

Working Paper 2026.2.4.12

- Vol. 2, No. 4

ỨNG DỤNG KHOA HỌC DỮ LIỆU DỰ BÁO LẠM PHÁT THỜI GIAN THỰC SỬ DỤNG DỮ LIỆU TỪ GOOGLE TRENDS

Tạ Minh Nhật[1], Trịnh Minh Nghĩa, Trần Văn Nhân

Sinh viên K63 Tiêu chuẩn Tài chính quốc tế – Khoa QTKD & Tài chính – kế toán

Phân hiệu Trường Đại học Ngoại thương, TP.HCM, Việt Nam

Nguyễn Nam Cường

Sinh viên K63 Tiêu chuẩn Kế toán – Kiểm toán – Khoa QTKD & Tài chính – kế toán

Phân hiệu Trường Đại học Ngoại thương, TP.HCM, Việt Nam

Ngô Tuấn Anh

Sinh viên K24 Khoa học Dữ liệu – Khoa Khoa học Ứng dụng

Trường Đại học Bách khoa - ĐHQG TP.HCM, TP.HCM, Việt Nam

Đặng Lê Quang

Giảng viên Bộ môn khoa học cơ bản

Phân hiệu Trường Đại học Ngoại thương, TPHCM, Việt Nam

Tóm tắt

[1] Tạ Minh Nhật, Email: k63.2411315010@ftu.edu.vn

Nghiên cứu này nhằm mục đích xây dựng, so sánh và lựa chọn mô hình học máy tối ưu cho việc dự báo lạm phát theo thời gian thực tại Việt Nam trong giai đoạn 2014–2024, đồng thời thử nghiệm tích hợp dữ liệu hành vi tìm kiếm trực tuyến trên Google Trends để đánh giá giá trị dự báo gia tăng của nó. Để đạt được điều này, nghiên cứu đánh giá các thuật toán học máy tiên tiến - bao gồm các công cụ ước lượng co rút tuyến tính như Ridge, LASSO, Elastic Net và các phương pháp dựa trên cây phi tuyến tính như Random Forest, XGBoost - so với các mô hình kinh tế lượng truyền thống. Một bộ từ khóa Google Trends được địa phương hóa, bám sát rổ Chỉ số Giá tiêu dùng của Việt Nam, đã được chọn lọc và tích hợp thông qua phương pháp U-MIDAS. Kết quả thực nghiệm từ khung kiểm định ngược đệ quy giả thời gian thực cho thấy các mô hình học máy dựa trên phương pháp co rút, đặc biệt là hồi quy Ridge, có hiệu suất vượt trội đáng kể so với các mô hình cơ sở truyền thống. Tuy nhiên, việc tích hợp dữ liệu Google Trends không mang lại sự cải thiện ổn định về độ chính xác dự báo so với các mô hình chỉ sử dụng dữ liệu vĩ mô, và các mô hình lai cũng không thể vượt qua hiệu suất của mô hình Ridge cơ sở. Cuối cùng, nghiên cứu chứng minh rằng việc áp dụng học máy có điều chuẩn vào dữ liệu kinh tế vĩ mô có thể làm giảm độ trễ công bố thông tin lên đến 75%, qua đó cung cấp một hệ thống cảnh báo sớm vững chắc nhằm hỗ trợ Ngân hàng Nhà nước Việt Nam, các doanh nghiệp và nhà đầu tư trong việc chủ động quản lý rủi ro kinh tế vĩ mô.

Từ khoá: dự báo thời gian thực, lạm phát, học máy, Google Trends, Việt Nam, hồi quy Ridge.

DATA SCIENCE APPLICATION PREDICTS REAL-TIME INFLATION USING DATA FROM GOOGLE TRENDS

Abstract

This study aims to construct, compare, and select the optimal machine learning model for real-time inflation forecasting (nowcasting) in Vietnam over the 2014–2024 period, while experimentally integrating online search behavior data (Google Trends) to assess its incremental predictive value. To achieve this, the research evaluates advanced machine learning algorithms - including linear shrinkage estimators (Ridge, LASSO, Elastic Net) and non-linear tree-based methods (Random Forest, XGBoost) - against traditional econometric models. A localized set of Google Trends keywords aligned with the Vietnamese Consumer Price Index (CPI) basket was curated and integrated using the U-MIDAS approach. Empirical results from a recursive pseudo-real-time backtesting framework reveal that shrinkage-based machine learning models, particularly Ridge

regression, significantly outperform traditional benchmark models. However, the integration of Google Trends data did not add a stable improvement in predictive accuracy over macro-only specifications, and hybrid models failed to improve upon the baseline Ridge model. Ultimately, the study demonstrates that applying regularized machine learning to macroeconomic data can reduce the information publication lag by up to 75%, providing a robust early warning system to support the State Bank of Vietnam, enterprises, and investors in proactive macroeconomic risk management.

Keywords: nowcasting, inflation, machine learning, Google Trends, Vietnam, Ridge regression

1. Introduction

In modern macroeconomics, inflation is not merely a price index but a core variable guiding monetary policy decisions. However, policy effectiveness is often constrained by data limitations and publication lags. The study by Heng et al. (2024), conducted in the context of Cambodia, indicates that despite increasing digitalization, developing economies still face challenges in accessing macroeconomic data, particularly regarding timeliness and publication delays. In Vietnam, inflation plays a central role in monetary policymaking, as the State Bank prioritizes price stability (Epstein et al., 2022), while official CPI data are released monthly on the 6th of the following month (GSO, 2024), creating delays that may hinder timely policy responses. Traditional econometric models such as ARIMA and VAR, though fundamental for time-series forecasting, have inherent limitations: ARIMA is predominantly linear and data-intensive (Khashei & Bijari, 2010), while VAR suffers from dimensionality and lag-selection issues as the number of variables increases (Nicholson, Matteson, & Bien, 2020), and struggles to capture complex nonlinear dynamics (Chan, 2022).

Previous studies also highlight the issue of data delays. Nguyen, Le, Hoang, Tran, and Giap (2024) show that export-related analysis in Vietnam is constrained by a "significant lag" between data generation and official release, limiting the responsiveness of conventional models. Ngo (2021) further demonstrates that traditional inflation measurement faces difficulties due to delayed price data collection, especially during the COVID-19 period, emphasizing the need for timely and accurate data for decision-making. Although the study proposes online price data as an alternative source, it focuses on price monitoring rather than real-time forecasting. Thus, a research gap remains in applying data science methods to nowcast and forecast inflation in real time in Vietnam, particularly using Google Trends data.

In the digital era, improving forecasting quality requires integrating nowcasting approaches with alternative data sources. From an information economics perspective, search behavior on Google Trends reflects public attention and inflation expectations, providing early signals before official CPI releases. Moreover, machine learning algorithms such as Ridge, LASSO, and Random Forest can extract signals from high-dimensional data, reducing overfitting and improving forecast accuracy compared to traditional linear models. However, existing empirical evidence is largely concentrated in developed economies rather than developing ones (M. Narita & Yin, 2018). In Vietnam, macroeconomic forecasting studies still rely heavily on lagged data, while the application of Google Trends remains mostly confined to financial and stock markets rather than macroeconomic indicators such as inflation (Bui & Nguyen, 2019; Dang et al., 2024; Nguyen & Pham, 2018), revealing a clear research gap in ML-based inflation forecasting. In increasingly complex macroeconomic environments, traditional econometric models often struggle to capture nonlinear and evolving relationships, whereas machine learning provides a flexible framework capable of identifying hidden nonlinear patterns and adapting to structural changes, making it a promising direction for improving inflation forecasting and real-time economic analysis (Goulet Coulombe et al., 2022). To address this gap, the study develops a multi-tier empirical framework to nowcast CPI in Vietnam over the period 2014–2024, comparing the performance of machine learning models with traditional benchmarks such as VAR and evaluating accuracy improvements from integrating localized Google Trends keywords. The findings are expected to lay the groundwork for an early warning system to support macroeconomic policymaking.

2. Theoretical framework

2.1. The role of Google Trends data and Machine Learning algorithms in the real-time forecasting (Nowcasting) mechanism

In macroeconomic management, the "publication lag" of indicators such as the Consumer Price Index (CPI) creates a significant information gap, often causing regulatory decisions to fall into a delayed response state. To address this issue, real-time forecasting (Nowcasting) methods are applied to estimate macroeconomic variables as they occur. The empirical research framework of this study is grounded in Information Economics (Stigler, 1961) and the theory of Rational Inattention (Sims, 2003). Under this perspective, individuals' information-seeking behavior is not random; rather, it constitutes a purposeful response to

economic risk. When inflationary pressure intensifies, consumers tend to reallocate their limited attention toward information related to living costs and household purchasing power. Building on this foundation, the study develops an “Attention-to-CPI” mechanism, through which Google Trends data are employed as a proxy reflecting public attention and, to some extent, inflation expectations in real time. Accordingly, the selection of localized keywords such as “gasoline prices,” “rice prices,” “pork prices,” and “inflation” is not arbitrary, but deliberately designed on the basis of Vietnam’s CPI basket structure and the search behavior of domestic consumers, with the aim of capturing early signals of price pressure before official CPI figures are released. As inflation rises, consumers tend to actively allocate their attention to searching for price-related information to manage household purchasing power risks. Therefore, internet search volume serves as a proxy variable reflecting inflation expectations, creating an "Attention-to-CPI" mechanism that provides an early leading indicator for the economy (Bracha & Tang, 2022).

Furthermore, to process high-frequency behavioral data (weekly/daily) and high-dimensional feature spaces, Machine Learning (ML) algorithms act as core tools. Unlike traditional econometric models, ML possesses the ability to automatically select features, avoid multicollinearity, and capture complex nonlinear relationships, thereby optimizing the extraction of signals from Google Trends to deliver the most accurate forecasts.

3. Research methods

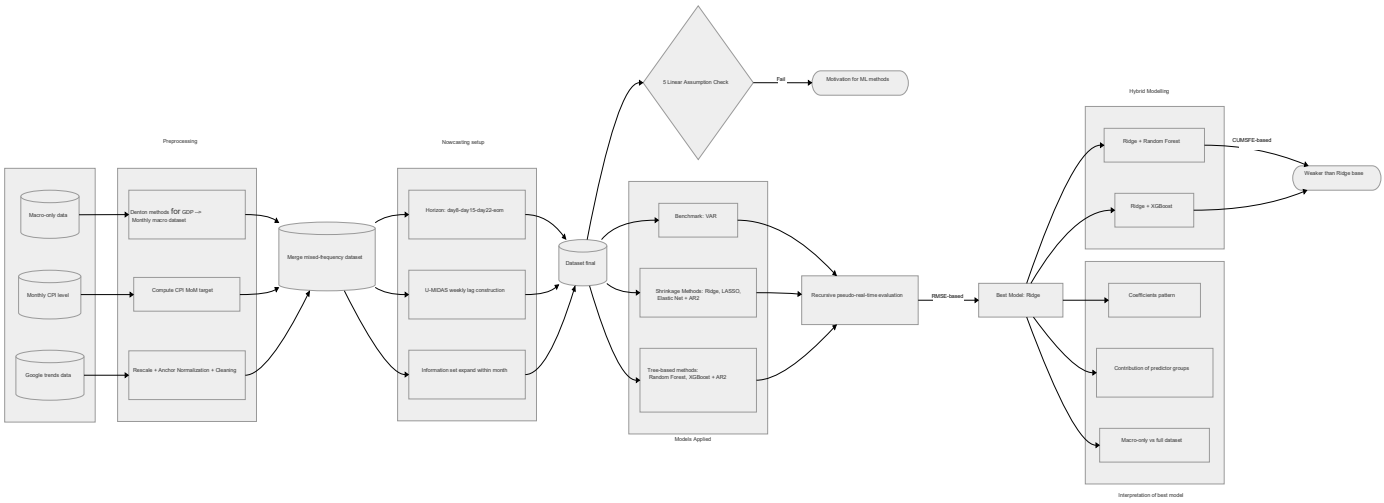


Figure 1: The empirical procedure for mixed-frequency data integration and machine learning model selection.

Table 1: A conceptual synthesis of the nowcasting framework components.

Component	Theoretical role	Data characteristics	Function in the nowcasting framework	Expected contribution
Macroeconomic variables	Fundamental economic drivers (demand, supply, monetary, external).	Low-frequency (monthly/quarterly), stable but lagged.	Core structural baseline for inflation dynamics.	Forecast stability, interpretability, and capturing persistent trends.
Google Trends search data	Proxy for public attention and inflation expectations.	High-frequency (weekly), timely, potentially noisy.	Complementary real-time signals preceding official data.	Enhance timeliness and early warning capacity at early horizons.
Mixed-frequency integration layer	Temporally align heterogeneous data sources.	Mixed schedules (weekly, monthly, quarterly).	Align frequencies via Denton-Cholette & U-MIDAS; standardize inputs.	Resolve frequency mismatch and build a unified predictor matrix.
Machine learning models	Signal extraction from high-dimensional, correlated spaces.	Robust to multicollinearity and nonlinearities.	Map integrated inputs to monthly CPI nowcasts.	Superior out-of-sample predictive accuracy vs. traditional models.

Component	Theoretical role	Data characteristics	Function in the nowcasting framework	Expected contribution
Recursive pseudo-real-time evaluation	Mimic actual policymaking; prevent look-ahead bias.	Time-ordered by horizon (Days 8, 15, 22, EOM).	Test model reliability under realistic data availability constraints.	Ensure empirical robustness and practical policy relevance.
Nowcasted inflation output	Final operational target.	Monthly target variable (delayed publication).	System endpoint integrating macro, search, and ML components.	Support proactive early warning and macroeconomic decision-making.

3.1. Sample data

In terms of data processing, due to the heterogeneous nature of the data sources involving different frequencies and scales, a rigorous preprocessing pipeline was implemented. First, to address the frequency mismatch between quarterly GDP and other monthly macroeconomic variables, we applied the Denton-Cholette method (Dagum & Cholette, 2006). This time-series disaggregation technique converts the quarterly GDP series into monthly data by minimizing the discrepancy between the generated monthly series and the original quarterly data. Second, all input variables were normalized using the StandardScaler technique. Because the macroeconomic and Google Trends variables have vastly different scales and units, this procedure transforms each data series to have a mean of 0 and a standard deviation of 1. This step is mandatory to prevent weight bias in machine learning algorithms, particularly regularization models like LASSO or Ridge (Hastie et al., 2009). Ultimately, our final dataset incorporates 16 independent search keywords alongside the aforementioned macroeconomic variables, yielding a total of 3,240 observations for the model estimation.

3.2. Model specification and evaluation methodology

3.2.1. The U-MIDAS Approach (Handling Mixed-Frequency Data)

In the context of real-time forecasting (nowcasting), one of the major challenges in incorporating behavioral data lies in the discrepancy in data frequency. Specifically, Google Trends search data are updated continuously on a weekly basis, whereas the target variable, the Consumer Price Index (CPI), is released only monthly. Directly aggregating weekly observations into a single monthly statistic would obscure week-to-week fluctuations, which may contain highly valuable information for early warning systems. Therefore, this study adopts the U-MIDAS approach to transform weekly signals into separate lagged features, thereby preserving the temporal structure of the information flow.

The high-frequency information block of indicator j at forecast origin t for forecast horizon h is represented by the following lag vector:

$$z_{j,t}^{(h)} = \left(z_{j,t,0}^{(h)}, z_{j,t,1}^{(h)}, \dots, z_{j,t,K-1}^{(h)}\right)$$

$z_{j,t}^{(h)}$: Denotes the vector summarizing the high-frequency search signals (Google Trends) of indicator/keyword *j*available at time t for forecast horizon h.

K: Denotes the number of weekly lags included in the model. In this study, $K = 8$, meaning that each keyword series is expanded into eight weekly lags.

$z_{j,t,0}^{(h)}$: Denotes the most recently available weekly signal at forecast horizon h.

$z_{j,t,1}^{(h)}, \dots, z_{j,t,K-1}^{(h)}$: Denote the progressively more distant weekly search observations from the recent past.

To address the mixed-frequency challenge, the U-MIDAS approach is employed to expand weekly Google Trends signals into distinct lagged variables, thereby preserving their temporal ordering and variability. Unlike traditional MIDAS frameworks that impose parametric restrictions on lag weights, U-MIDAS flexibly estimates a separate coefficient for each lag. This unrestricted approach is particularly advantageous for highly volatile behavioral data, enabling the model to precisely capture early inflationary signals. In this study, U-MIDAS functions primarily as a feature engineering step, mapping weekly search series into a unified input matrix alongside monthly macroeconomic indicators for subsequent machine learning algorithms. Furthermore, by strictly adhering to a pseudo-real-time setting—utilizing only data genuinely available at each forecast origin—this procedure not only retains critical time-flow information but also ensures a highly realistic and robust evaluation of the nowcasting system's practical performance.

3.2.2. Machine Learning Models Integrating Google Trend

To handle the high-dimensional predictor space comprising macroeconomic indicators and multiple weekly lags of Google Trends, this study employs Ridge Regression as the core machine learning model. This choice is motivated by the characteristics of the dataset: a large number of input variables, strong correlations among predictors, and a high level of noise arising from mixed-frequency data. Compared with nonlinear models such as Random Forest or XGBoost, Ridge is more suitable when the objective is to maintain forecast stability and to distribute information across many closely related variables.

The general form of the Ridge Regression model is defined by the following optimization problem:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \sum_{t=1}^{n} (y_t - \beta_0 - \mathbf{w}_t^{\top} \boldsymbol{\beta})^2 + \lambda \mathcal{P}(\boldsymbol{\beta}) \right\},$$

Where $\hat{\beta}$ is the estimated coefficient vector, $\mathbf{w}_t^{\top}$ is the input vector at time t, n is the number of observations, λ is the regularization parameter; and $\mathcal{P}(\boldsymbol{\beta})$ is the penalty function imposed on the regression coefficients.

In this study, the model input matrix includes: (i) monthly macroeconomic variables, (ii) lagged values of CPI, and (iii) weekly Google Trends features that have been transformed through the U-MIDAS approach into separate lagged variables. This allows the model to jointly incorporate traditional macroeconomic information and highly timely digital behavioral signals. Before training, the variables are standardized to ensure scale consistency and to improve the effectiveness of the regularization mechanism.

In summary, the machine learning model integrating Google Trends in this study is constructed in two steps. First, weekly search data are transformed into lagged features using U-MIDAS. Second, these features are combined with monthly macroeconomic variables in a Ridge Regression model. This approach makes it possible to exploit real-time signals from Google Trends while maintaining statistical stability in a high-dimensional data environment.

3.2.2.1.Ridge Scenario with Google Trends Integration

To quantify consumer inflation expectations, this scenario integrates weekly online search behavior data from Google Trends as independent variables. To mitigate spurious correlations and data noise often associated with automated web scraping (Lazer et al., 2014), a rigorous three-step manual keyword selection process was implemented.

First, the search scope was strictly aligned with the 11 main categories of the GSO's CPI basket to ensure a robust theoretical foundation reflecting actual consumption weights. Second, highly representative keywords were manually filtered based on their CPI weights and common Vietnamese search behaviors (e.g., "giá thịt lợn" for food, or "giá xăng" for transport). Finally, the search volumes were extracted using consistent parameters across the 11-year observation period (Location: Vietnam, Category: All, Type: Web Search), adhering to standard macroeconomic forecasting protocols (Choi & Varian, 2012).

Table 2: List of selected Google Trends variables

Category (Based on CPI basket / Macroeconomics)	**Keywork**	**Symbol**	**Economic Significance / Rationale for Selection**
Transport & Energy	"giá xăng"	GGT_XANG	Reflects consumers' direct concern regarding fuel costs. This highly volatile commodity group has a rapid and direct impact on transportation costs, triggering cost-push inflation.
	"đổ xăng"	GGT_DOXANG	
	"giá xăng hôm nay"	GGT_XANGHN	
	"giá dầu"	GGT_DAU	
Housing, electricity, water, fuel, and construction materials	"tiền điện"	GGT_DIEN	Represents essential monthly household living expenses. An increase in search volume reflects the pressure of living costs and expectations of price hikes for state-regulated commodities.
	"giá gas"	GGT_GAS	
	"giá gạo"	GGT_GAO	Food and foodstuffs constitute the largest weight in Vietnam's CPI basket. The search

Food and catering services	"giá heo hơi"	GGT_HEOHOI	volume for these keywords serves as a sensitive measure of the public's expectations regarding basic meal costs.
	"giá thịt lợn"	GGT_THITLON	
Gold & Foreign Currency	"giá vàng"	GGT_VANG	Although not directly calculated in core CPI, gold and foreign currencies act as safe-haven assets. High search frequencies reflect public sentiment concerning inflation risks and the depreciation of the VND (inflation expectations).
	"giá vàng sjc"	GGT_VANGSJC	
	"giá vàng hôm nay"	GGT_VANGHN	
	"tỷ giá"	GGT_TYGIA	
	"tỷ giá vietcombank"	GGT_TYGIAVCB	
Monetary Policy & Macroeconomics	"lãi suất"	GGT_LAISUAT	Directly reflects public interest in the macroeconomic environment and the State Bank of Vietnam's (SBV) tightening or easing monetary policies. This acts as the most direct proxy for general public sentiment and inflation expectations.
	"lạm phát"	GGT_LAMPHAT	

3.2.2.2.Ridge Macro-only Scenario (only use macroeconomic data)

To establish a rigorous benchmark for quantifying the actual information gain derived from online search data, this study specifies a forecasting scenario based purely on fundamental macroeconomic factors. Under this setting, the high-frequency Google Trends variable set is entirely omitted. The Ridge model is calibrated using only a suite of macroeconomic indicators—reflecting the broader economic climate—and autoregressive lags of inflation to capture inherent price persistence

Table 3: Selected macroeconomic variables

Variable	**Symbol**	**Description and Rationale**
Inflation Lags	*y_{t-1}, y_{t-2}*	1-month and 2-month lags of the CPI growth rate, incorporated to capture the persistence and inertia of inflation..
Industrial Production Index	*IIP*	**Data source**: FiinPro. IIP growth narrows the output gap, creating upward price pressure (Giannone et al., 2008). Since IIP input shocks often precede CPI changes (Gokmenoglu et al., 2015), nowcasting IIP improves real-time inflation forecasting accuracy (Furukawa et al., 2024).
Consumer Price Index	*CPI*	**Data source:** General Statistics Office of Vietnam (GSO). As the foundational metric for inflation, CPI reflects living costs and purchasing power erosion (Mankiw, 2022). It serves as a core forecasting variable and an "early warning system" for Central Banks to manage economic stability (Mchukwa et al., 2025).
Credit Growth	TANG_TRUONG_TIN_DUNG	**Data source:** Extracted from State Bank of Vietnam (SBV) reports.

		Credit growth drives inflation via demand-side stimulus. Rapid expansion increases liquidity, bolstering consumption and investment, shifting Aggregate Demand (AD) rightward. When Aggregate Supply (AS) faces capacity constraints, this excess demand elevates the general price level (Crowley, 2008).
M2 Money Supply	*M2*	**Data source:** Vietstock Rooted in the Quantity Theory of Money and Friedman's (1970) assertion that inflation is a "monetary phenomenon," M2 is a pivotal forecasting variable. Based on the Equation of Exchange ($M \times V = P \times Y$), the price level (P) rises when money supply growth (M) exceeds real output (Y). Empirical studies confirm M2 expansion as a primary inflation driver in Vietnam, with monetary shocks exerting persistent, long-term effects on prices (Nguyen et al., 2022).
Foreign Direct Investment	FDI_DK; FDI_GIAINGAN	**Data source:** General Statistics Office of Vietnam (GSO) FDI impacts inflation twofold. Short-term inflows stimulate investment, shifting Aggregate Demand (AD) rightward and triggering demand-pull inflation. Long-term, FDI enhances productivity and expands Aggregate Supply (AS), promoting price stability (Alfaro, 2003; Herzer & Nunnenkamp, 2013). Ultimately, FDI's net

		effect depends on the nation's absorptive capacity (OECD, 2008).
USD/VND Exchange Rate	TY_GIA	**Data source**: FiinPro. Exchange rates affect inflation mainly through exchange rate pass-through (ERPT): domestic currency depreciation raises import prices, generating cost-push inflation and shaping expectations (Krugman et al., 2018). This effect is stronger in more open economies (Campa & Goldberg, 2005).
Gross Domestic Product	GDP_DANH_NGHIA_Ty_VND; GDP_THUC_Ty_VND	**Data source:** FiinPro. GDP is a key proxy for aggregate demand and economic conditions in inflation forecasting. Theoretically, output shocks can generate price pressure through the demand-pull channel, and the GDP–inflation relationship may be nonlinear (Pollin & Zhu, 2006). Recent studies also stress combining real GDP with high-frequency data such as Google Trends for nowcasting early inflation signals (Proietti et al., 2021; Kohns & Bhattacharjee, 2022).

4. Experimental results and analysis

4.1. Evaluation Framework

This study employs a recursive pseudo-real-time out-of-sample evaluation approach. Unlike random sample splitting, this method replicates the exact information set that would have been available at each historical point in time, thereby eliminating look-ahead bias. As a result, the evaluation accurately reflects the model's genuine out-of-sample predictive ability rather than merely capturing its in-sample fit.

The recursive evaluation process begins in January 2021. At each forecasting point, the model is re-estimated using an expanding window of data, and only information available before the target month is utilized to generate the forecast (nowcast).

In terms of performance metrics, the Root Mean Squared Error (RMSE) is adopted as the primary evaluation criterion. This metric heavily penalizes large forecast errors because deviations are squared before averaging, making it particularly suitable for the stringent requirements of inflation forecasting. A lower RMSE value indicates a higher level of predictive accuracy.

4.2. Performance of Machine Learning Models versus Traditional Models

This section compares the benchmark VAR model with two groups of machine learning models: shrinkage methods (Ridge, LASSO, Elastic Net) and tree-based methods (Random Forest, XGBoost).

Table 4: Comparison of Forecasting Errors (RMSE) between Machine Learning and Benchmark Models

	Benchmark	Shrinkage Methods			Tree-based methods	
Horizon	VAR	Ridge	LASSO	ElasticNet	Random Forest	XGBoost
Day8	0.00571	0.00346	0.00392	0.00392	0.00427	0.00429
Day15	0.00564	0.00465	0.00455	0.00491	0.00433	0.00429
Day22	0.00571	0.00446	0.00440	0.00488	0.00461	0.00455
EOM	0.00558	0.00450	0.00477	0.00580	0.00472	0.00462
Overall	0.00566	0.00426	0.00441	0.00488	0.00448	0.00443

The recursive out-of-sample evaluation reveals a clear pattern of differentiation: machine learning models—particularly shrinkage-based methods (Ridge and LASSO)—deliver substantially superior forecasting performance compared with the traditional benchmark model (VAR) as well as nonlinear machine learning algorithms (XGBoost and Random Forest). Across the entire forecasting cycle, the Ridge model achieves the lowest RMSE at **0.00426**.

Other machine learning models follow with LASSO (**0.00441**), XGBoost (**0.00443**), Random Forest (**0.00447**), and Elastic Net (**0.00488**). In contrast, the VAR model records the highest forecasting error, reaching **0.00566**.

This considerable performance gap stems from the inherent characteristics of mixed-frequency macroeconomic forecasting datasets. As the number of predictors expands—often containing substantial noise and strong multicollinearity—the VAR model exhibits structural limitations because it must estimate a large number of parameters without any built-in regularization mechanism. Consequently, the performance of the traditional econometric model deteriorates rapidly when the signal-to-noise ratio is low.

In contrast, machine learning models—most notably Ridge Regression—address this challenge effectively through the use of a penalty function:

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \sum_{t=1}^{n} (y_t - \beta_0 - \mathbf{w}_t^{\top}\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \ (l_2 - penalty)$$

that continuously shrinks coefficient estimates. This regularization mechanism stabilizes the estimation process and mitigates the risk of overfitting. Moreover, the fact that tree-based algorithms (Random Forest and XGBoost), although outperforming VAR, still rank behind Ridge suggests that the inflation dynamics in Vietnam retain a fundamentally linear core structure. Tree-based models, which flexibly partition the data, may become overly sensitive when trained on relatively limited samples, potentially fitting short-term random noise rather than persistent economic signals. This tendency can lead to instability in recursive out-of-sample forecasting performance.

As can be seen that the empirical evidence provides strong statistical support for accepting Hypothesis H1. Machine learning algorithms—particularly regularized linear models—demonstrate more effective, stable, and accurate performance than traditional regression-based econometric models in the context of high-dimensional real-time inflation forecasting.

4.3. Evaluating the Contribution of Google Trends Data

To evaluate the role of search behavior, the study compares two Ridge specifications: a macro-only model and a full model that adds Google Trends variables.

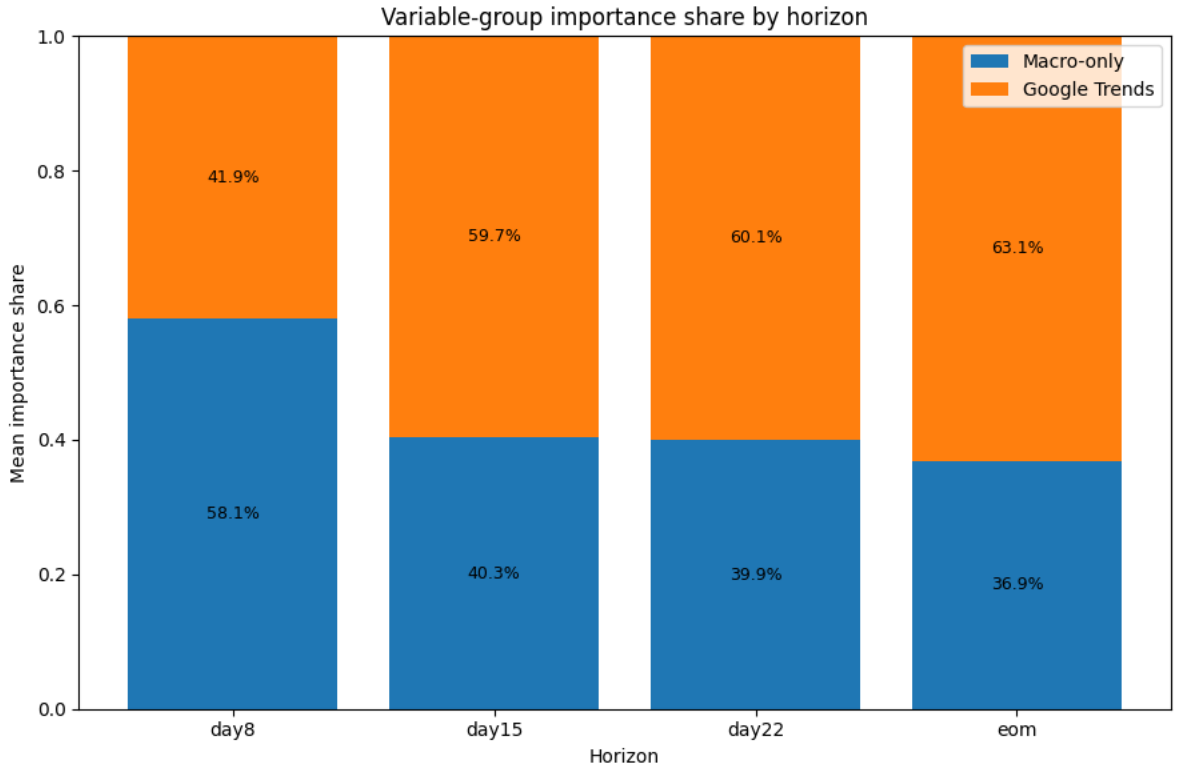


Figure 2: Variable-Group Importance Share by Forecasting Horizon

The variable-group importance analysis shows that Google Trends receives a non-trivial share of total coefficient weight, increasing from 41.9% at Day 8 to 63.1% at end-of-month. However, coefficient weight does not necessarily imply a stable gain in forecasting accuracy.

Table 5: Comparison of Forecasting Errors: Macro-only vs. Google Trends Integrated Models

Horizon_tag	Google Trends included	Macro-only
Day8	0.00346	0.00410
Day15	0.00465	0.00420
Day22	0.00446	0.00366
EOM	0.00450	0.00465
Overall	0.00426	0.00415

The RMSE comparison shows that the macro-only specification performs slightly better overall. The full model records an overall RMSE of 0.00426, while the macro-only version achieves 0.00415, implying that Google Trends increases average forecast error by about 2.9%.

Google Trends appears to add some useful information only at the earliest horizon (Day 8), but this advantage fades later in the month. A likely explanation is that search intensity reacts strongly to news cycles and temporary public attention, adding high-frequency noise that can dilute more stable macroeconomic signals.

4.4. Evaluation of the Hybrid Model Framework

The study also evaluates two hybrid extensions—Ridge + Random Forest and Ridge + XGBoost—to determine whether a second nonlinear stage can improve on the Ridge baseline by learning residual structure.

Table 6: Forecasting Performance (RMSE) of Hybrid Models versus Baseline Ridge Regression

	Benchmark	Hybrid Model	
Horizon_tag	Ridge	Ridge + RF	Ridge + XGBoost
Day8	0.00346	0.00428	0.00377
Day15	0.00465	0.00462	0.00473
Day22	0.00446	0.00483	0.00458
Eom	0.00450	0.00449	0.00467
overall	0.00426	0.00455	0.00444

The results indicate that the hybrid models do not improve upon Ridge. The baseline Ridge model achieves the lowest overall RMSE of 0.00426, compared with 0.00444 for Ridge + XGBoost and 0.00455 for Ridge + Random Forest.

CUMSFE paths lead to the same conclusion: neither hybrid specification delivers a persistent advantage over Ridge across forecasting horizons.

This suggests that once Ridge captures the dominant linear structure of inflation, the remaining residual component contains little systematic nonlinear information. Adding a second nonlinear learner therefore increases complexity without generating consistent out-of-sample gains.

4.5. Interpretation of the Optimal Model Structure and In-depth Analysis of the Macro-only Scenario

After rejecting the hybrid model specifications, the baseline Ridge Regression structure demonstrated overall superiority. However, to precisely identify the source of the model's predictive power and to make a final determination on whether to retain the behavioral data block (Google Trends), the study "opens the black box" of the model. This is achieved by analyzing the Ridge coefficients heatmap and directly comparing two scenarios: Ridge Full (combining Macro and Google Trends) and Ridge Macro-only (using only macroeconomic data).

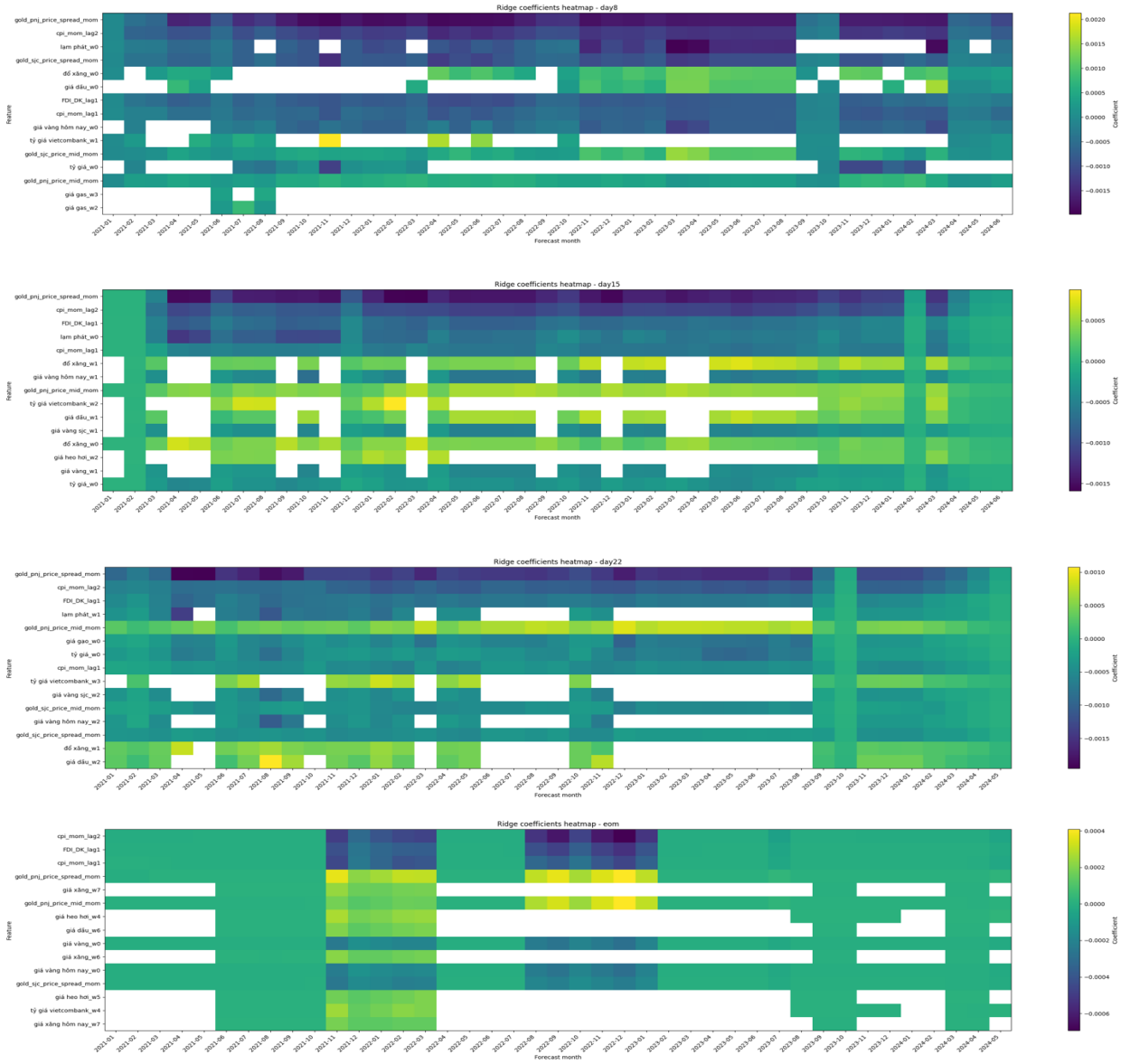


Figure 3: Recursive Evolution of Ridge Regression Coefficients Across Forecasting Horizons

The coefficients heatmap provides a visual representation of the contribution and stability of the predictors across each recursive evaluation origin. The chart reveals that core macroeconomic variables maintain highly stable predictive power, as evidenced by consistent color bands throughout the testing cycle. This reaffirms that inflation in Vietnam is driven by

a solid linear backbone. The Ridge algorithm excels at aggregating and appropriately assigning weights to these correlated macroeconomic signals without eliminating them.

To accurately quantify the actual value of the Google Trends variable set when placed alongside this robust macroeconomic framework, the study compares the forecasting errors between the full dataset scenario and the macro-only scenario.

Table 7: Comparison of Forecasting Errors (RMSE): Full Dataset vs. Macro-only Scenario

Horizon_tag	Google Trends included	Macro-only
Day8	0.00346	0.00410
Day15	0.00465	0.00420
Day22	0.00446	0.00366
EOM	0.00450	0.00465
Overall	0.00426	0.00415

The results indicate that the Ridge Macro-only scenario yields better overall forecasting performance than the Full model. The inclusion of a large volume of Google Trends lagged variables actually increases the average forecasting error by approximately 2.9%, suggesting that search-based data adds noise rather than meaningful predictive value.

A horizon-level analysis reveals that Google Trends provides a marginal useful signal only very early in the month (Day 8), where the Full scenario briefly outperforms the Macro-only scenario. However, this advantage quickly reverses from Day 15 onward, with the performance gap widening considerably by Day 22.

The root cause is that Google Trends introduces high-frequency but noisy variation into a forecasting problem primarily driven by macroeconomic fundamentals. Search behavior reacts to news cycles and short-term psychological shifts that do not reflect persistent inflationary pressures. Even with Ridge's regularization mechanism, excessive noisy predictors still dilute the value of core macro variables, whose signal is already captured more reliably by traditional indicators such as commodity prices, exchange rates, and monetary aggregates in this context.

General Conclusion: The Ridge model's superiority stems primarily from its ability to efficiently aggregate correlated macroeconomic information, rather than from any substantial value added by search data. The Ridge Regression structure using purely macroeconomic data

therefore represents the most ideal solution — a parsimonious, highly stable real-time inflation forecasting model that delivers optimal accuracy for practical policymaking in Vietnam.

4.6. Discussion

The empirical results demonstrate that regularized linear models, particularly Ridge Regression, significantly outperform traditional econometric methods such as VAR in high-dimensional environments. Contrary to initial hypotheses, integrating high-frequency Google Trends data or employing complex hybrid nonlinear algorithms — such as Random Forest and XGBoost — degraded overall predictive accuracy rather than enhancing it. Incorporating search-based variables introduced behavioral noise that diluted core macroeconomic signals, particularly over longer forecast horizons, while hybrid models combining linear and nonlinear components tended to overfit the residuals. These findings reaffirm that inflation dynamics in Vietnam are fundamentally driven by a stable linear backbone. By efficiently aggregating correlated macroeconomic variables without search-based noise, parsimonious macro-only models are not only more accurate but also capable of substantially reducing the publication lag associated with official statistics. Consequently, overly complex specifications incorporating alternative behavioral data prove counterproductive in this context, and simpler, well-regularized models remain the most effective approach for near-term inflation forecasting in emerging economies like Vietnam.

5. Conclusion

This study successfully developed a real-time inflation forecasting (nowcasting) framework for Vietnam, effectively reducing the official publication lag by 75%. The findings conclude that a Ridge Regression model relying exclusively on core macroeconomic indicators (Macro-only) is the optimal solution, achieving the lowest overall forecasting error (RMSE of 0.00415). By efficiently aggregating correlated macroeconomic data without the noise of high-frequency search inputs, this model provides a highly reliable, stable, and interpretable early warning system to support proactive monetary policymaking and economic risk management in Vietnam.

REFERENCES

Alfaro, L., & School, H. (2003). Foreign Direct Investment and Growth: Does the Sector Matter?

Bracha, A., & Tang, J. (2022). Inflation Levels and (In)Attention. The Review of Economic Studies. https://doi.org/10.1093/restud/rdae063

Bui, V. X., & Nguyen, H. T. (2019). Stock market activity and Google Trends: the case of a developing economy. Journal of Economics and Development, 21(2), 191–212.

Campa, J., Manuel, Goldberg, L., & González-Mínguez, J. (2005). Exchange rate pass-through to import prices in the euro area.

Chan, J. C. C. (2022). Large Hybrid Time-Varying Parameter VARs. Journal of Business & Economic Statistics, 1–34.

Chen, T., & Guestrin, C. (n.d.). XGBoost: A Scalable Tree Boosting System. https://doi.org/10.1145/2939672.2939785

Cholette, P. A., & Dagum, E. B. (2006). Benchmarking Time Series with Autocorrelated Survey Errors. International Statistical Review / Revue Internationale de Statistique, vol 62 no 3, p. 365.

Crowley, J. (2008). Credit Growth in the Middle East, North Africa, and Central Asia Region. IMF Working Papers, 08(184), 1.

Duc, D. T. V., Hoai, N. T., Nguyen, V. P., Nguyen, D. P., Anh, N. H., & Hai, H. H. (2024). Google Search Intensity and Stock Returns in Frontier Markets: Evidence from the Vietnamese Market. Economics and Business Review, 10(1).

Epstein, N., Gornicka, L., Musil, K., Nalban, V., & Ha, N. (2022). Quarterly Projection Model for Vietnam: A Hybrid Approach for Monetary Policy Implementation. IMF Working Papers, 2022(125), 1.

Friedman, M. (1970). A Theoretical Framework for Monetary Analysis. Journal of Political Economy, 78(6), 1385–1386.

Furukawa, K., Hisano, R., Minoura, Y., & Yagi, T. (2024). A nowcasting model of industrial production using alternative data and machine learning approaches. Japan and the World Economy, 71, 101271

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. Journal of Monetary Economics, 55(4), 665–676.

Gokmenoglu, K., Azin, V., & Taspinar, N. (2015). The Relationship between Industrial Production, GDP, Inflation and Oil Price: The Case of Turkey. Procedia Economics and Finance, 25, 497–503.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., & Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? Journal of Applied Econometrics. https://doi.org/10.1002/jae.2910

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition. Springer New York.

Heng, D. (2024). Nowcasting and Near-Term Forecasting Cambodia's Economy. IMF Working Papers, 2024(147), 1.

Herzer, D., & Nunnenkamp, P. (2013). Inward and outward FDI and income inequality: evidence from Europe. Review of World Economics, 149(2), 395–422.

Khashei, M., & Bijari, M. (2010). An artificial neural network (p,d,q) model for timeseries forecasting. Expert Systems with Applications, 37(1), 479–489.

Kohns, D., & Bhattacharjee, A. (2022). Nowcasting growth using Google Trends data: A Bayesian Structural Time Series model. International Journal of Forecasting. https://doi.org/10.1016/j.ijforecast.2022.05.002

Krugman, P. R., Obstfeld, M., & Melitz, M. J. (2018). International economics theory and policy. Uttar Pradesh, India Pearson.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. Science, 343(6176), 1203–1205.

Mankiw, N. G. (2022). Macroeconomics. Worth Publishers.

Mchukwa, E. W., Mbwambo, S. K., & Mchomvu, Z. B. (2025). Forecasting Consumer Price Index (CPI) - Empirical Analysis from Tanzania. Economic Insights – Trends and Challenges, 2025(1), 73–83.

Narita, F., & Yin, R. (2018). In Search of Information: IMF Working Papers, 18(286), 1.

Ngo, H. (2021, December 17). Using online prices for inflation estimation and pricing behavior research: a case study from Vietnamese multichannel retailers during the COVID-19 pandemic. Thesis.eur.nl. https://thesis.eur.nl/pub/61251

Nguyen, D. D., & Pham, M. C. (2018). Search-based Sentiment and Stock Market Reactions: An Empirical Evidence in Vietnam. The Journal of Asian Finance, Economics and Business, 5(4), 45–56

Nguyen, T. H., Le, M. T., Hoang, A. T., Tran, K. A., & Giap, C. N. (2024). Nowcasting Vietnam's Export Growth with Mixed Frequency Data. Revista de Gestão Social E Ambiental, 18(9), e06237.

Nguyen, V. A. T., Hoang, T. T., & Le, D. A. (2022). A study on the relationship between money supply and inflation in Vietnam from 2005 to 2021. Accounting, 8(4), 395–402.

Nicholson, W. B., Wilms, I., Bien, J., & Matteson, D. S. (2020). High Dimensional Forecasting via Interpretable Vector Autoregression. 21(166), 1–52.

OECD Benchmark Definition of Foreign Direct Investment FOURTH EDITION 2008.

POLLIN, R., & ZHU, A. (2006). Inflation and economic growth: a cross-country nonlinear analysis. Journal of Post Keynesian Economics, 28(4), 593–614.

Proietti, T., Giovannelli, A., Ottavio Ricchi, A Citton, Cristian Tegami, & Tinti, C. (2021). Nowcasting GDP and its components in a data-rich environment: The merits of the indirect approach. 37(4), 1376–1398.

Sims, C. A. (2003). Implications of rational inattention. Journal of Monetary Economics, 50(3), 665–690.

Stigler, G. J. (1961). The Economics of Information. Journal of Political Economy, 69(3), 213–225.

Thay đổi thời gian công bố số liệu CPI, GDP, GRDP. (2024). General Statistics Office of Vietnam. https://www.nso.gov.vn/su-kien/2024/06/thay-doi-thoi-gian-cong-bo-so-lieu-cpi-gdp-grdp/

Varian, H. R., & Choi, H. (2012). Predicting the Present with Google Trends. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.1659302