Uncertainty-Aware Spatial Mixture Classification of Imbalanced CO_2 Plume States: Application to the Sleipner 2019 Benchmark

Muhammad Amir Saeed
m.saeed@campus.unimib.it
University of Milano-Bicocca https://orcid.org/0000-0003-1650-8194
Sadia Saba
University of Milano-Bicocca https://orcid.org/0009-0006-3166-8953

Research Article

Keywords: cluster-weighted model, multinomial logit, two-parameter shrinkage, spatial classification, class imbalance, carbon dioxide storage

Posted Date: August 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10412995/v1

Additional Declarations: No competing interests reported.

Uncertainty-Aware Spatial Mixture Classification of Imbalanced CO_2 Plume States: Application to the Sleipner 2019 Benchmark

Muhammad Amir Saeed[1*] and Sadia Saba[1]

[1*]Department of Economics, Management and Statistics, University of Milano-Bicocca, Milan, Italy.

*Corresponding author(s). E-mail(s): m.saeed@campus.unimib.it;
Contributing authors: s.saba5@campus.unimib.it;

Abstract

Reliable plume-state classification is essential for monitoring geological carbon dioxide storage, but benchmark reservoir data are often spatially heterogeneous, severely imbalanced, and affected by collinearity among geological and petrophysical predictors. We develop a spatially informed, expectation-maximization-type cluster-weighted multinomial logit mixture for classifying binned plume states. Spatial coordinates enter the component-specific covariate distributions, while plume-state probabilities are represented by component-specific multinomial logit models. A two-parameter shrinkage update stabilizes local estimates under collinearity and small effective component sizes, and class-weighted estimation addresses class imbalance. The method is evaluated on a 200-metre binned table from the Sleipner 2019 carbon dioxide storage benchmark, containing 3,409 bins and three plume-state classes, with 87% of observations outside the plume. Performance is assessed using spatial-block cross-validation, leave-one-layer-out transfer, and an ablation analysis. The model identifies three interpretable, depth-organized latent regimes and provides posterior uncertainty for each bin. It achieves a cross-validated balanced accuracy of **0.561**, compared with **0.343** for an unweighted global multinomial logit model and **0.333** for a majority-class baseline. The ablation analysis shows that class weighting and spatial covariates drive most of the improvement. The latent mixture and shrinkage update provide smaller, statistically indistinguishable gains. Overall, the framework offers an interpretable and uncertainty-aware approach to minority plume-state recovery under severe imbalance, while the shrinkage update primarily stabilizes ill-conditioned local fits.

Keywords: cluster-weighted model, multinomial logit, two-parameter shrinkage, spatial classification, class imbalance, carbon dioxide storage

1 Introduction

Geological storage of carbon dioxide is a major technology for reducing industrial greenhouse-gas emissions while maintaining energy and industrial production systems. In operational storage projects, monitoring the spatial migration of injected CO_2 is central to evaluating containment, conformance and long-term risk. The Sleipner project in the Norwegian North Sea is one of the most studied industrial-scale offshore CO_2 storage sites. Since injection began in 1996, repeated seismic surveys and modelling studies have made Sleipner a reference case for analysing plume migration in a layered saline aquifer (Arts et al, 2004; Chadwick et al, 2004; Chadwick and Noy, 2010, 2015; Furre et al, 2024).

Open benchmark datasets create an opportunity to develop reproducible statistical methods for storage monitoring. The Sleipner 2019 Benchmark Model released through CO_2 DataShare contains simulation grids, petrophysical properties, well information, velocity maps, interpreted horizons and plume-related spatial information (CO2 DataShare, 2020). These data are scientifically valuable but statistically difficult. Observations are spatially arranged, often aggregated into grid cells or bins, and involve continuous predictors, categorical layer indicators, distances to geological or operational features, interpreted plume boundaries and uncertain response labels. Moreover, plume-state classes are typically imbalanced: most grid cells are outside the plume, while plume-boundary and plume-interior cells are the scientifically important minority classes.

A single global classifier is unlikely to represent such data adequately. Geological layering, structural trapping, depth variation and plume migration imply that the relationship between reservoir predictors and plume state may change across the storage formation. For example, depth, velocity, layer index, distance to the injection well and plume-boundary geometry may have different implications in lower reservoir units, intermediate compartments and upper accumulations. A model that assumes one global regression surface can therefore be both statistically restrictive and geologically hard to interpret.

Finite mixture models provide a natural approach to latent heterogeneity by assuming that observations arise from several unobserved sub-populations or regimes (Titterington et al, 1985; McLachlan and Peel, 2000; Frühwirth-Schnatter, 2006). In spatial reservoir analysis, a latent component may be interpreted as a statistical regime characterized by a particular distribution of covariates and a particular response mechanism. However, ordinary mixture models generally focus on the marginal distribution of observed variables and do not always provide a direct conditional model for a plume-state response.

Cluster-weighted models address this limitation by modelling the joint distribution of a response and covariates as a mixture of products of component-specific covariate densities and component-specific conditional response models (Gershenfeld, 1997; Gershenfeld et al, 1999; Ingrassia et al, 2012, 2014, 2015). This makes them attractive for binned reservoir analysis because each component can describe both a local reservoir-covariate profile and a local plume-state classification rule. The present paper builds on this idea by considering a spatially informed cluster-weighted model with multinomial logit response components.

Multinomial logit regression is a standard model for unordered multi-class responses (Luce, 1959; McFadden, 1974). In reservoir applications, however, component-specific multinomial logit estimates may be unstable because component effective sample sizes can be small and reservoir predictors are often collinear. This collinearity is not merely numerical; it reflects geological structure. Depth, layer position, velocity, petrophysical properties and distance measures may be correlated because they all arise from related depositional, structural and spatial processes. Removing correlated variables can discard scientifically meaningful information, while unregularized local estimation can inflate coefficients and degrade minority-class recovery.

Ridge-type, Liu-type and two-parameter estimators have been proposed to stabilize regression and discrete-choice models under multicollinearity (Hoerl and Kennard, 1970; Liu, 1993; Yang and Chang, 2010; Farghali et al, 2023). The contribution of this paper is to embed a component-specific two-parameter shrinkage update inside an EM-type spatially informed cluster-weighted multinomial logit mixture. In this setting, regularization is not presented as a new estimator by itself. Instead, it is a stabilizing mechanism for local multinomial logit updates within a heterogeneous spatial mixture model.

The paper is deliberately positioned as an applied methodological contribution. The aim is to combine three ideas that are usually treated separately: spatial mixture modelling for latent reservoir regimes, multinomial logit classification for plume-state responses and component-specific two-parameter shrinkage for local estimation under collinearity. The empirical analysis uses the Sleipner 2019 Benchmark Model to examine whether the framework can identify interpretable latent statistical regimes, improve minority plume-state recovery relative to unweighted baselines, and produce model-based posterior uncertainty maps.

The remainder of the paper is organized as follows. Section 2 reviews related work on mixture and cluster-weighted models, mixtures of multinomial logits, regularization under multicollinearity, and statistical learning for Sleipner. Section 3 states the contribution relative to existing methods. Section 4 defines the spatially informed cluster-weighted multinomial logit mixture and the component-specific two-parameter update. Section 5 presents the Sleipner application, including model selection, the latent regimes, conditioning and regime-specific shrinkage, apparent performance, spatial-block and leave-one-layer-out cross-validation, an ablation, and class-wise behaviour. Section 6 discusses interpretation and limitations, Section 7 concludes, and Section 8 outlines future work.

2 Related Literature

2.1 Finite mixture models and model-based classification

Finite mixture models are widely used for clustering, density estimation, model-based classification and latent heterogeneity analysis. Let $\boldsymbol{u}_i$ denote an observed vector. A finite mixture density has the form

$$f(\boldsymbol{u}_i; \Theta) = \sum_{g=1}^{G} \pi_g f_g(\boldsymbol{u}_i; \theta_g), \tag{1}$$

where G is the number of mixture components, $\pi_g > 0$, $\sum_{g=1}^{G} \pi_g = 1$, and f_g is the component density. Introducing a latent indicator $Z_i \in \{1, \ldots, G\}$ gives posterior membership probabilities that can be used for soft and hard classification. Mixture models are commonly estimated by the expectation–maximization algorithm (Dempster et al, 1977), and information criteria such as BIC and ICL are often used for model selection (Schwarz, 1978; Biernacki et al, 2000).

For reservoir data, mixture models are useful because the subsurface is spatially heterogeneous and layered. A mixture representation allows latent statistical regimes to be inferred from the data rather than imposed through deterministic geological rules. However, standard Gaussian mixtures are limited when the scientific target is a response variable such as plume state. They cluster covariates, but they do not necessarily explain how a categorical plume response changes conditionally on reservoir predictors.

2.2 Cluster-weighted models

Cluster-weighted models extend mixture modelling by factorizing the joint distribution of a response Y and covariates $\boldsymbol{X}$ as

$$p(y_i, \boldsymbol{x}_i) = \sum_{g=1}^{G} \pi_g f_g(\boldsymbol{x}_i; \theta_g) q_g(y_i \mid \boldsymbol{x}_i; \beta_g), \tag{2}$$

where f_g is the component-specific marginal distribution of the covariates and q_g is the component-specific conditional distribution of the response. This representation supports local statistical modelling because each component describes both a covariate profile and a response mechanism (Ingrassia et al, 2012, 2014, 2015). Generalized cluster-weighted models allow response distributions from the exponential family, making the approach relevant for binary, count and categorical response settings.

The present paper uses the cluster-weighted view because plume-state classification depends on both spatial/geological heterogeneity and conditional response behaviour. A latent component may represent, for example, a deeper background regime, a transition regime or a shallow accumulation regime. These should be interpreted as statistical regimes, not deterministic geological facies. Their scientific value lies in summarizing where the relationship between covariates and plume state differs across the reservoir.

2.3 Mixtures of multinomial logit models

Finite mixtures of multinomial logit models and related mixtures of generalized linear models have been used for clustering and classification with categorical outcomes (Grün and Leisch, 2008; Papastamoulis, 2023). In such models, different latent components have different multinomial regression coefficients. This allows the effect of covariates on category probabilities to vary across latent groups. In the reservoir setting, this is attractive because the same predictor may have different implications in different layers or spatial regimes.

However, mixture models with multinomial logit components can be numerically fragile. The M-step requires fitting weighted multinomial logit models, and the effective sample size for some components may be small. If the design matrix is collinear or if a minority response class is rare inside a component, the Hessian or information matrix may be ill-conditioned. Stabilized Newton or ridge-type updates are therefore often useful in practice.

Most closely related to the present work, Olobatuyi and Ariyo (2022) proposed a multinomial cluster-weighted model (MCWM) in which the conditional response distribution given the covariates is multinomial logit and estimation proceeds by an EM algorithm with iteratively reweighted least squares (EM-IRLS). Their formulation establishes identifiability for this model class and demonstrates strong clustering performance on high-dimensional data. The model proposed here shares this multinomial-logit cluster-weighted core but differs in three respects that are central to the reservoir-monitoring setting. First, spatial heterogeneity is captured by including the spatial coordinates and depth among the covariates of each component's density, so that the estimated regimes occupy distinct spatial regions of the reservoir; a fully spatial gating model, in the tradition of finite mixtures of regressions with concomitant variables (Ingrassia et al, 2015), is a natural generalization that we leave to future work.

Second, the local multinomial logit M-step is replaced by a component-specific two-parameter regularized update (Eq. (20)) that stabilizes estimation when local information matrices are ill-conditioned because of multicollinearity, rare classes or small effective component sizes; earlier mixture-of-logit work has used ridge-type or penalized updates Bashir and Carter (2010) but not a component-wise two-parameter estimator.

Third, the framework is developed for severely imbalanced, spatially binned plume-state data, with class-weighted fitting and balanced evaluation rather than ordinary clustering accuracy. The contribution is therefore the *integration* of these elements, not the multinomial-CWM idea itself, which is due to Olobatuyi and Ariyo (2022).

2.4 Regularization and two-parameter estimators under multicollinearity

Multicollinearity can make maximum likelihood estimates unstable in logistic and multinomial logit models. Ridge-type estimators introduce shrinkage through a positive diagonal adjustment to the information matrix (Hoerl and Kennard, 1970; Schaefer et al, 1984; le Cessie and van Houwelingen, 1992). Liu-type estimators introduce an additional shrinkage structure that can reduce variance under certain conditions (Liu, 1993). Two-parameter estimators combine ridge and Liu-type ideas and have been extended to generalized linear and multinomial logit settings (Yang and Chang, 2010; Farghali et al, 2023).

The present paper does not claim to derive the first two-parameter estimator for multinomial logit regression. Instead, it uses a two-parameter update as a component-specific stabilization step inside a spatially informed cluster-weighted multinomial logit mixture. This is important because the degree of collinearity and the amount of shrinkage required may vary across latent reservoir regimes.

2.5 Statistical learning for CO_2 storage and Sleipner data

The Sleipner site has been studied extensively using seismic monitoring, reservoir modelling, plume-geometry interpretation, history matching and uncertainty analysis (Arts et al, 2004; Chadwick et al, 2004; Chadwick and Noy, 2010, 2015; Furre et al, 2024). The Sleipner 2019 Benchmark Model provides a reproducible basis for method development because it includes a simulation grid, petrophysical properties, well data, horizons, velocity maps and interpreted plume-related information (CO2 DataShare, 2020). Machine-learning and statistical methods are increasingly used for CO_2 storage monitoring, including plume detection, uncertainty quantification and simulation-model calibration (Ahmadinia et al, 2020; Romero et al, 2023; Bruer et al, 2024).

A common limitation of purely predictive machine-learning approaches is that they may produce class labels without latent-regime interpretation or posterior uncertainty. The proposed model is designed to fill this gap by combining classification, latent reservoir-regime identification and uncertainty mapping in one probabilistic framework.

3 Contribution and Scope

The contribution of this paper is deliberately framed to avoid overstating novelty. The individual ingredients cluster-weighted modelling, mixtures of multinomial logits, spatial gating functions and two-parameter shrinkage estimators are established. The novelty lies in integrating these elements for imbalanced, spatially binned reservoir plume-state classification and in evaluating the resulting framework through spatial validation and ablation, rather than in proposing cluster-weighted multinomial logits or two-parameter shrinkage as standalone new methods.

The main contributions are as follows.

First, the paper formulates a spatially informed cluster-weighted multinomial logit mixture for binned reservoir plume-state classification. Each grid bin is represented by spatial coordinates, reservoir covariates, layer information and a categorical plume-state response. The model estimates latent statistical regimes and allows each regime to have its own covariate distribution and multinomial logit response mechanism. In the implementation used here, the mixing weights are constant and spatial information enters through the component covariate densities; the model should therefore be read as spatially informed rather than as a full spatial random-field or spatial gating model. This extends the multinomial cluster-weighted model of Olobatuyi and Ariyo (2022) by incorporating the spatial coordinates in the component covariate densities, component-specific two-parameter regularized local updates, and a class-weighted fitting and evaluation protocol for imbalanced plume-state data.

Second, the paper embeds a component-specific two-parameter regularized update into the local multinomial logit M-step. The update is intended to stabilize coefficient estimation when local information matrices are ill-conditioned because of multicollinearity, rare classes or small effective component sizes.

Third, the paper distinguishes between likelihood-based model fitting and class-weighted pseudo-likelihood fitting. Because class weighting changes the likelihood target, information criteria computed from weighted objectives are reported as pseudo-BIC and pseudo-ICL unless an unweighted observed-data likelihood is used.

Fourth, the paper emphasizes spatial validation. Apparent in-sample results are reported only as descriptive evidence. Predictive claims should be based on spatial-block cross-validation or held-out spatial/layer regions.

Fifth, the empirical application to the Sleipner 2019 Benchmark Model illustrates how the proposed framework can produce latent-regime maps, plume-state predictions, class-wise minority-plume recovery metrics and posterior uncertainty maps.

4 Methodology

4.1 Binned reservoir observations

Let the reservoir domain be partitioned into n spatial bins. For bin i, let $\boldsymbol{s}_i = (s_{i1}, s_{i2}, s_{i3})^\top$ denote spatial coordinates, such as easting, northing and depth. Let $\boldsymbol{x}_i \in \mathbb{R}^p$ denote continuous reservoir predictors, such as standardized depth, thickness, distance to the injection well, distance to interpreted plume boundary, velocity or transformed petrophysical quantities. Let $\boldsymbol{c}_i$ denote categorical predictors such as layer or stratigraphic interval.

The response is a plume-state class

$$Y_i \in \{1, \dots, J\}. \tag{3}$$

In the Sleipner application, $J = 3$ and the classes are outside plume, plume boundary and plume interior. If a bin aggregates many fine cells, the response may alternatively be represented by multinomial counts

$$\boldsymbol{y}_i = (y_{i1}, \dots, y_{iJ})^\top, \qquad N_i = \sum_{j=1}^{J} y_{ij}. \tag{4}$$

The single-label case is obtained by setting $N_i = 1$.

4.2 spatially informed cluster-weighted multinomial logit mixture

Assume that each bin belongs to one of G latent statistical reservoir regimes. Let $Z_i \in \{1, \ldots, G\}$ be the unobserved component label. The model factorizes the joint distribution of response and covariates as

$$p(y_i, \boldsymbol{s}_i, \boldsymbol{x}_i, \boldsymbol{c}_i; \Theta) = \sum_{g=1}^{G} \pi_g \, f_g(\boldsymbol{s}_i, \boldsymbol{x}_i, \boldsymbol{c}_i; \boldsymbol{\theta}_g) \, q_g(y_i \mid \boldsymbol{x}_i, \boldsymbol{c}_i; \boldsymbol{\beta}_g), \tag{5}$$

where $\pi_g > 0$, $\sum_{g=1}^{G} \pi_g = 1$ are the finite mixture weights, the component covariate density f_g is defined over the spatial coordinates $\boldsymbol{s}_i$ together with the reservoir covariates, and q_g is the component-specific multinomial logit response model. Because f_g carries the spatial coordinates, the components occupy distinct spatial regions even though the mixture weights are constant. The full parameter vector is

$$\Theta = \{\pi_1, \ldots, \pi_{G-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_G\}. \tag{6}$$

4.3 Mixing weights and the role of spatial covariates

The fitted model uses constant mixture weights, π_g, across bins. Spatial heterogeneity is introduced through the component covariate density f_g, which includes standardized easting, northing and depth together with the other reservoir covariates. Consequently, the posterior component probabilities vary spatially because the likelihood of the spatial-covariate vector varies by component, even though the prior mixing weights are constant. This differs from a spatial gating model, where π_g itself would be an explicit function of location. A natural extension is

$$\pi_g(s_i; \eta) = \frac{\exp\{\eta_g^\top h(s_i)\}}{\sum_{r=1}^{G} \exp\{\eta_r^\top h(s_i)\}}, \tag{7}$$

with spatial basis functions $h(s_i)$. We leave this extension to future work.

4.4 Component-specific covariate model

For mixed-type reservoir attributes, write $\boldsymbol{x}_i^{(c)}$ for continuous variables and $\boldsymbol{x}_i^{(d)}$ or $\boldsymbol{c}_i$ for discrete variables. A practical component model is

$$f_g(\boldsymbol{x}_i, \boldsymbol{c}_i; \boldsymbol{\theta}_g) = \phi_p(\boldsymbol{x}_i^{(c)}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \prod_{r=1}^{R} \prod_{m=1}^{M_r} \rho_{grm}^{\mathbb{I}(c_{ir}=m)}, \tag{8}$$

where ϕ_p is a multivariate Gaussian density and ρ_{grm} is the probability of category m for categorical predictor r in component g. Log-transformations should be applied to highly skewed predictors such as permeability before standardization. If heavy tails or outliers are substantial, a multivariate t density can be substituted for the Gaussian density.

4.5 Component-specific multinomial logit response model

Let $\boldsymbol{v}_i = (1, \boldsymbol{x}_i^\top, \boldsymbol{d}_i^\top)^\top$ be the design vector in the response model, where $\boldsymbol{d}_i$ contains dummy variables for selected categorical predictors. Conditional on $Z_i = g$, the multinomial logit probabilities are

$$p_{igj}^{(g)} = \Pr(Y_i = j \mid \boldsymbol{v}_i, Z_i = g) = \frac{\exp(\boldsymbol{v}_i^\top \boldsymbol{\beta}_{gj})}{1 + \sum_{\ell=1}^{J-1} \exp(\boldsymbol{v}_i^\top \boldsymbol{\beta}_{g\ell})}, \quad j = 1, \dots, J-1, \tag{9}$$

with baseline category probability

$$p_{igJ}^{(g)} = 1 - \sum_{j=1}^{J-1} p_{igj}^{(g)}. \tag{10}$$

For a binned count response,

$$q_g(\boldsymbol{y}_i \mid \boldsymbol{v}_i; \boldsymbol{\beta}_g) = \frac{N_i!}{\prod_{j=1}^{J} y_{ij}!} \prod_{j=1}^{J} \left(p_{igj}^{(g)}\right)^{y_{ij}}. \tag{11}$$

For single-label classification, $y_{ij} = \mathbb{I}(Y_i = j)$.

4.6 Observed-data likelihood and class-weighted pseudo-likelihood

The unweighted observed-data log-likelihood is

$$\ell(\Theta) = \sum_{i=1}^{n} \log \left[\sum_{g=1}^{G} \pi_g \, f_g(\boldsymbol{s}_i, \boldsymbol{x}_i, \boldsymbol{c}_i; \boldsymbol{\theta}_g) \, q_g(y_i \mid \boldsymbol{v}_i; \boldsymbol{\beta}_g) \right]. \tag{12}$$

When class imbalance is severe, class weights may be introduced in the conditional response part. Let w_j be the weight for class j. For a single-label response, the weighted conditional contribution is

$$q_g^{(w)}(Y_i \mid \boldsymbol{v}_i; \boldsymbol{\beta}_g) = \prod_{j=1}^{J} \left(p_{igj}^{(g)}\right)^{w_j \mathbb{I}(Y_i = j)}. \tag{13}$$

This produces a class-weighted pseudo-likelihood rather than the ordinary probability model. Therefore, information criteria computed from weighted objectives are reported as pseudo-BIC or pseudo-ICL. For formal likelihood-based model selection, the unweighted likelihood in Eq. (12) should also be reported.

4.7 E-step

At iteration t, compute the posterior probability that bin i belongs to component g:

$$\tau_{ig}^{(t)} = \frac{\pi_g^{(t)} f_g(\boldsymbol{s}_i, \boldsymbol{x}_i, \boldsymbol{c}_i; \boldsymbol{\theta}_g^{(t)})\, q_g(y_i \mid \boldsymbol{v}_i; \boldsymbol{\beta}_g^{(t)})}{\sum_{r=1}^{G} \pi_r^{(t)} f_r(\boldsymbol{s}_i, \boldsymbol{x}_i, \boldsymbol{c}_i; \boldsymbol{\theta}_r^{(t)})\, q_r(y_i \mid \boldsymbol{v}_i; \boldsymbol{\beta}_r^{(t)})}. \tag{14}$$

If class weights are used, the corresponding weighted conditional contribution can be used for fitting. Posterior probabilities should be interpreted relative to the fitted objective.

4.8 M-step for gates and covariate distributions

The mixture weights are updated by $\hat{\pi}_g = \frac{1}{n}\sum_{i=1}^{n} \tau_{ig}$. The covariate-model parameters are updated by weighted maximum likelihood. For the Gaussian part of Eq. (8),

$$\hat{\boldsymbol{\mu}}_g = \frac{\sum_{i=1}^{n} \tau_{ig} \boldsymbol{x}_i^{(c)}}{\sum_{i=1}^{n} \tau_{ig}}, \tag{15}$$

$$\hat{\boldsymbol{\Sigma}}_g = \frac{\sum_{i=1}^{n} \tau_{ig} (\boldsymbol{x}_i^{(c)} - \hat{\boldsymbol{\mu}}_g)(\boldsymbol{x}_i^{(c)} - \hat{\boldsymbol{\mu}}_g)^\top}{\sum_{i=1}^{n} \tau_{ig}}. \tag{16}$$

Small diagonal covariance regularization may be used to avoid singular estimates. For categorical variables,

$$\hat{\rho}_{grm} = \frac{\sum_{i=1}^{n} \tau_{ig} \mathbb{I}(c_{ir} = m)}{\sum_{i=1}^{n} \tau_{ig}}. \tag{17}$$

4.9 Component-specific two-parameter regularized multinomial logit update

For component g, let $\boldsymbol{\beta}_g$ collect all non-baseline multinomial logit coefficients. Its dimension is $r = (J-1)p_v$, where p_v is the number of variables in $\boldsymbol{v}_i$, including the intercept. In an IWLS approximation, define

$$\boldsymbol{A}_g = \tilde{\boldsymbol{V}}^\top \boldsymbol{W}_g \tilde{\boldsymbol{V}}, \qquad \boldsymbol{b}_g = \tilde{\boldsymbol{V}}^\top \boldsymbol{W}_g \boldsymbol{z}_g, \tag{18}$$

where $\tilde{\boldsymbol{V}}$ is the block design matrix, $\boldsymbol{W}_g$ is the working weight matrix incorporating posterior weights τ_{ig} and, where used, class weights, and $\boldsymbol{z}_g$ is the working response vector. The unregularized IWLS update is

$$\hat{\boldsymbol{\beta}}_{g,ML} = \boldsymbol{A}_g^{-1} \boldsymbol{b}_g. \tag{19}$$

When $\boldsymbol{A}_g$ is ill-conditioned, this update may be unstable. The proposed component-specific two-parameter update is

$$\hat{\boldsymbol{\beta}}_{g,TPE}(k_g, d_g) = (\boldsymbol{A}_g + \boldsymbol{I}_r)^{-1} (\boldsymbol{A}_g + d_g \boldsymbol{I}_r)(\boldsymbol{A}_g + k_g \boldsymbol{I}_r)^{-1} \boldsymbol{b}_g, \tag{20}$$

where $k_g \geq 0$, $0 \leq d_g \leq 1$, and $\boldsymbol{I}_r$ is the $r \times r$ identity matrix. The update includes useful special cases:

$$\hat{\boldsymbol{\beta}}_{g,TPE}(0,1) = \hat{\boldsymbol{\beta}}_{g,ML}, \tag{21}$$

$$\hat{\boldsymbol{\beta}}_{g,TPE}(k_g,1) = (\boldsymbol{A}_g + k_g \boldsymbol{I}_r)^{-1} \boldsymbol{b}_g, \tag{22}$$

$$\hat{\boldsymbol{\beta}}_{g,TPE}(0,d_g) = (\boldsymbol{A}_g + \boldsymbol{I}_r)^{-1} (\boldsymbol{A}_g + d_g \boldsymbol{I}_r) \hat{\boldsymbol{\beta}}_{g,ML}. \tag{23}$$

Thus, maximum likelihood, ridge-type and Liu-type updates are obtained as limiting cases.

Because Eq. (20) is a biased regularized update, it is not automatically a true maximum-likelihood M-step. We therefore describe the algorithm as an EM-type or regularized generalized EM algorithm. In implementation, a monotonicity check can be added: accept the regularized update only if the fitted objective increases; otherwise reduce the step size toward the previous coefficient vector or use the unregularized/ridge fallback update.

4.10 Choice of shrinkage parameters

The shrinkage parameters (k_g, d_g) are selected component by component. For predictive work, the preferred choice is nested spatial-block cross-validation over a grid $\mathcal{K} \times \mathcal{D}$. For computationally cheaper fitting, a local pseudo-BIC criterion may be used:

$$(\hat{k}_g, \hat{d}_g) = \arg \min_{k \in \mathcal{K},\, d \in \mathcal{D}} \left\{ -2Q_g \left(\hat{\boldsymbol{\beta}}_{g,TPE}(k,d) \right) + a_n \, \mathrm{df}_g(k,d) \right\}, \tag{24}$$

where Q_g is the weighted conditional log-likelihood contribution for component g, $a_n = \log(n)$ gives a BIC-type penalty, and

$$\mathrm{df}_g(k,d) = \mathrm{tr}\{\boldsymbol{S}_g(k,d)\}, \tag{25}$$

with

$$\boldsymbol{S}_g(k,d) = (\boldsymbol{A}_g + \boldsymbol{I}_r)^{-1} (\boldsymbol{A}_g + d\boldsymbol{I}_r)(\boldsymbol{A}_g + k\boldsymbol{I}_r)^{-1} \boldsymbol{A}_g. \tag{26}$$

The grid for k should include zero and a range of positive values, while the grid for d should cover $[0,1]$. All continuous predictors must be standardized so that shrinkage parameters are comparable across coefficients.

4.11 Approximate mean squared error

Let

$$\boldsymbol{M}_g(k,d) = (\boldsymbol{A}_g + \boldsymbol{I}_r)^{-1} (\boldsymbol{A}_g + d\boldsymbol{I}_r)(\boldsymbol{A}_g + k\boldsymbol{I}_r)^{-1} \boldsymbol{A}_g. \tag{27}$$

Then

$$\hat{\boldsymbol{\beta}}_{g,TPE}(k,d) = \boldsymbol{M}_g(k,d) \hat{\boldsymbol{\beta}}_{g,ML}. \tag{28}$$

Under the local quadratic approximation,

$$\mathrm{Cov}\{\hat{\boldsymbol{\beta}}_{g,TPE}(k,d)\} \approx \boldsymbol{M}_g(k,d) \boldsymbol{A}_g^{-1} \boldsymbol{M}_g(k,d)^\top, \tag{29}$$

while the approximate bias is

$$\text{Bias}\{\hat{\boldsymbol{\beta}}_{g,TPE}(k,d)\} \approx \{\boldsymbol{M}_g(k,d) - \boldsymbol{I}_r\}\boldsymbol{\beta}_g. \tag{30}$$

The approximate scalar mean squared error is therefore

$$\text{MSE}_g(k,d) = \text{tr}\left[\boldsymbol{M}_g(k,d)\boldsymbol{A}_g^{-1}\boldsymbol{M}_g(k,d)^\top\right] + \boldsymbol{\beta}_g^\top\{\boldsymbol{M}_g(k,d) - \boldsymbol{I}_r\}^\top\{\boldsymbol{M}_g(k,d) - \boldsymbol{I}_r\}\boldsymbol{\beta}_g. \tag{31}$$

In applications, $\boldsymbol{\beta}_g$ can be replaced by a pilot estimate. Eq. (31) clarifies the bias–variance trade-off: shrinkage reduces variance at the cost of bias, and is useful when variance reduction dominates.

4.12 Algorithm

Algorithm 1 Regularized spatially informed cluster-weighted multinomial logit mixture

Select candidate component numbers $G = 1, \ldots, G_{\max}$.
for each G **do**
    Initialize component labels using spatially constrained k-means, Gaussian mixtures or multiple random starts.
    Initialize $\pi_g^{(0)}, \boldsymbol{\theta}_g^{(0)}, \boldsymbol{\beta}_g^{(0)}$ for $g = 1, \ldots, G$.
    repeat
        E-step: compute posterior probabilities τ_{ig} using Eq. (14).
        M-step 1: update mixture weights $\hat{\pi}_g = \frac{1}{n}\sum_i \tau_{ig}$.
        M-step 2: update covariate-distribution parameters $\boldsymbol{\theta}_g$ by weighted maximum likelihood.
        M-step 3: compute IWLS quantities $\boldsymbol{A}_g$ and $\boldsymbol{b}_g$ for each component.
        Select (k_g, d_g) using local pseudo-BIC or nested spatial-block validation.
        Compute the two-parameter update in Eq. (20).
        Accept the update if the fitted objective increases; otherwise use a damped step or fallback ridge update.
        Evaluate the unweighted log-likelihood and, where applicable, the weighted pseudo-likelihood.
    until relative change is below ϵ or maximum iterations are reached.
    Store likelihood, pseudo-likelihood, BIC, pseudo-BIC, ICL, pseudo-ICL, and entropy.
    Store active-component diagnostics and validation metrics.
end for
Select the preferred model using validation performance, active-component diagnostics and interpretability.

4.13 Model selection and active-component diagnostics

For an unweighted fitted model with m_G free parameters, the Bayesian information criterion is

$$\mathrm{BIC}(G) = -2\,\ell(\hat{\Theta}_G) + m_G \log(n), \tag{32}$$

so that lower values indicate a better penalized fit. The integrated completed likelihood adds an entropy penalty,

$$\mathrm{ICL}(G) = \mathrm{BIC}(G) - 2\sum_{i=1}^{n}\sum_{g=1}^{G} \hat{\tau}_{ig} \log(\hat{\tau}_{ig}), \tag{33}$$

which favours solutions with well-separated components. If class weights are used in fitting, the corresponding quantities are labelled pseudo-BIC and pseudo-ICL.

An active-component diagnostic is used to avoid interpreting empty or negligible components. Let $n_g^{\mathrm{eff}} = \sum_{i=1}^{n} \hat{\tau}_{ig}$ be the effective size of component g. A component is considered active if

$$n_g^{\mathrm{eff}} > \max(30,\, 0.02\,n). \tag{34}$$

5 Empirical Study: Sleipner CO_2 Benchmark

5.1 Constructing the binned reservoir table

The empirical analysis uses the Sleipner 2019 Benchmark Model from CO_2 DataShare, which provides interpreted plume-boundary outlines, well-pick information, spatial coordinates and layer-specific reservoir descriptors. Our question is deliberately narrow: can a regularized spatially informed cluster-weighted multinomial logit mixture recover interpretable latent regimes and improve the detection of minority plume states in a spatially binned, strongly imbalanced grid, and which parts of the framework are responsible for any improvement?

The plume-boundary files for layers $L1, \ldots, L9$ were parsed into spatial polygons, a 200 m grid was generated per layer, and the plume-coverage fraction of each bin was computed. A representative point of each bin supplied spatial covariates (easting, northing, reference depth, layer index, distance to the injection well), with reference depths and thicknesses taken from the well picks. Continuous covariates were standardized before fitting. Feeder and chimney files were excluded because their coordinate columns could not be parsed reliably. We therefore do not define a feeder-proximal class and do not make claims about feeder-controlled plume behaviour.

Each bin was labelled outside-plume, plume-boundary or plume-interior. The final table held $n = 3409$ bins, dominated by the background class (Table 1). Fig. 1 shows the spatial footprint of each response class across the nine layers, illustrating how the boundary and interior classes occupy a small, spatially concentrated fraction of each layer relative to the dominant background class. With roughly 87% of bins outside the plume, ordinary accuracy is a misleading target, and we evaluate throughout with balanced accuracy, macro-F1, and class-wise recall.

Table 1 Distribution of response classes in the binned Sleipner modelling table.

Response class	Number of bins	Percentage
Outside plume	2965	86.98
Plume boundary	331	9.71
Plume interior	113	3.31
Total	3409	100.00

To counter the imbalance, the conditional response model was fitted with class weights formed as an inverse-power transform of the class frequencies (exponent 0.75), capped at 12, floored at 0.20 and rescaled to unit mean weight so the effective sample size is preserved. The resulting weights were 0.578, 2.99 and 6.24 for the outside, boundary and interior classes. Because these weights change the fitting objective, all weighted model-selection criteria are reported as pseudo-BIC and pseudo-ICL. Specifically, if n_j is the number of observations in class j, preliminary weights were computed as $w_j^* \propto n_j^{-0.75}$, truncated to the interval $[0.20, 12]$, and rescaled so that the sample average weight was one. These weights were used only in the conditional response part of the fitting objective.

5.2 Models compared

The proposed mixture was fitted for $G = 1, \ldots, 5$ with three random starts each. In the fitted implementation the component mixing weights are constant and spatial heterogeneity is carried by the spatial covariates inside each component's covariate density. The model is benchmarked against a graded set of alternatives so that the source of any improvement can be isolated: a majority-class baseline; a global multinomial logit (MLE); the same with class weights; a global two-parameter (TPE) multinomial logit; and the cluster-weighted mixture with, in turn, an unregularized (MLE) M-step, a ridge M-step and the proposed two-parameter M-step. A non-spatial mixture (covariate density without the horizontal coordinates) is included to test whether spatial position adds anything beyond depth and layer.

5.3 Selecting the number of latent regimes

The selection diagnostics are in Table 2. The weighted information criteria favour larger mixture candidates, with the best raw pseudo-BIC and pseudo-ICL values occurring among fits that contain empty or negligible components. Specifically, the $G = 4$ and $G = 5$ candidates have one and two empty components, respectively, and therefore cannot be interpreted at face value. Fig. 2 plots the pseudo-BIC and pseudo-ICL criteria in Table 2 across candidate mixture sizes, showing the sharp improvement from $G = 1$ to $G = 2$ and the subsequent plateau among the degenerate larger candidates. A strict rule admitting only fits with no empty components selects $G = 2$. Pruning the negligible components from the $G = 5$ candidate leaves three active regimes, each above the minimum effective-size threshold of $\max(30, 0.02n) = 68.2$ bins. We therefore report a three-active-component solution and keep $G = 2$ as a

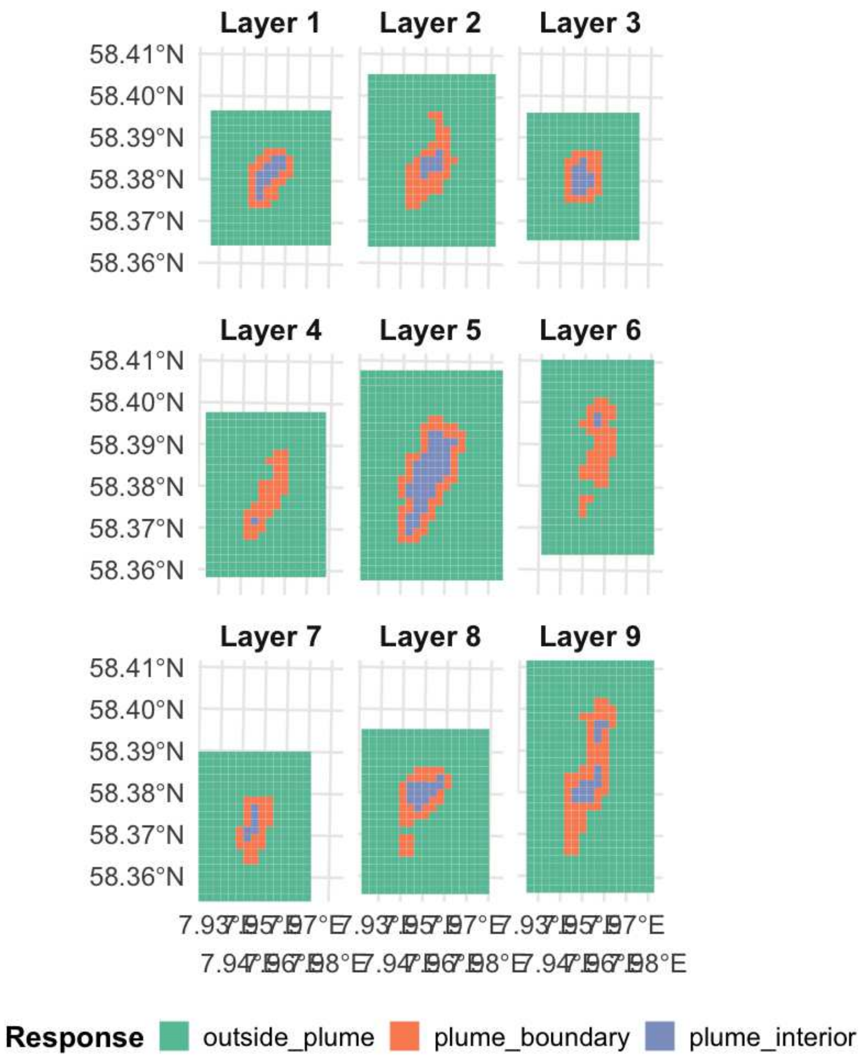

Fig. 1 Observed plume-state classes on the binned Sleipner grid, by layer.

conservative reference, and we are explicit that three is what survives pruning rather than a value any clean criterion selects directly.

Table 2 Model-selection diagnostics across candidate mixture sizes. Weighted criteria are pseudo-BIC and pseudo-ICL (class-weighted fitting changes the likelihood target). Lower is better.

Candidate G	Active G	Empty	Objective	Pseudo-BIC	Pseudo-ICL	Degenerate
5	3	2	46351	−89659	−89375	Yes
4	3	1	46350	−90269	−89982	Yes
2	2	0	44771	−88330	−88330	No
3	2	1	44771	−87720	−87720	Yes
1	1	0	36189	−71775	−71775	No

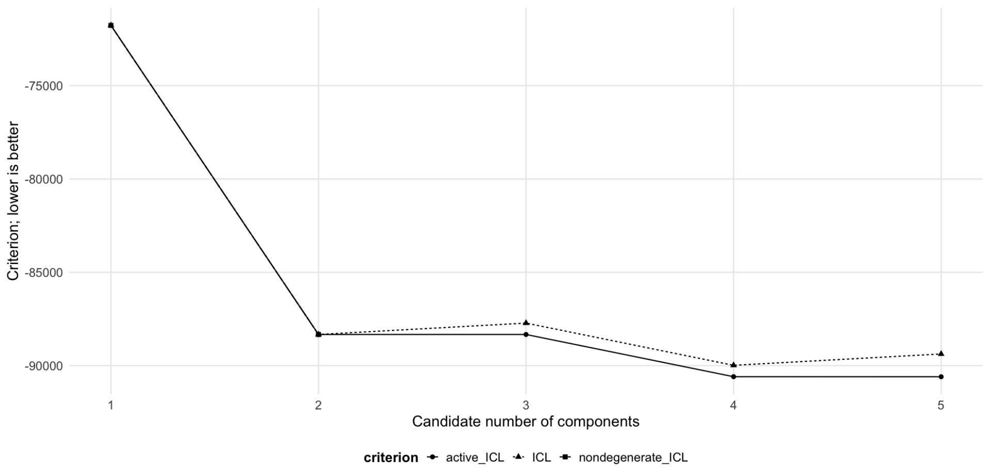

Fig. 2 Model-selection criteria across candidate mixture sizes.

5.4 The latent regimes

The three regimes separate mainly by depth and stratigraphic position (Table 3). One regime concentrates on the deeper stack (mean layer index 8.19, mean depth 0.954 km, mixing weight 0.352); the other two sit in shallower to intermediate layers with very similar depths (mean layer indices 3.81 and 3.69). The mixture has, in effect, partitioned the reservoir along a deep–shallow axis. Fig. 3 maps the three estimated regimes across layers, showing this deep-shallow partition spatially and confirming that regime 1 dominates the deeper layers while regimes 2 and 3 partition the shallower layers. This is consistent with the layered Utsira interval and is reassuring as a geological sanity check. At the same time, it indicates that the latent structure is partly re-expressing known stratigraphic organization rather than discovering wholly new geological units. These are statistical regimes, not facies; their practical value is the posterior membership probability, and therefore the uncertainty, attached to every bin. Fig. 4 maps this posterior uncertainty across the grid, showing that it concentrates along the spatial transitions between regimes.

Table 3 Active latent components after pruning. Effective n is the sum of posterior responsibilities.

Active	Raw	Weight	Effective n	Mean depth (km)	Mean layer
1	2	0.352	1201	0.954	8.19
2	4	0.236	805	0.863	3.81
3	5	0.411	1403	0.861	3.69

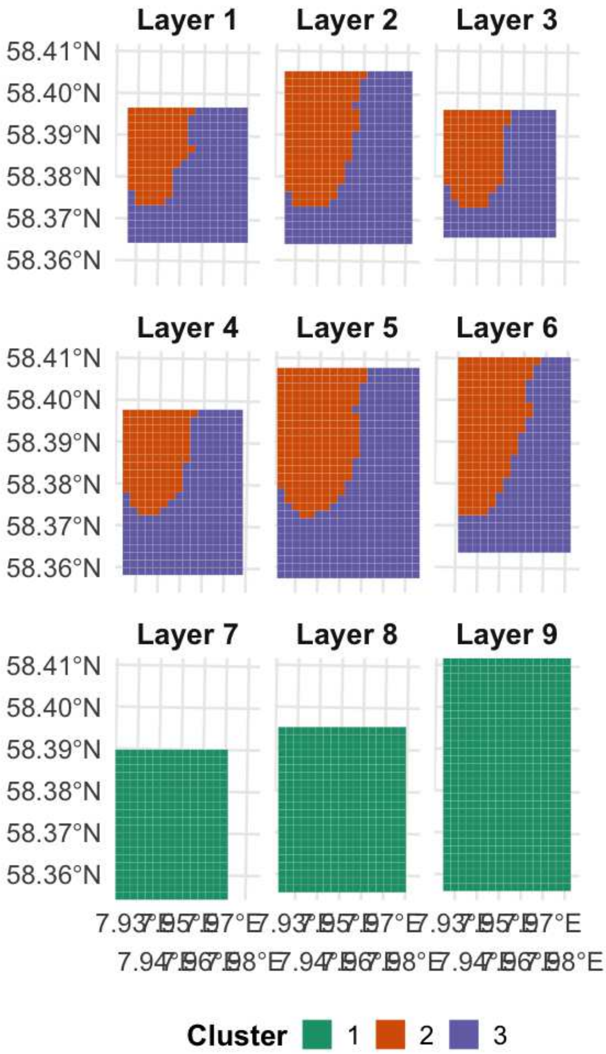

Fig. 3 Estimated latent reservoir regimes on the binned Sleipner grid, by layer.

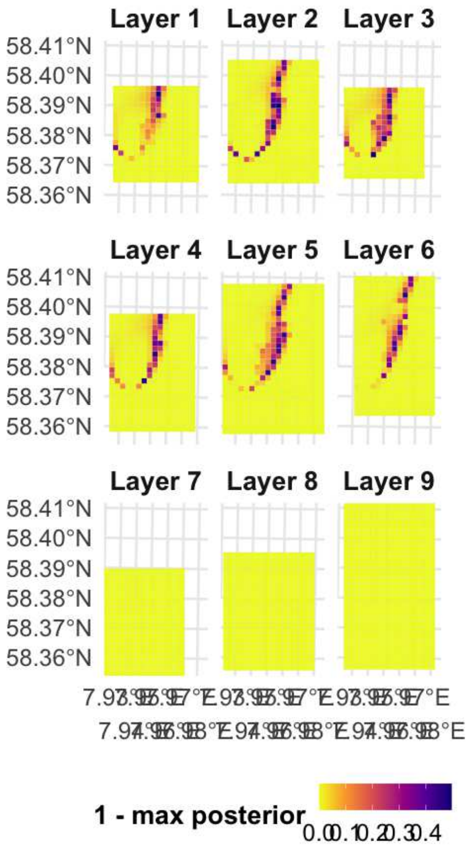

Fig. 4 Posterior uncertainty for latent-regime assignment (one minus the maximum posterior component probability).

5.5 Conditioning and regime-specific shrinkage

The two-parameter estimator is justified only if the local response models are ill-conditioned, so we measured it (Table 4). All three components are severely ill-conditioned, with condition numbers of order 10^9. Crucially, the selected shrinkage varies across regimes in the direction the method predicts: the most ill-conditioned component (1) is pulled toward strong shrinkage ($\hat{k} = 0.5$, $\hat{d} = 0.5$), whereas the better-conditioned components (2 and 3) remain close to maximum likelihood ($\hat{k} = 0.01$, $\hat{d} = 0.95$ and 0.80). This provides diagnostic evidence that the selected amount of regularization varies across latent regimes, with stronger shrinkage selected for the most ill-conditioned local fit.

Two cautions accompany this table. First, the extreme magnitude of cond(A_g), and especially the maximum VIF in component 1, is inflated by covariates that are constant or nearly constant in the present modelling table. In the final implementation, such variables should be removed before fitting or replaced with spatially varying petrophysical information. Second, as the cross-validation below shows, adaptive shrinkage stabilizes local estimation but does not produce a statistically distinguishable predictive gain on this benchmark.

Table 4 Local conditioning and selected shrinkage by active component. (Magnitudes of $\mathrm{cond}(\boldsymbol{A}_g)$ and VIF are inflated by two constant covariates; see text.)

Active component	Effective n	$\mathrm{cond}(\boldsymbol{A}_g)$	Max VIF	$\hat{k}_g$	$\hat{d}_g$
1	1201	5.2×10^9	5.7×10^8	0.50	0.50
2	805	2.7×10^9	125	0.01	0.95
3	1403	4.1×10^9	113	0.01	0.80

5.6 Apparent (in-sample) performance

We first report apparent performance (Table 5) as a diagnostic only; it is computed on the bins used for fitting, and with spatially autocorrelated data it is optimistic. The pattern is nonetheless instructive. The majority baseline wins ordinary accuracy (0.870) by never predicting a plume class, with balanced accuracy 0.333; the unweighted global models behave the same way (0.864/0.363). Class weighting lifts the global model to about 0.61 balanced accuracy, and the mixture to 0.732. Tellingly, global TPE and global MLE are identical here, and the two class-weighted global models differ only at the third decimal – a first indication, confirmed below, that two-parameter shrinkage contributes little. Fig. 5 displays these results graphically, contrasting ordinary accuracy with balanced accuracy and macro F1 across the fitted models.

Table 5 Apparent (in-sample) classification performance. Diagnostic only; claims rest on Table 7.

Model	Accuracy	Balanced acc.	Macro prec.	Macro recall	Macro F1	Brier	Log-loss
Proposed spatial CWM-MNL-TPE	0.847	0.732	0.603	0.732	0.648	0.214	0.380
Class-weighted global MNL-MLE	0.785	0.611	0.490	0.611	0.523	0.296	0.532
Class-weighted global MNL-TPE	0.785	0.609	0.489	0.609	0.522	0.296	0.532
Global MNL-TPE	0.864	0.363	0.499	0.363	0.367	0.189	0.347
Global MNL-MLE	0.864	0.363	0.495	0.363	0.367	0.189	0.347
Majority-class baseline	0.870	0.333	0.290	0.333	0.310	–	–

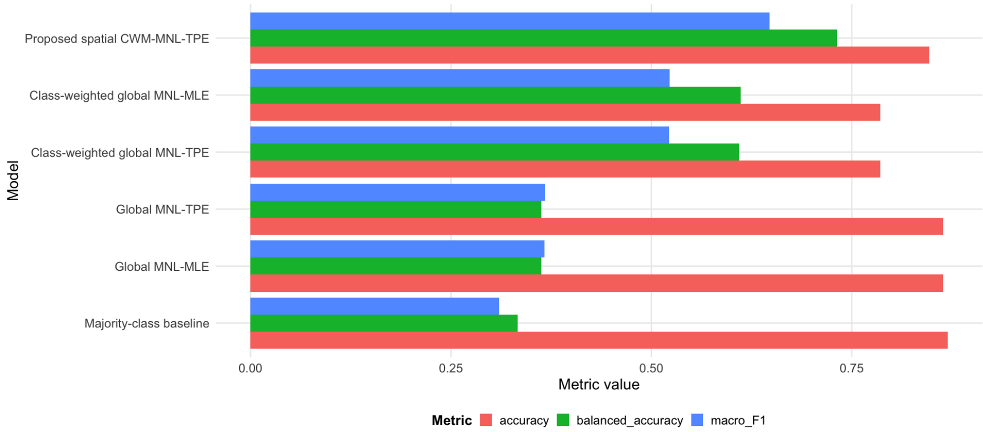

Fig. 5 Ordinary versus balanced performance across fitted models (apparent).

5.7 Spatial-block cross-validation

Because adjacent 200 m bins are correlated and plume classes are spatially concentrated, we evaluate generalization with spatial-block cross-validation: the easting–northing plane is tiled into contiguous 1 km blocks, all bins in a block are kept in one fold, and blocks are assigned so minority bins are spread across folds (Table 6). Mixture predictions use covariate-only gating, so the response label never enters the assignment of a test bin to a component; shrinkage was selected on the training folds only.

For a held-out bin, predicted class probabilities were computed as

$$\Pr(Y_i = j \mid s_i, x_i, c_i) = \sum_{g=1}^{G} \Pr(Z_i = g \mid s_i, x_i, c_i) \Pr(Y_i = j \mid x_i, c_i, Z_i = g), \tag{35}$$

where

$$\Pr(Z_i = g \mid s_i, x_i, c_i) = \frac{\hat{\pi}_g f_g(s_i, x_i, c_i; \hat{\theta}_g)}{\sum_{r=1}^{G} \hat{\pi}_r f_r(s_i, x_i, c_i; \hat{\theta}_r)}. \tag{36}$$

The observed response Y_i is not used in this test-time gating calculation.

Table 6 Class composition of the five spatial-block folds (1 km blocks, minority-stratified).

Fold	Outside plume	Plume boundary	Plume interior
1	628	102	62
2	672	78	13
3	589	49	13
4	551	55	14
5	525	47	11

The cross-validated comparison is in Table 7. Three things stand out, and we report them plainly. First, honest generalization is well below the apparent figure: the proposed model attains a balanced accuracy of 0.561 (SD 0.093) out of sample, against 0.732 in-sample, the difference being exactly the optimism that label-informed responsibilities introduce. Second, the proposed model is the best point estimate, but only marginally: the class-weighted global multinomial logit reaches 0.548–0.550 and the unregularized and ridge mixtures reach 0.547 and 0.550. With only five spatial folds, paired tests have limited power. A paired Wilcoxon comparison does not separate the proposed model from the class-weighted global model, the ridge mixture, or the unregularized mixture ($p = 0.31$–1.0). Third, the large gaps are elsewhere: the proposed model beats the majority baseline (0.333), the unweighted global model (0.343) and the non-spatial mixture (0.422) in every fold ($p = 0.0625$, the smallest attainable value at five folds). In other words, the balanced-class performance is driven by class weighting and by the spatial covariates, not by the latent mixture or the two-parameter shrinkage. Fig. 6 plots the cross-validated balanced accuracy with fold-wise standard deviations for each model, illustrating the overlap in error bars among the top-performing candidates.

Table 7 Spatial-block cross-validation (5 folds, 1 km blocks; covariate-only gating; training-fold shrinkage selection). "Fold" means SD for balanced accuracy in parentheses. CB = class-balanced. Best per column in bold.

Model	Balanced acc.	Macro F1	Recall (bdy)	Recall (int)	CB log-loss	CB Brier
Proposed spatial CWM-MNL-TPE	**0.561** (0.093)	**0.473**	0.394	**0.451**	1.27	**0.617**
Class-weighted global MNL-TPE	0.550 (0.085)	0.472	**0.430**	0.397	1.28	0.618
Spatial CWM-MNL-ridge	0.550 (0.089)	0.458	0.406	0.422	1.37	0.627
Class-weighted global MNL-MLE	0.548 (0.085)	0.470	0.421	0.400	1.32	0.619
Spatial CWM-MNL-MLE	0.547 (0.088)	0.455	0.392	0.422	1.38	0.626
Non-spatial CWM-MNL-TPE	0.422 (0.068)	0.364	0.174	0.306	**1.26**	0.693
Global MNL-MLE	0.343 (0.019)	0.330	0.037	0.010	1.99	0.889
Majority-class baseline	0.333 (0.000)	0.311	0.000	0.000	18.4	1.33

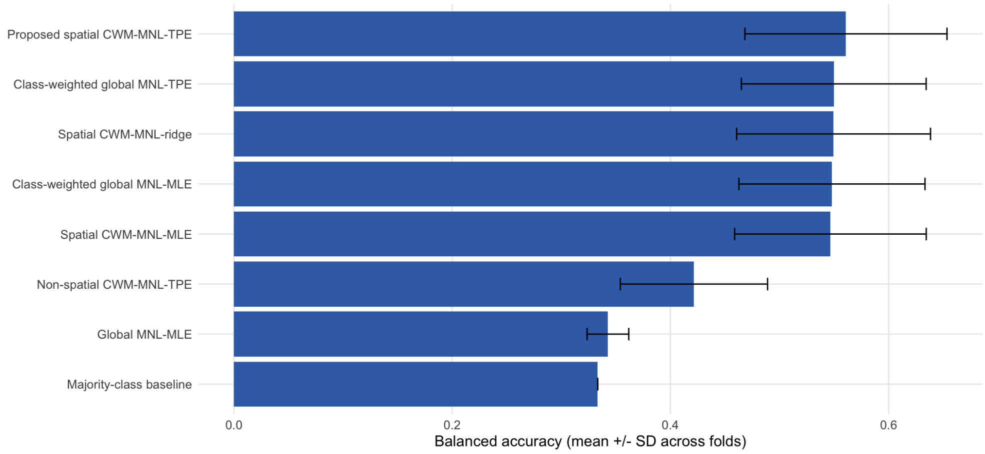

Fig. 6 Spatial-block cross-validated balanced accuracy by model (mean ± SD across folds).

A leave-one-layer-out analysis (Table 8) tests stratigraphic transfer and is, if anything, less favourable to the regularization. Held-out balanced accuracy falls to 0.490 for the proposed model and 0.507 for the unregularized mixture, with the class-weighted global models close behind (0.487–0.488). Transfer across unseen layers is genuinely hard, the class-balanced log-loss and Brier degrade for every model, and the two-parameter shrinkage provides no advantage in this setting; the plain mixture transfers best on this metric. We report this candidly because it bounds the strength of any predictive claim.

Table 8 Leave-one-layer-out performance (train on eight layers, predict the ninth; fold means).

Model	Balanced acc.	Macro F1	Recall (bdy)	Recall (int)
Spatial CWM-MNL-MLE	**0.507**	0.390	**0.521**	0.214
Proposed spatial CWM-MNL-TPE	0.490	0.393	0.459	0.222
Class-weighted global MNL-TPE	0.488	**0.402**	0.370	**0.293**
Class-weighted global MNL-MLE	0.487	0.401	0.367	0.293
Spatial CWM-MNL-ridge	0.457	0.345	0.414	0.222
Non-spatial CWM-MNL-TPE	0.403	0.268	0.459	0.222
Global MNL-MLE	0.363	0.323	0.026	0.111
Majority-class baseline	0.333	0.311	0.000	0.000

5.8 Ablation: where the improvement comes from

The ablation in Table 9 runs under the same folds and changes one element at a time, and it resolves the question the title raises. Moving from the unweighted global model to the class-weighted global model is the decisive step (balanced accuracy 0.343 → 0.548). Adding the latent mixture on top of class weighting moves balanced accuracy only from 0.548 to 0.547–0.561, a change within fold-to-fold noise. Within the mixture, ridge and two-parameter shrinkage raise balanced accuracy from 0.547 (MLE) to 0.550 and 0.561, again inside the noise band. Removing the horizontal coordinates, by contrast, is costly: the non-spatial mixture falls to 0.422. The conclusion we draw, and state, is that the balanced recovery of minority plume bins is driven by class weighting and spatial covariates; the mixture contributes interpretability and adaptive local stabilization rather than a measurable accuracy gain; and the two-parameter shrinkage yields the best point estimate but no statistically distinguishable improvement on this benchmark. Fig. 7 visualizes these ablation results across balanced accuracy, macro F1 and class-wise recall for each model variant.

Table 9 Ablation under spatial-block cross-validation. Each adjacent contrast isolates one component. Fold means.

Model variant	Balanced acc.	Macro F1	Recall (bdy)	Recall (int)	CB log-loss
Class-weighted global MNL-MLE	0.548	0.470	0.421	0.400	1.32
Class-weighted global MNL-TPE	0.550	0.472	0.430	0.397	1.28
Spatial CWM-MNL-MLE	0.547	0.455	0.392	0.422	1.38
Spatial CWM-MNL-ridge	0.550	0.458	0.406	0.422	1.37
Proposed spatial CWM-MNL-TPE	**0.561**	**0.473**	0.394	0.451	1.27
Non-spatial CWM-MNL-TPE	0.422	0.364	0.174	0.306	1.26

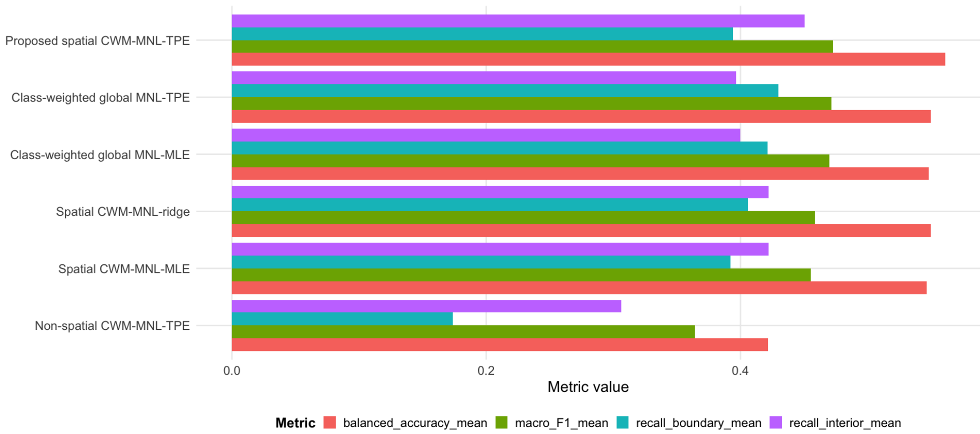

Fig. 7 Ablation under spatial-block cross-validation.

5.9 Class-wise behaviour

The class-wise breakdown of the selected model and its confusion matrix are in Tables 10–12. The model is strong on the background class and, importantly, recovers minority bins that the global models miss: in-sample recall is 0.647 for boundary and 0.673 for interior bins, and out of sample (Table 7) it holds 0.39–0.45 recall for those classes against essentially zero for the unweighted global model. Boundary precision is modest by construction, since the boundary is a transition category; the row-normalized matrix shows the errors are mostly confusions with the adjacent interior or outside classes rather than wholesale misses. Fig. 8 presents this row-normalized confusion matrix as a heatmap, and Fig. 9 compares class-wise recall across the majority baseline, the class-weighted global model and the proposed mixture. For monitoring-oriented screening, this error structure may be preferable to a classifier that misses plume-associated bins entirely, although the relative cost of false boundary alarms and missed plume bins should ultimately be specified by the monitoring objective.

Table 10 Class-wise performance of the selected model (in-sample).

Response class	Support	Predicted	Precision	Recall	F1
Outside plume	2965	2666	0.974	0.876	0.922
Plume boundary	331	580	0.369	0.647	0.470
Plume interior	113	163	0.466	0.673	0.551

Table 11 Confusion matrix for the selected model (in-sample).

Observed \ Predicted	Outside plume	Plume boundary	Plume interior
Outside plume	2597	334	34
Plume boundary	64	214	53
Plume interior	5	32	76

Table 12 Row-percent confusion matrix for the selected model (in-sample).

Observed \ Predicted	Outside plume	Plume boundary	Plume interior
Outside plume	87.59	11.26	1.15
Plume boundary	19.34	64.65	16.01
Plume interior	4.42	28.32	67.26

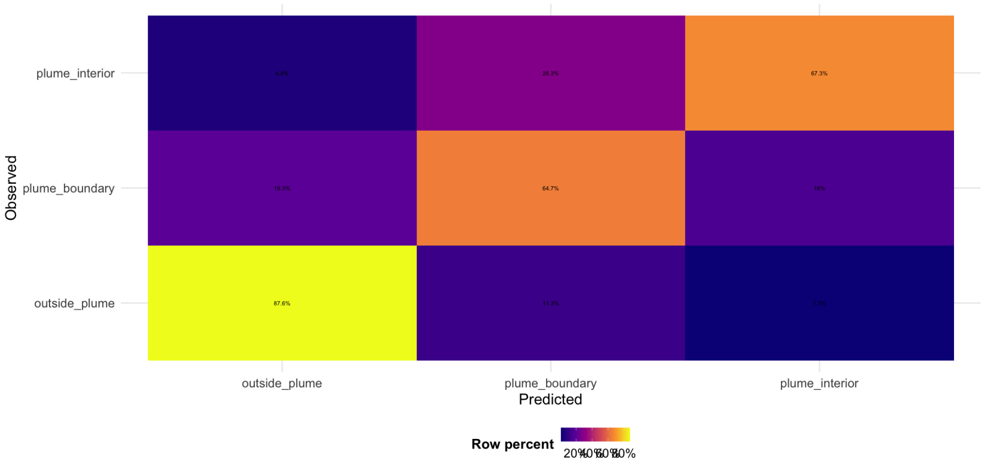

Fig. 8 Row-normalized confusion matrix for the selected model.

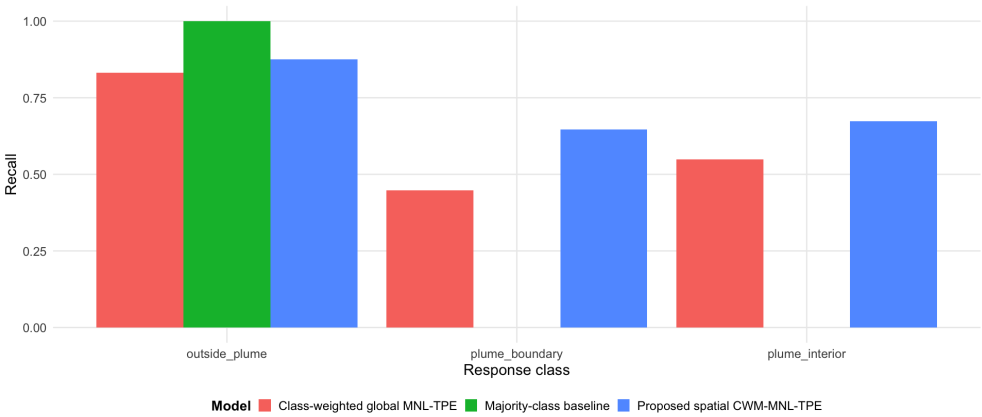

Fig. 9 Class-wise recall across the majority baseline, the class-weighted global model and the proposed mixture.

5.10 What the empirical study shows

The honest summary is more nuanced than "the proposed model wins." Under spatial-block cross-validation the proposed mixture has the best point estimate of balanced accuracy (0.561) and the best macro F1 (0.473), and it recovers a substantial share of the minority plume bins that the global baselines effectively ignore. But the margin over a class-weighted global multinomial logit (0.548–0.550) and over its own unregularized and ridge variants is small and not statistically separable at five folds, and under leave-one-layer-out the unregularized mixture transfers slightly better. The ablation locates the real drivers: class weighting (worth about 0.21 in balanced accuracy) and spatial covariates (whose removal costs about 0.14). The latent mixture adds interpretable regimes, model-based posterior uncertainty and adaptive local stabilization rather than a clearly measurable accuracy gain. The conditioning analysis suggests that the two-parameter shrinkage operates more strongly in the worse-conditioned regime, which is the behaviour the update was designed to produce, but its predictive payoff on this benchmark is modest.

6 Discussion

The Sleipner application makes a clean point about modelling strongly imbalanced spatial reservoir data, and it makes it in a way that is more measured than we had initially expected. A global multinomial logit, faced with a background class that fills 87% of the grid, simply predicts the background almost everywhere; its ordinary accuracy is high and its balanced accuracy is barely above chance. The single most effective remedy is not a richer model but a corrected objective: weighting the rare classes raises balanced accuracy from about 0.34 to about 0.55 on its own. Spatial information is the second lever, as the collapse of the non-spatial mixture shows. Against that backdrop, the latent mixture and the two-parameter shrinkage each add only a small, statistically indistinguishable increment to predictive accuracy.

We think this is still a worthwhile result, provided the contribution is framed for what it is. The mixture's value is not primarily a higher accuracy number; it is interpretability and model-based posterior uncertainty. The model returns three depth-organized regimes consistent with the Utsira architecture, a posterior membership probability for every bin, and uncertainty maps that flag exactly the transition zones a monitoring analyst would want to scrutinize. The two-parameter estimator's value, similarly, is mechanistic rather than headline-grabbing: the conditioning analysis shows that the local information matrices are severely ill-conditioned and that the selected shrinkage adapts across regimes, applying strong shrinkage where conditioning is worst and reverting to near-maximum-likelihood where it is not. That adaptivity stabilizes the local coefficients, which matters for interpreting the component-specific effects even when it does not move the aggregate score. A reader interested in robust local inference under collinearity should care about this even though a reader interested only in a leaderboard number would not.

Three methodological lessons carry beyond Sleipner. The first concerns evaluation: Ordinary accuracy and unbalanced aggregate probabilistic scores can reward predicting the majority class, so they are not sufficient for this problem. Balanced accuracy, macro F1, class-wise recall and class-balanced probabilistic scores are more appropriate yardsticks for minority plume-state recovery. The second concerns the gap between apparent and predictive performance: the in-sample balanced accuracy of 0.732 fell to 0.561 under honest spatial-block evaluation, and reporting the in-sample figure as a result would have overstated the model by a wide margin. The third concerns mixture reporting: raw information criteria here prefer fit with empty components, and pruning to effective-size-bearing components while reporting the strict non-degenerate model as a reference is a small discipline worth adopting in any spatial mixture study.

Several limitations bound the claims. The labels are constructed from interpreted plume polygons and bin-level rules, so the class distribution depends on bin size, interior threshold and boundary buffer, and a sensitivity sweep over those choices belongs in the final version. Feeder and leakage information was set aside because the coordinates could not be parsed, which both narrows the response and removes a class of scientifically interesting bins. The fitted model uses constant mixing weights with spatial covariates in the component density rather than the spatial gating function of the methodology; the two should be reconciled. The reported local conditioning is also inflated by two petrophysical covariates that are constant across the table and should be dropped or replaced. Finally, this is a statistical classifier built on interpreted boundaries, and it should be read alongside seismic interpretation and flow simulation rather than as a substitute.

7 Conclusion

We proposed and evaluated a spatially informed cluster-weighted multinomial logit mixture with component-specific two-parameter regularized updates for plume-state classification in imbalanced, binned reservoir data, and applied it to the Sleipner 2019 benchmark. The framework recovers three interpretable, depth-organized latent statistical regimes; attaches model-based posterior uncertainty to every bin; and, under spatial-block cross-validation, recovers a substantial fraction of the minority plume-boundary and plume-interior bins missed by unweighted baselines. It reaches a cross-validated balanced accuracy of 0.561, compared with 0.333 for the majority classifier and 0.343 for an unweighted global multinomial logit.

The candid finding is that this balanced recovery is driven primarily by class weighting and by the use of spatial covariates, while the latent mixture and the two-parameter shrinkage each contribute small, statistically indistinguishable increments on this benchmark. The mixture's enduring value is therefore interpretability and model-based posterior uncertainty rather than raw accuracy, and the two-parameter estimator's value is adaptive, regime-specific stabilization of local multinomial logit updates. On this benchmark, however, that stabilization does not translate into a statistically distinguishable predictive improvement over simpler class-weighted alternatives. Read this way, the contribution is an interpretable, uncertainty-aware and honestly validated framework for minority plume-state recovery under severe imbalance, rather than a claim that two-parameter regularization improves classification accuracy.

8 Future work

Four extensions would sharpen the conclusions. First, the response construction should be stress-tested by varying bin size (for example 100, 200 and 300 m), the interior threshold and the boundary buffer, reporting the range of outcomes rather than a single configuration. Second, the constant petrophysical covariates should be replaced with spatially varying porosity and permeability, the feeder and leakage information reinstated once coordinates are verified, and the spatial gating function implemented so that the code matches the stated model; each of these directly addresses a limitation above. Third, because the differences among the class-weighted variants are small relative to fold variance at five folds, the final evaluation should use ten spatial-block folds, which would also let the paired tests separate the proposed model from the weak baselines below the conventional threshold. Fourth, a controlled simulation study generating spatial bins from known regimes with tunable collinearity and imbalance would isolate the conditions under which two-parameter shrinkage measurably improves on ridge and maximum likelihood, putting the regularization claim on a footing that a single benchmark cannot provide; the natural application extension is to the time-lapse Sleipner surveys, allowing the latent regimes and plume-state probabilities to evolve across seismic vintages.

Acknowledgements

The authors thank the CO_2 Storage Data Consortium and CO_2 DataShare for providing public access to the Sleipner 2019 Benchmark Model. The authors acknowledge institutional support through a PhD scholarship from the Department of Economics, Management and Statistics, University of Milano-Bicocca.

Statements and Declarations

Funding

No external research grant or project-specific funding was received for this study.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

The Sleipner 2019 Benchmark Model used in this study is publicly available through the CO_2 Storage Data Consortium and CO_2 DataShare, subject to the terms of the original dataset licence.

Code availability

The analysis code and derived modelling materials are available from the corresponding author upon reasonable request, subject to the terms of the original dataset licence.

Materials availability

Not applicable.

References

Ahmadinia M, Shariatipour SM, Andersen O, et al (2020) Quantitative evaluation of the joint effect of uncertain parameters in CO2 storage in the Sleipner project, using data-driven models. International Journal of Greenhouse Gas Control 103:103180. https://doi.org/10.1016/j.ijggc.2020.103180

Arts R, Eiken O, Chadwick A, et al (2004) Monitoring of CO2 injected at Sleipner using time-lapse seismic data. Energy 29(9–10):1383–1392. https://doi.org/10.1016/j.energy.2004.03.072

Bashir S, Carter EM (2010) Penalized multinomial mixture logit model. Computational Statistics 25(1):121–141

Biernacki C, Celeux G, Govaert G (2000) Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(7):719–725. https://doi.org/10.1109/34.865189

Bruer G, Gahlot AP, Chow E, et al (2024) Seismic monitoring of CO2 plume dynamics using ensemble Kalman filtering. arXiv preprint arXiv:240905193 URL https://arxiv.org/abs/2409.05193

le Cessie S, van Houwelingen JC (1992) Ridge estimators in logistic regression. Journal of the Royal Statistical Society: Series C 41(1):191–201. https://doi.org/10.2307/2347628

Chadwick RA, Noy DJ (2010) History-matching flow simulations and time-lapse seismic data from the Sleipner CO2 plume. Petroleum Geology Conference Series 7:1171–1182. https://doi.org/10.1144/0071171

Chadwick RA, Noy DJ (2015) Underground CO2 storage: Demonstrating regulatory conformance by convergence of history-matched modelled and observed CO2 plume behavior using Sleipner time-lapse seismics. Greenhouse Gases: Science and Technology 5(3):305–322. https://doi.org/10.1002/ghg.1488

Chadwick RA, Arts R, Eiken O, et al (2004) 4D seismic imaging of an injected CO2 plume at the Sleipner field, central North Sea. In: 3D Seismic Technology: Application to the Exploration of Sedimentary Basins, vol 29. Geological Society of London, London, p 311–320

CO2 DataShare (2020) Sleipner 2019 benchmark model. Dataset, URL https://co2datashare.org/dataset/sleipner-2019-benchmark-model, accessed 26 June 2026

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B 39(1):1–38

Farghali RA, Qasim M, Kibria BMG, et al (2023) Generalized two-parameter estimators in the multinomial logit regression model: Methods, simulation and application. Communications in Statistics – Simulation and Computation 52(7):3327–3342. https://doi.org/10.1080/03610918.2021.1934023

Frühwirth-Schnatter S (2006) Finite Mixture and Markov Switching Models. Springer, New York

Furre AK, Eiken O, Alnes H, et al (2024) Sleipner 26 years: How well-established subsurface monitoring methods can demonstrate safe CO2 storage. Geoenergy https://doi.org/10.1144/geoenergy2024-015

Gershenfeld N (1997) Nonlinear inference and cluster-weighted modeling. Annals of the New York Academy of Sciences 808(1):18–24

Gershenfeld N, Schöner B, Metois E (1999) Cluster-weighted modelling for time-series analysis. Nature 397:329–332. https://doi.org/10.1038/16870

Grün B, Leisch F (2008) Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. Journal of Classification 25(2):225–247. https://doi.org/10.1007/s00357-008-9022-8

Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(1):55–67. https://doi.org/10.1080/00401706.1970.10488634

Ingrassia S, Minotti SC, Vittadini G (2012) Local statistical modeling via a cluster-weighted approach with elliptical distributions. Journal of Classification 29(3):363–401. https://doi.org/10.1007/s00357-012-9114-3

Ingrassia S, Minotti SC, Punzo A (2014) Model-based clustering via linear cluster-weighted models. Computational Statistics & Data Analysis 71:159–182. https://doi.org/10.1016/j.csda.2013.02.012

Ingrassia S, Punzo A, Vittadini G, et al (2015) The generalized linear mixed cluster-weighted model. Journal of Classification 32:85–113. https://doi.org/10.1007/s00357-015-9175-1

Liu K (1993) A new class of biased estimate in linear regression. Communications in Statistics – Theory and Methods 22(2):393–402. https://doi.org/10.1080/03610929308831027

Luce RD (1959) Individual Choice Behavior: A Theoretical Analysis. John Wiley & Sons, New York

McFadden D (1974) Conditional logit analysis of qualitative choice behavior. In: Zarembka P (ed) Frontiers in Econometrics. Academic Press, New York, p 105–142

McLachlan GJ, Peel D (2000) Finite Mixture Models. John Wiley & Sons, New York, https://doi.org/10.1002/0471721182

Olobatuyi K, Ariyo O (2022) Multinomial cluster-weighted models for high-dimensional data. arXiv preprint arXiv:220811221

Papastamoulis P (2023) Model based clustering of multinomial count data. Advances in Data Analysis and Classification https://doi.org/10.1007/s11634-023-00547-5

Romero J, Luiken N, Ravasi M (2023) Seeing through the CO2 plume: Joint inversion-segmentation of the Sleipner 4D seismic dataset. arXiv preprint arXiv:230311662 URL https://arxiv.org/abs/2303.11662

Schaefer RL, Roi LD, Wolfe RA (1984) A ridge logistic estimator. Communications in Statistics – Theory and Methods 13(1):99–113. https://doi.org/10.1080/03610928408828664

Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6(2):461–464. https://doi.org/10.1214/aos/1176344136

Titterington DM, Smith AFM, Makov UE (1985) Statistical Analysis of Finite Mixture Distributions. John Wiley & Sons, New York

Yang H, Chang X (2010) A new two-parameter estimator in linear regression. Communications in Statistics – Theory and Methods 39(6):923–934. https://doi.org/10.1080/03610920902807911