# Can Metadata Guide Variable Selection for Macroeconomic Forecasting?

Andrew R. Garcia* Marco Vega†

March 24, 2026

**Abstract**

This paper asks whether structural metadata from an institutional registry carries enough signal to guide variable selection in macroeconomic forecasting. We study this question using two complementary approaches. The first, Metadata $\varepsilon$-Greedy, is a stochastic search policy that uses permutation-invariant embeddings of registry metadata to guide a fixed-budget search over predictor subsets, with forecasting loss as the only feedback signal. The second, Metadata Bayes, performs variable selection entirely within metadata space: it constructs group-level priors from institutional descriptors, updates them via partial correlation with the target, and selects predictors through Thompson sampling, without ever evaluating a forecasting model during selection.

Both methods are evaluated on forecasting Peruvian headline CPI under two forecasters, a Vector Autoregression (VAR) and a Random Forest, and benchmarked against random search, greedy forward selection, LASSO, Bayesian Ridge, PCA, and a state-of-the-art Bayesian variable selection method. Metadata Bayes, despite never observing forecasting loss during selection, achieves out-of-sample accuracy competitive with all baselines including the Bayesian benchmark. Metadata $\varepsilon$-Greedy further improves on these results under the VAR during the COVID shock period. Together, the results suggest that registry metadata encodes enough economic structure to serve as a meaningful proxy for predictive relevance, complementing rather than replacing existing forecasting pipelines.

**Keywords:** Variable selection, Metadata, Economic forecasting, Bayesian methods, Stochastic search.

*Universidad de Ingeniería y Tecnología
†Banco Central de Reserva del Perú

# ¿Pueden los metadatos guiar la selección de variables en la predicción macroeconómica?

**Abstract**

Este trabajo investiga si los metadatos estructurales de un registro institucional contienen suficiente señal para guiar la selección de variables en la predicción macroeconómica. Abordamos esta pregunta mediante dos enfoques complementarios. El primero, Metadata $\varepsilon$-Greedy, es una política de búsqueda estocástica que utiliza representaciones invariantes a permutaciones de los metadatos del registro para guiar una búsqueda con presupuesto fijo sobre subconjuntos de predictores, utilizando la pérdida de predicción como única señal de retroalimentación. El segundo, Metadata Bayes, realiza la selección de variables completamente a partir de los metadatos: construye priors a nivel de grupo a partir de descriptores institucionales, los actualiza mediante correlaciones parciales con la variable objetivo y selecciona predictores mediante muestreo de Thompson, sin evaluar en ningún momento un modelo de predicción durante la selección.

Ambos métodos se evalúan en la predicción de la inflación general del Perú (CPI headline) utilizando dos modelos de predicción, un Vector Autorregresivo (VAR) y un Random Forest, y se comparan contra búsqueda aleatoria, selección hacia adelante greedy, LASSO, Bayesian Ridge, PCA y un método de selección de variables bayesiano de última generación. Metadata Bayes, a pesar de no observar la pérdida de predicción durante la selección, alcanza una precisión fuera de muestra comparable a todos los métodos base, incluyendo el benchmark bayesiano. Metadata $\varepsilon$-Greedy mejora adicionalmente estos resultados bajo el VAR durante el período de choque de COVID. En conjunto, los resultados sugieren que los metadatos del registro codifican suficiente estructura económica como para servir como un proxy significativo de relevancia predictiva, complementando, en lugar de reemplazar, los pipelines de predicción existentes.

**Palabras clave:** Selección de variables, Metadatos, Predicción económica, Métodos bayesianos, Búsqueda estocástica.

# 1 Introduction

Macroeconomic datasets maintained by central banks contain hundreds of monthly and quarterly indicators covering external accounts, fiscal aggregates, prices, credit, expectations, and activity. These rich information sets create the potential for improved forecasting performance, but they also create a practical challenge: the number of possible predictor subsets grows combinatorially with the size of the candidate pool, making exhaustive search infeasible. Traditional approaches such as judgmental variable selection or one-time penalized regressions often fail to adapt when data coverage, relevance, or structural conditions shift over time.

This paper asks a focused empirical question: does structural metadata from an institutional registry carry enough signal to guide variable selection for macroeconomic forecasting? Central bank registries typically associate each time series with categorical descriptors such as sector, source institution, publication area, and hierarchical classification. These descriptors are administrative rather than statistical, yet they encode meaningful economic structure. We investigate whether this structure can serve as a proxy for predictive relevance, and what the limits of that proxy are when selection is evaluated against out-of-sample forecasting performance.

We study this question through two complementary methods that exploit metadata in fundamentally different ways. The first, Metadata $\varepsilon$-Greedy, is a stochastic search policy that embeds registry metadata using permutation-invariant set encoders (Zaheer et al. 2017) and employs a small neural scoring model to guide a fixed-budget search over predictor subsets, with forecasting loss as the only feedback signal. The second, Metadata Bayes, operates entirely within metadata space: it constructs group-level Beta priors from institutional descriptors, updates them via partial correlation with the forecasting target, and selects predictors through Thompson sampling, without evaluating any forecasting model during selection. Comparing these two approaches — one that consults forecasting loss, one that does not — allows us to isolate the contribution of metadata structure at different stages of the selection pipeline and to examine directly what happens when variable selection is anchored to a held-out validation criterion versus when it is grounded entirely in training-data structure.

Both methods are evaluated on forecasting Peruvian headline CPI under two forecasters, a Vector Autoregression (VAR) and a Random Forest, and benchmarked against random search, greedy forward selection, LASSO, Bayesian Ridge, PCA, and the state-of-the-art Bayesian variable selection method of Jankowiak (2023). Stochastic methods are repeated across independent experiments to produce distributional results; we report both distributional statistics and the single best specification selected by validation minimization, which mirrors the operational choice available to practitioners.

The methods introduced here are variable-selection layers and are therefore applicable upstream of any forecasting model. The relevant question is not whether metadata-guided selection outperforms shrinkage or model-averaging approaches at forecasting; it is whether registry metadata contains enough signal to make variable selection more stable, more transparent in its rationale, and less dependent on computationally intensive procedures. The results speak to this directly. Metadata Bayes, which never evaluates a forecasting model during selection, achieves out-of-sample accuracy competitive with the Bayesian variable selection benchmark of Jankowiak (2023) while completing its selection procedure in seconds. Its selection rationale is directly legible in terms of the institutional registry categories that practitioners already use to organize their data. At the same

time, Metadata $\varepsilon$-Greedy — which does consult forecasting loss — does not improve consistently over random search. This contrast is itself a finding: it suggests that in settings with high missingness and collinearity-sensitive forecasting models, anchoring selection to a held-out validation criterion introduces instability that metadata structure alone can avoid. We examine this mechanism carefully and discuss its implications for variable selection practice in institutional settings.

# 2 Related Work

The literature relevant to this study spans four areas: forecasting with large macroeconomic panels, Bayesian variable selection, stochastic and heuristic search for model selection, and representation learning for set-structured inputs.

## Macroeconomic Forecasting and Variable Selection

Large-panel forecasting has motivated a substantial literature on dimensionality reduction and regularization. Factor-based approaches compress large information sets through principal components and targeted predictors (Stock and Watson, 2002; Bai and Ng, 2008). Bayesian VAR methods achieve tractability by imposing shrinkage priors and structured regularization within multivariate autoregressive systems (Giannone et al. 2012; Koop, 2013). More broadly, machine learning has become increasingly prominent in macroeconomic and financial forecasting as a flexible alternative to purely linear specifications (Medeiros et al. 2019; Athey, 2019), with recent work illustrating how predictive gains can arise from nonlinear feature interactions at scale (Gu et al. 2020).

These approaches improve tractability in high-dimensional settings by combining fixed predictor pools with shrinkage, regularization, or one-shot dimension reduction. Our study differs by asking whether the institutional metadata attached to each series can itself inform which predictors are worth selecting, prior to or independently of any statistical estimation, and by studying the conditions under which metadata-based selection criteria generalize reliably out of sample.

## Bayesian Variable Selection

Bayesian approaches to variable selection place priors over inclusion indicators and use posterior inference to identify relevant predictors. Spike-and-slab priors and their continuous relaxations are well-established in this tradition (Mitchell and Beauchamp, 1988; George and McCulloch, 1993). Scaling these methods to very large predictor pools has remained challenging, but recent work has made substantial progress: Jankowiak (2023) introduces a method for Bayesian variable selection in up to one million dimensions using scalable MCMC and subset-based likelihood approximations. We use this method as a state-of-the-art benchmark and find that Metadata Bayes, which never evaluates a forecasting model during selection, achieves competitive out-of-sample accuracy, suggesting that metadata structure encodes information that partially substitutes for explicit statistical variable selection.

### Stochastic and Heuristic Search in High-Dimensional Choice Spaces

Variable selection in large predictor spaces often requires stochastic or heuristic exploration. Classical techniques include random search, simulated annealing, and genetic algorithms. In machine learning, fixed-budget stochastic search has been studied extensively in hyperparameter optimization, where random search and Bayesian optimization provide effective baselines (Bergstra and Bengio, 2012; Snoek et al. 2012). Metadata $\varepsilon$-Greedy draws on these ideas by combining an epsilon-greedy mechanism with Bayesian updating of variable inclusion probabilities, guided by metadata rather than raw predictive features.

### Representation Learning for Sets and Metadata

DeepSets (Zaheer et al. 2017) introduced permutation-invariant architectures that map sets into fixed-length feature vectors, which we use to represent predictor subsets through their metadata compositions. Related ideas appear in meta-learning and AutoML, where structured side information guides exploration toward promising configurations (Vanschoren, 2018; Drori et al. 2019), and in recommendation systems, where heterogeneous metadata improves retrieval beyond raw interaction signals (Liu et al. 2019). We adapt these ideas to the macroeconomic setting by embedding subsets through institutional registry descriptors rather than time-series statistics.

### Positioning

This study sits at the intersection of macroeconomic forecasting, Bayesian variable selection, and representation learning for structured metadata. Its contribution is empirical: we show that institutional registry metadata, typically treated as auxiliary bookkeeping information, encodes enough economic structure to meaningfully guide variable selection, and we characterize the conditions under which metadata-based selection criteria generalize out of sample. The two methods we introduce differ in complexity and in the information they consume, and their comparative performance, including an honest negative result for the more complex method, constitutes the central finding of the paper.

## 3 Methodology

This section describes the two variable selection methods evaluated in this study. Both operate exclusively on structural metadata associated with each macroeconomic series; neither uses raw time-series values to construct proposals. The methods differ in how they exploit metadata. Metadata Bayes performs selection entirely within metadata space using Bayesian inference on group structure. Metadata $\varepsilon$-Greedy instead uses metadata embeddings to guide a stochastic search over predictor subsets, with forecasting loss as the feedback signal. The shared infrastructure, including the metadata registry, preprocessing, and encoding steps, is described first, followed by a description of each method.

## 3.1 Metadata Registry and Preprocessing

All variables in the study come from the central statistical registry maintained by the Banco Central de Reserva del Perú (BCRP). Each time series is associated with a fixed set of categorical descriptors, including its publication area, source institution, thematic category, hierarchical series description, and group-level classifications. These metadata fields constitute the only information used by both selection methods.

Several registry fields contain hierarchical labels encoded as character strings with forward-slash separators (e.g., `"Producción / Manufactura / Metales"`). We decompose these into separate categorical attributes by splitting on the slash delimiter, producing a consistent hierarchy of metadata levels (`lvl1`, `lvl2`, `lvl3`), each treated as an independent categorical feature. Fields without internal structure are left unchanged. The metadata dimensions are not assumed to be orthogonal: overlap across descriptors is expected and reflects how institutional registries are organized.

Not all metadata fields carry equal information about the structure of the series panel. To assess this empirically, we compute the mutual information between each categorical field and a correlation-based clustering of the full series panel. The clustering is obtained by applying $k$-means ($k = 10$) to the rows of the pairwise correlation matrix $R = \mathrm{corr}(X)$, where $X$ denotes the $T \times N$ panel of macroeconomic series. The results, reported in Table 1, show that *Grupo_de_serie_lvl1* dominates with a mutual information score of 0.71, far ahead of *Categoría de serie* (0.25), *Fuente* (0.24), and *Grupo_de_serie_lvl2* (0.19). Based on this audit, we retain the four fields with the highest scores as the active metadata dimensions for both methods. The label-to-integer mappings for these retained fields are reported in Appendix B.

| Metadata Field | Mutual Information | Unique Values |
|---|---|---|
| Grupo_de_serie_lvl1 | 0.706 | 128 |
| Categoría de serie | 0.245 | 23 |
| Fuente | 0.235 | 30 |
| Grupo_de_serie_lvl2 | 0.190 | 23 |
| Área que publica | 0.148 | 12 |
| Grupo de publicación | 0.123 | 9 |

Table 1: Mutual information between each metadata field and a correlation-based clustering of the series panel. Fields above the horizontal rule are retained as active metadata dimensions.

## 3.2 Metadata Bayes

Metadata Bayes performs variable selection entirely within metadata space. It never evaluates a forecasting model during selection; the only inputs are the institutional metadata descriptors and the correlation structure between candidate series and the forecasting target measured on the training window. The procedure organizes candidate predictors into metadata groups, maintains a Beta prior over the inclusion probability of each group, updates that prior using partial correlation as a likelihood signal, and finally selects variables via Thompson sampling.

**Prior initialization.** For each metadata group $g$, a Beta distribution $\text{Beta}(\alpha_g, \beta_g)$ is initialized using the mean absolute correlation of the group's variables with the forecasting target:

$$\alpha_g = \lambda \cdot \bar{\rho}_g + 1, \qquad \beta_g = \lambda \cdot (1 - \bar{\rho}_g) + 1,$$

where $\bar{\rho}_g$ is the mean absolute correlation of group $g$'s variables with the target and $\lambda$ is a prior strength hyperparameter controlling how much weight the metadata prior receives relative to the likelihood updates.

**Posterior update.** The prior is refined through $P$ sequential passes. In each pass $p$, the method computes the mean absolute partial correlation of each group's variables with the target, controlling for variables already selected in the previous pass. This partial correlation score $s_g^{(p)} \in [0, 1]$ serves as a likelihood signal and updates the Beta parameters as

$$\alpha_g \leftarrow \alpha_g + w_p \cdot s_g^{(p)}, \qquad \beta_g \leftarrow \beta_g + w_p \cdot (1 - s_g^{(p)}),$$

where $w_p = n_{\text{obs}}/p$ is a pass weight that decays across passes to down-weight later updates.

**Selection.** After the final pass, $k$ groups are selected via Thompson sampling: $N_{\text{TS}}$ draws of $\theta_g \sim \text{Beta}(\alpha_g, \beta_g)$ are taken for all groups simultaneously, the top-$k$ groups by $\theta_g$ are identified in each draw, and the most frequently selected combination is chosen. One variable is then drawn from each selected group, weighted by partial correlation with the target and penalized proportionally to the variable's missingness rate. All experiments use $P = 5$ passes, $\lambda = 5$, $k = 4$, and $N_{\text{TS}} = 200$.

Algorithm 1 summarizes the full procedure.

**Algorithm 1:** Metadata Bayes

**Input:** Candidate variables with metadata groups; training window $\mathcal{D}_{\text{train}}$; target series $y$; parameters $k$, $P$, $\lambda$, $N_{\text{TS}}$.

**Output:** Selected variable subset $S^*$.

**Initialize:** for each group $g$, compute $\bar{\rho}_g$ (mean absolute correlation with $y$ on $\mathcal{D}_{\text{train}}$) and set $\alpha_g = \lambda\bar{\rho}_g + 1$, $\beta_g = \lambda(1 - \bar{\rho}_g) + 1$.

Set $S_0 = \emptyset$.

**for** $p = 1$ **to** $P$ **do**

  Compute $s_g^{(p)}$ = mean absolute partial correlation of group $g$ with $y$, controlling for variables in $S_{p-1}$;

  Set $w_p = n_{\text{obs}}/p$;

  $\alpha_g \leftarrow \alpha_g + w_p \cdot s_g^{(p)}$;

  $\beta_g \leftarrow \beta_g + w_p \cdot (1 - s_g^{(p)})$;

  **for** $n = 1$ **to** $N_{\text{TS}}$ **do**

    Draw $\theta_g \sim \text{Beta}(\alpha_g, \beta_g)$ for all $g$;

    Record top-$k$ groups by $\theta_g$;

  $\mathcal{G}_p \leftarrow$ most frequent top-$k$ combination across draws;

  $S_p \leftarrow$ one variable per group in $\mathcal{G}_p$, weighted by partial correlation $\times$ (1 − miss rate);

**return** $S^* = S_P$

### 3.3 Metadata $\varepsilon$-Greedy

Metadata $\varepsilon$-Greedy is a stochastic search policy that uses forecasting loss as a feedback signal to guide exploration over predictor subsets. Unlike Metadata Bayes, it requires running a forecasting model at each trial, which makes it substantially more expensive but allows it to adapt directly to predictive performance rather than correlation structure alone.

**Metadata encoding.** Let $v_i$ denote a candidate predictor and let $m_i = (m_i^{(1)}, \ldots, m_i^{(d)})$ denote its metadata vector, where each component corresponds to a categorical descriptor from the active registry fields. A field-specific embedding layer converts each categorical value into a dense vector, and a linear projection maps the concatenated embeddings into a shared latent space, producing $\phi(m_i) \in \mathbb{R}^h$. No raw time-series values or positional indices enter the encoder. For a selected subset $S$ of $k$ predictors, a permutation-invariant representation is formed through a Deep Set aggregator (Zaheer et al. 2017):

$$\Psi(S) = \rho\left(\sum_{v_i \in S} \phi(m_i)\right),$$

where $\rho$ is a small two-layer MLP. Sum pooling ensures the representation is invariant to the ordering of variables within the subset.

**Preference scoring.** A preference neural network (PNN) maps the Deep Set representation $\Psi(S)$ concatenated with the observed validation loss to variable-level inclusion scores via a two-tower architecture. A set tower processes the concatenated input into a shared representation $z \in \mathbb{R}^H$, and an item tower scores each candidate variable through a learned embedding matrix $E \in \mathbb{R}^{V \times H}$:

$$\text{score}_j = \sigma\left(z \cdot E_j^\top + b_j\right), \quad j = 1, \ldots, V.$$

The PNN is trained using a relative ranking objective: trials are ranked by validation loss, each trial is assigned a soft target in $[0.1, 0.9]$ proportional to its rank, and the PNN is updated to predict these targets via binary cross-entropy on the scores of the selected variables.

**Search loop.** The search proceeds in two stages. During the warmup stage, $W$ subsets of $k$ predictors are drawn uniformly at random, the forecasting model is evaluated on each, and trial embeddings and losses are stored to populate the history. The PNN is initialized on the first warmup trial. During the guided stage, at each iteration the algorithm follows an $\varepsilon$-greedy policy: with probability $\varepsilon$ it draws a subset at random (exploration), and with probability $1 - \varepsilon$ it selects the $k$ variables with the highest PNN preference scores (exploitation), with a repeat-decay penalty that reduces scores for variables appearing frequently in recent selections. The PNN is retrained every $U$ iterations using a sliding window of recent trials. All experiments use $W = 10$ warmup trials, 50 guided iterations, $k = 4$, $\varepsilon = 0.15$, $U = 6$, and a window of 30 trials for PNN updates.

Algorithm 2 summarizes the procedure.

**Algorithm 2:** Metadata $\varepsilon$-Greedy

**Input:** Candidate variables with metadata; forecasting model $f$; parameters $k$, $W$, $T$, $\varepsilon$, $U$.

**Output:** Best evaluated subset $S^*$.

**Initialize:** empty trial history $\mathcal{H}$; PNN $G$ uninitialized.

**Warmup stage:**

**for** $w = 1$ **to** $W$ **do**

- Draw $S_w$ of $k$ variables uniformly at random;
- Evaluate $\ell_w = \text{RMSE}(f, S_w, \mathcal{D}_{\text{val}})$;
- Compute $\Psi(S_w)$ via metadata encoder and Deep Set;
- Append $(\Psi(S_w), \ell_w, S_w)$ to $\mathcal{H}$;
- **if** $w = 1$ **then**
  - Initialize PNN $G$;

**Guided stage:**

**for** $t = 1$ **to** $T$ **do**

- **if** rand() $< \varepsilon$ **then**
  - Draw $S_t$ of $k$ variables uniformly at random;
  - `/* Exploration */`
- **else**
  - Compute score$_j$ for all $j$ via PNN $G$;
  - Apply repeat-decay penalty to recently selected variables;
  - $S_t \leftarrow$ top-$k$ variables by score;
  - `/* Exploitation */`
- Evaluate $\ell_t = \text{RMSE}(f, S_t, \mathcal{D}_{\text{val}})$;
- Compute $\Psi(S_t)$ and append $(\Psi(S_t), \ell_t, S_t)$ to $\mathcal{H}$;
- **if** $t \bmod U = 0$ **then**
  - Retrain $G$ on sliding window of $\mathcal{H}$ using relative ranking loss;

**return** $S^* = \arg\min_{S \in \mathcal{H}} \ell$

## 3.4 Computational Complexity

The two methods differ substantially in computational cost. Metadata Bayes requires no forecasting model evaluations during selection. Its dominant cost is computing partial correlations across the candidate pool, which scales as $\mathcal{O}(P \cdot N \cdot P_{\text{passes}})$ where $P$ is the number of candidate series, $N$ is the length of the training window, and $P_{\text{passes}} = 5$ is fixed. Thompson sampling adds a negligible $\mathcal{O}(G \cdot N_{\text{TS}})$ term where $G$ is the number of metadata groups. In practice, the full selection procedure completes in seconds.

Metadata $\varepsilon$-Greedy is dominated by the cost of $T = 50$ forecasting model evaluations, each requiring fitting and evaluating the forecasting model on the training and validation windows. PNN retraining every $U = 6$ trials adds an $\mathcal{O}((T/U) \cdot W_{\text{window}} \cdot h)$ term where $h$ is the hidden dimension and $W_{\text{window}} = 30$ is the training window size for the PNN. The dominant cost is therefore $\mathcal{O}(T \cdot C_f)$ where $C_f$ is the per-trial forecasting cost.

For comparison, the state-of-the-art Bayesian variable selection method of Jankowiak (2023) requires $\mathcal{O}(S \cdot k^2)$ conditional posterior inclusion probability computations per MCMC iteration, where $S$ is the active subset size and $k$ is the number of included

variables, with the full statistical model evaluated at each step. While Subset wTGS is highly optimized for large $P$, it still requires repeated likelihood evaluations under the full generative model. Metadata Bayes avoids this entirely by substituting partial correlation for likelihood evaluation, achieving competitive out-of-sample accuracy at a fraction of the computational cost.

### 3.5 Forecasting Models, Target Variable, and Data Partitioning

Both selection methods are evaluated under two forecasting models representing complementary paradigms in applied macroeconomic work. The first is a Vector Autoregression (VAR) with one lag, which serves as a linear benchmark sensitive to multicollinearity and therefore conservative about the benefits of variable selection. The second is a Random Forest with 400 trees and maximum depth 16, representing a nonlinear alternative. Random Forest models also receive binary missingness masks as additional features, whereas the VAR is estimated using only the imputed series.

The forecasting target in all experiments is monthly headline CPI inflation for Lima Metropolitana, as published in the BCRPData registry (code PN01271PM; monthly percentage change). Forecast accuracy is evaluated using one-step-ahead root mean squared error:

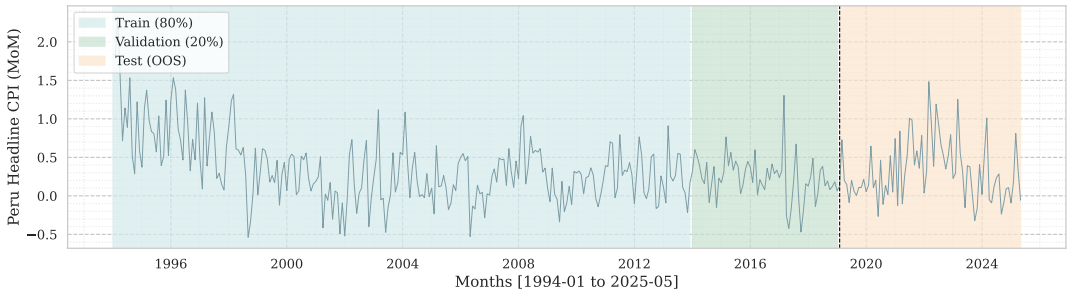
$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(y_t - \hat{y}_t)^2}. \tag{1}$$

Historical data are divided into a training window (1994–2015) used for model estimation, a validation window (2016–2019) used as the selection objective for Metadata $\varepsilon$-Greedy and for final specification choice across stochastic runs, and an out-of-sample window (2020–2025) reserved exclusively for evaluation. A COVID shock subperiod (2021–2023) is reported separately to assess robustness under structural change. Neither selection method accesses out-of-sample data at any point. Figure 1 illustrates the partition.

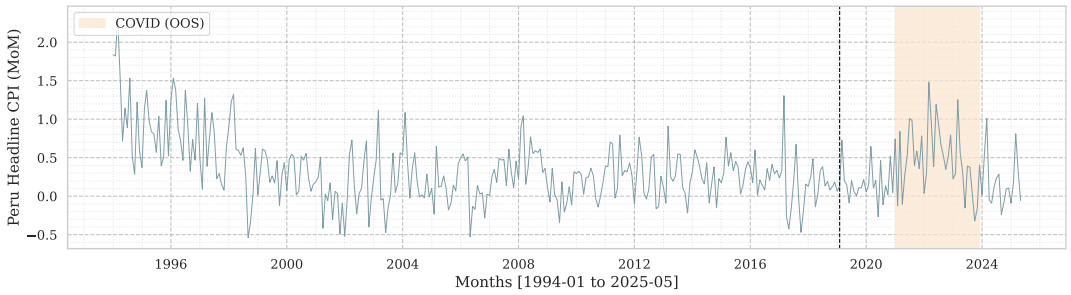
## 4 Results and Discussion

Table 2 reports validation RMSE and out-of-sample RMSE for all methods under both forecasting models across the full evaluation window (2020–2025) and the COVID shock period (2021–2023). For stochastic methods, results are averaged across 60 independent experimental runs. Each run consists of a complete variable-selection procedure followed by validation and out-of-sample evaluation using the same data splits. Random Search performs 60 trials per run. Metadata $\varepsilon$-Greedy performs 50 search trials preceded by 10 warmup trials. Metadata Bayes performs five posterior update passes. The Bayesian Variable Selection benchmark (Jankowiak, 2023) uses 1,000 burn-in iterations followed by 2,000 sampling iterations.

For Random Forest, the Greedy Forward baseline is applied to a candidate pool prefiltered to 500 series by correlation with the target due to computational constraints, whereas all other methods operate on the full pool of 3,030 series. In addition to the standard Metadata Bayes results, Table 2 also reports a validation-selected configuration in which the run with the lowest validation loss across the 60 runs is identified and then evaluated on held-out data. In this configuration the validation window is used only for run selection rather than for variable selection itself.




(a) Train (1994–2015), Validation (2016–2019), and out-of-sample (2020–2025).




(b) COVID-19 shock window (2021–2023).

Figure 1: Partition of Peruvian headline CPI into training, validation, and out-of-sample windows.

A characteristic of the candidate pool that bears directly on the interpretation of the results is reported in Table 3. The median series in the Peruvian registry has 28.6% of its observations missing, and more than half of all series exceed the 25% threshold. This reflects the institutional history of a registry that has grown organically over several decades, with series beginning and ending at different points and reporting schedules evolving over time. Series with more than 50% missing observations are excluded from the candidate pool. Remaining gaps are imputed using a centered rolling median with fallback to the unconditional median. Binary missingness masks are retained and passed to the Random Forest as additional features but are not used by the VAR. Imputation introduces series-specific noise that is correlated within each series yet orthogonal to the true dynamics of the forecasting target. The consequences of this for validation-based selection are discussed below.

## 4.1 Forecast Performance and the Limits of Validation-Based Selection

The most diagnostically informative result in Table 2 is the COVID out-of-sample RMSE of $0.69 \pm 0.23$ for Random Search under VAR. The standard deviation of 0.23 indicates that outcomes varied dramatically across independent runs, from near-competitive to substantially worse than nearly every other method in the table. A similar pattern, though less pronounced, appears for Metadata $\varepsilon$-Greedy under VAR, which produces a COVID RMSE of $0.48 \pm 0.09$.

Both methods select their final variable set by taking the argmin of validation loss across trials. When a large number of candidate subsets are evaluated against the same

| Method | Random Forest | | | VAR | | |
|---|---|---|---|---|---|---|
| | Val | Test Full | Test Covid | Val | Test Full | Test Covid |
| Random Search | 0.24 ± 0.01 | 0.36 ± 0.03 | 0.47 ± 0.03 | 0.28 ± 0.01 | 0.38 ± 0.05 | 0.69 ± 0.23 |
| Metadata $\varepsilon$-Greedy | 0.26 ± 0.01 | 0.38 ± 0.04 | 0.47 ± 0.03 | 0.29 ± 0.00 | 0.40 ± 0.06 | 0.48 ± 0.09 |
| Metadata Bayes | 0.27 ± 0.02 | 0.36 ± 0.03 | 0.45 ± 0.03 | 0.29 ± 0.01 | 0.36 ± 0.02 | 0.43 ± 0.03 |
| Metadata Bayes (Val Selected) | 0.23 | 0.32 | 0.41 | 0.28 | 0.35 | 0.42 |
| Bayesian Variable Selection | 0.29 ± 0.00 | 0.35 ± 0.00 | 0.38 ± 0.00 | 0.35 ± 0.02 | 0.37 ± 0.01 | 0.38 ± 0.01 |
| Greedy Forward | 0.26 | 0.35 | 0.42 | 0.27 | 0.49 | 0.62 |
| LASSO | 0.27 | 0.40 | 0.43 | 0.32 | 0.41 | 0.44 |
| Bayesian Ridge | 0.29 | 0.41 | 0.43 | 0.40 | 0.41 | 0.43 |
| PCA | 0.48 | 0.41 | 0.43 | 0.31 | 0.37 | 0.40 |

Table 2: Validation and out-of-sample RMSE across variable selection methods. Stochastic methods are reported as mean ± standard deviation across 60 independent runs. Deterministic baselines are single evaluations.

Table 3: Candidate pool missingness statistics after preprocessing. The sample is filtered to monthly series starting before January 1994 with at most 50% missing observations, near-zero variance below threshold, and zero-value ratio below 50%.

| Statistic | Peru |
|---|---|
| Series count | 3,030 |
| Observations | 377 |
| Mean missingness | 24.2% |
| Median missingness | 28.6% |
| Series >25% missing | 54.9% |

held-out window, the minimum validation loss will tend to be achieved by a specification that partially fits the noise structure of that window rather than the underlying signal. This phenomenon is well documented in the model selection literature and motivates practical corrections such as the one-standard-error rule (Breiman et al. 1984). In the present setting it is compounded by the high missingness rates in the candidate pool. Imputed series introduce noise that can align with the validation window by chance, and under VAR this alignment is particularly consequential: the model has no built-in mechanism for handling near-collinear predictors, and a subset of mutually correlated imputed series can produce coefficient estimates that fit the validation period and then deteriorate sharply out of sample.

Metadata $\varepsilon$-Greedy does not resolve this problem. Its neural scoring model is trained to concentrate search toward low-validation subsets, which in this setting means directing exploration toward precisely the region of the search space most susceptible to the overfitting tendency described above. The guidance signal is misaligned with the objective of out-of-sample generalization. This accounts for the absence of a consistent improvement over Random Search in Table 2, and in some configurations for performance that is modestly worse. These results are reported in full; the failure mode is itself informative about the conditions under which metadata-guided search of this form is and is not appropriate.

The Bayesian Variable Selection method of Jankowiak (2023) shows a markedly different pattern. Its out-of-sample performance is stable across both evaluation windows and both forecasting models, with near-zero variance under Random Forest ($0.35 \pm 0.00$ full, $0.38 \pm 0.00$ COVID) and low variance under VAR ($0.37 \pm 0.01$ full, $0.38 \pm 0.01$

COVID). This stability reflects the design of the method. Instead of optimizing a held-out predictive loss, it computes posterior inclusion probabilities under a generative model, identifying variables that plausibly explain the observed responses given a regularizing prior over coefficient magnitudes. Selection is based on the likelihood of the training data and is therefore less susceptible to the overfitting tendency that affects validation-based methods. The comparison is not fully symmetric, however. Bayesian Variable Selection and the search-based methods address different inferential problems, so the performance difference should be interpreted as evidence about which selection approach generalizes more reliably in this setting rather than as a general claim about the superiority of Bayesian inference over predictive search.

Metadata Bayes was motivated by this observation. It anchors the selection step in the partial correlation structure of metadata groups within the training window instead of using scores from a held-out window. This separates the selection criterion from the validation argmin problem. No forecasting model is evaluated during selection. The method instead asks which metadata groups are structurally associated with the forecasting target, as indicated by their partial correlations in the training data. The results are consistent with this design. Under VAR, Metadata Bayes achieves a mean out-of-sample RMSE of $0.36 \pm 0.02$ on the full window and $0.43 \pm 0.03$ during the COVID period. The reduction in cross-experiment variance relative to Random Search is substantial. The COVID standard deviation falls from 0.23 to 0.03, an 87% reduction that reflects the stability of a selection criterion that does not depend on a particular held-out window. Under Random Forest, Metadata Bayes achieves $0.36 \pm 0.03$ on the full window, matching the mean performance of Random Search while maintaining stable variance across experiments.

Bayesian Variable Selection remains the strongest method in overall performance across most configurations. Metadata Bayes closes much of the gap while operating at a fraction of the computational cost. Its selection procedure requires only partial correlation calculations on the training window and completes in seconds, whereas the Bayesian method requires thousands of MCMC iterations with full model evaluation at each step. The validation-selected Metadata Bayes configuration, which chooses the run with the lowest validation loss across independent experiments and then evaluates it on held-out data, achieves 0.32 (full) and 0.41 (COVID) under Random Forest and 0.35 (full) and 0.42 (COVID) under VAR. These results are competitive with the Bayesian benchmark on the full window. In this configuration the validation window is used only to select among completed runs rather than to guide variable selection. This explains its greater stability relative to the search-based methods. It remains a post-hoc step and should be interpreted accordingly.

Among the deterministic baselines, Greedy Forward selection is competitive under Random Forest (0.35 full, 0.42 COVID) but deteriorates markedly under VAR (0.49 full, 0.62 COVID), consistent with the collinearity sensitivity of that model when presented with greedily accumulated series from a high-missingness pool. LASSO performs moderately across both models. Bayesian Ridge and PCA are the weakest baselines overall, with Bayesian Ridge performing particularly poorly under VAR.

## 4.2 Selected Predictor Sets

Table 4 reports the predictors selected by the best-performing configurations of Bayesian Variable Selection and Metadata Bayes. The selected variables are economically inter-

| Code | Variable Name |
|---|---|
| *Random Forest* | |
| *Bayesian Variable Selection* | |
| PN01282PM | IPC No Transables (var% mensual) |
| PN01372PM | Productos No Transables – Alimentos |
| PN01373PM | Productos No Transables – Servicios |
| PN01314PM | IPC sin Alimentos y Bebidas |
| *Metadata Bayes* | |
| PN01273PM | IPC (variación porcentual 12 meses) |
| PN01296PM | Inflación Subyacente – Bienes |
| PN01303PM | Inflación Subyacente – Servicios – Educación |
| PN01314PM | IPC sin Alimentos y Bebidas (var% mensual) |
| *VAR* | |
| *Bayesian Variable Selection* | |
| PN01282PM | IPC No Transables (var% mensual) |
| PN01280PM | IPC Transables (var% mensual) |
| PN09817PM | IPC No Subyacente (var% mensual) |
| PN01381PM | Productos No Transables – Servicios – Otros Servicios |
| *Metadata Bayes* | |
| PN01273PM | IPC (variación porcentual 12 meses) |
| PN01282PM | IPC No Transables (var% mensual) |
| PN01297PM | Inflación Subyacente – Bienes – Alimentos y Bebidas |
| PN39523PM | IPC Importado (índice Dic. 2021 = 100) |

Table 4: Predictors selected by the best-performing Bayesian Variable Selection and Metadata Bayes configurations under each forecasting model.

pretable components of the CPI structure, including transable and non-transable inflation indices as well as measures of core inflation. This alignment with standard inflation decompositions used in central bank analysis suggests that the selection procedures are capturing meaningful economic structure rather than arbitrary statistical correlations. Predictor sets for the deterministic baseline methods are reported in Appendix A.

## 4.3 Ablation: Posterior Update Passes

Table 5 examines how the number of posterior update passes in Metadata Bayes affects forecasting performance in the VAR model. Each configuration is evaluated over 30 independent runs. With a single pass, the method reduces to computing marginal partial correlations between each metadata group and the target. In this baseline configuration, the model achieves an out-of-sample RMSE of $0.40 \pm 0.03$ on the full evaluation window and $0.48 \pm 0.04$ during the COVID period.

Introducing additional passes progressively conditions the selection criterion on variables chosen in earlier steps. Performance improves monotonically through five passes, where the out-of-sample RMSE falls to $0.36 \pm 0.01$ on the full window and $0.43 \pm 0.02$ during the COVID period. These correspond to reductions of approximately 10% and 11% relative to the single-pass baseline.

Cross-experiment variance also declines as the number of passes increases. The stan-

| Passes | Val | Test Full | Test Covid |
|---|---|---|---|
| 1 | 0.30 ± 0.02 | 0.40 ± 0.03 | 0.48 ± 0.04 |
| 2 | 0.29 ± 0.01 | 0.37 ± 0.02 | 0.44 ± 0.03 |
| 3 | 0.29 ± 0.01 | 0.36 ± 0.02 | 0.43 ± 0.03 |
| 4 | 0.29 ± 0.01 | 0.36 ± 0.01 | 0.44 ± 0.02 |
| 5 | 0.29 ± 0.01 | **0.36 ± 0.01** | **0.43 ± 0.02** |
| 8 | 0.29 ± 0.01 | **0.36 ± 0.01** | 0.43 ± 0.02 |
| 10 | 0.29 ± 0.01 | 0.36 ± 0.02 | 0.43 ± 0.03 |
| *Reference* | | | |
| BVS | 0.35 ± 0.02 | 0.37 ± 0.01 | 0.38 ± 0.01 |

Table 5: Ablation study on the number of posterior update passes in Metadata Bayes (VAR model, 30 independent runs per configuration). Each pass recomputes partial correlations while conditioning on predictors selected in previous passes, progressively refining the Beta posterior over metadata groups. Results are reported as mean ± standard deviation across runs. BVS is included as a reference.

dard deviation of the COVID RMSE decreases from 0.041 with one pass to 0.018 with five passes, a reduction of 56%. This stabilization reflects a progressively more concentrated posterior: successive conditioning steps remove redundant predictors and sharpen the group-level inclusion probabilities. Beyond five passes, performance plateaus and variance increases slightly. This behavior arises from the progressive downweighting scheme $w_p = N/(p+1)$, which reduces the influence of later passes and limits the amount of new information incorporated at each update. Consequently, additional passes provide little further benefit. Based on this analysis, all reported Metadata Bayes results use five posterior update passes. Even the single-pass configuration already exhibits substantially lower cross-experiment variance than the search-based methods, indicating that the primary stability gains arise from the metadata-based selection criterion itself, while the multi-pass refinement mainly improves predictive accuracy.

## 4.4 Discussion and Limitations

The results clarify the conditions under which metadata-based selection is useful. The main challenge in this setting is not dimensionality alone, but the instability of validation-based selection when the candidate pool contains many incomplete series and the forecasting model is sensitive to collinearity. In such environments, optimizing held-out predictive loss can lead to unstable variable sets and poor generalization. Methods that separate the selection step from validation performance behave more reliably. Bayesian Variable Selection addresses this through a full generative framework, while Metadata Bayes uses the structural information contained in the institutional registry.

A central result of the paper is that metadata-based selection can approach the performance of a state-of-the-art Bayesian variable selection method while requiring only a fraction of the computational cost. This suggests that the categorical descriptors attached to each series in a central bank registry contain meaningful economic structure that can guide variable selection. The advantage does not come from replacing statistical inference with metadata, but from using metadata to anchor the selection process in stable structural groupings rather than noisy validation signals. The mutual informa-

tion analysis in Section 3.1 supports this interpretation. The dominant metadata field, *Grupo_de_serie_lvl1*, carries substantially more information about the correlation structure of the panel than any other descriptor, indicating that the registry taxonomy reflects genuine economic organization rather than purely administrative classification.

The results also highlight several limitations. Metadata Bayes does not consistently outperform Bayesian Variable Selection, and its run-to-run variance, although much smaller than that of search-based methods, remains noticeable. One likely explanation is that the metadata grouping is a coarse representation of the conditional dependence structure of the panel. Registry metadata organizes series into broad institutional or thematic categories, but it cannot distinguish between series within the same category that have different relationships with the forecasting target. Richer metadata representations or the inclusion of additional institutional descriptors could improve the precision of the selection process. Another possibility is to combine metadata-based priors with a full Bayesian variable selection framework. The Beta posteriors produced by Metadata Bayes could serve as informative priors to initialize the MCMC procedure of Jankowiak (2023), which may reduce burn-in in settings where registry metadata is informative.

The performance of Metadata $\varepsilon$-Greedy provides a useful negative result. The neural scoring model successfully concentrates the search toward subsets with low validation error, but this objective is unreliable in the present setting. When validation performance is itself unstable, focusing the search on the validation argmin can amplify overfitting rather than improve generalization. Whether alternative training objectives, such as stability-oriented regularization or objectives that explicitly account for validation-window noise, could make better use of the metadata embedding remains an open question.

Finally, the empirical evaluation is limited to a single forecasting target and a single institutional registry. The BCRP registry contains relatively rich and well-structured metadata compared with many statistical databases, so performance may differ in environments where metadata is sparse or inconsistently defined. Evaluating the approach across additional targets, countries, and institutional registries would provide a clearer picture of its general applicability. In practical applications, institutions often impose additional filters on the candidate pool, such as restricting attention to stationary transformations or to particular thematic groups. These filters can be incorporated directly in the input set without altering the metadata-based selection procedures.

# 5 Conclusion

This paper investigated whether structural metadata from an institutional registry carries enough signal to guide variable selection for macroeconomic forecasting. The question is motivated by a practical observation: central bank registries associate each time series with rich categorical descriptors ( sector, source, publication area, hierarchical classification) that encode genuine economic organization, yet this information is rarely exploited in formal variable selection procedures. We developed two methods that use this metadata in different ways and evaluated them against a range of baselines, including a state-of-the-art Bayesian variable selection method, on the task of forecasting Peruvian headline CPI.

The results support a qualified positive answer to the motivating question, but the path to that answer is instructive. Metadata $\varepsilon$-Greedy, the more complex of the two proposed methods, does not produce consistent improvements over random search. The

issue is not that metadata lacks useful information. Rather, the method uses metadata to concentrate the search on subsets with low validation error. In a setting where many candidate series are incomplete and the forecasting model is sensitive to collinearity, the validation argmin becomes an unreliable selection criterion. Improving the efficiency of this optimization therefore does not improve generalization. It simply concentrates the overfitting.

Metadata Bayes avoids this problem by fully separating selection from forecasting evaluation. Instead of looking at validation performance, it asks which metadata groups are structurally related to the forecasting target. This relationship is measured through their partial correlation structure in the training data, and variables are selected using Thompson sampling from the resulting Beta posteriors. Because the selection step never uses the held-out window, the method avoids the validation argmin problem entirely.

The results show that this design choice matters. Metadata Bayes reduces COVID-period cross-experiment variance by 87% relative to random search under VAR, approaches the out-of-sample performance of Bayesian Variable Selection (Jankowiak, 2023) across both forecasting models, and does so in seconds rather than requiring thousands of MCMC iterations with full model evaluation at every step. The broader implication is that institutional metadata, often treated as administrative bookkeeping, contains enough economic structure to partially substitute for explicit statistical inference about variable relevance, and in some configurations it can do so quite effectively. The mutual information audit reported in Section 3.1 provides independent evidence: the dominant metadata field carries substantially more information about the correlation structure of the panel than any other descriptor, suggesting that the registry taxonomy reflects real economic organization. When this structure is used through an appropriate selection criterion based on training data rather than held-out loss, the result is variable selection that is stable, computationally light, and interpretable in terms of the institutional categories practitioners already use to organize their data.

Several directions follow naturally from these findings. One immediate possibility is to combine Metadata Bayes with a full Bayesian variable selection procedure. The Beta posteriors produced during the metadata-based selection stage could serve as informative priors to warm-start the MCMC procedure of Jankowiak (2023), which may reduce the burn-in period in institutional settings where the registry is well organized. Another avenue concerns the richness of the metadata representation. The four fields retained here were chosen through a mutual information audit, but more granular descriptors or representations that capture temporal and relational structure among series could improve the selection process further. The evaluation presented here also focuses on a single target and a single registry, so applying the approach to other countries, other institutional databases, and multiple forecasting objectives would help clarify how broadly the method generalizes.

For central banks and other policy institutions, the practical message is straightforward. The metadata already attached to their datasets contains useful information for variable selection, and this information can be used without collecting additional data, running expensive computations, or modifying the forecasting models already in place. Metadata Bayes provides a transparent and institutionally interpretable procedure for doing so, with a selection logic that can be explained directly through the registry categories that organize the data. Whether the objective is to reduce reliance on ad hoc filtering, improve robustness during volatile periods, or make the selection process more systematic and reproducible, metadata-based selection offers a practical starting point.

# References


Athey, Susan (2019). "The Impact of Machine Learning on Economics". In: *The Economics of Artificial Intelligence: An Agenda*. Ed. by Ajay Agrawal, Joshua Gans, and Avi Goldfarb. University of Chicago Press, pp. 507–547. URL: https://www.nber.org/books-and-chapters/economics-artificial-intelligence-agenda/impact-machine-learning-economics.

Bai, Jushan and Serena Ng (2008). "Forecasting economic time series using targeted predictors". In: *Journal of Econometrics* 146.2, pp. 304–317.

Bergstra, James and Yoshua Bengio (2012). "Random Search for Hyper-Parameter Optimization". In: *J. Mach. Learn. Res.* 13, pp. 281–305. URL: https://api.semanticscholar.org/CorpusID:15700257.

Breiman, Leo et al. (1984). *Classification and Regression Trees*. Wadsworth International Group. DOI: 10.1201/9781315139470.

Drori, Iddo et al. (2019). "AutoML using Metadata Language Embeddings". In: *ArXiv* abs/1910.03698. URL: https://api.semanticscholar.org/CorpusID:203952124.

George, Edward I. and Robert E. McCulloch (1993). "Variable selection via Gibbs sampling". In: *Journal of the American Statistical Association* 88, pp. 881–889. DOI: 10.1080/01621459.1993.10476353.

Giannone, Domenico, Michele Lenza, and Giorgio E. Primiceri (2012). "Prior Selection for Vector Autoregressions". In: *Review of Economics and Statistics* 97, pp. 436–451. URL: https://api.semanticscholar.org/CorpusID:14849387.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu (2020). "Empirical Asset Pricing via Machine Learning". In: *The Review of Financial Studies* 33.5, pp. 2223–2273. DOI: 10.1093/rfs/hhaa009. URL: https://doi.org/10.1093/rfs/hhaa009.

Jankowiak, Martin (2023). "Bayesian Variable Selection in a Million Dimensions". In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Ed. by Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent. Vol. 206. Proceedings of Machine Learning Research. PMLR, pp. 253–282. URL: https://proceedings.mlr.press/v206/jankowiak23a.html.

Koop, Gary (2013). "Forecasting with Medium and Large Bayesian VARs". In: *Journal of Applied Econometrics* 28, pp. 177–203. URL: https://api.semanticscholar.org/CorpusID:6504643.

Liu, Tianqiao et al. (2019). "Recommender Systems with Heterogeneous Side Information". In: *The World Wide Web Conference*. URL: https://api.semanticscholar.org/CorpusID:86825364.

Medeiros, Marcelo C et al. (2019). "Machine learning and big data in macroeconomics and finance". In: *Journal of Economic Surveys* 33.1, pp. 1–26.

Meinshausen, Nicolai and Peter Bühlmann (2010). "Stability selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4, pp. 417–473. DOI: https://doi.org/10.1111/j.1467-9868.2010.00740.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2010.00740.x.

Mitchell, Toby J. and John J. Beauchamp (1988). "Bayesian Variable Selection in Linear Regression". In: *Journal of the American Statistical Association* 83, pp. 1023–1032. DOI: 10.1080/01621459.1988.10478694.

Mullainathan, Sendhil and Jann Spiess (2017). "Machine Learning: An Applied Econometric Approach". In: *Journal of Economic Perspectives* 31, pp. 87–106. URL: https://api.semanticscholar.org/CorpusID:157481740.

Snoek, Jasper, H. Larochelle, and Ryan P. Adams (2012). "Practical Bayesian Optimization of Machine Learning Algorithms". In: *Neural Information Processing Systems*. URL: https://api.semanticscholar.org/CorpusID:632197.

Stock, James H and Mark W Watson (2002). "Forecasting using principal components from a large number of predictors". In: *Journal of the American Statistical Association* 97.460, pp. 1167–1179.

Vanschoren, Joaquin (2018). "Meta-Learning: A Survey". In: *Automated Machine Learning*. URL: https://api.semanticscholar.org/CorpusID:52938664.

Zaheer, Manzil et al. (2017). "Deep Sets". In: *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. Curran Associates, Inc., pp. 3391–3401. URL: http://papers.nips.cc/paper/6931-deep-sets.pdf.

| Code | Variable Name |
|---|---|
| *LASSO* | |
| PN00214MM | Liquidez Total (millones S/) |
| PN01155MM | Capitalización Bursátil (millones S/) |
| RD12970DM | Producción minera – Hierro (tm.f) |
| RD12950DM | Desembarque de anchoveta – Total (toneladas) |
| *Bayesian Ridge* | |
| PN00111MM | Activos Internos Netos (millones S/) |
| PN00586MM | Pasivos con el Exterior de Largo Plazo |
| PN38805BM | Molibdeno – Precio (¢US$ por libra) |
| PN00559MM | Depósitos de Encaje (MN) |
| *PCA* | |
| F1 | Principal Component 1 |
| F2 | Principal Component 2 |
| F3 | Principal Component 3 |
| F4 | Principal Component 4 |

Table 6: Predictor sets used by deterministic baseline methods.

# A Predictor Sets for Deterministic Baselines

This appendix reports the predictor sets used by the deterministic baseline methods. Each method selects four predictors from the candidate pool using the training sample. Because these procedures are deterministic, the same predictor sets are used for both forecasting models.

# B Metadata Registry (Active Fields)

This appendix reports the categorical metadata fields used by the selection procedures. As described in Section 3.1, an information audit based on mutual information with the correlation structure of the panel was used to identify the most informative metadata descriptors.

Only the four fields with the highest mutual information values (Table 1) are retained as active metadata dimensions. The tables below list the mapping from textual labels to the integer identifiers used by the metadata encoder.

Table 7: Metadata registry: Categoría de serie

| **Label** | **ID** |
|---|---|
| Balanza comercial | 0 |
| Banco Central de Reserva | 1 |
| Caja del tesoro | 2 |
| Cotizaciones internacionales | 3 |
| Empresas bancarias | 4 |
| Expectativas Empresariales | 5 |
| Expectativas Macroeconómicas | 6 |
| Exportaciones e importaciones | 7 |
| Gastos | 8 |
| Indicadores de coyuntura | 9 |
| Inflación | 10 |
| Ingresos | 11 |
| Mercado de capitales | 12 |
| Operaciones de las empresas bancarias | 13 |
| Otras cuentas monetarias | 14 |
| PBI por sectores | 15 |
| Precios y tarifas | 16 |
| Producción | 17 |
| Remuneraciones y empleo | 18 |
| Resultado económico | 19 |
| Sector público | 20 |
| Sistema financiero | 21 |
| Sistemas de pagos | 22 |
| Sociedades creadoras de depósito | 23 |
| Tasas de interés | 24 |
| Tasas de interés internacionales | 25 |
| Tipo de cambio de otras divisas | 26 |
| Tipo de cambio nominal | 27 |
| Tipo de cambio real | 28 |
| Términos de intercambio | 29 |

Table 8: Metadata registry: Fuente

| Label | ID |
|---|---|
| BCRP | 0 |
| BCRP, COES, Sunat, Ministerio de Energía y Minas, empresas cementeras | 1 |
| BCRP, FMI, Reuters | 2 |
| BCRP, SBS, Reuters, Datatec | 3 |
| BCRP, Sunat | 4 |
| BCRP, Sunat, Zofratacna, Banco de la Nación | 5 |
| BCRP, Sunat, Zofratacna, Banco de la Nación, empresas | 6 |
| BCRP, Sunat, empresas | 7 |
| BVL, Cavali | 8 |
| El Peruano | 9 |
| Encuesta de Expectativas Macroeconómicas | 10 |
| FMI, Reuters | 11 |
| FMI, Reuters, SBS | 12 |
| INEI | 13 |
| INEI, Ministerio de Agricultura | 14 |
| INEI, Ministerio de Agricultura, Ministerio de Energía y Minas, Ministerio de la Producción | 15 |
| INEI, Ministerio de Energía y Minas | 16 |
| INEI, Ministerio de la Producción | 17 |
| INEI, Osinergmin | 18 |
| MISSING | 19 |
| Ministerio de Comercio Exterior y Turismo | 20 |
| Ministerio de Economía y Finanzas | 21 |
| Ministerio de Economía y Finanzas, BCRP, Banco de la Nación | 22 |
| Ministerio de Economía y Finanzas, Banco de la Nación | 23 |
| Ministerio de Economía y Finanzas, Banco de la Nación, Sunat | 24 |
| Ministerio de Economía y Finanzas, Banco de la Nación, Sunat, EsSalud, sociedades de beneficencia pública, empresas estatales | 25 |
| Ministerio de Economía y Finanzas, Banco de la Nación, Sunat, EsSalud, sociedades de beneficencia pública, gobiernos locales, instituciones públicas | 26 |
| Ministerio de Energía y Minas | 27 |
| Ministerio de Trabajo y Promoción del Empleo | 28 |
| Ministerio de la Producción | 29 |
| Osinergmin | 30 |
| Reuters | 31 |
| Reuters, Creed Rice y Oryza (para el arroz tailandés) | 32 |
| SBS | 33 |
| SBS, BCRP, FMI, Reuters | 34 |
| SMV, BVL, Cavali, Ministerio de Economía y Finanzas | 35 |

Continued on next page

| Label | ID |
|---|---|
| Sunat | 36 |
| Sunat-Aduanas, se excluyen los valores extremos | 37 |
| empresas bancarias | 38 |
| empresas bancarias, SBS | 39 |
| empresas financieras | 40 |
| entidades del sistema financiero | 41 |

Table 9: Metadata registry: Grupo_de_serie_lvl1

| Label | ID |
|---|---|
| Afiliados activos del Sistema Privado de Pensiones por departamentos (número) | 0 |
| Afiliados activos del Sistema Privado de Pensiones por departamentos (var% anual) | 1 |
| Ahorro del sistema financiero | 2 |
| Arribos a los establecimientos de hospedaje según departamentos (número) | 3 |
| Balanza comercial | 4 |
| Balanza comercial (variación porcentual) | 5 |
| Bolsa de Valores de Lima | 6 |
| Bonos (millones S/) | 7 |
| Bonos (millones US$) | 8 |
| Bonos del sector privado por moneda y plazo | 9 |
| CCE: cheques recibidos y rechazados a nivel nacional | 10 |
| CCE: cheques recibidos y rechazados y transferencias de crédito en ME | 11 |
| CCE: cheques recibidos y rechazados y transferencias de crédito en ME (estructura porcentual) | 12 |
| CCE: cheques recibidos y rechazados y transferencias de crédito en MN | 13 |
| CCE: cheques recibidos y rechazados y transferencias de crédito en MN (estructura porcentual) | 14 |
| Cotizaciones de productos (promedio del periodo) | 15 |
| Crédito al sector privado de las sociedades creadoras de depósito, por tipo de crédito | 16 |
| Crédito al sector privado de las sociedades creadoras de depósito, por tipo de crédito y por monedas | 17 |
| Crédito de las empresas bancarias al sector privado en ME | 18 |
| Crédito de las empresas bancarias al sector privado en MN | 19 |
| Crédito de las sociedades creadoras de depósito al sector privado | 20 |
| Crédito de las sociedades creadoras de depósito al sector privado (fin de periodo) | 21 |

Continued on next page

| Label | ID |
|---|---|
| Crédito del sistema bancario al sector privado | 22 |
| Crédito del sistema bancario al sector privado (fin de periodo) | 23 |
| Crédito del sistema financiero al sector privado | 24 |
| Crédito del sistema financiero al sector privado (fin de periodo) | 25 |
| Crédito directo del sistema financiero al sector privado por departamentos | 26 |
| Crédito neto al sector público de las sociedades creadoras de depósito | 27 |
| Cuentas monetarias de las empresas bancarias | 28 |
| Cuentas monetarias de las sociedades creadoras de depósito | 29 |
| Cuentas monetarias del Banco Central de Reserva del Perú | 30 |
| Cuentas monetarias del Banco de la Nación | 31 |
| Cuentas monetarias del sistema bancario | 32 |
| Cuentas monetarias del sistema financiero | 33 |
| Depósitos del sector privado, por tipo de depósito y por monedas | 34 |
| Depósitos en el sistema financiero por departamentos | 35 |
| Desembarque de anchoveta para harina y aceite por departamento y puerto (toneladas) | 36 |
| Emisión primaria y multiplicador (millones S/) | 37 |
| Emisión primaria y multiplicador (var% 12 meses) | 38 |
| Emisión primaria y multiplicador (var% mensual) | 39 |
| Empleo Mensual en Lima Metropolitana | 40 |
| Empleo informal | 41 |
| Encaje de las empresas bancarias (promedios diarios) | 42 |
| Encaje de las empresas bancarias en ME | 43 |
| Encaje de las empresas bancarias en MN | 44 |
| Encaje, depósitos overnight y liquidez por institución en ME | 45 |
| Encaje, depósitos overnight y liquidez por institución en MN | 46 |
| Expectativas empresariales por zonas | 47 |
| Expectativas empresariales sectoriales | 48 |
| Expectativas empresariales totales | 49 |
| Expectativas macroeconómicas | 50 |
| Exportaciones de productos no tradicionales | 51 |
| Exportaciones de productos tradicionales | 52 |
| Exportaciones de productos tradicionales (precios) | 53 |
| Exportaciones de productos tradicionales (volumen) | 54 |
| Exportaciones por Departamento (Valores FOB en millones US$) | 55 |
| Exportaciones por grupo de actividad económica | 56 |
| Exportaciones por grupo de productos | 57 |
| Exportaciones por grupo de productos (estructura porcentual) | 58 |
| Exportaciones por grupo de productos de Amazonas (Valores FOB en millones US$) | 59 |

Continued on next page

| Label | ID |
|---|---|
| Exportaciones por grupo de productos de Ancash (Valores FOB en millones US$) | 60 |
| Exportaciones por grupo de productos de Apurimac (Valores FOB en millones US$) | 61 |
| Exportaciones por grupo de productos de Arequipa (Valores FOB en millones US$) | 62 |
| Exportaciones por grupo de productos de Ayacucho (Valores FOB en millones US$) | 63 |
| Exportaciones por grupo de productos de Cajamarca (Valores FOB en millones US$) | 64 |
| Exportaciones por grupo de productos de Callao (Valores FOB en millones US$) | 65 |
| Exportaciones por grupo de productos de Cusco (Valores FOB en millones US$) | 66 |
| Exportaciones por grupo de productos de Huancavelica (Valores FOB en millones US$) | 67 |
| Exportaciones por grupo de productos de Huanuco (Valores FOB en millones US$) | 68 |
| Exportaciones por grupo de productos de Ica (Valores FOB en millones US$) | 69 |
| Exportaciones por grupo de productos de Junín (Valores FOB en millones US$) | 70 |
| Exportaciones por grupo de productos de La Libertad (Valores FOB en millones US$) | 71 |
| Exportaciones por grupo de productos de Lambayeque (Valores FOB en millones US$) | 72 |
| Exportaciones por grupo de productos de Lima (Valores FOB en millones US$) | 73 |
| Exportaciones por grupo de productos de Loreto (Valores FOB en millones US$) | 74 |
| Exportaciones por grupo de productos de Madre de Dios (Valores FOB en millones US$) | 75 |
| Exportaciones por grupo de productos de Moquegua (Valores FOB en millones US$) | 76 |
| Exportaciones por grupo de productos de Pasco (Valores FOB en millones US$) | 77 |
| Exportaciones por grupo de productos de Piura (Valores FOB en millones US$) | 78 |
| Exportaciones por grupo de productos de Puno (Valores FOB en millones US$) | 79 |
| Exportaciones por grupo de productos de San Martín (Valores FOB en millones US$) | 80 |
| Exportaciones por grupo de productos de Tacna (Valores FOB en millones US$) | 81 |

Continued on next page

| Label | ID |
|---|---|
| Exportaciones por grupo de productos de Tumbes (Valores FOB en millones US$) | 82 |
| Exportaciones por grupo de productos de Ucayali (Valores FOB en millones US$) | 83 |
| Exportaciones por grupo de productos sin ubigeo (Valores FOB en millones US$) | 84 |
| Flujo de caja del tesoro público | 85 |
| Forwards de monedas de las empresas bancarias con el público (millones US$) | 86 |
| Forwards y swaps de monedas de las empresas bancarias (millones US$) | 87 |
| Forwards y swaps de monedas interbancarios (millones US$) | 88 |
| Fuentes de la emisión primaria (millones S/) | 89 |
| Gastos del gobierno central (millones S/ 2007) (descontinuada) | 90 |
| Gastos del gobierno central (millones S/) | 91 |
| Gastos del gobierno central (millones de soles diciembre 2021) | 92 |
| Gastos no financiero de gobiernos regionales | 93 |
| Gastos no financieros de gobiernos locales por departamentos | 94 |
| Gastos no financieros del gobierno general (millones S/ 2007) (descontinuada) | 95 |
| Gastos no financieros del gobierno general (millones S/) | 96 |
| Gastos no financieros del gobierno general (millones de soles diciembre 2021) | 97 |
| Importaciones de fertilizantes (precios FOB US$/TM) | 98 |
| Importaciones por Aduana (Valores FOB en millones US$) | 99 |
| Importaciones según uso o destino económico | 100 |
| Indicadores de coyuntura | 101 |
| Indicadores de las empresas bancarias | 102 |
| Indicadores de riesgo para países emergentes: EMBIG | 103 |
| Indicadores indirectos de la tasa de utilización de la capacidad instalada del sector manufacturero | 104 |
| Inflación de socios comerciales | 105 |
| Ingresos corrientes del gobierno central (millones S/) | 106 |
| Ingresos corrientes del gobierno central en términos reales (millones S/ 2007) (descontinuada) | 107 |
| Ingresos corrientes del gobierno central en términos reales (millones de soles diciembre 2021) | 108 |
| Ingresos corrientes del gobierno general (millones S/ 2007) (descontinuada) | 109 |
| Ingresos corrientes del gobierno general (millones S/) | 110 |
| Ingresos corrientes del gobierno general (millones de soles diciembre 2021) | 111 |
| Ingresos recaudados por SUNAT | 112 |
| Ingresos tributarios recaudados por SUNAT | 113 |

Continued on next page

| Label | ID |
|---|---|
| Inversión bruta fija de gobiernos locales por departamentos (millones S/) | 114 |
| Inversión bruta fija de los gobiernos regionales (millones S/) | 115 |
| Inversión bruta fija del gobierno nacional por departamentos (millones S/) | 116 |
| Liquidez de las empresas bancarias | 117 |
| Liquidez de las empresas bancarias (var% 12 meses) | 118 |
| Liquidez de las empresas bancarias (var% mensual) | 119 |
| Liquidez de las sociedades creadoras de depósito | 120 |
| Liquidez de las sociedades creadoras de depósito (fin de periodo) | 121 |
| Liquidez del Banco de la Nación | 122 |
| Liquidez del Banco de la Nación (var% 12 meses) | 123 |
| Liquidez del Banco de la Nación (var% mensual) | 124 |
| Liquidez del sistema bancario | 125 |
| Liquidez del sistema bancario (fin de periodo) | 126 |
| Liquidez del sistema bancario (promedio del período) | 127 |
| Liquidez del sistema financiero | 128 |
| Liquidez del sistema financiero (fin de periodo) | 129 |
| Liquidez internacional del BCRP | 130 |
| Medios de pago distintos al efectivo, cajeros y banca virtual: Monto de las operaciones en ME (millones US$) | 131 |
| Medios de pago distintos al efectivo, cajeros y banca virtual: Monto de las operaciones en MN (millones S/) | 132 |
| Medios de pago distintos al efectivo, cajeros y banca virtual: Número de operaciones en ME (miles) | 133 |
| Medios de pago distintos al efectivo, cajeros y banca virtual: Número de operaciones en MN (miles) | 134 |
| Monto nominal de certificados y depósitos del Banco Central (millones S/) | 135 |
| Obligaciones de las sociedades creadoras de depósito con el sector público | 136 |
| Operaciones cambiarias del BCRP con las empresas bancarias (millones US$) | 137 |
| Operaciones del gobierno central (millones S/) | 138 |
| Operaciones del gobierno central en términos reales (millones S/ 2007) (descontinuada) | 139 |
| Operaciones del gobierno central en términos reales (millones de soles diciembre 2021) | 140 |
| Operaciones del sector público no financiero (millones S/ 2007) (descontinuada) | 141 |
| Operaciones del sector público no financiero (millones S/) | 142 |
| Operaciones del sector público no financiero (millones de soles diciembre 2021) | 143 |
| Operaciones en ME de las empresas bancarias (millones US$) | 144 |

Continued on next page

| Label | ID |
|---|---|
| Otras divisas | 145 |
| Pagos a través del LBTR, sistema de liquidación multibancaria de valores y CCE | 146 |
| Pagos de alto y bajo valor (millones S/) | 147 |
| Pagos de alto y bajo valor (millones de operaciones) | 148 |
| Permanencia promedio en los establecimientos de hospedaje (número de días) | 149 |
| Pernoctaciones en los establecimientos de hospedaje (número) | 150 |
| Precios de productos sujetos al sistema de franjas de precios (US$ por toneladas) | 151 |
| Producción agropecuaria (miles de toneladas) | 152 |
| Producción agropecuaria (variación porcentual interanual) | 153 |
| Producción de electricidad por departamento | 154 |
| Producción de productos mineros según departamentos | 155 |
| Producción manufacturera (variación porcentual interanual) | 156 |
| Producción manufacturera (índice 2007 = 100) | 157 |
| Producción minera e hidrocarburos (miles de unidades recuperables) | 158 |
| Producción minera e hidrocarburos (variación porcentual interanual) | 159 |
| Producción pesquera (miles de toneladas) | 160 |
| Producción pesquera (variación porcentual interanual) | 161 |
| Producto bruto interno y demanda interna (variación porcentual interanual) | 162 |
| Producto bruto interno y demanda interna (índice 2007 = 100) | 163 |
| Puestos de trabajo e ingresos del sector formal | 164 |
| Remuneraciones | 165 |
| Repos del Banco Central y depósitos públicos (millones S/) | 166 |
| Saldo de los certificados de depósito del BCRP (millones S/) | 167 |
| Saldo de obligaciones domésticas de las empresas bancarias en MN por institución | 168 |
| Saldo de obligaciones internas de las empresas bancarias en ME por institución | 169 |
| Sistema LBTR: transferencias en MN y ME | 170 |
| Sistema LBTR: transferencias en MN y ME (estructura porcentual) | 171 |
| Sistema privado de pensiones | 172 |
| Swaps de monedas de las empresas bancarias con el público (millones US$) | 173 |
| Tasa de encaje de las empresas bancarias | 174 |
| Tasa de encaje de las empresas bancarias total | 175 |
| Tasas de interés activas promedio de las cajas municipales de ahorro y crédito por modalidad (% términos efectivos anuales) | 176 |
| Tasas de interés activas promedio de las cajas rurales de ahorro y crédito por modalidad (% términos efectivos anuales) | 177 |

Continued on next page

| Label | ID |
|---|---|
| Tasas de interés activas promedio de las empresas bancarias por modalidad (términos efectivos anuales) | 178 |
| Tasas de interés activas y pasivas promedio de las empresas bancarias en ME (términos efectivos anuales) | 179 |
| Tasas de interés activas y pasivas promedio de las empresas bancarias en MN (términos efectivos anuales) | 180 |
| Tasas de interés de bonos del gobierno peruano | 181 |
| Tasas de interés de los Certificados de Depósito del BCRP | 182 |
| Tasas de interés del Banco Central de Reserva | 183 |
| Tipo de cambio | 184 |
| Tipo de cambio de las principales monedas | 185 |
| Tipo de cambio nominal (S/ por canasta) | 186 |
| Tipo de cambio real bilateral del Perú respecto a países latinoamericanos (promedio del período) | 187 |
| Términos de intercambio de comercio exterior (var% 12 meses) | 188 |
| Términos de intercambio de comercio exterior (var% acumulada) | 189 |
| Términos de intercambio de comercio exterior (var% mensual) | 190 |
| Términos de intercambio de comercio exterior (índice 2007 = 100) | 191 |
| Variación RIN del BCRP (millones US$) | 192 |
| Venta de energía eléctrica por departamento | 193 |
| Índice de precios Lima Metropolitana (var% 12 meses) | 194 |
| Índice de precios Lima Metropolitana (var% acumulada) | 195 |
| Índice de precios Lima Metropolitana (var% mensual) | 196 |
| Índice de precios Lima Metropolitana (índice 2009 = 100) (descontinuada) | 197 |
| Índice de precios Lima Metropolitana (índice Dic.2021 = 100) | 198 |
| Índice de precios al consumidor Lima Metropolitana: clasificación sectorial (variación porcentual) | 199 |
| Índice de precios al consumidor Lima Metropolitana: clasificación transables | 200 |
| Índice del tipo de cambio real (base 2009=100) | 201 |
| Índice del tipo de cambio real (var% 12 meses) | 202 |
| Índice del tipo de cambio real (var% mensual) | 203 |
| Índices reales de precios de combustibles y de tarifas de servicios públicos (2010 = 100) | 204 |
| índices de Precios al Consumidor Lima Metropolitana (Índice Dic.2021=100) | 205 |

Table 10: Metadata registry: Grupo_de_serie_lvl2

| Label | ID |
|---|---|
| Desestacionalizados | 0 |
| Impuesto General a las Ventas interno según departamento | 1 |
| ME (%) | 2 |
| MISSING | 3 |
| MN (%) | 4 |
| Promedio móvil tres meses (miles de personas) | 5 |
| Promedio móvil tres meses (porcentaje) | 6 |
| Sector Privado (millones S/) | 7 |
| Sector Privado (millones US$) | 8 |
| Sector Público (millones S/) | 9 |
| Sector Público (millones US$) | 10 |
| empresas bancarias | 11 |
| empresas financieras | 12 |
| fin de periodo (S/ por US$) | 13 |
| fin de periodo (millones S/) | 14 |
| fin de periodo (millones US$) | 15 |
| fin de periodo (var% 12 meses) | 16 |
| fin de periodo (var% mensual) | 17 |
| impuesto a la renta según departamento | 18 |
| no transables (variación porcentual) | 19 |
| por institución (millones S/) | 20 |
| por institución (millones US$) | 21 |
| promedio del periodo (S/ por US$) | 22 |
| promedio del período (S/ por UM) | 23 |
| promedio del período (UM por US$) | 24 |
| promedio del período (var% 12 meses) | 25 |
| promedio del período (var% mensual) | 26 |
| tributos aduaneros según departamento | 27 |
| tributos internos según departamento | 28 |
| valores FOB (millones US$) | 29 |