High-Frequency Retail Pricing Data as an Inflation Indicator: Comovement and Forecasting in the Eurozone

Didrik Liu
didliu@kth.se

Glossary

HICP – Harmonised Index of Consumer Prices. The official inflation measure used across the Eurozone, published monthly by Eurostat.
CPI – Consumer Price Index. A national measure of price changes for a basket of goods and services. HICP is the EU-standardized equivalent.
MoM – Month-over-month. Percentage change from one month to the next.
ESL – Electronic Shelf Label.
Jevons Index – A price index computed as the geometric mean of price relatives. Unweighted, so all items contribute equally.
Nowcasting – Forecasting the present or very near future, before official statistics are published.
MIDAS – Mixed Data Sampling. A regression framework for using high-frequency data (e.g. daily) to predict a low-frequency target (e.g. monthly).
U-MIDAS – Unrestricted MIDAS. Each high-frequency observation enters as a separate regressor without imposing a polynomial lag structure.
XGBoost – Extreme Gradient Boosting.
SARIMA – Seasonal Autoregressive Integrated Moving Average. A univariate time series model that captures trend, seasonality, and autocorrelation.
Random walk – A baseline forecast where next period's value equals this period's value. Also called the naïve forecast.
RMSE – Root Mean Squared Error. Penalizes large errors more than small ones.
MAE – Mean Absolute Error. Average of absolute prediction errors.
DM test – Diebold-Mariano test. Tests whether two forecast models have significantly different accuracy.
OOS – Out-of-sample. Predictions on data not used during training.
Expanding window – Evaluation scheme where the training set grows by one period at each step.
Pooled model – A model trained on all countries combined, learning shared patterns.
Bagged model – A model trained separately per region then aggregated. Short for bootstrap aggregating.
Menu costs – The cost a firm incurs when changing a price. ESL technology reduces these to near zero.
Calvo pricing – A model where firms reset prices with a fixed probability each period.
COICOP – Classification of Individual Consumption According to Purpose. The HICP product classification system.
ECB – European Central Bank. Sets monetary policy for the Eurozone.
Eurostat – The statistical office of the EU.
LOCO – Leave-one-country-out cross-validation.
Pearson r – Correlation coefficient measuring linear association, ranging from –1 to +1.

Contents

1 Introduction ... 4
1.1 Background ... 4
1.1.1 Measuring Inflation: CPI and HICP ... 4
1.1.2 Alternative High-Frequency Data Sources ... 5
1.1.3 The State of Inflation Forecasting ... 6
1.1.4 Mixed-Data Sampling (MIDAS) ... 8
1.2 Problem Formulation ... 9
1.3 Purpose ... 10
1.4 Research Questions ... 10
1.5 Delimitations ... 10
1.6 Contributions ... 11
1.7 Thesis Outline ... 11
2 Theoretical Framework ... 12
2.1 Laspeyres vs. Jevons ... 12
2.2 Cross-Country Co-Movement in Retail Prices ... 12
2.3 Nominal Rigidities and the Menu-Cost Channel ... 13
2.4 Gradient Boosting and XGBoost ... 13
2.5 Statistical Testing: The Diebold-Mariano Test ... 13
2.6 Related Work ... 14
3 Methodology ... 15
3.1 The ESL Dataset ... 15
3.2 Jevons Index Construction ... 17
3.3 Macroeconomic Features ... 18
3.4 Mixed-Frequency Nowcasting Framework ... 18
3.4.1 U-MIDAS Feature Construction ... 18
3.4.2 Real-Time Updating ... 19
3.4.3 Evaluation and Benchmarking ... 19
3.4.4 Hyperparameter Tuning and Overfitting Safeguards ... 22
3.4.5 Data Validation Methodology ... 23
3.4.6 Use of AI ... 23
4 Results ... 24
4.1 RQ1: Co-movement Between ESL Data and Official Inflation ... 24
4.1.1 Summary — RQ1 ... 25
4.2 RQ2: ESL-Based Inflation Nowcasting ... 25
4.2.1 Overall Results ... 25
4.2.2 Per-Country Analysis ... 28
4.2.3 Summary — RQ2 ... 28
4.3 RQ3: Regional and Spatial Factors ... 29
4.3.1 Spatial Feature Set ... 29
4.3.2 Overall Results ... 29
4.3.3 Per-Country Spatial Effects ... 30

4.3.4 Leave-One-Country-Out Robustness 30
4.3.5 Summary — RQ3 31
5 Discussion 32
5.1 What the RMSE Improvement Actually Means 32
5.2 The Ablation Paradox 33
5.3 Why Spatial Features Do Not Help 34
5.4 Shelf Prices, Transaction Prices, and What We Are Actually Measuring . 34
5.5 Coverage Bias and Representativeness 35
5.6 The Inflation Regime Problem 35
5.7 Relation to Prior Work 36
5.8 Limitations 37
5.9 Ethical Considerations 38
6 Conclusion 39
Bibliography 40
7 Appendix 42

1 Introduction

1.1 Background

Inflation has re-emerged as a critical concern for the global economy. After decades of relative stability, the Eurozone recently experienced record-high price volatility, with annual inflation reaching peaks that challenged both policymakers and business leaders. In this volatile environment, timely and accurate information is paramount. However, traditional official statistics—such as the Harmonized Index of Consumer Prices (HICP)-suffer from inherent publication lags and a lack of spatial granularity. These national aggregates often mask significant regional disparities. For instance, at the peak of the 2022 inflation surge, annual HICP rates within the Eurozone varied from roughly 7 percent in France to over 20 percent in the Baltic states (European Parliament, 2023).

To bridge this information gap, economists have increasingly turned to alternative high-frequency data. Cavallo and Rigobon (Cavallo & Rigobon, 2016), through the "Billion Prices Project," demonstrated that online micro-price data shares a high co-movement with official CPI and can effectively identify price stickiness and structural breaks. Yet, while online prices offer speed (Cavallo, 2017), they lack a critical dimension: the space dimension. Most physical transactions still occur in brick-and-mortar stores, where local competitive dynamics—such as the density of competitors or regional supply chain shocks—influence pricing decisions.

This thesis proposes a novel approach to inflation "nowcasting" by leveraging a high-frequency, regionally-tagged Electronic Shelf Label (ESL) dataset sourced from a major European ESL cloud platform. By combining the micro-price index construction methods of Cavallo and Rigobon (Cavallo & Rigobon, 2016) with modern high-dimensional machine learning frameworks, this study investigates whether daily retail pricing data — augmented with macroeconomic indicators — can serve as an early warning system for broader inflationary trends.

Since ESL coverage is mainly concentrated in grocery retail, this thesis focuses specifically on food inflation — the HICP food subindex — rather than headline inflation. This scope is made explicit here and carried through the analysis.

1.1.1 Measuring Inflation: CPI and HICP

Inflation is broadly defined as the rate of change in the price level of goods and services typically purchased by households. While the concept is straightforward, its measurement is complex and relies on distinct methodologies designed for different economic objectives.

The most widely recognized measure is the Consumer Price Index (CPI). National statistical agencies calculate CPI by tracking the prices of a fixed "basket of goods" that represents the spending habits of an average consumer. This basket is weighted based on expenditure shares. For example, if households spend more of their income on housing than on entertainment, housing costs will have a proportionally larger impact on the final index.

In the context of the Eurozone, the primary metric is the Harmonised Index of Consumer Prices (HICP). Unlike national CPIs, which may vary in methodology (e.g., how they treat owner-occupied housing), the HICP provides a standardized measure to allow for cross-country comparisons within the European Union. It is strictly a "pure price index" (Laspeyres-type), meaning it measures the changing cost of a fixed basket over time without attempting to measure the cost of maintaining a constant level of utility (a Cost-of-Living Index). The HICP is decomposed into several subindices by COICOP classification: food and non-alcoholic beverages (CP01), energy (NRG), services (SERV), and industrial goods (IGD), among others. This decomposition is important because different subindices respond to different economic drivers — food prices are sensitive to agricultural supply shocks and weather, while energy prices respond to geopolitical events and commodity markets.

The traditional collection process involves manual surveys and physical store visits on a monthly or bi-monthly basis (Eurostat, 2024). This manual approach renders the resulting index a historical record rather than a real-time monitoring tool, with publication lags of two to four weeks after the reference period.

Constructing these indices involves several critical statistical challenges. First, statistical offices must select "representative items" (e.g., a specific type of milk or smartphone) to act as proxies for broader categories, since tracking every single product is infeasible. Second, quality adjustment is required: when products improve (e.g., a computer becomes faster), a price increase may reflect better quality rather than pure inflation, and agencies use hedonic adjustments to strip out these quality improvements. Third, substitution bias affects traditional fixed-weight indices, which overstate inflation because they do not fully account for consumers switching to cheaper alternatives when prices rise (e.g., buying chicken instead of beef). While some indices use geometric means to mitigate this, it remains a persistent challenge. The ESL-based approach used in this study differs from official HICP in several ways: it draws on real-time cloud logs rather than manual surveys, covers the full product range of participating stores rather than a pre-selected basket, reports with near real-time lag (less than 24 hours vs. 2 to 4 weeks), and provides regional (NUTS-2) spatial resolution rather than national aggregates.

1.1.2 Alternative High-Frequency Data Sources

Research has demonstrated that "Big Data" sources can circumvent the publication lags of official statistics to provide early signals of turning points in consumer prices. Three main categories of alternative data have emerged.

Web-scraped online prices constitute the most established alternative source. The Billion Prices Project (Cavallo & Rigobon, 2016) pioneered this approach by collecting daily prices from online retailers across multiple countries, constructing real-time price indices that showed high co-movement with official CPI and could identify structural breaks weeks before official releases. Cavallo (Cavallo, 2017) further demonstrated that online and offline prices are identical approximately 72 percent of the time for large multi-channel retailers, supporting the use of online data as a proxy for physical retail prices. However, online prices lack spatial metadata — they are typically national or retailer-wide, with no information about which specific store or region a price applies to.

Scanner data from point-of-sale systems provide transaction-level records of quantities and prices. Several national statistical offices have begun incorporating scanner data into their official CPI calculations (Białek & Beręsewicz, 2021), with the Netherlands, Norway, Sweden and Switzerland among the early adopters. Scanner data offers comprehensive coverage of all products sold through participating retailers but is typically available only with a delay (weekly or monthly) and is subject to strict confidentiality agreements that limit research use.

Electronic Shelf Labels (ESL) represent a newer and largely unexplored data source. ESL systems — digital price tags deployed in brick-and-mortar stores — are connected to centralized cloud platforms that log every price change with precise timestamps and regional geolocation. Unlike scanner data, which records completed transactions, ESL data captures the *intent* to change prices — the retailer's pricing decision — before any consumer responds. This distinction is economically meaningful: ESL data reflects supply-side pricing behavior, while scanner data reflects the interaction of supply and demand. A leading ESL provider in Europe operates a cloud platform processing price updates from thousands of stores across multiple countries, generating a dataset of over 32 billion price-change events. This thesis is the first — to the authors' knowledge — to exploit ESL data for inflation nowcasting.

A critical property of ESL data that distinguishes it from other sources is that the cloud platform logs price *changes*, not daily price snapshots. Items whose prices remain stable are invisible in the logs. This selection bias has important methodological implications: naive day-over-day price chaining produces severe upward volatility bias, because the sample on any given day is systematically enriched with items undergoing price adjustments. Addressing this bias requires month-over-month matching of prices rather than daily chaining, a methodological choice discussed in detail in the Methodology section.

1.1.3 The State of Inflation Forecasting

To understand the context for this study's forecasting approach, it is essential to review the current standard methods used by central banks and economic institutions. These generally fall into three categories: univariate time-series models, structural economic models, and emerging machine learning approaches.

Univariate and time-series benchmarks. Despite the complexity of the global economy, simple univariate models remain surprisingly difficult to beat. The "Naïve" Random Walk model (which assumes future inflation will equal current inflation) and Autoregressive (AR) models are frequently used as benchmarks. A seminal study by Atkeson and Ohanian (Atkeson & Ohanian, 2001) demonstrated that simple random walk models often outperform complex multivariate models in out-of-sample forecasting, particularly over horizons beyond one quarter. While accurate in stable periods, these models are "atheoretical" — they blindly extrapolate past trends without understanding underlying economic drivers, making them slow to react to sudden structural shocks like the 2022 energy crisis.

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model extends the basic AR framework by incorporating seasonal patterns and differencing for non-stationarity. SARIMA remains a standard benchmark in inflation forecasting because it captures the well-documented seasonal patterns in food prices (harvest cycles, holiday demand) and the persistence of inflation dynamics. However, SARIMA requires a sufficiently long and continuous time series to estimate its parameters reliably — a requirement that proves problematic when working with the relatively short ESL data coverage period available for this study.

Structural and Phillips Curve models. Central banks, such as the ECB and the Federal Reserve, traditionally rely on structural models like the Phillips Curve, which posits a trade-off between unemployment (or economic slack) and inflation. The reliability of the Phillips Curve has been heavily debated. Stock and Watson (Stock & Watson, 2007) argued that its predictive power has diminished significantly since the "Great Moderation" of the 1980s. These models rely on "unobserved" variables like the Output Gap or NAIRU (Non-Accelerating Inflation Rate of Unemployment), which are difficult to estimate in real-time and are subject to large revisions.

Machine learning for inflation forecasting. Recently, researchers have begun applying Machine Learning (ML) to inflation forecasting with promising results. Medeiros et al. (Medeiros et al., 2021) conducted a comprehensive comparison of ML methods for U.S. inflation forecasting using 150–200 predictors and found that Random Forest models systematically outperform traditional linear benchmarks by capturing complex non-linearities in the data. Their results demonstrated that the benefits of ML are most pronounced in data-rich environments where the number of potential predictors exceeds what traditional regression can handle.

Garcia, Medeiros, and Vasconcelos (Garcia et al., 2017) extended this line of research to an emerging-market context (Brazil), showing that high-dimensional models such as LASSO and Random Forest outperform expert surveys in real-time inflation forecasting. Their work is particularly relevant because it demonstrated ML's advantage in environments with high short-term volatility — a characteristic shared by the Eurozone inflation dynamics examined in this thesis.

In the specific context of mixed-frequency data, Lundgren and Wicktor (Lundgren & Wicktor, 2023) applied XGBoost to nowcast U.S. inflation using a combination of monthly macroeconomic data and daily financial variables. Their study found that XGBoost was the top-performing model for identifying inflation turning points in real-time, outperforming both traditional MIDAS regressions and Random Forest. This finding motivates the choice of XGBoost as the primary ML algorithm in this thesis.

The primary ML algorithm used in this thesis, XGBoost (Extreme Gradient Boosting) (Chen & Guestrin, 2016), is a scalable tree-boosting system that has become the dominant algorithm for structured/tabular data in applied ML. XGBoost builds an ensemble of decision trees sequentially, where each new tree corrects the residual errors of the previous ensemble. Its built-in regularization (L1 and L2 penalties on leaf weights, maximum tree depth, minimum child weight) makes it particularly suitable for settings with limited training data and high-dimensional feature spaces — both characteristics of the inflation nowcasting problem studied here. Figure 1 illustrates the sequential boosting process.

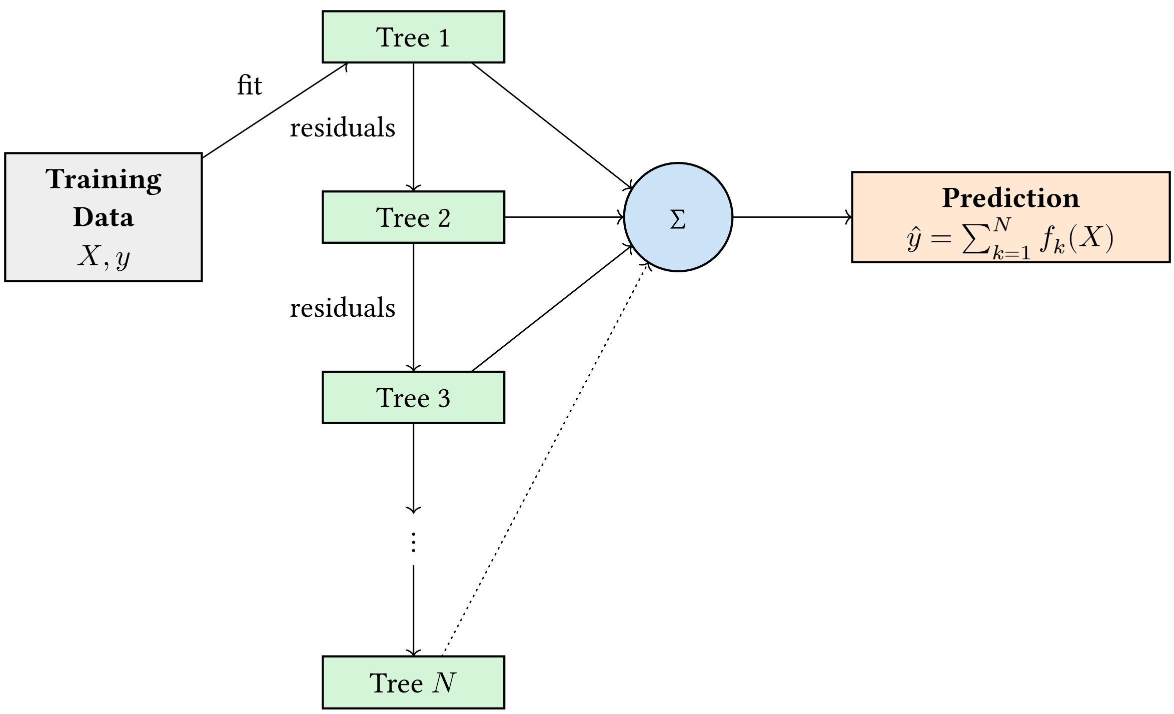

Figure 1: XGBoost sequential boosting. Each tree is fitted on the residual errors of the previous ensemble. The final prediction is the sum of all trees. Regularization (L1/L2 penalties, tree depth limits) prevents overfitting.

1.1.4 Mixed-Data Sampling (MIDAS)

A key methodological challenge in inflation nowcasting is bridging the frequency gap between high-frequency indicators (daily prices) and the low-frequency target (monthly HICP). The Mixed-Data Sampling (MIDAS) framework, introduced by

Ghysels, Santa-Clara, and Valkanov (Ghysels et al., 2004), addresses this by allowing direct regression of low-frequency variables on high-frequency regressors without requiring pre-aggregation.

In its original form, MIDAS imposes a parametric weighting function (such as Almon or Beta polynomials) on the high-frequency lags to reduce dimensionality. The Unrestricted MIDAS (U-MIDAS) variant relaxes this constraint by treating each high-frequency lag as an independent regressor, which is feasible when the number of high-frequency observations per low-frequency period is manageable (e.g., 22 trading days per month). When combined with ML methods that handle high-dimensional feature spaces natively (such as XGBoost), U-MIDAS becomes particularly attractive: the model can learn which specific daily lags are most informative without imposing a parametric decay structure.

This thesis adopts the U-MIDAS approach, where each of the last 22 daily Jevons Index log-returns within a month becomes a separate feature. This preserves the full intra-month temporal structure, allowing the model to detect patterns such as end-of-month price adjustments or early-month promotional resets that a simple monthly average would obscure.

1.2 Problem Formulation

Despite the advancements in using machine learning and alternative data for economic forecasting, a significant gap remains in the literature. Existing research typically falls into two categories that leave complementary blind spots:

Online micro-price studies utilize vast datasets of web-scraped prices (e.g., The Billion Prices Project (Cavallo & Rigobon, 2016)) but lack geospatial metadata, treating the market as a single non-spatial entity. These studies demonstrate that alternative price data can track official CPI, but they do not attempt to *forecast* inflation using ML models trained on the micro-price dynamics.

National macro-forecasting studies such as Garcia et al. (Garcia et al., 2017) and Medeiros et al. (Medeiros et al., 2021) demonstrate that high-dimensional ML models can outperform expert surveys and traditional econometric baselines. However, these models rely on aggregate national indicators (unemployment, interest rates, commodity prices) and ignore the heterogeneous behavior of local markets. They also operate at monthly frequency, discarding the intra-month dynamics that high-frequency data sources could provide.

The intersection — using *physical retail micro-price data* at *daily frequency* within an *ML-based mixed-frequency framework* — remains largely unexplored. Lundgren and Wicktor (Lundgren & Wicktor, 2023) demonstrated the potential of combining daily financial data with ML for inflation nowcasting, but their daily variables were financial market indicators (bond yields, commodity prices), not retail prices. No prior study, to the authors' knowledge, has used Electronic Shelf Label data — which captures the pricing decisions of physical retailers in real time — as an input to inflation nowcasting models.

This gap motivates the central question of this thesis: can daily ESL pricing data, processed through a MIDAS-XGBoost framework, improve inflation nowcasting relative to traditional baselines? And if so, does augmenting the ESL signal with macroeconomic indicators yield additional gains?

1.3 Purpose

The primary purpose of this thesis is to evaluate whether high-frequency Electronic Shelf Label (ESL) pricing data can improve short-term inflation nowcasting relative to traditional econometric baselines. Using a tournament framework, the study benchmarks a MIDAS-XGBoost architecture — which ingests daily Jevons Index log-returns as unrestricted regressors — against naïve and SARIMA baselines. The study further investigates whether regional and cross-country spatial dependencies in ESL pricing data carry additional predictive signal for inflation forecasting.

1.4 Research Questions

To fulfill this purpose, the study addresses three research questions:

RQ1: Co-movement Between ESL Data and Official Inflation Does the ESL-derived Jevons Index exhibit statistically significant co-movement with the official HICP food subindex, and can it serve as a meaningful proxy for retail food price dynamics?

RQ2: ESL-Based Inflation Nowcasting Can a MIDAS-XGBoost model using daily Electronic Shelf Label pricing data forecast the monthly HICP food subindex, and how accurately does it perform compared to traditional baselines (naïve random walk and SARIMA)?

RQ3: Regional and Spatial Factors Do regional and cross-country spatial factors — price dynamics in neighboring regions and countries — improve inflation nowcasting accuracy beyond temporal-only models?

1.5 Delimitations

- **Data Scope**: The study is limited to product-level pricing data from an ESL cloud platform, covering predominantly grocery retail products (food and fast-moving consumer goods, though product categories are not identified in the data). It does not cover the full CPI basket (services, energy, and housing are excluded from the ESL data, though they enter as macro features).
- **Geographical Scope**: The analysis covers six Eurozone countries (France, Belgium, Italy, Spain, Finland, and Estonia) where sufficient store density exists.
- **Forecasting Horizon**: The study focuses on one-month-ahead nowcasting, aligning with the monthly publication cycle of official HICP.

1.6 Contributions

This thesis makes three contributions to the literature on inflation nowcasting:

1. **Novel data source and validation**: It is the first study to use Electronic Shelf Label (ESL) cloud data — real-time logs of pricing decisions in physical retail stores — for inflation analysis, and validates that the resulting Jevons Index co-moves significantly with official HICP food inflation across six Eurozone countries.
2. **MIDAS-XGBoost framework**: It combines the U-MIDAS mixed-frequency approach with XGBoost gradient boosting, preserving daily intra-month price dynamics as unrestricted regressors rather than pre-aggregating to monthly frequency.
3. **Spatial signal evaluation**: It systematically tests whether cross-country spatial dependencies in retail pricing data carry predictive signal for aggregate inflation, using both pooled and bagged ensemble architectures to control for pseudo-replication.

1.7 Thesis Outline

The remainder of this thesis is organized as follows. Section 2 presents the theoretical framework, covering index number theory, cross-country price co-movement, gradient boosting, and related work. Section 3 describes the methodology, including data collection, feature engineering, model specifications, and the evaluation framework. Section 4 reports the empirical results for all three research questions. Section 5 discusses the findings, compares with related work, and addresses limitations. Section 6 concludes with a summary and directions for future research.

2 Theoretical Framework

2.1 Laspeyres vs. Jevons

Bridging the gap between individual price quotes and aggregate indices requires a method of aggregation. This research employs Index Number Theory — specifically the Jevons Index — to construct a price index from ESL micro-price data.

Official CPI typically utilizes a Laspeyres-type formula, which uses fixed base-period weights (Eurostat, 2024). However, for high-frequency micro-data where expenditure weights are unavailable in real-time, the Jevons Index is preferred (Cavallo & Rigobon, 2016). The Jevons formula computes the geometric mean of individual items' price ratios ($\frac{p_t}{p_{\{t-1\}}}$), which partially accounts for consumer substitution bias by giving equal weight to each item's price change rather than weighting by expenditure shares.

2.2 Cross-Country Co-Movement in Retail Prices

The spatial econometrics literature (Anselin, 1988; LeSage & Pace, 2009) provides formal models for geographic dependencies, where the outcome at location i depends on outcomes at neighboring locations. LeSage and Pace (LeSage & Pace, 2009) demonstrate that spatial lags can improve regional forecasting when applied to contiguous regions sharing similar economic structures. However, the setting in this thesis differs substantially from the classical spatial framework.

The ESL data covers six Eurozone countries, each represented by Jevons indices aggregated to the NUTS-2 region level from one or more ESL-equipped retailers, with the HICP target defined at the country level. With only six cross-sectional units at the national level, formal spatial econometric methods — which rely on large panels and well-defined contiguity structures — are not applicable. Instead, this thesis tests a simpler hypothesis: whether the price dynamics observed in other countries' ESL data carry predictive signal for the focal country's HICP outcome.

The theoretical motivation for such cross-country co-movement rests on economic linkages: trade in goods transmits input cost shocks across borders, shared supply chains propagate disruptions simultaneously, and monetary policy (e.g., ECB rate decisions) affects all Eurozone members. These are economic linkage mechanisms, distinct from the geographic proximity described by Tobler's First Law (Tobler, 1970) or the store-level competitive effects modeled by Hotelling (Hotelling, 1929).

The cross-country features constructed in this thesis — average Jevons links of other countries, price gaps, and dispersion measures — are best understood as testing whether common shocks or economic linkages across Eurozone retail markets carry predictive signal, rather than as a test of spatial autocorrelation in the classical sense. If these features fail, it does not imply that spatial price dynamics are uninformative for inflation — only that country-level averages from six heterogeneous retailers do not capture them.

2.3 Nominal Rigidities and the Menu-Cost Channel

Why retail price changes should carry information about inflation rests on the theory of nominal rigidities. In the Calvo (Calvo, 1983) framework, firms reset prices with a fixed probability each period; the New Keynesian Phillips Curve treats this staggering as a key source of inflation persistence. State-dependent models attribute the rigidity to menu costs (Levy et al., 1997) — the operational cost of changing a price tag. This matters directly for ESL data: by reducing the cost of a price change to a near-zero digital signal, ESL technology raises the effective reset probability $(1 - \theta)$, so ESL-equipped stores adjust prices more frequently and the resulting index is more volatile than official HICP. The empirical relevance of this channel is revisited in Section 5.1.

2.4 Gradient Boosting and XGBoost

Gradient boosting is an ensemble learning method that builds a predictive model by sequentially adding weak learners (typically decision trees) that correct the errors of the existing ensemble. Given a loss function $L(y, \hat{y})$, the method constructs an additive model $F(x) = \sum_{m=1}^{M} f_m(x)$, where each new tree f_m is fitted to the negative gradient (pseudo-residuals) of the loss with respect to the current prediction.

XGBoost (Chen & Guestrin, 2016) extends the basic gradient boosting framework with several innovations relevant to this study. First, it incorporates L1 (Lasso) and L2 (Ridge) regularization on leaf weights, which penalizes model complexity and reduces overfitting — critical when training data is limited to 14–20 months per region. Second, it uses a second-order Taylor approximation of the loss function, enabling more efficient optimization. Third, it handles missing values natively by learning optimal default split directions, which is useful when some daily lags are zero-padded due to varying numbers of trading days.

The key hyperparameters in this study are: maximum tree depth (controlling model complexity), number of estimators (ensemble size), learning rate (shrinkage per step), and regularization strengths (α and λ). The bagged variant uses shallow trees (depth 2) with 30 estimators per region, while the pooled variant's hyperparameters are selected via grid search on the initial training window (see Section 3.4.4).

2.5 Statistical Testing: The Diebold-Mariano Test

To formally compare forecast accuracy between models, this study employs the Diebold-Mariano (DM) test (Diebold & Mariano, 1995). Given two sequences of forecast errors $e_{1,t}$ and $e_{2,t}$ from competing models, the test evaluates whether the expected difference in squared errors is zero:

$$d_t = e_{1,t}^2 - e_{2,t}^2, \quad H_0 : E[d_t] = 0$$

The test statistic is the mean of d_t divided by its standard error, which follows a standard normal distribution asymptotically. A significant positive d indicates that model 2 has lower squared forecast errors (is more accurate) than model 1. The

DM test is particularly suitable for comparing nested models (e.g., temporal vs. macro-augmented) because it does not require the forecast errors to be normally distributed or serially uncorrelated.

2.6 Related Work

This study draws on four main strands of literature (see Table 6 in the Appendix for a summary). Cavallo and Rigobon (Cavallo & Rigobon, 2016) established that web-scraped micro-prices can anticipate CPI turning points, validating alternative price data for inflation measurement. Garcia et al. (Garcia et al., 2017) and Medeiros et al. (Medeiros et al., 2021) demonstrated that ML methods — particularly Random Forest and LASSO — systematically outperform traditional linear benchmarks in data-rich inflation forecasting environments. Most directly, Lundgren and Wicktor (Lundgren & Wicktor, 2023) showed that XGBoost combined with daily financial variables in a MIDAS framework is the top performer for CPI nowcasting. This thesis extends their approach by replacing financial market indicators with physical retail ESL data and adding a cross-country spatial dimension.

3 Methodology

3.1 The ESL Dataset

The raw data is a log of individual price-change events generated by an ESL cloud platform, comprising events across multiple European grocery retailers. A defining property of the data is that it records price changes rather than daily price snapshots: an observation is written only when a price is updated, so items whose prices remain stable do not appear in the logs on a given day. This event-log structure, rather than a balanced panel of daily prices, motivates the matching and aggregation choices described in the Jevons Index Construction section.

Each event carries an item identifier, a regional identifier, a timestamp, and the updated price. The item identifier is used only to match an item to itself across consecutive periods so that a price relative can be computed; it is never retained in the output. The regional identifier locates each price change at the NUTS-2 level, which is the finest spatial unit used in this study. The data contains no consumer-side information — no transactions, quantities, or shopper identities — so it captures the supply-side pricing decision (the price posted on the shelf) and nothing about demand. It likewise contains no product-category labels, so the focus on food is an inference from the grocery-retail composition of the participating stores rather than a category filter applied to the data.

From these events the analysis constructs a panel keyed by region and month. Each region-month observation carries 31 ESL-derived features — the 22 daily Jevons log-returns within the month and 9 aggregate statistics summarising them — together with 6 lagged macroeconomic features drawn from public Eurostat and ECB sources. The forecasting target, the month-over-month change in the official HICP food subindex, is defined at the country level. Because several NUTS-2 regions within a country share the same country-level target, the panel contains many more region-month rows than independent target observations; this pseudo-replication is addressed by the bagged architecture described in Section 3.4.3.

The target the models predict is the month-over-month change in the HICP food subindex, shown per country in Figure 2. The series is noisy and close to mean-zero, with markedly different volatility across markets — Estonia and Belgium swing far more than France — and no dominant within-country trend. This is a difficult target, and its volatility profile foreshadows the per-country results in Section 4.2, where the largest model gains appear precisely in the most volatile markets.

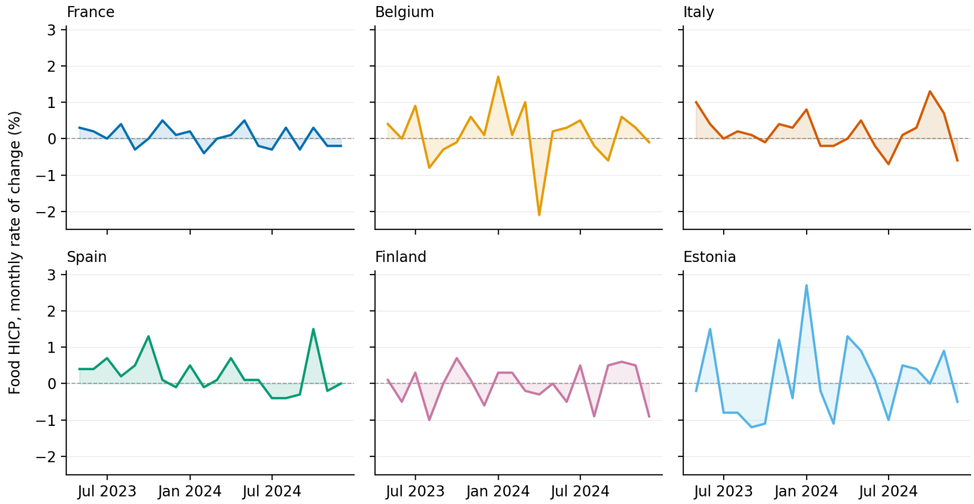

Figure 2: The forecasting target by country: month-over-month change in the HICP food subindex (CP011), May 2023 – December 2024, on a shared scale. Source: Eurostat (`prc_hicp_mmor`). The series is volatile and largely mean-zero; Estonia and Belgium show the widest month-to-month swings, while France is comparatively stable.

The ESL data used in this study was anonymized and aggregated prior to analysis to comply with the EU General Data Protection Regulation (GDPR) and the contractual confidentiality obligations owed to the data provider and its retail customers. Several layers of anonymization were applied:

No individual item identification. The events include an item identifier used solely to match items across time when computing price relatives; it is never retained in any output. The Jevons Index is a geometric mean across all matched items within a region, producing an aggregate that cannot be reverse-engineered to recover individual product prices.

No store-level identification. The events carry only an item identifier and a NUTS-2 regional identifier; no store identifiers, coordinates, or other store metadata are present in the data accessed for this study. All spatial analysis therefore operates at the NUTS-2 level, and no individual store can be isolated.

No consumer data. Unlike scanner or loyalty-card data, ESL logs contain no information about consumer transactions, purchase quantities, or shopper identities. The data captures only the supply-side pricing decision — the price displayed on the shelf — with no demand-side component.

These anonymization measures mean that the study operates on aggregated price indices rather than micro-level price records. While this sacrifices some granularity (e.g., category-level or brand-level analysis is not possible), it ensures full GDPR compliance and protects the commercial interests of the participating retailers and their customers. The trade-off is acceptable for the research objective of nowcasting the national HICP food subindex, which is itself a macro-level aggregate.

3.2 Jevons Index Construction

The study constructs two versions of the Jevons Index from ESL price-change logs, each serving a different purpose.

Daily Jevons links form the primary input to the nowcasting models. For each store-day, a Jevons link is computed as the geometric mean of price relatives between each item's current and previous closing price:

$$J_{s,t} = \prod_{i=1}^{n_{s,t}} \left(\frac{p_{i,t}}{p_{i,t-1}} \right)^{1/n_{s,t}}$$

where $n_{s,t}$ is the number of matched items in store s on day t. These daily links are aggregated to the region level by weighting each store proportionally to its matched item count. Concretely, the region-level log-return is the item-count-weighted mean of the store-level log-returns, $\sum_s \left(n_{s,t}/N_t\right) \cdot \log\left(J_{s,t}\right)$, where $N_t = \sum_s n_{s,t}$; stores contributing more matched price observations therefore carry more weight, so the aggregate is not distorted by small stores with few logged changes. The Jevons link J_t is a ratio centered around 1.0 (e.g., 1.003 for a 0.3% price increase). The log transformation $\log(J_t)$ is applied to center these values around zero, ensure symmetry (so that equal-magnitude increases and decreases have equal absolute values), and enable additivity (log-returns over multiple days can be summed rather than multiplied). The resulting daily log-returns within each month become the U-MIDAS regressors used by the XGBoost models — each trading day's log-return is treated as a separate feature, preserving intra-month price dynamics that monthly aggregation would discard.

Monthly Jevons Index is used for the validation analysis (comparing with official HICP). Because the ESL platform logs only price *changes* — items with stable prices are invisible — daily chaining over longer horizons accumulates volatility bias. The monthly index therefore uses month-over-month direct matching: for each item observed in both month $m - 1$ and month m, the last recorded price in each month is taken as the closing price. The closing price is preferred over the within-month average because it reflects the retailer's current pricing decision rather than blending old and new prices. Moreover, computing a meaningful average would require knowing how long each price was in effect, but since the ESL platform logs only price *changes* (not daily snapshots), the duration between events cannot be reliably estimated for items with stable prices. The closing price sidesteps this by using the last observed value, which is always well-defined. Three filters reduce noise: (1) a stable basket requirement, retaining only items present in at least 3 months; (2) tight price-ratio bounds $[0.8, 1.25]$ to exclude promotional spikes and data errors; and (3) store-weighted aggregation. The chained index is $P_m = P_{m-1} \times J_m$, initialized at $P_0 = 100$.

3.3 Macroeconomic Features

Alongside the ESL-derived features, the study draws six macroeconomic indicators from public Eurostat and ECB sources, all entered at a one-month lag (t− 1). These comprise the HICP subindices for energy (NRG), services (SERV), and industrial goods (IGD), the ECB main refinancing rate and its one-month lag, and the producer price index (PPI). Energy and industrial-goods prices are included because food inflation is partly driven by input costs — fuel, packaging, and transport — that these subindices proxy. The services subindex captures the broader inflation regime and second-round wage effects, while the ECB rate reflects the monetary-policy stance that conditions all Eurozone prices. Only lagged values are used: contemporaneous HICP subindices are published with the same delay as the food-subindex target, so including them would be circular for a real-time nowcast. The ECB refinancing rate is entered as a step-function approximation of the prevailing policy rate rather than from official daily records.

3.4 Mixed-Frequency Nowcasting Framework

To bridge the temporal gap between daily Electronic Shelf Label (ESL) logs and the monthly publication cycle of official inflation figures, this study adopts the Mixed-Data Sampling (MIDAS) framework. This approach addresses the problem when high-frequency indicators (daily price relatives) are available before the low-frequency target (monthly HICP).

3.4.1 U-MIDAS Feature Construction

The study utilizes the Unrestricted MIDAS (U-MIDAS) approach to regress the monthly HICP target on daily ESL-based signals without pre-aggregating daily data into monthly averages, which would discard valuable intra-month volatility (see Figure 3 for an illustration). Each daily Jevons log-return within a month is treated as an independent regressor, yielding 22 daily lag features per monthly observation. These are complemented by 9 aggregate statistics computed over the same daily values: mean, standard deviation, minimum, maximum, range, and skewness of the daily log-returns, plus the number of trading days, average number of stores, and average matched items in the month. Together, these form a 31-feature vector per country-month observation. Unlike traditional MIDAS, which imposes a parametric weighting function (e.g., Almon or Beta polynomials) to reduce dimensionality, U-MIDAS lets the ML model determine the importance of each daily lag directly. This is feasible because the number of daily observations per month (22) is small enough for XGBoost to handle without a parametric constraint, while still allowing the model to capture non-linear interactions between specific intra-month price movements and the monthly HICP outcome.

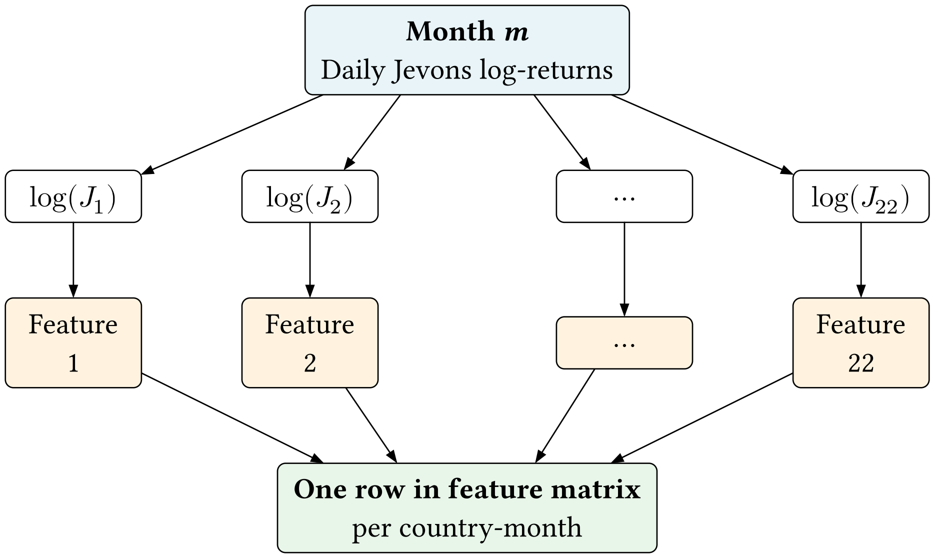

Figure 3: U-MIDAS feature construction: each daily log-return within a month becomes a separate regressor, preserving intra-month price dynamics.

3.4.2 Real-Time Updating

The primary advantage of this framework is its ability to perform real-time updates. As daily ESL logs arrive from the cloud platform, the information set (I_t) is updated dynamically. This mirrors the methodology of Lundgren and Wicktor (Lundgren & Wicktor, 2023), where the inclusion of high-frequency data allowed for the identification of inflation turning points with significantly lower latency than traditional macroeconomic models.

3.4.3 Evaluation and Benchmarking

The "Tournament Framework" evaluates predictive performance using an expanding-window out-of-sample scheme. The first six months serve as the initial training window; each subsequent month is predicted one step ahead. Performance is measured by Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) across all country-month predictions.

The out-of-sample count follows directly from the panel: six countries, each contributing the months after the six-month initial training window, net of months with insufficient ESL coverage, yields 78 country-month predictions in total. The 114 country-months used for the RQ1 validation reduce to these 78 out-of-sample predictions once the six-month initial training window is removed from each country.

The models are chosen to answer a series of increasingly specific questions. The naïve baseline asks whether any model is needed at all. SARIMA tests whether the HICP series' own history is sufficient. The macro-only XGBoost asks whether publicly available indicators already capture the signal. The ESL-only XGBoost isolates the contribution of the novel ESL data. Finally, the combined model tests whether the two sources complement each other. This layered design ensures that any improvement attributed to ESL data is measured against progressively stronger baselines.

The tournament evaluates five model tiers:

1. **Naïve**: Last month's observed HICP food subindex MoM change (random walk baseline).
2. **SARIMA**: Fitted on the full continuous HICP series from Eurostat to ensure a fair statistical baseline.
3. **MIDAS-XGBoost (ESL-only)**: XGBoost trained on 31 ESL-derived features (22 U-MIDAS daily lags + 9 aggregate statistics). A pooled variant trains one model on all regions simultaneously; a bagged variant trains one weak learner per region and averages predictions to address pseudo-replication.
4. **XGBoost (macro-only)**: XGBoost trained on macroeconomic features only, without any ESL-derived inputs. This ablation isolates the predictive contribution of macro indicators and provides the counterfactual needed to assess whether ESL data adds value beyond publicly available macroeconomic series.
5. **MIDAS-XGBoost (ESL + macro)**: The ESL-only model augmented with lagged macroeconomic features: one-month-lagged HICP subindices for energy, services, and industrial goods, the ECB main refinancing rate (known in real-time), and lagged producer prices where available. Only lagged (t−1) macro features are used, since contemporaneous HICP subindices are published with the same delay as the HICP food subindex target itself. By comparing this combined model against both the ESL-only and macro-only variants, the ablation quantifies the marginal contribution of each data source.

The bagged architecture addresses a key concern: all regions within a country share the same HICP target, so a pooled model trained on region-level data implicitly overweights countries with more regions. Regional bagging — training one shallow XGBoost (depth 2, 30 estimators) per region and averaging — eliminates this pseudo-replication bias.

Model	Type	Description
Naïve	Baseline	Last month's HICP food subindex MoM (random walk)
SARIMA	Statistical	Fitted on full continuous Eurostat HICP series
Pooled temporal	ML	XGBoost on all regions: 22 U-MIDAS lags + 9 aggregate stats
Pooled macro-only	ML	XGBoost on 6 lagged macro features only (no ESL data)
Pooled +macro	ML	Temporal + 6 lagged macro features
Bagged temporal	Ensemble	One weak XGBoost per region, predictions averaged
Bagged +macro	Ensemble	Bagged temporal + macro features

Table 1: Model specifications. Temporal models use 31 ESL-derived features; macro models add 6 lagged macroeconomic indicators.

Table 1 summarizes the model specifications. The temporal feature set consists of 31 features: 22 U-MIDAS daily lags (log-returns of the Jevons link for the last 22 trading days within each month) and 9 aggregate statistics (mean, standard deviation, minimum, maximum, range, skewness, number of trading days, average stores, and average matched items). These nine statistics summarise the shape of the intra-month price-change distribution that the 22 daily lags represent individually: central tendency (mean), spread (standard deviation, range, min, max), asymmetry (skewness), and three coverage measures (number of trading days, average stores, average matched items). The coverage measures matter because the number of logged price changes varies across months; they let the model down-weight months whose Jevons link rests on a thin sample. The SARIMA(1,1,1) specification follows the standard parsimonious choice for a differenced, mean-reverting monthly series and avoids over-fitting seasonal terms on a short sample.

The macro-augmented variants add 6 lagged macroeconomic features: one-month-lagged HICP subindices for energy (NRG), services (SERV), and industrial goods (IGD), the ECB main refinancing rate and its one-month lag, and lagged producer prices (PPI). Only lagged (t−1) macro variables are included; contemporaneous HICP subindices are excluded because they are published with the same delay as the

HICP food subindex target, making their inclusion nearly circular for nowcasting purposes.

3.4.4 Hyperparameter Tuning and Overfitting Safeguards

With short training windows (14–20 months per region) and a high feature-to-observation ratio (up to 37 features for limited country-months in the pooled model), overfitting is a primary concern. Several safeguards are employed.

Grid search with nested cross-validation. Hyperparameters are selected via grid search over maximum tree depth $\in \{2, 3, 4, 5\}$, number of estimators $\in \{30, 50, 100, 200\}$, learning rate $\in \{0.01, 0.05, 0.1\}$, and L2 regularization strength $\lambda \in \{0.1, 1.0, 5.0\}$ — a total of 144 configurations. To prevent information leakage, hyperparameter selection uses only the initial training window (the first six months of data) with a 3-fold time-series-aware cross-validation (`TimeSeriesSplit`). The selected hyperparameters are then held fixed across all subsequent expanding-window steps, ensuring that no future data influences model selection. This approach avoids the computational cost of re-running grid search at every forecast origin while maintaining strict temporal separation between tuning and evaluation data.

Regularization. XGBoost's built-in L1 and L2 penalties on leaf weights constrain model complexity. The bagged variant further restricts tree depth to 2 with only 30 estimators per region, producing deliberately weak learners that are unlikely to memorize noise.

Multiple random seeds. Each model configuration is trained across multiple random seeds to assess sensitivity to stochastic initialization (tree subsampling, feature subsampling). The reported RMSE values are averaged across seeds, and the standard deviation across runs is monitored. If performance varies substantially across seeds, this signals instability rather than genuine predictive signal.

Comparison across architectures. The pooled and bagged architectures serve as mutual consistency checks. If a feature set improves the pooled model but not the bagged model (or vice versa), this asymmetry suggests the improvement is an artifact of the specific training regime rather than a robust signal.

In addition to the expanding-window evaluation, a leave-one-country-out (LOCO) cross-validation is performed as a robustness check. For each of the six countries, a model is trained on the remaining five and evaluated on the held-out country. LOCO tests whether the learned relationships generalize across national markets — a necessary condition for the pooled model to be considered a genuine cross-country nowcasting tool rather than an artifact of in-sample fitting. Because each held-out country is entirely absent from training, LOCO also reveals whether feature sets (temporal, spatial, macro) contribute transferable signal or merely capture country-specific noise.

3.4.5 Data Validation Methodology

Before using the Jevons Index as input features for the nowcasting models, the study tests whether it co-moves with official Eurostat HICP. The ESL data originates from grocery retailers whose product mix is predominantly food and fast-moving consumer goods, though the data does not include product category labels. The HICP food subindex is used as the benchmark, as it most closely corresponds to the product categories sold in the observed stores. The monthly Jevons Index used for this validation is constructed as described in the Jevons Index Construction section above.

3.4.6 Use of AI

Large language models (LLMs) were used as assistive tools at several stages of this thesis. **In the planning phase:** LLMs assisted with creating the timeline of the project and structured sprints to ensure every step delivers a measurable result.

In the data engineering phase: LLMs assisted with writing and debugging SQL queries for BigQuery and Python code for data processing pipelines.

In the modeling phase: they were used to draft boilerplate code for the tournament framework, troubleshoot library-specific errors, and review feature engineering logic. In the writing phase, LLMs assisted with structuring sections, improving language clarity, and formatting references.

All LLM-generated code was reviewed, tested, and validated by the author before inclusion. No LLM was used to generate research findings, interpret results, or draw conclusions - these reflect the author's own analysis and judgment. The models and experimental design were conceived independently; LLMs served as productivity tools analogous to code editors or documentation search engines.

4 Results

This section presents the empirical findings organized around the three research questions. Section 4.1 addresses RQ1 by testing whether the ESL-derived Jevons Index co-moves with official HICP. Section 4.2 addresses RQ2 by evaluating how accurately MIDAS-XGBoost models forecast the monthly HICP food subindex. Section 4.3 addresses RQ3 by testing whether cross-country spatial factors improve nowcasting accuracy.

4.1 RQ1: Co-movement Between ESL Data and Official Inflation

Table 2 reports the Pearson correlations between the monthly Jevons MoM and HICP food subindex MoM for each country.

Country	**Pearson *r***	***p*-value**	**Months**
France (FR)	+0.40	0.08	20
Belgium (BE)	+0.49	0.03	20
Italy (IT)	−0.05	0.87	16
Spain (ES)	+0.38	0.10	20
Finland (FI)	+0.42	0.07	20
Estonia (EE)	+0.31	0.21	18
Pooled	**+0.43**	**< 0.01**	**114**

Table 2: Pearson correlation between Jevons MoM and HICP food subindex MoM, May 2023 – Dec 2024.

The pooled correlation of $r = +0.43$ ($p < 0.01$) indicates statistically significant co-movement across all six countries. Belgium ($r = +0.49$) and Finland ($r = +0.42$) exhibit the strongest positive correlations, followed by France ($r = +0.40$), Spain ($r = +0.38$), and Estonia ($r = +0.31$). Italy ($r = -0.05$) shows no co-movement, likely due to low coverage for Italian stores during the sample. Italy is retained in the pooled analysis for completeness and as a transparency check — its inclusion makes the pooled correlation a conservative estimate — but its country-level results should be read as uninformative rather than as evidence against the method. Individual country-level tests generally do not reach significance at $\alpha = 0.05$ due to limited sample sizes, though Belgium approaches significance ($p = 0.03$). Comparison against the all-items index (CP00) yielded a weaker pooled correlation, confirming that the ESL-derived index tracks the food subindex more closely than headline HICP.

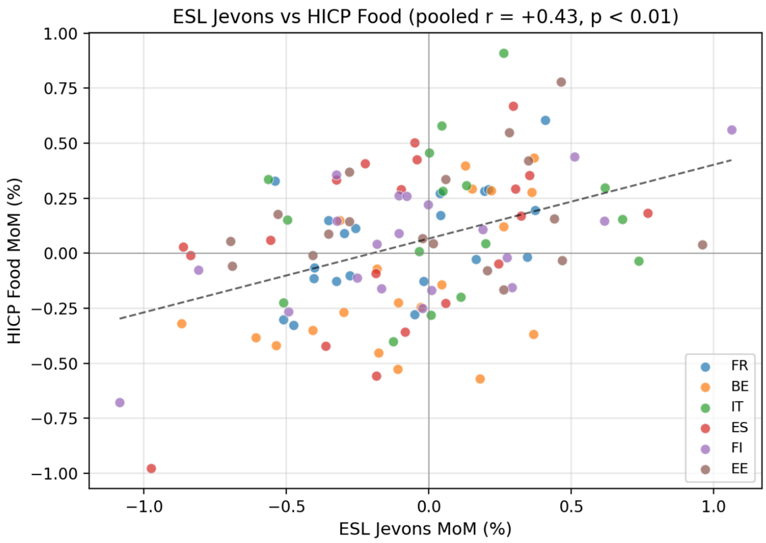

Figure 4: Scatter plot of ESL Jevons MoM vs. HICP food subindex MoM across all study countries. Each point represents one country-month. The dashed line shows the pooled linear fit ($r = +0.43$, $p < 0.01$).

Statistical co-movement is a necessary precondition, not the end goal: it establishes that the Jevons index and official food inflation are driven by a shared component rather than being unrelated series. A significant positive correlation means the index contains information about the HICP target that a forecasting model can, in principle, exploit. RQ1 therefore licenses the nowcasting exercise in RQ2 — without demonstrated co-movement, any predictive performance there would more likely reflect spurious fitting than genuine signal.

4.1.1 Summary — RQ1

The ESL-derived Jevons Index shows statistically significant co-movement with the HICP food subindex ($r = 0.43$, $p < 0.01$) across 114 country-months in six Eurozone countries. Five of six countries show positive correlations, with Belgium and Finland the strongest. Italy is the exception, showing no co-movement. These results establish that the Jevons Index captures meaningful retail price dynamics, supporting its use as the primary feature source in the nowcasting models that follow.

4.2 RQ2: ESL-Based Inflation Nowcasting

4.2.1 Overall Results

Table 3 reports the out-of-sample RMSE across all predicted country-months.

Scope	Naïve	SARIMA	Pooled ESL	Pooled macro	Pooled ESL+ macro	Bagged
Pooled	0.0114	0.0087	**0.0073**	0.0074	**0.0073**	0.0076
FR	0.0060	**0.0034**	0.0039	0.0041	0.0039	0.0038
BE	0.0090	0.0096	0.0068	0.0067	0.0068	**0.0063**
IT	0.0067	0.0057	**0.0053**	**0.0053**	**0.0053**	**0.0053**
ES	0.0080	0.0071	**0.0053**	0.0055	0.0054	0.0059
FI	0.0110	0.0071	0.0071	0.0071	**0.0070**	0.0074
EE	0.0204	0.0143	0.0121	0.0122	**0.0121**	0.0130

Table 3: Out-of-sample RMSE by model and country (lower is better). Bold values indicate the best model per row. 78 country-months across six Eurozone countries.

Figure 5 compares out-of-sample RMSE across all models, and Figure 6 plots predicted against actual values.

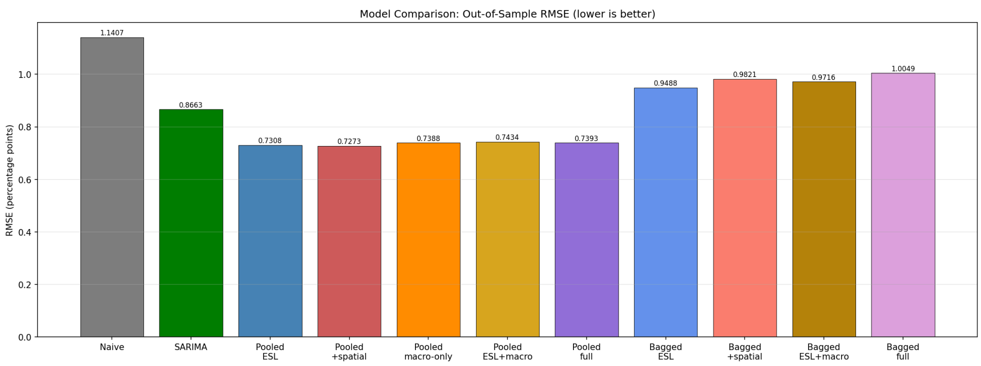

Figure 5: Out-of-sample RMSE comparison across all tournament models (lower is better).

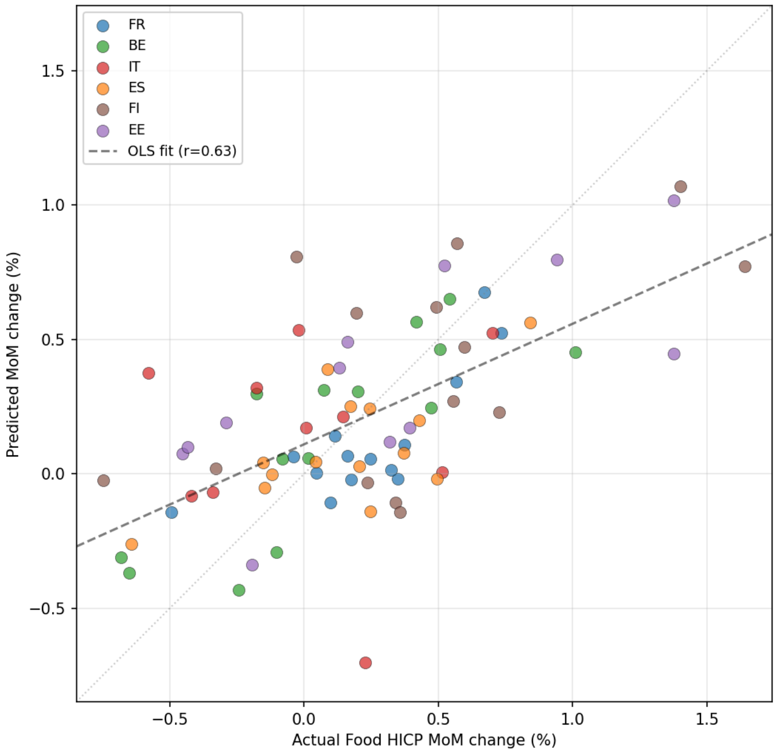

Figure 6: Predicted vs. actual Food HICP MoM change (%) for the pooled ESL+macro model. Each point represents one country-month. The dashed line shows the OLS fit ($r = 0.63$, $p < 0.001$), indicating meaningful directional tracking.

Several findings emerge from Table 3, Figure 5, and Figure 6:

MIDAS-XGBoost substantially outperforms the naïve baseline. The pooled temporal (ESL-only) XGBoost achieves a pooled RMSE of 0.0073, representing a 35.9% reduction relative to the naïve random walk (RMSE = 0.0114). This improvement is consistent across all six Eurozone countries, with the largest gains in Estonia (—40.5%), Finland (—35.7%), and Spain (—33.3%).

SARIMA provides a competitive but weaker baseline. When fitted on the full continuous HICP series from Eurostat, SARIMA achieves a pooled RMSE of 0.0087, a 24.1% improvement over the naïve baseline. SARIMA outperforms all pooled ML variants in France (RMSE = 0.0034 vs. 0.0039 for pooled ESL-only), where the long and stable HICP history favors univariate statistical models. However, SARIMA underperforms in countries with more volatile inflation dynamics (Belgium, Estonia).

ESL-only and macro-only achieve comparable accuracy. The macro-only model (RMSE = 0.0074, —35.2% vs. naïve) performs comparably to the ESL-only model (RMSE = 0.0073, —35.9%). This near-parity is notable given that the macro-only model uses just 6 lagged features versus 31 ESL-derived features, suggesting that lagged HICP subindices and the ECB rate carry substantial predictive information for the HICP food subindex.

Combining ESL and macro features yields marginal gains. The ESL+macro model (RMSE = 0.0073, −35.8%) performs comparably to the ESL-only model, suggesting that the macro features add little incremental signal once the ESL features are included.

Bagged models are invariant to feature augmentation. All bagged model variants produce identical predictions (RMSE = 0.0076, −33.1%), regardless of whether macro, spatial, or combined features are included. However, this invariance likely reflects the model's limited capacity rather than the features' lack of signal: with only 14–20 training observations per region and shallow trees (depth = 2, 30 estimators), the weak learners can only exploit the 1–2 strongest temporal features, leaving no capacity to evaluate additional inputs.

4.2.2 Per-Country Analysis

At the country level, the ESL-based models show the strongest improvements in markets with more volatile HICP food dynamics. Estonia (naïve RMSE = 0.0204) sees a 40.5% reduction with the pooled temporal model, Finland (naïve RMSE = 0.0110) sees a 35.7% reduction, and Spain sees a 33.3% reduction.

France presents an interesting case: SARIMA achieves the lowest RMSE (0.0034), outperforming all pooled ML variants (best: 0.0039). This suggests that for countries with stable, mean-reverting HICP dynamics, the autoregressive structure of monthly HICP itself is highly informative.

Estonia exhibits the highest absolute RMSE across all models, reflecting the Baltic region's more volatile HICP food dynamics. Despite this, the pooled ESL+macro model achieves a 40.5% improvement over the naïve baseline — the largest per-country improvement — suggesting that the ML framework is particularly valuable in volatile markets where simple baselines perform poorly.

4.2.3 Summary — RQ2

The MIDAS-XGBoost framework with daily U-MIDAS lags outperforms the naïve random walk baseline, achieving a 35.9% RMSE reduction in out-of-sample inflation nowcasting across 78 country-months. The improvement is robust across six Eurozone countries and holds for both pooled and bagged model architectures. An ablation analysis reveals that ESL-only and macro-only models achieve comparable accuracy (−35.9% and −35.2% vs. naïve, respectively). Combining ESL and macro features yields marginal additional gains (−35.8%). SARIMA, when properly fitted on continuous HICP series, achieves a 24.1% improvement — competitive but inferior to the ML approach in most countries. However, the extent to which this improvement reflects genuine directional nowcasting versus a better estimate of the unconditional mean cannot be fully resolved with the available sample size.

4.3 RQ3: Regional and Spatial Factors

4.3.1 Spatial Feature Set

To test whether regional and cross-country price dynamics carry additional predictive signal, the temporal models are augmented with 6 cross-country spatial features:

- **Cross-country Jevons average**: The mean Jevons link of all other countries in the same month (contemporaneous and lagged $t-1$).
- **Cross-country dispersion**: Standard deviation of Jevons links across other countries.
- **Own-country lagged Jevons**: The focal country's Jevons link at $t-1$ (autoregressive term).
- **Country price gap**: Difference between the focal country's Jevons link and the cross-country average (contemporaneous and lagged $t-1$).

These features capture two hypothesized mechanisms: *contagion* (inflation spreading from one country to another through trade and supply chains) and *convergence* (countries with higher price levels adjusting downward toward the cross-country mean). The spatial models are evaluated in pooled and bagged configurations, and additionally combined with macro features in a "full" specification.

4.3.2 Overall Results

Table 4 reports the out-of-sample RMSE for all model variants, isolating the effect of spatial augmentation.

Scope	Pooled temporal	Pooled +spatial	Pooled +full	Bagged temporal	Bagged +spatial
Pooled	0.0073	0.0075	0.0074	0.0076	0.0076
FR	0.0039	0.0040	0.0040	0.0038	0.0038
BE	0.0068	0.0069	0.0069	0.0063	0.0063
IT	0.0053	0.0054	0.0054	0.0053	0.0053
ES	0.0053	0.0055	0.0055	0.0059	0.0059
FI	0.0071	0.0073	0.0072	0.0074	0.0074
EE	0.0121	0.0125	0.0124	0.0130	0.0130

Table 4: Effect of spatial augmentation on out-of-sample RMSE (lower is better). **Temporal** uses the 31 ESL-derived features only; **+spatial** adds the 6 cross-country spatial features; **+full** adds both the spatial and the 6 lagged macro features. **Pooled** trains one model across all regions; **Bagged** trains one weak learner per region and averages. Bagged +spatial is identical to bagged temporal because the per-region learners assign no weight to the cross-country features.

Cross-country spatial features slightly degrade performance. The pooled spatial model (RMSE = 0.0075) performs worse than the temporal-only model (0.0073), a +2.7% increase that is not statistically significant (Diebold-Mariano $p = 0.23$) but consistent across all six countries. Adding spatial features introduces noise dimensions that the model overfits to in-sample.

Spatial features do not complement other feature sets. When spatial and macro features are combined (the "full" model, RMSE = 0.0074), performance remains slightly below the temporal-only baseline. This pattern indicates that cross-country Jevons links carry no incremental signal beyond what the temporal features already capture.

Bagged models do not use spatial features. The bagged temporal and bagged spatial models produce identical predictions (RMSE = 0.0076). However, this should not be interpreted as evidence against spatial features specifically — the bagged architecture's limited capacity (depth = 2, 14–20 observations per region) prevents it from exploiting *any* additional features beyond the dominant temporal lags, including macro features that the pooled model finds useful.

4.3.3 Per-Country Spatial Effects

At the country level, the spatial features consistently hurt or have no effect in the pooled specification:

- All six countries show equal or worse RMSE when spatial features are added. Estonia (+3.3%) and Finland (+2.8%) see the largest degradation, while Italy and Spain show smaller increases. No country benefits from spatial augmentation.
- No country shows consistent spatial improvement across both the pooled and bagged specifications.

4.3.4 Leave-One-Country-Out Robustness

Table 5 reports leave-one-country-out (LOCO) cross-validation results, where for each hold-out country a model is trained on the remaining five countries. This tests whether spatial features help when predicting a country the model has never seen.

Hold-out	Temporal	+Spatial	Macro-only	+Macro	+Full
FR	0.0032	**0.0032**	0.0032	0.0034	0.0034
BE	0.0060	0.0060	0.0059	**0.0059**	0.0060
IT	**0.0040**	0.0040	0.0041	0.0041	0.0041
ES	0.0050	0.0051	0.0050	**0.0050**	0.0050
FI	0.0065	0.0065	0.0065	**0.0064**	0.0065
EE	0.0109	0.0109	0.0109	**0.0109**	0.0109
Average	0.0060	0.0060	**0.0059**	**0.0059**	0.0060

Table 5: Leave-one-country-out RMSE by feature set. Temporal and macro-only features generalize equally well on average.

The LOCO results confirm the expanding-window findings. Spatial features degrade cross-country generalization in the majority of hold-out countries. The temporal, macro-only, and ESL+macro models achieve comparable average RMSE, while spatial augmentation increases average error.

Notably, the macro-only model generalizes well across countries, suggesting that lagged macro features capture transferable cross-country signal. The ESL+macro model shows mixed LOCO results, suggesting country-specific interactions between ESL and macro features.

4.3.5 Summary — RQ3

Regional and cross-country spatial factors do not improve inflation nowcasting accuracy. Adding cross-country spatial features (average Jevons link, price gaps, dispersion) slightly degrades pooled RMSE by +2.7%, consistent across all six countries. The bagged architecture produces identical predictions with and without spatial features, though this reflects the model's limited capacity rather than a definitive test of the spatial hypothesis. No country benefits from spatial augmentation in either the pooled or bagged evaluation.

5 Discussion

5.1 What the RMSE Improvement Actually Means

Returning to the overarching question posed in Section 1.3 — *can daily ESL pricing data, processed through a MIDAS-XGBoost framework, improve inflation nowcasting relative to traditional baselines, and does augmenting it with macroeconomic indicators yield additional gains?* — the answer to the first part is a qualified yes: the framework reduces RMSE by 35.9% over the naïve baseline and outperforms or matches SARIMA in five of six countries. The answer to the second part is no: macro augmentation adds no reliable gain over ESL data alone, and the two sources reach comparable accuracy independently. The practical value of ESL therefore lies less in raw accuracy than in timing — it delivers a signal of comparable quality weeks earlier than lagged macro indicators.

The RMSE improvement over the naïve baseline is real, but it requires careful interpretation. A large part of the gain may come from the model learning to predict a value closer to the unconditional mean of HICP food inflation rather than from tracking the direction of month-to-month changes. The naïve random walk repeats last month's value, so any model that shrinks predictions toward the sample mean will score a lower RMSE in a mean-reverting series, even without learning the actual dynamics.

This does not invalidate the result. Predicting a better constant is still useful in practice, and the scatter plot in Figure 6 shows that the model does achieve positive directional correlation ($r = 0.63$) between predicted and actual values. But it does mean that the 35.9% headline number should not be read as proof that the model has fully learned to nowcast inflation from ESL data. The truth is likely somewhere in between: the model captures some genuine signal from the daily price dynamics, but also benefits from mean-reversion in ways that are hard to disentangle with only 78 out-of-sample observations.

The U-MIDAS feature structure is designed to help with directional tracking. By feeding in each of the 22 daily Jevons log-returns as separate regressors rather than collapsing them into a monthly average, the model can in principle learn which parts of the month matter most, whether prices tend to move early and then stabilize, or if there is a late-month acceleration. The aggregate statistics (mean, standard deviation, range, skewness) act as a safety net for months where the number of trading days varies. Whether the model is actually exploiting this temporal structure or simply using the aggregate features to estimate the mean is an open question that would require a longer evaluation window to answer.

From a theoretical perspective, ESL technology connects to the menu cost literature introduced in Section 2.3. In the Calvo (Calvo, 1983) framework, firms reset prices with some fixed probability each period. Traditional shelf labels make price changes costly (Levy et al., 1997), which keeps that probability low. ESL technology removes this friction, so ESL-equipped stores adjust prices more freely, producing a more volatile price distribution. The Jevons Index reflects this: it is considerably more volatile month-over-month than official HICP. In a Calvo framework, this corresponds to a higher reset probability $(1 - \theta)$. The daily U-MIDAS lags can in principle observe the *within-month dynamics* of price resetting, but whether the model exploits this or defaults to a mean estimate depends on how much data is available relative to the number of features.

One result worth highlighting is France, where SARIMA beats the pooled ML models (RMSE 0.0034 vs. 0.0039 for pooled ESL-only), though the bagged model comes close at 0.0038. This is consistent with the mean-reversion interpretation: French food HICP has a stable, mean-reverting pattern, and SARIMA is specifically designed to model that structure. When the target is that predictable from its own history, there is less room for high-frequency data to add value. The ESL-based models show their largest gains in countries with more volatile dynamics like Estonia and Finland, where the naïve baseline is weaker and there is more room for improvement, but also more room for mean-reversion effects.

5.2 The Ablation Paradox

The ablation result is also consistent with this interpretation. ESL-only and macro-only models perform almost identically (DM $p = 0.28$), and combining them does not help. If both sources are primarily helping the model converge on a better central estimate, then adding more features just gives the model 37 parameters to estimate from a limited number of observations. The lagged HICP subindices and the ECB rate provide a macro-level anchor, while the ESL Jevons links provide a micro-level anchor from retail price movements. Both channels converge on a similar estimate, and the extra parameters add noise rather than precision.

The practical implication is still relevant regardless of this interpretation. ESL data is available in near real-time, within 24 hours of a price change. The macro features are lagged by a full month since contemporaneous HICP subindices are published with the same delay as the target itself. In a real deployment, the ESL signal would be the first one available each month, and the macro features could potentially refine the forecast once they are released. Even if both sources are primarily helping to estimate the mean, having the ESL estimate available weeks earlier is operationally valuable.

The bagged models tell a simpler story: all bagged variants produce the same predictions regardless of which features are included (RMSE 0.0076, about 33% below naïve). With only 16 to 20 training observations per region and shallow trees (depth 2, 30 estimators), the weak learners can only pick up the strongest temporal lags and ignore everything else. This is useful as a robustness check on the temporal baseline, but it cannot tell us anything about whether spatial or macro features carry signal.

5.3 Why Spatial Features Do Not Help

The spatial null result directly answers RQ3, and it is worth understanding why the hypothesis did not hold up. Cross-country Jevons averages, price gaps, and dispersion measures were designed to capture two mechanisms: contagion (inflation spreading across borders via trade and supply chains) and convergence (countries with higher price levels adjusting back toward the mean). Neither mechanism shows up in the data.

There are a few reasons for this. The Jevons Index computed from a limited set of ESL-equipped retailers is a noisy proxy for a country's overall price level. When you compute cross-country correlations in a noisy signal, the measurement error tends to dominate, and the model ends up fitting to spurious patterns that do not generalize out of sample. On top of that, the six countries in the sample represent quite different retail markets with different competitive structures and regulatory environments, so a common spatial relationship is unlikely to hold across all of them. And finally, with only about 16 to 20 out-of-sample months per country, there is simply not enough data to estimate cross-country dynamics reliably.

It should be noted that this study tests *cross-country* spatial relationships using ESL data with limited retailer coverage per country. This is a much harder setting than the within-country regional analysis that the spatial econometrics literature (LeSage & Pace, 2009) has shown to work well. In those studies, contiguous regions share similar economic structures and the spatial lags are computed from comprehensive price surveys. The cross-country setting here, with different retail markets, product mixes, and regulatory contexts, represents a much coarser test of the spatial hypothesis. In hindsight, it is not surprising that cross-country Jevons averages do not predict each other.

5.4 Shelf Prices, Transaction Prices, and What We Are Actually Measuring

All of the results above need to be interpreted in light of a basic measurement issue: ESL data captures *posted shelf prices*, not the prices consumers actually pay. Official HICP tries to measure transaction prices, including temporary promotions, loyalty card discounts, multi-buy offers, and coupons (Eurostat, 2024). The ESL system only sees the number on the digital tag before any of those discounts are applied.

This matters because promotional activity in European food retail is widespread. Loyalty card programs are operated by all the major retailers in this study, and a large share of transactions happen at promotional prices. So the Jevons Index likely overstates price levels relative to what consumers actually pay. Promotional cycles (weekly deals, end-of-month clearances) also inject systematic noise into the daily Jevons links that has nothing to do with actual inflation trends. There is a risk that the U-MIDAS features are partly picking up on these promotional patterns rather than genuine price movements.

During disinflationary periods, this gap could be especially problematic. Retailers might keep posted shelf prices stable while quietly increasing the depth or frequency of promotions, meaning the effective price falls but the ESL index does not register it.

The pooled correlation of $r = 0.43$ between the Jevons MoM and the HICP food subindex is consistent with this picture: the ESL data tracks the general direction of price movements, but diverges from the actual price level because of the promotional layer it cannot see. The higher volatility of the Jevons Index relative to HICP, visible in Figure 4, likely reflects a mix of genuine high-frequency price dynamics and promotional noise.

Despite all this, the model still achieves a 35.9% RMSE reduction. The posted shelf price is clearly an imperfect proxy, but it still carries enough information about the direction and timing of inflation movements to be useful. XGBoost may be learning to filter out some of the promotional noise through the ensemble of daily lags and aggregate statistics, extracting a trend component that tracks official inflation.

5.5 Coverage Bias and Representativeness

ESL adoption is concentrated among large, technology-oriented retail chains. The data therefore reflects pricing behavior at *large format supermarkets* in mostly urban and suburban areas. Small independent retailers, traditional markets, and hard-discount chains are largely absent.

This is not a minor gap. Hard-discount chains like Aldi and Lidl have become major players across Western Europe, and their pricing strategies (everyday low prices, minimal promotions) look very different from traditional supermarkets. In Southern European countries like Italy, independent grocers and traditional markets still account for a meaningful share of food retail. These excluded segments may behave quite differently from what the ESL data shows, especially during supply shocks or competitive price wars when large chains and discounters may move in different directions.

The $r = 0.43$ correlation with HICP food is consistent with this partial coverage. The ESL signal captures a meaningful share of retail price movements, but not all of them. Improving representativeness would require incorporating data from multiple retailers per country, including discounters and smaller chains, which is a natural direction for future work.

5.6 The Inflation Regime Problem

The study period (May 2023 to December 2024) is not a normal period for inflation. It covers the tail end of the post-pandemic price surge and the subsequent disinflation as supply chains normalized and the ECB tightened monetary policy (European Parliament, 2023). Food HICP in the Eurozone went from elevated year-over-year rates in early 2023 to near-zero by late 2024.

This raises a real question about whether the results would hold in a different environment. The 35.9% RMSE improvement may partly reflect the fact that the HICP food subindex was on a predictable downward trajectory during this period, which the daily lag structure can capture by detecting the gradual deceleration of price increases within each month. In a stable low-inflation regime (say 2015 to 2019), daily Jevons links would show less variance and a lower signal-to-noise ratio, which could reduce the ML model's edge over simpler baselines.

The "rockets and feathers" asymmetry (Peltzman, 2000) adds another layer of complexity. During the inflationary phase, retailers were actively raising prices, generating frequent and directionally aligned price changes in the ESL logs. During disinflation, prices fall more slowly than they rose. Since the ESL data only captures items that *change* price, this asymmetry shows up as a reduction in the volume of logged events rather than as explicit price decreases for stable items. This could explain why the Jevons Index stays more volatile than HICP even as official inflation declines: the index is being computed from a shrinking and increasingly non-representative sample of items that are still being repriced.

Testing whether the model generalizes across inflation regimes would require a much longer time series spanning inflationary, disinflationary, and stable periods. The roughly 20 months available per country are not enough for that.

5.7 Relation to Prior Work

The 36% RMSE reduction over the naïve baseline is broadly in line with what Cavallo and Rigobon (Cavallo & Rigobon, 2016) showed for online price indices, though their data achieved correlations above 0.90 with official CPI. Our ESL-based index shows a lower correlation ($r = 0.43$) due to narrower product coverage and the selection bias inherent in change-logging systems. But even with this weaker raw correlation, the daily frequency and U-MIDAS feature engineering allow the model to extract more predictive signal than a simple correlation would suggest.

The choice of XGBoost is consistent with Lundgren and Wicktor (Lundgren & Wicktor, 2023), who found it to be the best performer for CPI nowcasting with mixed-frequency data. The improvement magnitude (36%) is also in line with Medeiros et al. (Medeiros et al., 2021), who reported similar gains for Random Forest models on U.S. inflation data.

The spatial null result contrasts with the spatial econometrics literature, where spatial autoregressive models have improved regional forecasts (LeSage & Pace, 2009). But the settings are quite different: our spatial features are cross-country averages from limited ESL coverage, not within-region lags from comprehensive surveys. Regional spatial lags within a country could not be tested here because all regions share the same country-level HICP target.

5.8 Limitations

Shelf price vs. transaction price. The ESL data captures posted shelf prices, not effective transaction prices after loyalty discounts, promotions, and coupons. Official HICP aims to measure transaction prices. This means the Jevons Index is better understood as a proxy for the *direction* of price movements rather than a precise measure of consumer-facing inflation.

Coverage bias. ESL-equipped stores are predominantly large supermarkets. Small independents, traditional markets, and hard-discount retailers are absent from the data. Their pricing dynamics likely differ from those of the observed stores, and this limits how well the Jevons Index represents overall retail price movements.

Pricing strategy versus inflation. Not every logged price change reflects an underlying inflationary movement. Retailer-specific strategies — competitive repositioning, category-management decisions, dynamic or zone-based pricing, and promotional scheduling — generate price changes that are unrelated to aggregate cost pressure. The Jevons index cannot separate these strategy-driven changes from inflation-driven ones, which adds noise to the signal and is a further reason the index diverges from official HICP beyond the shelf-versus-transaction-price gap discussed in Section 5.4.

Short time series. The evaluation window provides only about 16 to 20 out-of-sample months per country. This limits the power of statistical tests (all DM tests have $p > 0.05$) and contributes to the ablation paradox where adding features hurts rather than helps, because 37 features cannot be reliably estimated from so few observations. The study period also coincides with a specific inflation regime (post-pandemic disinflation), and performance may not generalize to other macroeconomic conditions.

Asymmetric price adjustment. The "rockets and feathers" pattern means the model may perform differently in inflationary vs. disinflationary phases. During disinflation, ESL logs are dominated by items repriced less aggressively while stable items remain invisible. This regime-dependence cannot be assessed with the current data.

Coarse spatial granularity. The spatial features capture cross-country dynamics, which is a rough level of geographic aggregation. Within-country regional spatial lags could not be tested because all regions in a country share the same HICP target. Access to regional price indices (e.g. NUTS-2 level HICP) would enable a finer-grained spatial analysis.

Limited macro features. Only lagged (t−1) macroeconomic indicators are used, since contemporaneous HICP subindices are published with the same delay as the target. This limits the macro feature set to 6 variables. The ECB refinancing rate is entered as a step function approximation rather than from official daily records.

Pseudo-replication. Regional ESL data is matched against country-level HICP targets, so the region-month observations contain a much smaller number of truly independent target observations (country-months). This inflates the apparent training set size and the bagged architecture only partially addresses it.

5.9 Ethical Considerations

This study uses anonymized and aggregated pricing data from an ESL cloud platform. No personal consumer data, transaction records, or individually identifiable information is processed. Only region-level aggregations are used, and retailer identities are anonymized, consistent with the GDPR and contractual confidentiality obligations described in Section 3.1.

The macroeconomic features are sourced from publicly available Eurostat and ECB databases. From a societal perspective, better inflation nowcasting tools could benefit policymakers through faster monetary policy responses and consumers through greater price transparency. At the same time, if deployed commercially, such tools could create information asymmetries — for instance, enabling large retailers to anticipate competitors' pricing strategies — which raises concerns about market transparency and tacit coordination that would need to be weighed before any commercial deployment. The models developed here are research prototypes and are not intended for direct commercial use without further validation.

6 Conclusion

This thesis investigates whether Electronic Shelf Label pricing data can be used for short-term inflation nowcasting in the Eurozone.

On RQ1, the ESL-derived Jevons Index shows statistically significant co-movement with the HICP food subindex ($r = 0.43, p < 0.01$) across 114 country-months in six Eurozone countries, establishing that ESL pricing data captures meaningful retail food price dynamics.

On RQ2, a MIDAS-XGBoost model using 31 daily ESL-derived features reduces RMSE by 35.9% relative to the naïve random walk across 78 country-months, matching or outperforming SARIMA in five of six countries (with France as the exception). However, the extent to which this improvement reflects genuine directional nowcasting versus a better estimate of the unconditional mean cannot be fully resolved with the available sample size. An ablation analysis shows that ESL-only and macro-only models achieve comparable accuracy (35.9% and 35.2% RMSE reduction), meaning ESL data provides value on par with lagged macroeconomic indicators while being available in near real-time.

On RQ3, cross-country spatial features slightly degrade nowcasting accuracy (+2.7% pooled RMSE), consistently across all six countries. Leave-one-country-out cross-validation confirms this. The result likely reflects the difficulty of extracting spatial signal from noisy retailer-level price indices across heterogeneous markets rather than a general rejection of spatial price dynamics.

The validated comovement between the Jevons Index and HICP food inflation is an encouraging foundation. With a longer time series, broader retailer coverage, and more countries, the MIDAS-XGBoost framework is positioned to move beyond mean-reversion and capture directional dynamics directly; the binding constraint in this study is data quantity, not signal quality.

These findings carry implications for two audiences. For policymakers, a validated real-time food-price signal offers central banks and statistical offices an early read on inflation turning points weeks ahead of official releases, with clear value for the timing of monetary-policy decisions. For retail management, the same infrastructure that produces ESL logs could support internal price-monitoring and competitive benchmarking — though, as noted in the ethical discussion, a tool that lets large chains anticipate competitors' pricing also raises concerns about market transparency that would need to be weighed before any commercial deployment.

Future work should focus on extending the evaluation window across multiple inflation regimes, incorporating additional retailers per country to improve representativeness, exploring within-country regional spatial lags where regional price indices are available, and evaluating alternative ML architectures that may better capture the sequential structure of daily price data.

Bibliography

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers.

Atkeson, A., & Ohanian, L. E. (2001). Are Phillips curves useful for forecasting inflation?. *Quarterly Review*, *25*(Winter), 2–11.

Białek, J., & Beręsewicz, M. (2021). Scanner data in inflation measurement: from raw data to price indices. *Statistical Journal of the IAOS*, *37*(2), 547–571. https://doi.org/10.3233/SJI-210816

Calvo, G. A. (1983). Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics*, *12*(3), 383–398. https://doi.org/10.1016/0304-3932(83)90060-0

Cavallo, A. (2017). Are Online and Offline Prices Similar? Evidence from Large Multi-channel Retailers. *American Economic Review*, *107*(1), 283–303. https://doi.org/10.1257/aer.20160542

Cavallo, A., & Rigobon, R. (2016). The Billion Prices Project: Using Online Prices for Measurement and Research. *Journal of Economic Perspectives*, *30*(2), 151–178. https://doi.org/10.1257/jep.30.2.151

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599

European Parliament. (2023). *Inflation in the euro area* [Technical report].

Eurostat. (2024). *HICP Methodology* [Technical report]. https://ec.europa.eu/eurostat/web/hicp/methodology

Garcia, M. G. P., Medeiros, M. C., & Vasconcelos, G. F. R. (2017). Real-time inflation forecasting with high-dimensional models: The case of Brazil. *International Journal of Forecasting*, *33*(3), 679–693. https://doi.org/10.1016/j.ijforecast.2017.02.002

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2004). *The MIDAS Touch: Mixed Data Sampling Regression Models* [Technical report].

Hotelling, H. (1929). Stability in Competition. *The Economic Journal*, *39*(153), 41–57.

LeSage, J., & Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Chapman, Hall/CRC. https://doi.org/10.1201/9781420064254

Levy, D., Bergen, M., Dutta, S., & Venable, R. (1997). The Magnitude of Menu Costs: Direct Evidence from Large U.S. Supermarket Chains. *The Quarterly Journal of Economics*, *112*(3), 791–825.

Lundgren, G., & Wicktor, N. (2023). *Nowcasting U.S. Inflation Using Mixed Frequency Real-Time Data* [Master's thesis]. http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=9121796&fileOId=9121804

Medeiros, M. C., Vasconcelos, G. F. R., Veiga, Á., & Zilberman, E. (2021). Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods. *Journal of Business & Economic Statistics*, *39*(1), 98–119. https://doi.org/10.1080/07350015.2019.1637745

Peltzman, S. (2000). Prices Rise Faster than They Fall. *Journal of Political Economy*, *108*(3), 466–502. https://doi.org/10.1086/262126

Stock, J. H., & Watson, M. W. (2007). Why Has U.S. Inflation Become Harder to Forecast?. *Journal of Money, Credit and Banking*, *39*(s1), 3–33. https://doi.org/10.1111/j.1538-4616.2007.00014.x

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, *46*(Supplement), 234–240. https://doi.org/10.2307/143141

7 Appendix

Study	Data Strategy	Core Algorithm	Temporal Approach	Key Finding
Cavallo & Rigobon (2016) (Cavallo & Rigobon, 2016)	Scraped daily online prices from multi-channel retailers.	Jevons Index: unweighted geometric mean of price relatives.	Daily chained index from a base day.	Online micro-data can anticipate CPI turning points by weeks.
Garcia et al. (2017) (Garcia et al., 2017)	90+ macro variables (emerging market context).	Random Forest & LASSO benchmarked against AR.	Rolling-window real-time evaluation.	LASSO/RF systematically beat traditional linear benchmarks.
Medeiros et al. (2021) (Medeiros et al., 2021)	150–200 predictors for US inflation.	Random Forest; also tested Ridge, LASSO, Boosting.	Direct strategy: separate model per horizon h.	Random Forest dominates by capturing non-linearities.
Lundgren & Wicktor (2023) (Lundgren & Wicktor, 2023)	Monthly macro + daily financial variables.	XGBoost compared to U-MIDAS and RF.	Daily nowcasting with real-time updates.	XGBoost is top performer for CPI turning points.

Table 6: Summary of key related studies on inflation forecasting with alternative data and ML methods.