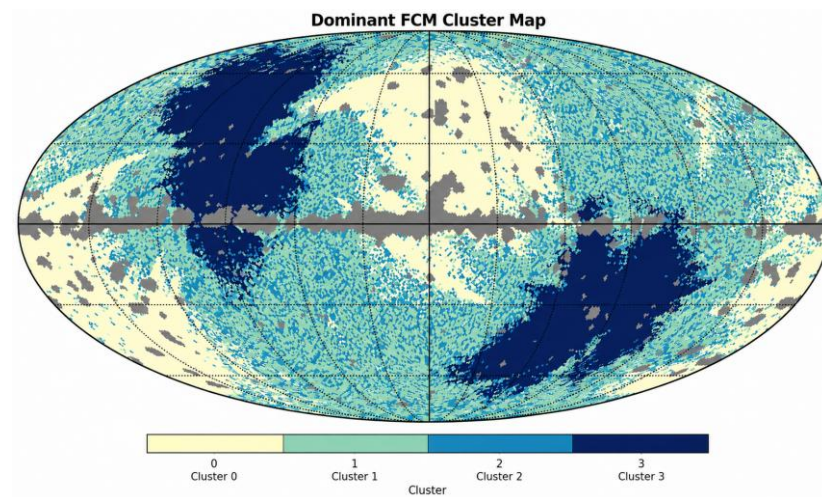


Figure 11. Dominant FCM cluster assignment map.

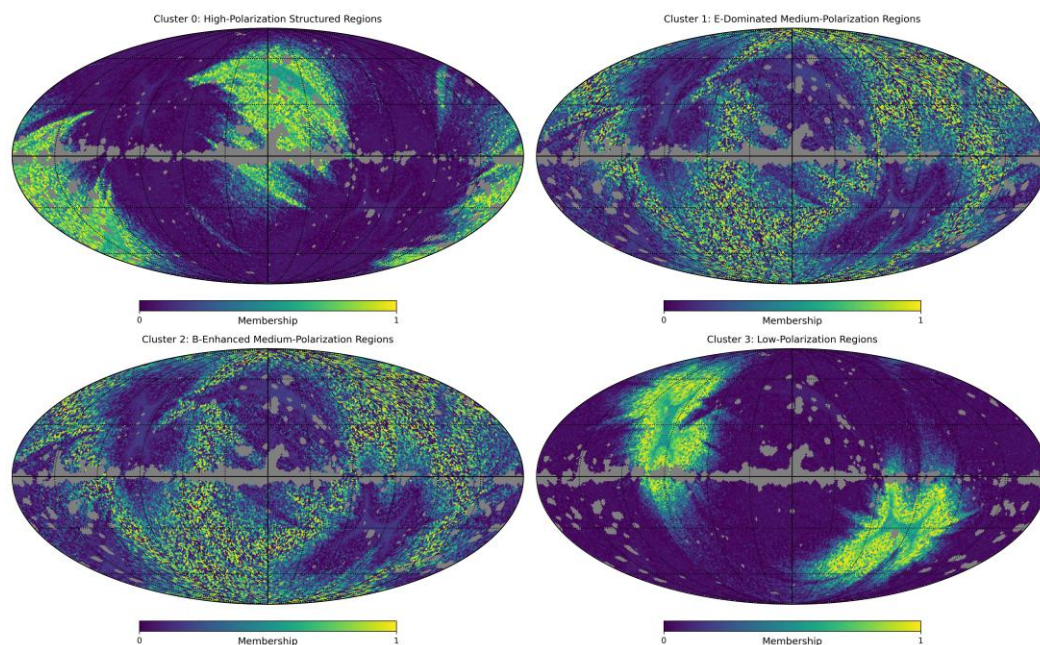


Figure 12. FCM membership maps for the four polarization regimes.

4.5. Ambiguity and Entropy Analysis of Fuzzy Memberships

The fuzzy nature of the FCM segmentation allows the uncertainty of each regional assignment to be analyzed through its maximum membership value, $u_{\max}$. The thresholds used to define clear, moderate, and ambiguous regions were adopted as operational interpretability criteria rather than as learned physical boundaries. Regions were classified into three ambiguity levels: clear when $u_{\max} \geq 0.80$, moderate when $0.60 \leq u_{\max} < 0.80$, and ambiguous when $u_{\max} < 0.60$. These categories distinguish regions strongly associated with one cluster from transitional regions with mixed membership profiles, Table 7.

Table 7. Distribution of regions by ambiguity level.

Ambiguity Level	Number of Regions	Fraction
clear	12,327	0.2738
moderate	15,494	0.3442
ambiguous	17,199	0.3820

Only 27.38% of the regions were clearly assigned to a single polarization regime, while 34.42% were moderate and 38.20% were ambiguous. This confirms that a substantial fraction of the valid sky cannot be adequately described by a rigid cluster assignment.

Figure 13 shows the spatial distribution of ambiguity levels. Ambiguous regions appear across extended areas of the valid sky. These regions can be interpreted as transition zones where the regional polarization descriptors do not strongly favor a single centroid.

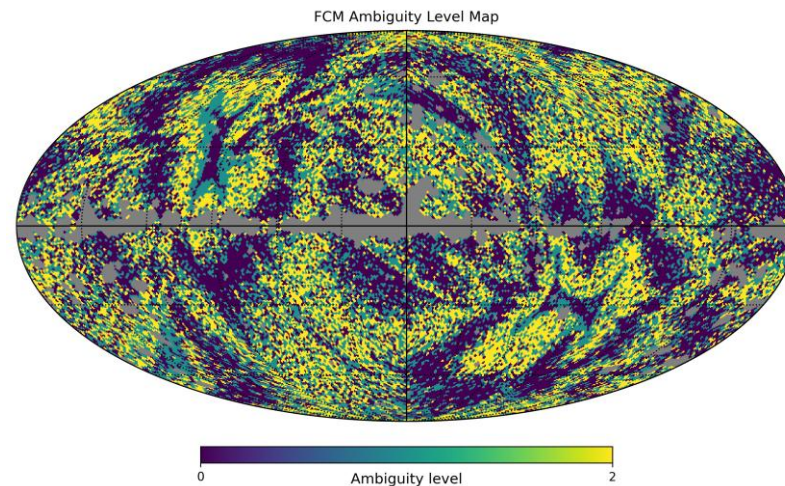


Figure 13. FCM ambiguity level map.

The map shows regions classified as clear, moderate, or ambiguous according to their maximum FCM membership.

To complement this categorical classification, normalized membership entropy was computed for each region, Figures 14 and 15. Entropy provides a continuous measure of how distributed the membership vector is across clusters: low values indicate concentrated membership, while high values indicate mixed or transitional profiles. The normalized entropy had a mean value of 0.622, a standard deviation of 0.226, and a median value of 0.685, confirming the presence of both clearly defined and highly mixed regions.

The map shows the normalized entropy of the FCM membership vector for each valid region. Higher values indicate greater membership dispersion across clusters.

The selected physical descriptors were then compared across ambiguity levels. As shown in Table 8, their mean values were very similar for clear, moderate, and ambiguous regions. This indicates that ambiguity is not driven by systematically higher or lower polarization amplitude, stronger *E*-mode energy, or a clearly different *B/E* balance. Instead, ambiguity appears to arise from the multivariate position of a region relative to several cluster centroids.

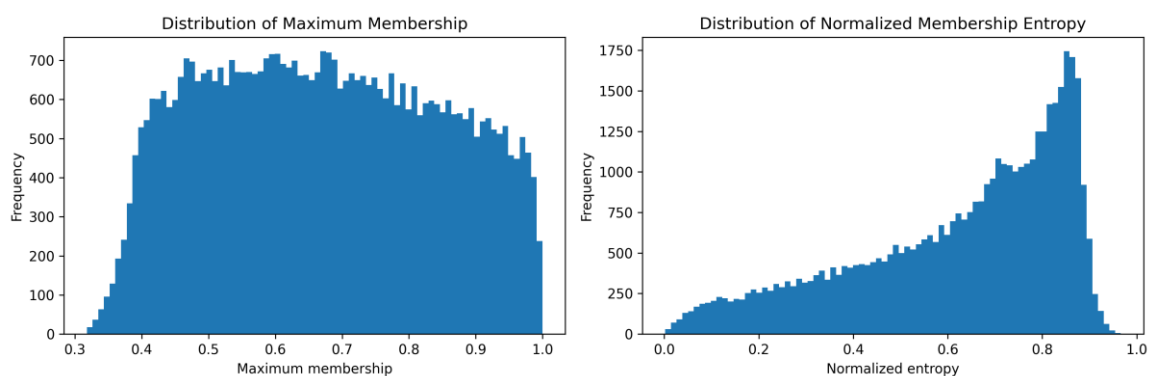


Figure 14. Distributions of maximum membership and normalized membership entropy.

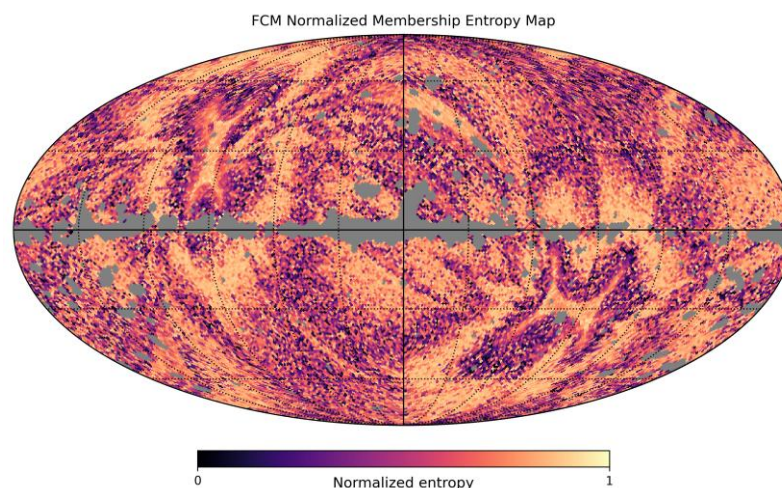


Figure 15. Normalized membership entropy map.

Table 8. Mean regional features by ambiguity level.

Ambiguity Level	P_{mean}	P_{std}	E_{energy}	$\log(B/E)$	u_{max}	H_{norm}
clear	34.99	18.30	824.45	−0.0479	0.8926	0.3075
moderate	35.22	18.44	840.49	−0.0475	0.6977	0.6444
ambiguous	35.10	18.37	828.02	−0.0487	0.4891	0.8262

The fuzzy metrics validate this interpretation. Clear regions have high maximum membership and low normalized entropy, whereas ambiguous regions have lower maximum membership and higher entropy. Therefore, ambiguous regions should not be interpreted as a separate fifth polarization regime, but as transitional or mixed regions where the selected descriptors combine characteristics of several clusters.

This ambiguity analysis is relevant for the subsequent XAI and perturbation stages, since regions with distributed memberships are natural candidates for sensitivity analysis. In this sense, ambiguity is not treated as a weakness of the model, but as an informative property of the fuzzy regional structure of the CMB.

4.6. Supervised Membership Modeling with XGBoost

After the fuzzy segmentation stage, a supervised modeling layer was introduced to evaluate whether the FCM memberships could be accurately reconstructed from the selected regional polarization features. For this purpose, one XGBoost regression model was trained for each fuzzy cluster.

The objective of this stage was to approximate the continuous membership functions generated by Fuzzy C-Means. Each model therefore learned the mapping between the selected regional descriptors and the degree of membership associated with one specific cluster.

The regional dataset was divided into training and test subsets using an 80/20 random split with a fixed random seed (random_state = 42). Since the final regional dataset contained 45,020 regions, approximately 36,016 regions were used for training and 9004 regions for testing. No separate validation subset was used because the XGBoost models were not optimized through an extensive hyperparameter search; the same fixed configuration was applied to the four membership-regression models. Stratified sampling was not applied because the supervised task was formulated as a regression problem over continuous fuzzy membership degrees rather than as a classification problem over discrete labels.

The predictive performance of each XGBoost model was evaluated using R^2 , MAE, and RMSE. The results are summarized in Table 9 and Figure 16.

Table 9. XGBoost performance metrics for fuzzy membership prediction.

Cluster	Cluster_Name	R^2	MAE	RMSE
0	High-Polarization Structured Regions	0.997060	0.010378	0.014755
1	<i>E</i> -Dominated Medium-Polarization Regions	0.983661	0.025108	0.033189
2	<i>B</i> -Enhanced Medium-Polarization Regions	0.984745	0.025030	0.032878
3	Low-Polarization Regions	0.996945	0.009461	0.014319

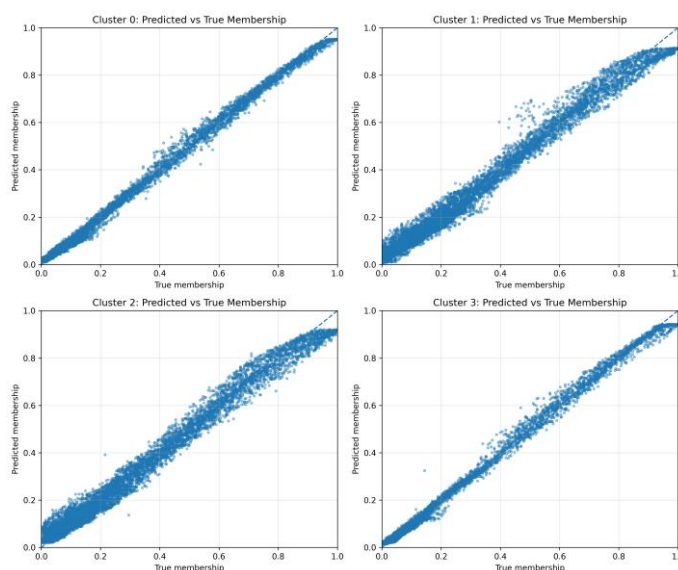
**Figure 16.** XGBoost performance metrics by cluster.

The panels show R^2 , MAE, and RMSE for the four XGBoost regression models trained to predict FCM membership degrees.

The results show very high predictive performance for all four membership models. The R^2 values ranged from 0.9837 to 0.9971, indicating that the selected regional variables explain nearly all the variability in the fuzzy memberships. The low MAE and RMSE values further confirm that the XGBoost models closely approximate the membership functions obtained from Fuzzy C-Means.

Clusters 0 and 3 achieved the best performance, with R^2 values above 0.996 and the lowest error levels. This is consistent with their interpretation as more clearly defined extreme regimes, namely high-polarization structured regions and low-polarization regions. By contrast, Clusters 1 and 2 showed slightly lower, although still very high, predictive performance. This is expected because both correspond to medium-polarization regimes and are mainly differentiated by the relative balance between *B*- and *E*-mode energy, which leads to greater overlap in feature space.

Figure 17 provides a visual validation of the supervised approximation by comparing predicted and true FCM memberships for the four clusters.

**Figure 17.** Predicted versus true FCM memberships for the XGBoost models.

Each panel compares the membership values obtained from Fuzzy C-Means with the membership values predicted by the corresponding XGBoost regression model. The dashed diagonal represents perfect agreement.

In all four clusters, the points are concentrated close to the diagonal line, indicating strong agreement between the original FCM memberships and the predicted values, $\hat{u}_j \approx u_j$. The alignment is particularly strong for Clusters 0 and 3, whereas Clusters 1 and 2 show slightly greater dispersion, especially at intermediate membership values. This pattern is again consistent with the stronger overlap between the two medium-polarization regimes.

Overall, the XGBoost layer provides a reliable supervised approximation of the fuzzy memberships. This result is important because the SHAP analysis in the next stage is applied to the XGBoost models. Therefore, before interpreting feature contributions, it is necessary to verify that these supervised models faithfully reproduce the membership structure generated by Fuzzy C-Means. The high predictive performance obtained here supports the use of XGBoost as an auxiliary explanatory model within the proposed FAS-XAI framework.

4.7. SHAP-Based Interpretation of Fuzzy Membership Models

After validating the XGBoost membership models, SHAP values were computed to explain how each regional feature contributed to the predicted membership of each fuzzy cluster. This stage provides the explainable layer of the proposed FAS-XAI framework, allowing the fuzzy segmentation to be interpreted in terms of physically meaningful CMB polarization descriptors.

The SHAP analysis was applied separately to each XGBoost regression model. The analysis explains the continuous membership degree of each region to each polarization regime. Global feature importance was quantified using the mean absolute SHAP value, which measures the average magnitude of each feature contribution to the predicted membership, Figure 18. The direction of the effect was then analyzed using SHAP beeswarm plots.

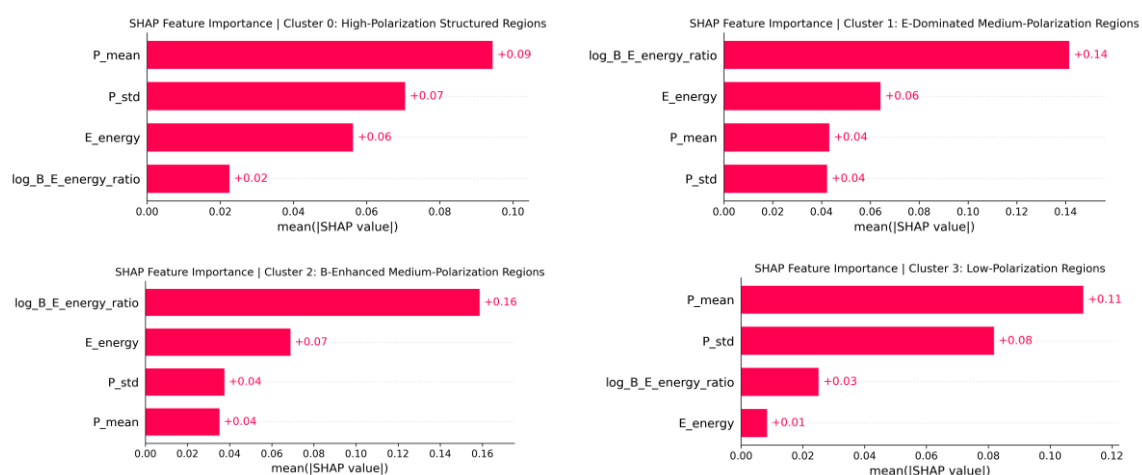


Figure 18. SHAP feature importance for the four fuzzy polarization regimes.

The panels show the mean absolute SHAP value of each feature for the XGBoost models associated with the four FCM clusters. Higher values indicate stronger average contribution to the predicted membership degree.

The global SHAP importance patterns were consistent with the physical interpretation of the FCM centroids, Table 10. For Cluster 0, corresponding to high-polarization structured regions, the most relevant variables were P_{mean} , P_{std} , and E_{energy} . This

indicates that membership to this regime is mainly explained by polarization amplitude, internal variability, and E -mode energy.

For Cluster 3, corresponding to low-polarization regions, the dominant variables were P_{mean} and P_{std} . Therefore, this regime is primarily governed by amplitude-related descriptors. By contrast, Clusters 1 and 2 were mainly explained by $\log(B/E)$, showing that the main distinction between the two medium-polarization regimes is not total polarization amplitude, but the relative balance between B - and E -mode energy.

Table 10. Summary of dominant SHAP features by fuzzy cluster.

Cluster	Physical Interpretation	Dominant SHAP Features	Main Explanatory Pattern
0	High-polarization structured regions	$P_{\text{mean}}, P_{\text{std}}, E_{\text{energy}}$	High amplitude, high variability, high E -mode energy
1	E -dominated medium-polarization regions	$\log(B/E), E_{\text{energy}}$	Relative dominance of E -mode energy
2	B -enhanced medium-polarization regions	$\log(B/E), E_{\text{energy}}$	Higher relative B -mode contribution
3	Low-polarization regions	$P_{\text{mean}}, P_{\text{std}}$	Low amplitude and low variability

To analyze the direction of the feature effects, SHAP beeswarm plots were generated for each cluster, Figure 19.

Each panel shows the SHAP value distribution for the XGBoost model associated with one FCM cluster. Positive SHAP values increase the predicted membership to the corresponding cluster, whereas negative SHAP values reduce it. The color scale represents the original feature value.

The beeswarm plots confirmed the directionality of the global SHAP results. For Cluster 0, high values of P_{mean} , P_{std} , and E_{energy} increased membership, confirming its interpretation as a high-polarization structured regime. For Cluster 3, low values of P_{mean} and P_{std} increased membership, consistent with its low-polarization profile.

For the two medium-polarization regimes, $\log(B/E)$ had opposite effects. In Cluster 1, low values of $\log(B/E)$ increased membership; since $\log(B/E) < 0$ implies $B/E < 1$, this indicates relative dominance of E -mode energy. In Cluster 2, high values of $\log(B/E)$ increased membership, indicating a relatively higher B -mode contribution. This interpretation should remain cautious: a higher relative B -mode contribution does not imply detection of primordial B -modes, but may reflect lensing, residual foregrounds, noise, leakage effects, or complex polarization structure.

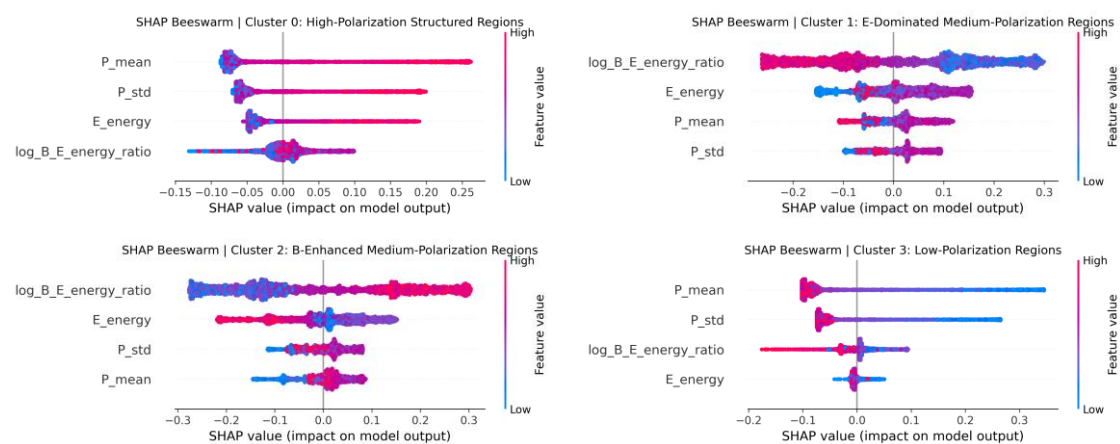


Figure 19. SHAP beeswarm plots for the four fuzzy polarization regimes.

Overall, the SHAP analysis confirms that the fuzzy memberships are governed by interpretable physical descriptors. The high- and low-polarization regimes are mainly explained by amplitude-related variables, whereas the two medium-polarization regimes are separated by the relative B/E balance. Thus, the proposed FAS-XAI framework not only segments the CMB polarization field into fuzzy regions but also identifies the physical variables driving each regional membership pattern.

4.8. Counterfactual Perturbation and Stability Analysis

After constructing an explainable fuzzy segmentation of the regional CMB polarization field, a second analytical phase was conducted to evaluate the stability of the obtained structure under controlled perturbations. The objective was to generate controlled counterfactual scenarios, rather than full observational uncertainty simulations, in which small variations were introduced into physically meaningful regional descriptors, and to assess their effect on fuzzy memberships and dominant-cluster assignments.

Two perturbation experiments were considered. The first introduced multiplicative perturbations into the polarization-amplitude variables P_{mean} and P_{std} . The second introduced additive perturbations into the relative balance variable $\log(B/E)$. In both cases, the original FCM centroids were kept fixed, so that the original clusters acted as reference polarization regimes. Therefore, changes in membership could be attributed to the perturbation itself rather than to a reconfiguration of the clustering solution.

For each perturbed scenario, new memberships U'_i were computed and compared with the original memberships U_i . Stability was evaluated using three indicators: total membership change, dominant-cluster switching, and change in fuzzy entropy.

Figure 20 compares the effect of both perturbation types across low, medium, and high perturbation levels.

The panels compare the mean membership change, dominant-cluster switch rate, and mean entropy change under low, medium, and high perturbation levels. For P , the levels correspond to 1%, 3%, and 5% multiplicative perturbations; for $\log(B/E)$, they correspond to additive perturbations of 0.01, 0.03, and 0.05.

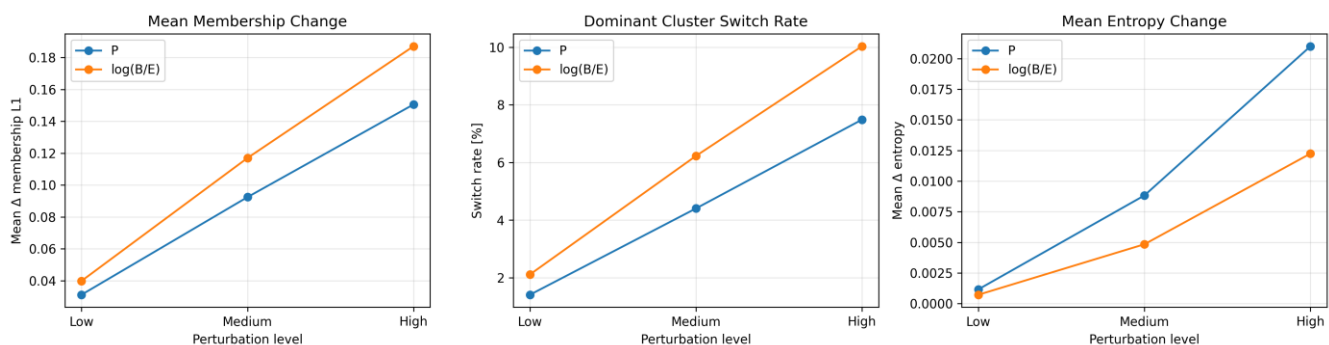


Figure 20. Comparison between P and $\log(B/E)$ perturbations.

The mean membership change increased progressively in both experiments, indicating a coherent response to increasing perturbation strength. However, $\log(B/E)$ perturbations produced consistently larger membership changes than P perturbations. At the high perturbation level, the mean membership change was 0.1506 for P and 0.1871 for $\log(B/E)$.

The dominant-cluster switch rate followed the same pattern. For P , the switch rate increased from 1.41% at the low level to 7.49% at the high level. For $\log(B/E)$, it increased from 2.11% to 10.03%. This indicates that the relative balance between B - and E -mode energies is a particularly sensitive dimension of the regional FAS-XAI feature space.

The entropy response was complementary. Perturbations in P produced a larger mean increase in entropy, especially at the high level, where ΔH reached 0.0210 compared with 0.0122 for $\log(B/E)$. Thus, variations in polarization amplitude tend to increase fuzzy ambiguity, whereas variations in $\log(B/E)$ more directly induce changes in the dominant regime.

To analyze specific transitions between dominant clusters, normalized transition matrices were computed for the high perturbation scenarios: 5% perturbation on P and 0.05 perturbation on $\log(B/E)$, Figure 21.

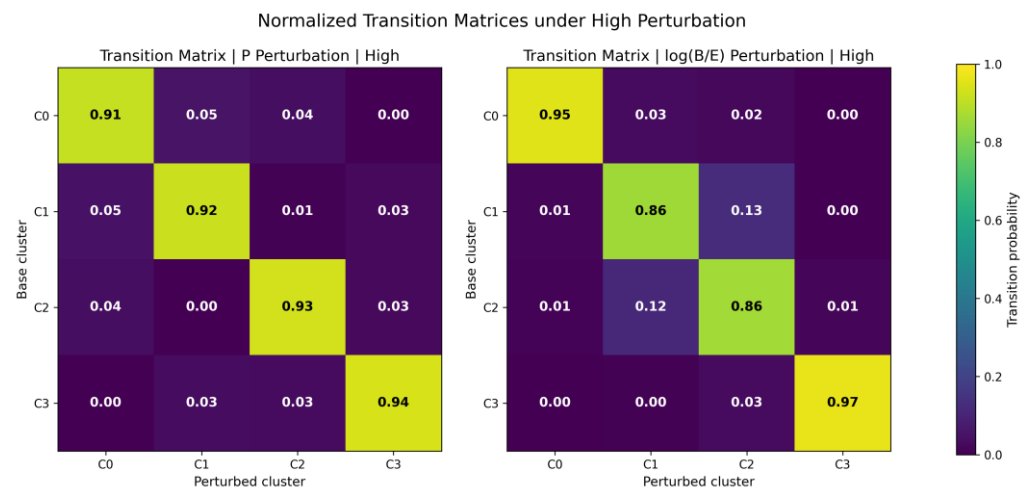


Figure 21. Normalized transition matrices under high perturbation.

The matrices show the probability of moving from each baseline cluster to each perturbed cluster under the high perturbation scenarios for P and $\log(B/E)$.

Under high P perturbation, the transition matrix showed a clearly dominant diagonal. The stability probabilities were 0.9108 for C0, 0.9196 for C1, 0.9332 for C2, and 0.9388 for C3. These values indicate that the four regimes were largely preserved under amplitude perturbation. Off-diagonal transitions were moderate and distributed across several clusters, suggesting smooth displacements between polarization-amplitude regimes rather than a single dominant transition pathway.

Under high $\log(B/E)$ perturbation, the extreme regimes C0 and C3 were highly stable, with stability probabilities of 0.9514 and 0.9661, respectively. In contrast, the intermediate regimes C1 and C2 were more sensitive, with stability probabilities of 0.8584 and 0.8602. This confirms that perturbations in the relative B/E balance mainly affect the boundary between E -dominated medium-polarization regions and B -enhanced medium-polarization regions, consistently with the SHAP results.

Figure 22 summarizes the cluster-wise stability under the strongest perturbation scenarios.

The figure shows the probability that each baseline cluster remains assigned to the same dominant cluster under high P and $\log(B/E)$ perturbations.

All clusters preserved high stability probabilities, confirming that the fuzzy regional structure is globally robust. The lowest stability values appeared for clusters C1 and C2 under $\log(B/E)$ perturbation, consistent with their interpretation as intermediate regimes separated mainly by the relative B/E balance. In contrast, the extreme regimes C0 and C3 remained highly stable, reinforcing the robustness of the high- and low-polarization structures.

Overall, the perturbation analysis shows that the regional CMB polarization structure obtained by FAS-XAI combines global robustness with local sensitivity. Perturbations in P mainly affect the amplitude dimension and tend to increase fuzzy ambiguity, while

perturbations in $\log(B/E)$ primarily affect the relative balance between E - and B -mode energies and induce more frequent transitions between the two intermediate regimes.

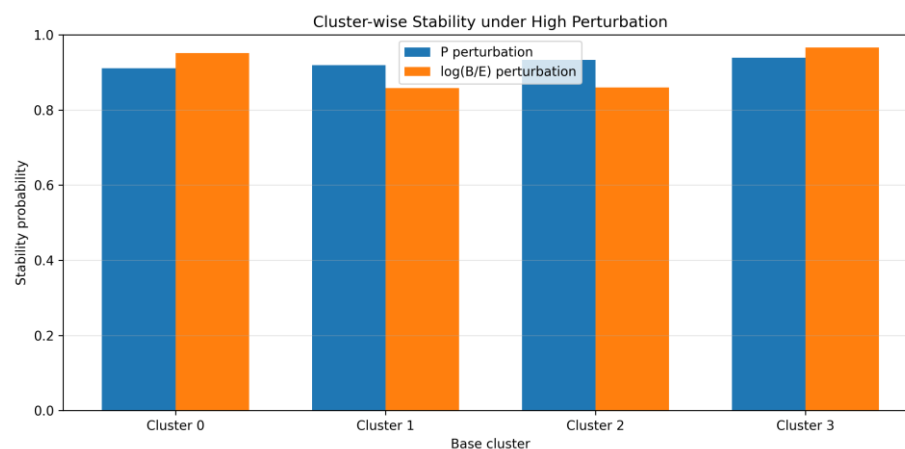


Figure 22. Cluster-wise stability under high perturbation.

From a counterfactual perspective, these results should be interpreted as controlled sensitivity scenarios rather than as alternative cosmological realizations. They quantify how stable or fragile the observed regional polarization structure is with respect to local changes in amplitude and E/B balance. The dominant regional structure remains largely preserved under small perturbations, while local transitions identify regions that are more sensitive, transitional, or ambiguous.

4.9. Sensitivity to the Valid-Region Threshold

As an additional internal robustness check, the sensitivity of the fuzzy segmentation to the valid-region threshold was evaluated. The baseline analysis retained regions with a valid-pixel fraction greater than or equal to 0.80 according to the Planck polarization mask. To assess whether the resulting regimes depended critically on this threshold, the FCM segmentation was repeated using alternative thresholds of 0.70, 0.90, and 0.95, while keeping the same four-feature space and the same number of clusters.

Table 11 summarizes the results. The number of retained regions varied only moderately, from 45,308 regions at the 0.70 threshold to 44,165 regions at the 0.95 threshold. The Fuzzy Partition Coefficient and Partition Entropy remained almost unchanged across thresholds. The mean maximum membership and normalized entropy were also highly stable. Moreover, the agreement with the baseline solution remained above 0.996 in all cases, with Adjusted Rand Index values above 0.990 and Normalized Mutual Information values above 0.982.

Table 11. Internal robustness check under alternative valid-region thresholds.

Valid-Region Threshold	Regions	FPC	PE	Mean Max. Membership	Mean Entropy	Agreement	ARI	NMI
0.70	45,308	0.5425	0.8618	0.6714	0.6216	0.9984	0.9957	0.9911
0.80	45,020	0.5425	0.8618	0.6714	0.6216	1.0000	1.0000	1.0000
0.90	44,566	0.5426	0.8615	0.6715	0.6215	0.9971	0.9919	0.9848
0.95	44,165	0.5427	0.8614	0.6715	0.6214	0.9965	0.9906	0.9825

These results indicate that the fuzzy regional structure is not driven by the specific valid-region threshold used in the baseline analysis. The main regimes remain stable under different levels of mask strictness. Nevertheless, this analysis should be interpreted as

an internal robustness check with respect to mask-controlled regional selection, not as a full external validation using alternative component-separation products, alternative masks, or Monte Carlo noise realizations.

5. Discussion and Future Works

5.1. Discussion

This study proposed an explainable fuzzy framework for the regional analysis of CMB polarization using the Planck PR3 SMICA IQU product. The results show that the regional polarization field can be represented through a compact set of physically interpretable variables: P_{mean} , P_{std} , E_{energy} , and $\log(B/E)$. These variables allowed the identification of four fuzzy polarization regimes: high-polarization structured regions, E -dominated medium-polarization regions, B -enhanced medium-polarization regions, and low-polarization regions.

A central result is that the fuzzy segmentation does not produce a rigid partition of the sky. Instead, the CMB polarization field appears as a continuous regional structure, with many regions showing partial membership to several clusters. The ambiguity and entropy analysis reinforces this interpretation: a substantial fraction of regions was classified as moderate or ambiguous, indicating transitional structures rather than sharply separated classes. In this sense, ambiguity should not be interpreted as a failure of the clustering procedure, but as an informative property of the regional polarization field.

The XGBoost membership models achieved $R^2 > 0.98$ for all clusters, showing that the selected regional variables can accurately reconstruct the fuzzy memberships. However, this does not validate the cosmological reality of the clusters by itself. Rather, it indicates that the FCM membership structure is internally consistent and can be approximated by nonlinear functions of the selected features.

The SHAP analysis provided an interpretable explanation of the fuzzy regimes. The high- and low-polarization clusters were mainly governed by P_{mean} and P_{std} , whereas the two medium-polarization regimes were primarily differentiated by $\log(B/E)$. This supports the physical interpretability of the segmentation. Nevertheless, the B -enhanced cluster must be interpreted cautiously: a relatively higher B -mode contribution does not imply detection of primordial B -modes, since it may reflect lensing, residual foregrounds, noise, leakage effects, or limitations associated with preprocessing and masking.

The perturbation analysis showed that the fuzzy regional structure is globally stable under small controlled variations, while also revealing local sensitivity. Perturbations in $\log(B/E)$ produced more frequent transitions between the E -dominated and B -enhanced medium-polarization regimes, confirming that the relative E/B balance is a sensitive dimension of the regional feature space. Perturbations in P , in contrast, mainly increased fuzzy ambiguity.

These counterfactual results should be interpreted methodologically rather than cosmologically. The perturbed configurations are not evidence of alternative universes or physically realized cosmological histories; they are controlled sensitivity experiments designed to assess how robust the observed regional structure is to small variations in selected variables.

Several limitations must be acknowledged. The analysis is based on a single Planck component-separation product, SMICA, and should be tested against alternative Planck products or independent component-separation methods. The results also depend on the selected mask, the valid-region threshold, and the regional aggregation scale $NSIDE_{\text{region}} = 64$. In addition, the selected features summarize regional behavior through relatively simple statistics, which improve interpretability but may discard information about spatial texture, angular correlations, or multipole-dependent polarization structure.

Finally, the regional B/E balance must be understood as a descriptive feature of the processed maps, not as a direct cosmological measurement. The proposed framework is therefore better interpreted as an exploratory and interpretable regional analysis tool, complementary to standard cosmological inference based on power spectra and parameter estimation.

Overall, the results suggest that the CMB polarization field can be studied as a fuzzy regional system: globally stable, locally transitional, and explainable through physically meaningful descriptors.

5.2. Future Work

Future work should first evaluate the robustness of the proposed framework using alternative CMB products and component-separation methods. Applying the same pipeline to other Planck maps, such as NILC, SEVEM, or Commander, would help determine whether the identified fuzzy regimes remain stable across different foreground-cleaning strategies.

A second extension should analyze the dependence of the results on masking choices and regional resolution. The present work used the Planck polarization mask and a regional grid with $NSIDE_{\text{region}} = 64$. Future analyses should test different valid-region thresholds, alternative masks, and several regional resolutions to assess whether the fuzzy regimes are intrinsic to the polarization structure or partly influenced by methodological choices.

Additional physically motivated features should also be explored. The current feature set was intentionally compact and interpretable, but future work could incorporate angular power information, local gradients, spatial texture descriptors, wavelet-based features, or multipole-band energy measures. These variables may capture aspects of the polarization field not fully represented by regional means, standard deviations, and energy ratios.

The perturbation analysis could be expanded by perturbing the original Stokes parameters Q and U , simulating structured foreground residuals, or introducing noise models consistent with instrumental uncertainties. This would provide a more realistic assessment of the stability of the fuzzy segmentation under observational and systematic effects.

A further direction would be to compare the detected fuzzy regimes with known astrophysical and cosmological structures. For example, high-ambiguity or B -enhanced regions could be contrasted with Galactic foreground templates, lensing maps, dust polarization maps, or known systematic residuals. This would help clarify whether the identified regimes reflect cosmological signal, residual contamination, or a mixture of both.

From the XAI perspective, future work should move beyond global SHAP summaries and explore local explanations for individual regions or groups of regions. Local SHAP profiles could help explain why specific areas are ambiguous, why some regions switch clusters under perturbation, or why certain regions remain especially stable.

Finally, the framework could be adapted to other cosmological and astrophysical datasets, including CMB temperature anomalies, lensing convergence maps, large-scale structure surveys, exoplanet candidate vetting, or galaxy morphology studies. In this sense, the present work can be understood as a methodological proof of concept for applying FAS-XAI to high-dimensional astrophysical fields where uncertainty, gradual transitions, and interpretability are central analytical challenges.

6. Conclusions

This study presented an explainable fuzzy framework for the regional analysis of CMB polarization using the Planck PR3 SMICA IQU map. The proposed approach integrates regional feature extraction, Fuzzy C-Means clustering, XGBoost membership modeling, SHAP-based explainability, and controlled perturbation analysis to characterize polarization regimes across the valid sky.

Starting from the non-inpainted Stokes parameters Q and U , together with the Planck polarization mask, the analysis constructed physically meaningful variables, including the polarization amplitude P and the scalar E - and B -mode maps. The original $NSIDE = 2048$ maps were aggregated into a regional HEALPix grid with $NSIDE_{\text{region}} = 64$, allowing high-resolution polarization information to be summarized into spatially coherent regions.

After masking and valid-region filtering, the final feature space was reduced to four interpretable descriptors, $P_{\text{mean}}, P_{\text{std}}, E_{\text{energy}}, \log(B/E)$. These variables captured complementary aspects of the polarization field: mean polarization amplitude, internal variability, E -mode energy, and the relative balance between B - and E -mode energies.

Fuzzy C-Means identified four regional polarization regimes: high-polarization structured regions, E -dominated medium-polarization regions, B -enhanced medium-polarization regions, and low-polarization regions. The fuzzy formulation was appropriate because many regions did not belong exclusively to a single group. Instead, the membership maps revealed gradual transitions and partial overlap between regimes, supporting the interpretation of CMB polarization as a continuous regional structure rather than a rigid partition of the sky.

The XGBoost models accurately reconstructed the FCM memberships, with $R^2 > 0.98$ for all clusters, enabling the use of SHAP explanations. SHAP showed that the high- and low-polarization regimes are mainly governed by polarization amplitude and variability, whereas the two medium-polarization regimes are primarily separated by the relative B/E balance. Therefore, the fuzzy clusters are explainable through physically meaningful descriptors.

The perturbation analysis showed that the regional structure is globally stable under small controlled variations, while retaining localized sensitivity. Perturbations in P_{mean} and P_{std} mainly increased fuzzy ambiguity, whereas perturbations in $\log(B/E)$ induced more frequent transitions between the two intermediate regimes. Even under the strongest perturbation scenarios, the dominant cluster structure was largely preserved.

Overall, the proposed FAS-XAI framework provides a structured way to analyze CMB polarization as a fuzzy, interpretable, and stability-aware regional system. It does not replace standard cosmological inference based on power spectra or parameter estimation but offers a complementary exploratory tool for identifying regional patterns, explaining membership structure, and evaluating sensitivity to controlled perturbations.

The main contribution of this work is methodological: it demonstrates that fuzzy clustering and explainable artificial intelligence can be integrated to study CMB polarization while accounting for uncertainty, gradual transitions, and physical interpretability. This opens a path toward broader applications of FAS-XAI in cosmology and astrophysics, particularly in contexts where complex fields cannot be adequately described by rigid classifications alone.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/math14132269/s1>. The repository includes a Python script used to reproduce the complete analysis pipeline.

Funding: This research received no external funding.

Data Availability Statement: Publicly available data were used in this study. The Planck PR3 Legacy CMB Maps are accessible through the European Space Agency Planck Legacy Archive at: https://esdcdoi.esac.esa.int/doi/html/data/astronomy/planck/CMB_Maps.html (accessed on 15 May 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. European Space Agency. PR3 Legacy CMB Maps, Version PR2 and PR3; European Space Agency: Paris, France, 2018. <https://doi.org/10.5270/esa-zfx8b4s>.
2. Planck Collaboration; Aghanim, N.; Akrami, Y.; Ashdown, M.; Aumont, J.; Baccigalupi, C.; Ballardini, M.; Banday, A.J.; Barreiro, R.B.; Bartolo, N.; et al. Planck 2018 results—VI. Cosmological parameters. *Astron. Astrophys.* **2020**, *641*, A6. <https://doi.org/10.1051/0004-6361/201833910>.
3. Hu, W.; White, M. A CMB polarization primer. *New Astron.* **1997**, *2*, 323–344. [https://doi.org/10.1016/S1384-1076\(97\)00022-5](https://doi.org/10.1016/S1384-1076(97)00022-5).
4. Ducout, A.; Bouchet, F.R.; Colombi, S.; Pogosyan, D.; Prunet, S. Non-Gaussianity and Minkowski functionals: Forecasts for Planck. *Mon. Not. R. Astron. Soc.* **2012**, *429*, 2104–2126. <https://doi.org/10.1093/mnras/sts483>.
5. Marín Díaz, G. FAS-XAI: Fuzzy and Explainable AI for Interpretable Vetting of Kepler Exoplanet Candidates. *Mathematics* **2025**, *13*, 1–32. <https://doi.org/10.3390/math13233796>.
6. Bezdek, J.C.; Ehrlich, R.; Full, W. FCM: The fuzzy c-means clustering algorithm. *Comput. Geosci.* **1984**, *10*, 191–203. [https://doi.org/10.1016/0098-3004\(84\)90020-7](https://doi.org/10.1016/0098-3004(84)90020-7).
7. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794. <https://doi.org/10.1145/2939672.2939785>.
8. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4768–4777.
9. Caldeira, J.; Wu, W.L.K.; Nord, B.; Avestruz, C.; Trivedi, S.; Story, K.T. DeepCMB: Lensing reconstruction of the cosmic microwave background with deep neural networks. *Astron. Comput.* **2019**, *28*, 100307. <https://doi.org/10.1016/j.ascom.2019.100307>.
10. Hortua, H.J.; Volpi, R.; Marinelli, D.; Malago, L. Parameter estimation for the cosmic microwave background with Bayesian neural networks. *Phys. Rev. D* **2020**, *102*, 103509. <https://doi.org/10.1103/PhysRevD.102.103509>.
11. Petroff, M.A.; Addison, G.E.; Bennett, C.L.; Weiland, J.L. Full-sky Cosmic Microwave Background Foreground Cleaning Using Machine Learning. *Astrophys. J.* **2020**, *903*, 104. <https://doi.org/10.3847/1538-4357/abb9a7>.
12. Wang, G.-J.; Shi, H.-L.; Yan, Y.-P.; Xia, J.-Q.; Zhao, Y.-Y.; Li, S.-Y.; Li, J.-F. Recovering the CMB Signal with Machine Learning. *Astrophys. J. Suppl. Ser.* **2022**, *260*, 13. <https://doi.org/10.3847/1538-4365/ac5f4a>.
13. Puglisi, G.; Bai, X. Inpainting Galactic Foreground Intensity and Polarization Maps Using Convolutional Neural Networks. *Astrophys. J.* **2020**, *905*, 143. <https://doi.org/10.3847/1538-4357/abc47c>.
14. Casas, J.M.; Bonavera, L.; González-Nuevo, J.; Baccigalupi, C.; Cueli, M.M.; Crespo, D.; Goitia, E.; Santos, J.D.; Sánchez, M.L.; de Cos, F.J. CENN: A fully convolutional neural network for CMB recovery in realistic microwave sky simulations. *Astron. Astrophys.* **2022**, *666*, A89. <https://doi.org/10.1051/0004-6361/202243450>.
15. Thorne, B.; Knox, L.; Prabhu, K. A generative model of galactic dust emission using variational autoencoders. *Mon. Not. R. Astron. Soc.* **2021**, *504*, 2603–2613. <https://doi.org/10.1093/mnras/stab1011>.
16. Farsian, F.; Krachmalnicoff, N.; Baccigalupi, C. Foreground model recognition through Neural Networks for CMB B-mode observations. *J. Cosmol. Astropart. Phys.* **2020**, *2020*, 17. <https://doi.org/10.1088/1475-7516/2020/07/017>.
17. Guzman, E.; Meyers, J. Reconstructing cosmic polarization rotation with ResUNet-CMB. *J. Cosmol. Astropart. Phys.* **2022**, *2022*, 30. <https://doi.org/10.1088/1475-7516/2022/01/030>.
18. Ciuca, R.; Hernandez, O.F. A Bayesian framework for cosmic string searches in CMB maps. *J. Cosmol. Astropart. Phys.* **2017**, *2017*, 28. <https://doi.org/10.1088/1475-7516/2017/08/028>.
19. Novaes, C.P.; Bernui, A.; Ferreira, I.S.; Wuensche, C.A. A neural-network based estimator to search for primordial non-Gaussianity in Planck CMB maps. *J. Cosmol. Astropart. Phys.* **2015**, *2015*, 64. <https://doi.org/10.1088/1475-7516/2015/09/064>.
20. Ciuca, R.; Hernandez, O.F.; Wolman, M. A convolutional neural network for cosmic string detection in CMB temperature maps. *Mon. Not. R. Astron. Soc.* **2019**, *485*, 1377–1383. <https://doi.org/10.1093/mnras/stz491>.
21. Novaes, C.P.; Bernui, A.; Ferreira, I.S.; Wuensche, C.A. Searching for primordial non-Gaussianity in Planck CMB maps using a combined estimator. *J. Cosmol. Astropart. Phys.* **2014**, *2014*, 18. <https://doi.org/10.1088/1475-7516/2014/01/018>.
22. Wadekar, D.; Thiele, L.; Hill, J.C.; Pandey, S.; Villaescusa-Navarro, F.; Spergel, D.N.; Cranmer, M.; Nagai, D.; Anglés-Alcázar, D.; Ho, S.; et al. The SZ flux-mass (Y-M) relation at low-halo masses: Improvements with symbolic regression and strong constraints on baryonic feedback. *Mon. Not. R. Astron. Soc.* **2023**, *522*, 2628–2643. <https://doi.org/10.1093/mnras/stad1128>.
23. Gupta, N.; Reichardt, C.L. Mass Estimation of Galaxy Clusters with Deep Learning. I. Sunyaev-Zel'dovich Effect. *Astrophys. J.* **2020**, *900*, 110. <https://doi.org/10.3847/1538-4357/aba694>.

24. Gorce, A.; Douspis, M.; Salvati, L. Retrieving cosmological information from small-scale CMB foregrounds II. The thermal Sunyaev Zel'dovich effect. *Astron. Astrophys.* **2022**, *662*, A122. <https://doi.org/10.1051/0004-6361/202243351>.
25. Bonjean, V. Deep learning for Sunyaev-Zel'dovich detection in Planck. *Astron. Astrophys.* **2020**, *634*, A81. <https://doi.org/10.1051/0004-6361/201936919>.
26. Thiele, L.; Wadekar, D.; Hill, J.C.; Battaglia, N.; Chluba, J.; Villaescusa-Navarro, F.; Hernquist, L.; Vogelsberger, M.; Anglés-Alcázar, D.; Marinacci, F. Percent-level constraints on baryonic feedback with distortion measurements. *Phys. Rev. D* **2022**, *105*, 083505. <https://doi.org/10.1103/PhysRevD.105.083505>.
27. Ocampo, I.; Canas-Herrera, G.; Nesseris, S. Neural Networks for cosmological model selection and feature importance using Cosmic Microwave Background data. *J. Cosmol. Astropart. Phys.* **2025**, *2025*, 4. <https://doi.org/10.1088/1475-7516/2025/02/004>.
28. Zaldarriaga, M.; Seljak, U. All-sky analysis of polarization in the microwave background. *Phys. Rev. D* **1997**, *55*, 1830–1840. <https://doi.org/10.1103/physrevd.55.1830>.
29. Gorski, K.M.; Hivon, E.; Banday, A.J.; Wandelt, B.D.; Hansen, F.K.; Reinecke, M.; Bartelmann, M. HEALPix: A Framework for High-Resolution Discretization and Fast Analysis of Data Distributed on the Sphere. *Astrophys. J.* **2005**, *622*, 759–771. <https://doi.org/10.1086/427976>.
30. Kamionkowski, M.; Kosowsky, A.; Stebbins, A. Statistics of cosmic microwave background polarization. *Phys. Rev. D* **1997**, *55*, 7368–7388. <https://doi.org/10.1103/physrevd.55.7368>.
31. Marín Díaz, G. Supporting Reflective AI Use in Education: A Fuzzy-Explainable Model for Identifying Cognitive Risk Profiles. *Educ. Sci.* **2025**, *15*, 923. <https://doi.org/10.3390/educsci15070923>.
32. You, C.; Wen, M. The entropy of fuzzy vectors. *Comput. Math. Appl.* **2008**, *56*, 1626–1633. <https://doi.org/10.1016/j.camwa.2008.03.014>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.