UCLouvain
FUCaM MONS

LOUVAIN
School of Management

Louvain School of Management

Macro-Financial Signals and On-Chain Fundamentals in Bitcoin Predictability

Assessing the Incremental Value of U.S. Indicators for Weekly Returns (2013–2025)

Author: Haitam Belguenani
Supervisor(s): Corentin Vande Kerckhove & Mikael Petitjean
Academic Year 2025-2026
Master [120] in Business Engineering, Specialized Finality in Business Analytics

PRACTICAL DOS AND DONTS OF GENERATIVE AI

LSM adopts the following practical guidelines developed by Prof. Benoît Gailly[3] in its "hitchhiker's guide to a Master Thesis":

DONTs	DOs
Delegate with no or limited understanding	**Think** first by yourself about the problem
Always use free and/or popular tools	**Select** the right source(s) of assistance
Share confidential, sensitive or proprietary content	**Provide** information regarding your context, motivations and expectations
Look for (or expect) definitive answers	**Ask** for suggestions and inspirations; iterate
Take for granted and integrate blindly	**Review** and challenge output and sources
Claim or make people believe that you did everything yourself	**Disclose** that you have been helped (and how)

DECLARATION

The declaration below is **mandatory and must appear on the first page (after the front page) of your master's thesis**.

STATEMENTS ON YOUR USE OF GENERATIVE AI

Confirm the following statements:

Statements on responsible use of generative AI	Your answer
The content of my/our master's thesis is my/our original output, even if I/we used generative AI to help prepare it.	☒Yes ☐No
I/we master the theories, tools and methods that I/we use for the thesis.	☒Yes ☐No
I/we verified that the output presented in the thesis is accurate and valid, and that it contains no hallucination, no false reference, etc.	☒Yes ☐No
I/we acknowledge that I/we master the final output and that I am/we are able to explain it and answer questions about it.	☒Yes ☐No
I/we made sure not to provide generative AI tools with access to confidential data or information	☒Yes ☐No

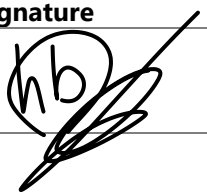
[3] Gailly, Benoît (2025) *The hitchhiker's guide to a Master Thesis supervised by Prof. Benoit Gailly at Louvain School of Management*. UCLouvain, LSM.

UCLouvain

LOUVAIN
School of Management

SIGNATURE

Sign your declaration:

	Author last name, first name(s)	Date	Signature
1	Belguenani, Haitam	05/08/2026	
2			

NOTE

During the preparation of this master's thesis, I, Haitam Belguenani, utilized the following AI tools for the purposes described below:

1. *Academic writing and paraphrasing assistance:* Claude (Anthropic) was used to propose and reformulate written passages across the thesis, in order to improve clarity, argument structure and academic register.

2. *LaTeX typesetting and bibliography integration assistance:* Claude (Anthropic) was used to assist with the typesetting and formatting of the thesis document in LaTeX,including syntax for mathematical expressions, tables and figures, and with the integration of the bibliography (Zotero import, biblatex/biber configuration, citation formatting in APA style).

3. *Code assistance and debugging:* Claude Code (Anthropic) was used to assist with the structuring, optimisation, and debugging of the Python codebase developed for the data pipeline and empirical analyses.

4. *Document review and consistency checking:* Claude (Anthropic) and NotebookLM (Google) were used to support the consultation and synthesis of academic sources during the literature review, and to check the internal consistency of the thesis (cross-references, citation accuracy, structural numbering) across chapters.

5. *English translation assistance:* Claude (Anthropic) and DeepL were used to assist with the translation of the thesis from French to English.

6. *Project management and process tracking:* Claude (Anthropic), via the Cowork interface, was used to maintain a record of the thesis's progress and decisions across working sessions.

After using these tools, I diligently reviewed and edited all content produced. I take full responsibility for the final content presented in this thesis. By signing this declaration, I affirm that the content of this master's thesis reflects my original work, augmented by the responsible use of AI.

Abstract

Bitcoin's institutionalization has turned its exposure to U.S. macro-financial conditions into a portfolio question. A substantial literature links monetary policy, inflation, the dollar, and equity factors to Bitcoin returns, but establishes that link in sample, with the benefit of hindsight. This thesis asks a narrower question: does the macro-financial block improve real-time forecasts of weekly Bitcoin log-returns beyond crypto-native variables, and does the answer vary across market regimes? Four nested specifications (baseline, crypto-native, macro-financial, combined) are estimated identically under a linear ARIMAX and a gradient-boosted tree model on 678 weekly observations spanning 2013 to 2025. A 156-week rolling window, whose structure is frozen before any access to results, yields 510 out-of-sample forecasts per configuration, compared through small-sample-corrected Diebold-Mariano tests and a Model Confidence Set, and evaluated over four event-based regimes fixed before any access to results. The answer is negative in all four. Adding macro-financial information never improves accuracy: it significantly degrades the linear forecast (DM = -2.52, $p = 0.012$) and leaves the tree forecast statistically unchanged ($p = 0.359$). The Model Confidence Set eliminates exactly one configuration, the linear combined model. No specification outperforms a rolling historical mean. In-sample detection of a macro-crypto link does not survive the cost of estimating it out of sample.

Acknowledgements

This thesis is coming to an end and it has been quite a journey. Between the doubts, the 2026 World Cup, the productive sleepless nights, the surges of confidence followed by renewed doubts, this work has taught me an enormous amount about the topic explored here, while also serving as a genuine first introduction to academic research. But beyond the content itself, it is on a personal level that it has pushed me to grow the most: learning to manage stress and deadlines, learning to question my own biases, learning to challenge myself. Lessons that will stay with me well beyond this thesis.

Taken as a whole, this year has been remarkably rich. The first semester took me on an Erasmus exchange to Tokyo, into an environment completely foreign to me, far from my usual points of reference, alone with myself. Looking back, it stands out as the best experience of my life: I met remarkable people from all walks of life, discovered a culture I fell in love with, and, above all, faced fears and limits I once believed were insurmountable, only to overcome them.

The second semester took on an entirely different tone: my first steps into the professional world at EY, and with them, my first taste of the rigor and intensity of corporate life. Doubts were not in short supply, but I come out of it more mature and confident about what lies ahead.

These five years of higher education are already drawing to a close, and time has gone by faster than expected. I look back on this period with real fondness, and above all, on the people who inspired me and pushed me to become the best version of myself. I hope what comes next will live up to it.

I would first like to thank my two supervisors, Professor Corentin Vande Kerckhove and Professor Mikael Petitjean, for their kindness, their availability, their rigor, and their advice. I was fortunate to take some of your courses during my studies, and I am grateful to have had your guidance throughout this thesis.

I would also like to thank all my friends, those I met at university, those I met on the other side of the world, and those who have been by my side since childhood. I shared this journey with some of you, and without you, this stage would have been far harder and far more monotonous.

A special thought goes to my family, my mother, my father, and my brother, who carried my doubts alongside me and were there for me in the moments I needed it most.

Finally, one last thought for myself: thank you for never giving up on yourself, despite the doubts and the obstacles. The limits exist only in your head. Break them, keep dreaming, and go conquer the world.

Contents

1 Introduction

Originally designed as a peer-to-peer electronic cash system meant to operate outside central banking (Nakamoto, 2008), Bitcoin is now accessible through spot exchange-traded funds, which hold the coins themselves instead of futures contracts. Approved by the U.S. Securities and Exchange Commission in January 2024, these funds are held on that basis in institutional portfolios. An asset conceived to escape monetary policy has become an asset whose sensitivity to monetary policy is an open empirical question. This shift places Bitcoin inside the same allocation decisions, risk budgets, and reporting requirements as conventional financial assets, and it turns its response to U.S. macro-financial conditions into a portfolio question rather than a curiosity.

Two strands of the literature answer that question differently. The first documents links between Bitcoin and macro-financial conditions: monetary policy announcements, inflation, the dollar, and equity market factors have been found to co-move with, or to Granger-cause, Bitcoin returns (Corbet et al., 2020; Nakagawa & Sakemoto, 2023; Palazzi et al., 2026). The second explains Bitcoin through dynamics of its own, using network activity and market variables such as hash rate, active addresses, trading volume, and momentum (Balcilar et al., 2017; Liu & Tsyvinski, 2021). Both bodies of work are persuasive. The studies cited here also share a feature: they establish their results in sample, over the completed period, with the benefit of hindsight.

Establishing a relationship and forecasting are distinct exercises. A variable may be significantly associated with returns over a completed sample and still fail to improve a forecast made in real time, when its coefficient has to be estimated from past data alone. The literature has tested existence extensively, and a smaller body of work has tested out-of-sample predictability directly (Berger & Koubová, 2024; Yae & Tian, 2022). To our knowledge, none has tested whether the macro-financial block adds anything out of sample to a model that already contains crypto-native variables. This is the gap the present work addresses. The conditioning on market regimes that follows is a property of the design rather than a second claim to novelty.

The research question is therefore the following:

> *In which market regimes do U.S. macro-financial indicators provide incremental predictive value for Bitcoin's weekly log-returns beyond crypto-native variables over the period 2013 to 2025?*

Four nested specifications were estimated to answer it: a baseline containing past returns only, a crypto-native specification, a macro-financial specification, and a combined specification. Each was applied identically to two functional families, a linear ARIMAX model and a gradient-boosted tree model, over 678 weekly observations from January 2013 to December 2025. Using two families ensures that the conclusions do not rest on a single modeling choice. Estimation proceeds through a fixed rolling window of 156 weeks whose structural choices were frozen on

the first window, before any access to results, and whose coefficients are re-estimated at every step, yielding 510 out-of-sample forecasts per configuration. Predictive accuracy is compared using the Diebold-Mariano test with a small-sample correction and a Model Confidence Set. Four regimes, dated from monetary and market events and fixed before any access to results, serve as post-hoc evaluation labels, and a 2026 hold-out sample was evaluated a single time.

The remainder of the thesis is organized as follows. Section 2 reviews the literature on macro-financial and crypto-native determinants of Bitcoin returns and states the gap in detail. Section 3 presents the data, the variables, and the empirical design, including the freezing protocol and the inference procedure. Section 4 reports the results, and Section 5 discusses them, together with the limitations of the design and directions for further research. Section 6 concludes.

2 Literature Review

Estimating the relationship between Bitcoin and its determinants over an entire sample assumes that this relationship is stable. The literature assembled in this section repeatedly contradicts this assumption. The influence of macro-financial conditions on Bitcoin returns has been examined extensively. How that influence varies with the state of the market has received limited attention. A third question remains unanswered: whether this influence provides information that the crypto-native block does not already contain.

Section 2.1 positions Bitcoin as a financial asset and documents the instability of its relationship with traditional markets. Sections 2.2 and 2.3 take up the two blocks of determinants in turn, macro-financial and then crypto-native, together with the channels through which they operate. Section 2.4 brings together the work devoted to the time variation of these relationships and to market regimes. Section 2.5 contrasts the two families of models employed, econometric and machine learning. Section 2.6 formalizes the resulting research gap and positions the contribution of this thesis.

2.1 Bitcoin as a Financial Asset

Bitcoin is originally defined as a peer-to-peer payment network, solving the double-spending problem without a centralized trusted entity (Nakamoto, 2008). Its actual use has departed from this. Analysis of transactions recorded on the blockchain establishes that holders use it mainly as an investment asset, and that its transactional use remains marginal (Baur, Hong, & Lee, 2018). This finding sets the object of this section: Bitcoin returns and their determinants. It also raises a question the literature has left open: which asset class Bitcoin belongs to.

The answers provided do not converge. A first characterization places Bitcoin between gold and the U.S. dollar. Modeled with GARCH and EGARCH, its volatility displays hedging capacities close to those of gold, together with a significant sensitivity to the Federal Reserve policy rate (Dyhrberg, 2016). A replication published in the same journal reaches the opposite conclusion. On the original sample as on an extended sample, Bitcoin returns form a series distinct from those of gold and the dollar, in their volatility as in their correlations (Baur, Dimpfl, & Kuck, 2018). A second characterization, more cautious, applies the typology of Baur and McDermott (2010), which distinguishes the hedge (permanent negative correlation), the safe haven (negative correlation in crisis periods) and the diversifier (positive but weak correlation). On 2011-2015 data and through a DCC-GARCH model, Bitcoin fails the hedge test, partially meets the diversifier test, and behaves as a safe haven in a single case, against extreme weekly declines in Asian equity indices (Bouri et al., 2017).

The meaning of these verdicts changes with the period studied. During the March 2020 crash, Bitcoin falls in near synchronization with the S&P 500, and its inclusion in an equity portfolio

increases the portfolio's downside risk (Conlon & McGee, 2020). Both the "digital gold" narrative and the partial diversifier status measured five years earlier predicted the opposite. Bouri et al. (2017) estimate a DCC-GARCH model on several asset classes, whereas Conlon and McGee (2020) study a single episode against the S&P 500 alone. The period therefore remains one possible explanation for the gap between their conclusions, among others. The properties attributed to Bitcoin relative to traditional markets nevertheless vary from one study to another, and the most opposed verdicts concern the most distant periods. Bitcoin's financial status thus appears to vary with the state of the market.

2.2 Macro-Financial Determinants of Bitcoin Returns

The question of a link between macroeconomic conditions and Bitcoin's valuation rests on an explicit theoretical foundation. In a general equilibrium model where a cryptocurrency without intrinsic value coexists with a fiat currency, the equilibrium price of the cryptocurrency depends directly on the monetary policy conducted by the central bank (Schilling & Uhlig, 2019). The relationship between macro conditions and Bitcoin is therefore a theoretical prediction to be tested, rather than an empirical correlation to be documented. This link also sets its own limit. Most of the variation in Bitcoin returns is driven by extrinsic volatility, that is, by price movements with no counterpart in fundamentals (Biais et al., 2023). Any model built on observable variables therefore addresses a minority of the variance, which sets realistic expectations for the explanatory power of the specifications estimated in Section 3.

This prediction is supported empirically by evidence on returns themselves. A Dynamic Factor Model aggregating around a hundred U.S. macroeconomic indicators finds that common macro factors significantly predict Bitcoin returns at a quarterly horizon, while indicators taken in isolation are not significant (Nakagawa & Sakemoto, 2023). This finding defines the bet made by the present thesis. Variables selected for their transmission channel and evaluated at weekly frequency may succeed where generic indicators fail at quarterly frequency. The selection rests on four economic channels, each presented below together with the evidence available for it.

The first channel is the discount rate. With no cash flows of its own, Bitcoin's value rests entirely on the discounting of an expected future flow, so that a rise in rates raises the opportunity cost of holding it and compresses valuations (Schilling & Uhlig, 2019). The Federal Reserve policy rate captures this channel directly. Its relationship to Bitcoin returns is confirmed at monthly frequency on a 2010-2025 sample (Abid, 2026), and the uncertainty surrounding FOMC decisions has a documented negative effect on those same returns (Shaikh, 2020).

The value factor (HML) of Fama and French (1993) operates through the same mechanism, expressed in the cross-section of the equity market. An asset whose value rests entirely on expected flows is a long-duration, growth-profile asset, and should load negatively on a factor that rewards value stocks. Bitcoin is the extreme case, since its entire value lies in expectations.

The sign of the prediction is therefore fixed before estimation. No direct evidence supports it in the literature surveyed here. The mechanism justifies its inclusion without guaranteeing predictive power: duration accounts for a contemporaneous comovement, whereas this thesis tests a relationship lagged by one week.

The second channel is the inflation hedge. The fixed issuance cap of twenty-one million units sustains a store-of-value narrative against anticipated monetary dilution, and makes the consumer price index the variable for this channel. A time-frequency wavelet approach documents a positive short-term association (L. Wang et al., 2023). This evidence remains indirect, since it concerns a contemporaneous association rather than a lagged relationship.

The third channel is risk appetite. The excess return of the equity market (Rm-Rf) provides its aggregate measure, with inflows dominating in a "risk-on" regime and outflows in a "risk-off " regime. A positive effect of the U.S. equity market on Bitcoin returns is reported by Sihombing et al. (2025), in a low-circulation journal, and confirmed by Abid (2026). On U.S. sectoral data from 2011 to 2023, Granger causality runs from all equity sectors to Bitcoin returns (Valadkhani, 2024).

The size factor (SMB) captures the same appetite within the equity market, through the rotation toward its riskiest and least liquid segment, which is also the segment most held by retail investors. Bitcoin belongs to this segment through its return characteristics, while sharing none of the accounting determinants of small capitalizations. SMB is the only variable in the macro-financial block backed by direct evidence, since it predicts cryptocurrency returns over the 2016-2020 subperiod (Bianchi et al., 2023), a regime dependence that Section 2.4 takes up again.

Rm-Rf, SMB and HML are the three factors of Fama and French (1993), and their use here requires two clarifications. The factors entering the macro-financial block are the equity factors, not their crypto counterparts: the three-factor model built for that market retains market, size and momentum, with no value factor (Liu et al., 2022), and its momentum factor belongs to the crypto-native block examined in Section 2.3. This thesis borrows the factorial legitimacy of that framework, not its factors. The exposure of crypto returns to traditional equity factors is also weak (Liu & Tsyvinski, 2021), but this measure concerns a contemporaneous loading. Including these factors tests Bitcoin's growing integration into equity markets over the recent regimes, without presupposing a stable exposure across the whole period.

The fourth channel is the dollar and global liquidity. Almost all Bitcoin trading is quoted in dollars, and a strong dollar signals a tightening of global financial conditions that puts mechanical pressure on assets denominated in that currency. The DXY index is the variable for this channel, and no direct evidence supports it in the literature surveyed here.

The six variables of the block therefore rest on justifications of unequal strength, and this section presents them accordingly. The policy rate and the price index rest on a solid mechanism and

on evidence, direct for the first, indirect for the second. SMB is the only variable in the block backed by direct evidence, while its mechanism rests on an analogy of characteristics. HML and the DXY show the opposite profile, a solid mechanism and no evidence, which makes them testable bets.

A substantial part of this literature has taken Bitcoin's volatility as its object rather than its returns. A comparison of GARCH-MIDAS models augmented with exogenous variables ranks global real economic activity first among the predictors of that volatility, while the VIX and the dollar's effective exchange rate show no significant predictive power on a 2013-2019 sample (Walther et al., 2019). This distinction matters for what follows. A macro indicator can be informative for volatility yet uninformative for returns, and most available results carry over poorly from one object to the other.

Two candidates are set aside. The S&P 500 carries information already absorbed by Rm-Rf. The VIX overlaps with the risk-aversion dimension of the retained factors, and Walther et al. (2019) report no significant predictive power for it, although on volatility rather than returns.

Geographical coverage is a separate decision. Bitcoin is quoted in dollars on almost all venues, the Federal Reserve steers the global liquidity to which it is exposed, and the institutional access structure created by the approval of spot ETFs is American. All four channels set out above therefore run through U.S. variables, and global real economic activity, dominant for volatility, falls outside this scope. The question posed here concerns the contribution of U.S. indicators specifically, not that of macro-financial conditions in general.

These four channels describe forces external to Bitcoin. A second research tradition explains its returns through variables generated by the network itself, which is the object of the following section.

2.3 Crypto-Native and On-Chain Determinants

This second research tradition falls into two sub-families. On-chain fundamentals, publicly observable on the blockchain, comprise the hash rate and active addresses. Crypto market factors, trading volume and momentum, describe the internal dynamics of the market. The two sub-families enter the design on different grounds. The former enter through their mechanism, no study having established them as systematic price factors. The latter enter through their documented empirical regularity, which is the standard route by which a factor is established in finance.

The first mechanism is production cost. As with any commodity, the price of Bitcoin should gravitate around its marginal cost of production, determined by the computing power that miners devote to the network (Hayes, 2019). The hash rate, the total computing power securing the network, is the observable proxy retained for it. Mining difficulty, a second candidate proxy, is

excluded from the design: being a mere algorithmic adjustment of the hash rate, it is mechanically redundant with it, with a correlation close to unity. The second mechanism concerns the use value of the network. The number of active addresses approximates the intensity of actual Bitcoin usage, whose value grows with adoption, and the traceability of transactions on the blockchain makes this measure available without an intermediary (Baur, Hong, & Lee, 2018).

The empirical evidence on these fundamentals is mixed. The hash rate and mining difficulty come out positive and significant on Bitcoin returns (Sihombing et al., 2025), active addresses and the hash rate are retained as predictors in a high-performing machine learning model (Nagula & Alexakis, 2022), and the relevance of technological determinants alongside economic ones is confirmed (W. Chen et al., 2021). Against this, the most systematic study in the field concludes, over a wide range of candidate factors, that production factors, mining costs included, fail to predict cryptocurrency returns, while factors specific to the crypto market do predict them (Liu & Tsyvinski, 2021).

This verdict supports the design of the present thesis rather than contradicting it, on two counts. The factor these authors validate is directly operationalized in the crypto-native block. Their negative verdict is unconditional, obtained over their full sample, whereas the studies cited above make significance depend on the period, the frequency, and the specification. The regime-based design supplies the conditional test their framework lacked.

Market factors rank by the strength of their evidence. Momentum comes first. It is the factor that Liu and Tsyvinski (2021) validate precisely where they reject production factors, and it is then confirmed as a systematic factor of the crypto market by a three-factor model built for that market (Liu et al., 2022). This double validation makes it the best-established variable in the block. It is operationalized as a cumulative return over recent weeks, with the horizon set in Section 3.

Trading volume comes next, on narrower evidence. A quantile approach documents that it predicts Bitcoin returns in intermediate market regimes, this power disappearing in strongly bullish and strongly bearish markets alike (Balcilar et al., 2017). This predictability weakens once tested out of sample: in a systematic comparison of daily predictors, trading volume and investor attention, well established in sample, yield no significant out-of-sample predictability (Yae & Tian, 2022). Volume therefore enters the block as a regularity whose robustness the present thesis tests, rather than as an established predictor.

The result of Balcilar et al. (2017) foreshadows the conditional logic of the present thesis. Even before macro-financial and crypto-native variables are compared, the predictability of a crypto-native variable proves to depend on the state of the market. The temporal stability of these relationships is the object of the following section.

2.4 Time-Varying Relationships and Market Regimes

The preceding sections have documented a fact that is uncomfortable for any unconditional approach: correlations between Bitcoin and traditional markets are unstable (Section 2.1), and the significance of predictors, macro and crypto-native alike, varies across studies, periods, and frequencies (Sections 2.2 and 2.3). If the contemporaneous dependence structure between Bitcoin and its environment varies with market conditions, predictive relationships derived from the same joint distribution may well be unstable too. This is the hypothesis the present thesis tests formally, and one that the recent literature makes increasingly plausible.

The existence of regime shifts in Bitcoin dynamics is documented by several studies. Estimated with Markov-switching GARCH models, two-regime and three-regime specifications significantly outperform single-regime models, both in fit and in risk forecasting, so that ignoring regimes biases estimation (Ardia et al., 2019). These breaks have an economic interpretation, since they map onto changes in policy rates and onto central banks' quantitative easing programs (Corbet et al., 2017). A regime-switching model further indicates that the effect of economic policy uncertainty on Bitcoin returns is itself conditional on the regime (Shaikh, 2020).

Beyond the existence of regimes, predictability itself proves unstable. In a dynamic model averaging framework, the SMB factor predicts cryptocurrency returns over the 2016-2020 subperiod alone, a regime dependence detected by a design that was not built to test for it (Bianchi et al., 2023). The same phenomenon appears for volatility, where the predictive power of macroeconomic indicators is markedly stronger in expansions and collapses in recessions and in high-volatility periods (J. Wang et al., 2023). The complementary pattern is also observed, exogenous factors delivering their largest improvement in volatility forecasting in bear markets (Walther et al., 2019). These results are obtained on different objects, frequencies, and methods, and their convergence suggests that instability is a structural property of the macro-Bitcoin link rather than the artifact of a particular specification.

This evidence nonetheless remains indirect with respect to the question posed here. It bears on volatility rather than returns (Ardia et al., 2019; Corbet et al., 2017; Walther et al., 2019; J. Wang et al., 2023), or on variables outside the framework retained here, such as economic policy uncertainty (Shaikh, 2020). The result of Bianchi et al. (2023) is the exception. Since the SMB factor belongs to the macro-financial block examined here, its predictive power confined to 2016-2020 constitutes the only direct evidence of regime dependence available for one of the design's variables. For the rest, instability is established in general terms, and remains unestablished for the specific relationship the present thesis examines.

The result most directly connected to the question of the present thesis is put forward by Palazzi et al. (2026). On daily data over 2014-2025, macro-financial variables acquire their explanatory power after 2019 and become the most informative during the monetary tightening episode of 2022-2023, while network fundamentals dominate the earlier period. This sequencing, funda-

mentals first and macro second, matches the event-based periodization adopted here, a convergence all the more notable in that it is obtained through independent periodization methods: statistical break-point detection in their work, exogenous anchoring on macro-monetary events in the present thesis. At the level of co-movements, this post-2020 shift also appears in the fourfold increase in the correlation between Bitcoin volatility and S&P 500 volatility after the onset of the pandemic (Iyer, 2022), evidence based on volatility. The relevant question is therefore no longer whether macro conditions predict Bitcoin, but when. These authors address it through in-sample Granger causality, without a formal test of incremental value and without out-of-sample validation.

This conditionality extends to the crypto-native block. Volume predicts returns in intermediate regimes alone (Balcilar et al., 2017), and Bitcoin's safe-haven role depends on the nature and duration of the critical periods considered (Aryan et al., 2025). Regime dependence is therefore documented on both sides of the comparison the present thesis organizes. It is never documented as a formal test of incremental value, as the following section shows.

2.5 Modeling Approaches: Econometric and Machine Learning Methods

The comparison between a linear econometric model and a nonlinear machine learning model is not an isolated innovation of the present thesis. Palazzi et al. (2026) already contrast Granger causality, a linear test, with transfer entropy, a nonlinear measure of informational dependence, because these two families capture different relational structures. The comparison between ARIMAX and XGBoost adopted here follows the same logic, transposed from causality testing to out-of-sample forecasting. ARIMAX assumes a linear and additive relationship: each predictor receives a single coefficient whose value is independent of the level of the other predictors, and the model sums these contributions. XGBoost proceeds by stacking threshold-based decision rules, which allows the effect of one predictor to vary with the value taken by the others, without imposing a functional form a priori. If the two models agree on the incremental value of the macro-financial block, the result is robust to the structure of the underlying relationship; if they disagree, that disagreement is itself informative.

The use of machine learning to predict Bitcoin returns from economic and technological determinants has direct precedents, set out in Section 2.3, where these models improve forecasts relative to standard econometric benchmarks (W. Chen et al., 2021; Nagula & Alexakis, 2022). A comparable superiority is reported against a benchmark OLS regression, on a variable selection derived from explicit economic theories, with the caveat that the dependent variable there is the price level of Bitcoin (Erfanian et al., 2022). This advantage also appears in a strictly univariate design, where the information set reduces to the history of returns (Berger & Koubová, 2024). The gap observed in this body of work therefore comes from the functional form of the model rather than from the information used, which justifies treating the choice of model and

the choice of variables as two distinct dimensions.

This advantage does not survive the move out of sample, however, and the distinction must be stated precisely. In sample, a model with more parameters fits mechanically better, since each additional parameter reduces the residuals by construction. The ranking of models there reflects their parameter count rather than their ability to forecast. Out of sample, that ranking reverses. On daily cryptocurrency returns, no machine learning method, from the LASSO to random forests and boosting, achieves significant out-of-sample predictability on Bitcoin, and the most heavily parameterized methods are penalized the most (Yae & Tian, 2022). The same ordering appears within neural networks, where deep architectures and LSTM layers fail to improve forecasts relative to a simple recurrent network (Berger & Koubová, 2024). When the signal-to-noise ratio is low, each additional parameter serves first to reproduce the noise of the estimation window, noise that is absent from the following window.

This mechanism extends beyond the case of Bitcoin. On the equity risk premium, almost all predictors that are significant in sample fail to beat the historical mean in real-time forecasting (Welch & Goyal, 2008), and weak predictability is restored only by imposing economic constraints on the forecasts (Campbell & Thompson, 2008). The difficulty encountered on Bitcoin returns therefore reflects a general property of forecasting problems with a low signal-to-noise ratio rather than a singularity of this market.

These findings govern two choices in the design adopted here. The two model families enter on equal footing, each giving access to a relational structure of its own. And the comparison between specifications bears on out-of-sample forecast errors rather than on in-sample fit. None of the studies surveyed in this section proceeds in this way on nested specifications, which is what the following section formalizes.

2.6 Research Gap and Positioning

The preceding review establishes a cumulative finding. Bitcoin's financial status is unstable (Section 2.1), its macro-financial and crypto-native determinants are each documented without ever being confronted (Sections 2.2 and 2.3), this instability extends to predictive relationships (Section 2.4), and the two model families used to capture them have left this conditionality outside their validation framework (Section 2.5). The established results are as follows. Macro-financial variables predict Bitcoin returns conditionally on the regime, their explanatory power rising after 2019 and peaking during the monetary tightening of 2022-2023 (Palazzi et al., 2026; L. Wang et al., 2023). Crypto-native variables predict across all regimes, while their individual significance remains itself conditional on the state of the market (Balcilar et al., 2017). The two blocks coexist, and neither substitutes for the other.

To our knowledge, no study surveyed here has tested the incremental predictive value of the macro-financial block beyond an already present crypto-native block, within an out-of-sample

validation framework. The closest works each depart from it on one specific point. Palazzi et al. (2026) document the evolution of Granger causality, linear and nonlinear alike, between macro-financial variables and Bitcoin returns across statistically identified regimes, but Granger causality is an in-sample property, and exploiting that relationship offers no guarantee of improved forecasts. The comparison of a linear and a nonlinear model is also documented (Berger & Koubová, 2024), at a constant information set, whereas the question posed here varies the information across two functional forms. The most systematic comparison of out-of-sample predictors examines them one by one without ever organizing them into nested blocks, and the set it declares exhaustive contains no macro-financial variable (Yae & Tian, 2022).

The design rests on three decisions. Its scope is set by two functional forms, linear and nonlinear, and by four regimes defined on macro-monetary events, which partition the evaluation period rather than the estimation sample. Its object is the comparison between a crypto-native specification alone and a combined specification, which tests an information contribution rather than a bivariate causality. Its protocol is out-of-sample validation by a fixed-size rolling window, together with a significance test on the differences in predictive performance between nested specifications.

The answer matters beyond the academic debate. If the contribution of the macro-financial block concentrates in certain regimes, a manager exposed to crypto-assets knows when to monitor U.S. monetary conditions and when to set them aside. If that contribution is zero everywhere, the information remains useful, since it rules out a set of signals that are costly to track.

3 Methodology

3.1 General Framework

This section turns the research question into an empirical protocol in which every choice is either fixed ex ante or selected on the first training window and then frozen, never adjusted on out-of-sample results. The setting is a univariate time series with exogenous regressors: the dependent variable is the weekly Bitcoin log return, $y_t = \ln(P_t/P_{t-1})$, measured at Friday's close. At each date, the task is to predict y_t using only the information available at the end of week $t-1$.

The protocol combines two dimensions. Four nested specifications isolate the predictive contribution of each block (Section 3.3); each is estimated with a linear model and with a nonlinear model, so that any conclusion holds whatever the form of the relationship (Section 3.4). Crossing them yields eight configurations, evaluated out of sample on a rolling window and grouped ex post by market regime (Section 3.5), then subjected to a two-level inference procedure (Section 3.6). Section 3.7 sets out the limitations of the protocol.

Two principles govern the design throughout and are invoked at each stage.

Anti-look-ahead. No information dated later than the prediction date enters the estimation, whether through the data, the preprocessing, or the specification choices. Without this discipline, measured performance would overstate what is actually attainable: the model would appear to predict returns from information it would have lacked in real time.

Ex ante freeze. The structure of the model is fixed once and for all, before any evaluation, as are the number and the order of the statistical tests: the backtest serves to evaluate the protocol, never to adjust it (López de Prado, 2018). Adjusting the structure or multiplying tests after observing the results amounts to selecting on the noise of the test sample, a form of data snooping that artificially inflates reported performance; chance is controlled formally by the Model Confidence Set (Section 3.6.3).

3.2 Data and Variables

Table 1 summarizes the eleven series used, their block, their source, and the transformation that ensures their stationarity. Each variable is justified individually in Sections 3.2.1 to 3.2.4; all explanatory variables enter at $t-1$.

Table 1: Summary of variables

Variable	Block	Source	Transformation
BTC log return (target)	Target	Coin Metrics (PriceUSD)	$\ln(P_t/P_{t-1})$, Friday close
Fed policy rate	Macro-financial	FRED (DFEDTARU)	Rate gap (deviation from 52-week MA)
CPI	Macro-financial	FRED (CPIAUCNS, NSA)	Year-over-year change
DXY	Macro-financial	Yahoo Finance	Weekly log change
$R_m - R_f$	Macro-financial	Kenneth French Library	Return (none)
SMB	Macro-financial	Kenneth French Library	Return (none)
HML	Macro-financial	Kenneth French Library	Return (none)
Hash rate	Crypto-native	Coin Metrics	Weekly log change (mean)
Active addresses	Crypto-native	Coin Metrics	Weekly log change (mean)
Volume	Crypto-native	Coin Metrics	Weekly log change (sum)
Momentum	Crypto-native	Derived from y	Sum of log returns from $t-1$ to $t-n$ ($n=12$)

3.2.1 Target Variable and Sample

The dependent variable is the weekly Bitcoin log return, $y_t = \ln(P_t/P_{t-1})$, where P_t is the Friday closing price (UTC), taken from the `PriceUSD` field of Coin Metrics ("Coin Metrics", 2026). Aggregated across several platforms, this rate avoids venue-specific anomalies, and the same provider supplies the on-chain variables, ensuring a common timestamp.

The weekly frequency is a deliberate trade-off: a daily step introduces microstructure noise with no predictive return at the horizon considered, while a monthly step would reduce the sample to roughly 150 points, too few for out-of-sample validation by regime. The Friday close also aligns Bitcoin with the clock of the macro-financial series.

The sample covers 2013–2025, the lower bound being dictated by the maturity of the on-chain data: before 2013, hash rate and active addresses are too sparse to form a continuous series. It contains 678 weekly observations, from January 4, 2013 to December 26, 2025, of which 510 serve for out-of-sample evaluation after the burn-in described in Section 3.5.2.

Stationarity of the target holds by construction, a log return being the first difference of a log price, and is confirmed empirically: ADF -15.22 ($p < 0.001$), KPSS 0.26 ($p \geq 0.10$). Details in Table 10.

3.2.2 Macro-Financial Block (6 Variables)

The macro-financial block brings together six variables representing monetary, inflation, exchange-rate, and equity market conditions in the United States. Their full economic justification, the four transmission channels to Bitcoin, is established in Section 2.

The three Fama and French factors, $R_m - R_f$, SMB, and HML, come from the Kenneth French Data Library (Fama & French, 1993) and belong to the risk appetite channel. $R_m - R_f$ is the equity market premium, SMB (*small minus big*) the size factor, and HML (*high minus low*) the value factor; the respective strength of their justification is discussed in Section 2. These three factors are already portfolio returns: stationary by nature, they enter without transformation, rebuilt at weekly frequency by compounding the daily factors. They enter at $t-1$ with no additional publication lag: a portfolio factor can be computed in real time from public closing prices, and the library's recompilation delay is a delay in dissemination rather than in the existence of the information (unlike the CPI).

The Fed policy rate represents the price of money and the stance of monetary policy: a tightening raises discount rates and compresses risk appetite, the channel through which it can weigh on a speculative asset such as Bitcoin. It enters as a deviation from its trend level (*rate gap*): $\text{FedGap}_t = \text{rate}_t - \text{MA}_{52}(\text{rate})_t$, the current rate minus its moving average over the preceding 52 weeks. This transformation preserves the reading in levels, a positive gap signalling a policy

tighter than the recent norm, while removing the near unit root of the raw level, which would expose the predictive regression to Stambaugh bias (Stambaugh, 1999). The moving average is strictly backward-looking.

The CPI captures the inflation regime, whose relevance rests on the recurring thesis of Bitcoin as an inflation hedge, or "digital gold", a hypothesis that the sign of this coefficient will put to the test. Inflation is measured by the year-over-year change in the non-seasonally-adjusted (NSA) CPI: seasonally adjusted series are revised ex post, their seasonal factors being re-estimated as new data arrive, which would introduce information unavailable in real time, whereas the year-over-year change removes seasonality by construction and makes a strongly trending raw index stationary, while remaining interpretable as the inflation rate.

One caveat is accepted: the year-over-year change is persistent, owing to the twelve-month overlap, to the point of failing the stationarity tests (Section 3.2.5), which exposes it to Stambaugh bias; the exception is documented and justified rather than corrected.

The DXY measures the strength of the dollar against a basket of currencies, to which Bitcoin, denominated in USD, is sensitive. The level of the index is a near random walk; it therefore enters as a weekly log change, stationary and interpretable as the movement of the dollar.

3.2.3 Crypto-Native Block (4 Variables)

The crypto-native block brings together four variables drawn from the Bitcoin network and market. The hash rate measures the aggregate computing power securing the network, active addresses count the distinct addresses involved in a transaction each day, and volume measures trading; all three come from Coin Metrics ("Coin Metrics", 2026) at daily frequency. Momentum, the fourth variable, is defined in Section 3.3.3.

Weekly aggregation follows the nature of each quantity: volume, which is a flow, is summed over the week; the hash rate, which is a rate, and active addresses, which form a count with repetitions, are averaged, since summing addresses would double-count those active on several days.

Once aggregated, these three series, strongly trending and growing at a near exponential pace, enter as weekly log changes, that is, as growth rates: the transformation makes them stationary and stabilizes the variance, and no predictive hypothesis bears on the level of the network, the signal being its growth rate. Their behavior under the stationarity tests is examined in Section 3.2.5.

3.2.4 Composition: Discarded Alternatives

Three variables were considered and then dropped. The S&P 500 and the VIX are redundant with the Fama-French factors, for the reasons set out in Section 2. Mining difficulty is dropped because of its near perfect redundancy with the hash rate, the two series measuring the same reality up to a scale factor (correlation ≈ 0.99).

These removals serve parsimony: under a 156-week window, every superfluous variable consumes degrees of freedom without distinct informational content. The correlation matrix of the retained predictors (Table 9) confirms that no strong residual redundancy remains: no pair exceeds 0.58 (FedGap / CPI, as expected), the other notable correlations being addresses / volume (0.41) and $R_m - R_f$ / SMB (0.25).

3.2.5 Preprocessing, Stationarity, and Variable Validation

All series are aligned on a common weekly grid indexed on the Friday close.

Lower-frequency series, namely the monthly CPI, are carried forward at their last published value (*forward-fill*), respecting the publication lag. Interpolation between monthly points is ruled out: it would use the following month's value before its release, introducing a leak of future information. Its counterpart, a CPI that is nearly constant within the month and of limited weekly predictive content, is accepted as preferable to a manufactured variation. Over the sample, the modal gap between two releases is four weeks; the single eight-week episode, from October 24 to December 12, 2025, corresponds to the actual interruption of BLS publication and is handled without correction, in line with the principle.

The stationarity of each transformed variable is checked by joint ADF and KPSS tests. Of the ten variables, seven are stationary under both tests. The other three, the CPI in year-over-year change and the hash rate and active addresses in log change, are the object of an explicitly documented exception rather than of a further transformation: the mechanism of each, and the supporting diagnostic, are set out in Table 10.

Descriptive statistics and the correlation matrix of the predictors appear in Tables 8 and 9.

A final descriptive check, run once on the full sample, examines whether the constructed variables carry the expected predictive sign. It is a sanity diagnostic rather than a selection filter: no variable was added, removed, or transformed on its basis, and it is reported for transparency as the only step of the protocol that consults the full sample. Under predictive timing, that is, a variable at $t-1$ correlated with the return at t, the correlations are all positive and consistent with the expected signs: momentum $+0.10$ to $+0.13$, hash rate $+0.03$, active addresses $+0.07$, volume $+0.05$. They are well below the corresponding contemporaneous correlations (momentum $+0.33$ to $+0.41$; addresses $+0.28$), a gap that is expected and reported in Table 11: it is the

honest reminder that the weekly predictability of Bitcoin is modest.

3.3 Nested Specifications

3.3.1 The Four Specifications

The identification strategy rests on four nested specifications, each containing the previous one as a special case. For the econometric family, they are written as restrictions on a single general model:

$$y_t = \alpha + \sum_{i=1}^{k} \phi_i \cdot y_{t-i} + \beta_C' \cdot X_{t-1}^C + \beta_M' \cdot X_{t-1}^M + \varepsilon_t \tag{1}$$

where X^C and X^M denote the crypto-native and macro-financial blocks defined in Section 3.2. The four specifications correspond to four sets of restrictions: Specification 1 (Baseline) imposes $\beta_C = \beta_M = 0$; Specification 2 (Crypto) imposes $\beta_M = 0$; Specification 3 (Macro-financial) imposes $\beta_C = 0$; Specification 4 (Combined) is the unconstrained model. Table 2 summarizes this nesting.

Table 2: The four nested specifications

Specification	Lags of y	Crypto-native	Macro-financial
1. Baseline	✓		
2. Crypto	✓	✓	
3. Macro-financial	✓		✓
4. Combined	✓	✓	✓

Note: central comparison, Spec 2 → Spec 4 (incremental value of the macro-financial block).
Secondary comparisons, Spec 1 → 2, Spec 1 → 3, Spec 3 → 4.

For XGBoost, each specification receives exactly the same columns as its ARIMAX counterpart, with only the functional form left free. The nesting is therefore informational, in the sense of nested variable sets, and inference proceeds through the comparison of forecasts (Section 3.6) rather than through restriction tests.

3.3.2 The Central Comparison

The four specifications yield four nested comparisons, each answering a different question. Three are secondary. Specification 1 to 2 measures the value of the crypto-native block alone; 1 to 3, the one of the macro-financial block alone; 3 to 4, the symmetric increment of crypto-native beyond macro-financial. The fourth is central and is set out below.

The central comparison of this thesis is the move from Specification 2 to Specification 4. It

reflects the asymmetry of the research question: the point is not to determine which block performs best, but whether macro-financial information adds something that crypto-native information leaves out.

3.3.3 Temporal Structure of the Variables

The temporal structure is deliberately parsimonious: the explanatory variables enter with their value at $t-1$ alone. Adding multiple lags of each explanatory variable would multiply the number of parameters without theoretical justification and, under a 156-week window with ten explanatory variables, would bring the model close to the floor of observations per variable (Section 3.5.2).

Only the dependent variable is lagged. The number of lags k is selected by the Bayesian information criterion (BIC) on the first training window, over the grid $k \in \{1,2,3,4\}$, and then frozen (retained value: $k = 2$, minimum BIC -177.66; details in Table 12). It is applied identically to both families, as the autoregressive order for ARIMAX and as lag columns for XGBoost, so that the comparison between families bears on the functional form rather than on the information available.

Momentum is defined as the sum of Bitcoin log returns from $t-1$ to $t-n$, which by additivity is exactly $\ln(P_{t-1}/P_{t-n-1})$, and operationalizes the factor validated by Liu and Tsyvinski (2021). It complements the lags under an explicit division of labor: the k lags form a short memory with free coefficients, while momentum compresses a medium memory into a single parameter.

The window n is selected by BIC over the grid $n \in \{8,12\}$, which rules out by construction any value at or below 4: a momentum of 4 combined with $k = 4$ would be an exact linear combination of the lags. The retained value is $n = 12$, but the BIC margin over $n = 8$ is minimal (0.08 point): this choice is documented as a limitation rather than as a clear tie-break (Table 13). Its effect on the central comparison is attenuated by construction, momentum entering identically in Specifications 2 and 4, so that the choice of n bears mainly on the absolute level of the RMSEs rather than on the measurement of the incremental macro value. The partial overlap between lags and momentum whenever $n > k$ is accepted and documented.

3.4 Models

3.4.1 Linear Specification (ARIMAX)

The linear family is an autoregressive model with exogenous variables: the target y_t is regressed on its own k lags, with k frozen by BIC on the first window (Section 3.3.3), and on the explanatory blocks measured at $t-1$. The order of differencing is set to $d = 0$, log returns being already stationary by construction.

The moving-average order is set to $q = 0$ by design rather than by selection. A moving-average term is a past forecast error, a quantity internal to the model that a feature-based learner such as XGBoost cannot receive as input. Allowing $q > 0$ would give ARIMAX information unavailable to XGBoost, and any performance gap would then mix functional form with information. Both families therefore remain equally blind to any moving-average structure, and the comparison bears on functional forms at equal information rather than on the best model attainable within each family.

The model is thus written as an ARX, the label ARIMAX being retained with q explicitly constrained. The coefficients are re-estimated at each step of the rolling window, while the order k remains frozen. Estimation relies on the `statsmodels` library (Seabold & Perktold, 2010).

3.4.2 Nonlinear Specification (XGBoost)

The nonlinear family is a gradient boosting of decision trees (T. Chen & Guestrin, 2016). It receives exactly the same columns as the ARX of its specification, with only the functional form left free (informational nesting, Section 3.3.1).

The search grid is deliberately restricted, as a mitigation of overfitting rather than a computational constraint: maximum depth $\in \{2,3,4\}$, learning rate $\in \{0.01, 0.05, 0.1\}$, row and column subsampling fixed at 0.8. These bounds contain model complexity and limit its variance on series with a low signal-to-noise ratio (López de Prado, 2018, ch. 7): trees of depth 2 to 4 capture only low-order interactions, suited to a faint signal; the retained learning rates, which are standard, impose a slow descent that early stopping offsets; subsampling at 0.8 introduces enough randomness to decorrelate the trees while preserving most of the information per iteration.

The number of trees is determined by early stopping on a chronological split internal to the window (120 weeks for fitting, 36 for validation). The hyperparameter setting is selected per specification, each specification competing at its best given its own information set, and then frozen on the first window. The four retained settings are very close, all with a learning rate of 0.1 and between 2 and 6 trees only, which rules out the risk that per-specification tuning biases the comparison of functional forms. This very small number of trees directly reflects how faint the weekly signal is: beyond that point, early stopping detects overfitting on the validation window. The details and the grid appear in Table 14.

Bagging (Random Forest), a lower-variance alternative, is set aside in favor of boosting's ability to extract a weak conditional signal; deep architectures are ruled out for lack of data.

Comparing the two families separates two possible sources of predictive power, a correctly specified linear relationship or a genuine nonlinearity. This comparison is a robustness check rather than the central contribution of the thesis, which remains the comparison between speci-

fications.

3.4.3 Interpretability

For the ARX, interpretation proceeds through the coefficients: each variable receives a single one, whose sign and magnitude give its overall effect, assumed constant from one week to the next.

Since XGBoost has no coefficients, interpretation there proceeds ex post through Shapley values (SHAP) (Lundberg & Lee, 2017), which assign to each variable, for each prediction, its marginal contribution to the outcome: a local contribution, specific to each week, where the coefficient gives a single overall effect.

These two readings serve only as diagnostics, on the full specification (Spec 4), to identify which variables weigh most and in which regime. Neither is used as a selection filter: selection remains theory-driven, so as to preserve the integrity of the comparison between nested specifications.

A shared caveat: with correlated variables, momentum and lags or co-moving macro variables, and a weak signal, attribution allocates credit in a partly arbitrary way, whether it proceeds through coefficients or through SHAP, and should be read with caution.

3.5 Market Regimes and Training Architecture

3.5.1 The Four Market Regimes

The sample is partitioned into four regimes defined by dated macro-financial events rather than by a statistical procedure. Each regime denotes a period in which Bitcoin dynamics, volatility and factor sensitivity, are structurally distinct. One point is decisive: the regimes are evaluation labels. They group the out-of-sample forecasts ex post for the conditional analysis; they never partition the training data.

- **R1, Maturation** (start of series → March 13, 2020): a market that is still thin and weakly institutionalized.
- **R2, Monetary expansion** (March 20, 2020 → March 11, 2022): a regime opened by the Federal Reserve's emergency action of March 15, 2020 (rate cut to 0–0.25% and massive asset purchases), which floods markets with liquidity.
- **R3, Tightening and FTX shock** (March 18, 2022 → January 5, 2024): opened by the first hike of the tightening cycle (FOMC of March 16, 2022), then marked by the collapse of FTX.

- **R4, Institutionalization** (January 12, 2024 $\rightarrow$ end of 2025): opened by the SEC's approval of the eleven spot Bitcoin ETFs (January 10, 2024).

The choice of event-based rather than statistical regimes (of the Markov-switching type) is deliberate. Regimes dated from primary sources are interpretable and, above all, **exogenous**: they are not estimated from the returns one seeks to predict, which rules out any circularity. Their robustness is tested by shifting the boundaries by ± 4 weeks (Section 3.6.5).

3.5.2 Fixed-Size Rolling Window

The architecture rests on a single model, continuously re-estimated on a fixed-size rolling window: for each prediction at week t, the model is trained on the interval $[t-156, t-1]$, and the window then slides by one week.

A start-up cost, the burn-in, is paid once at the beginning of the series: 168 weeks, that is, 12 weeks of feature construction, momentum requiring a history, and 156 weeks for the first window. The first out-of-sample prediction occurs at week 169 on March 25, 2016.

Three arguments justify this choice.

Feasibility. A model estimated regime by regime is structurally impossible without look-ahead: at the start of a regime, no observation of it is yet available. The fixed-size rolling window avoids this dilemma and ensures that all regimes' predictions rest on the same training volume.

Observation floor. The richest specification contains, in the worst case of the grid, 15 parameters, a constant, 4 lags and 10 explanatory variables: $156/15 \approx 10.4$ observations per parameter, above the threshold of 10 observations per parameter, a convention borrowed from the events-per-variable literature in logistic regression (Peduzzi et al., 1996) and used here as a prudential floor rather than an established optimum for either family.

Absence of a maturity gradient. An expanding window would train the first predictions on 156 weeks and the last on nearly 670, causing forecast quality to drift for a reason internal to the protocol. Since the regimes are chronological, the regime effect would be confounded with the effect of training size. Fixed size removes this gradient (López de Prado, 2018, ch. 12).

3.5.3 Choice of W, Fixed Ex Ante, and Sensitivity

The value $W = 156$ is fixed ex ante and is never adjusted on the results. It trades off two opposing forces. A short window trains the model on recent conditions, close to the regime being predicted, and preserves local adaptation. A longer window would bring more data, useful to XGBoost, but would mix in outdated dynamics and dilute the conditional signal. At 156 weeks ($\approx$ 3 years), the window exceeds the duration of the three recent regimes and therefore crosses their boundaries, while remaining short enough to adapt.

Sensitivity to this choice is tested on the grid $W \in \{130, 156, 208\}$, reported for both model families. It is a robustness display rather than a selection criterion. If the conclusions hold across the three values, the choice is shown to be robust; if they diverge, that divergence is reported as a result rather than resolved by retaining the most favorable value.

3.5.4 Development and Hold-Out Window

The development phase (2013 $\rightarrow$ end of 2025) carries all of the work: the rolling window produces there the 510 out-of-sample predictions analyzed by regime (R1: 208 points; R2: 104; R3: 95; R4: 103), from March 25, 2016 to December 26, 2025, and it is where the freezing of the structure, all the robustness checks, and the exploration of the inference take place.

The hold-out phase (year 2026) reserves a final slice evaluated once. The frozen pipeline, same rolling window and same structure, is applied there to data that were never consulted during development and could therefore influence no choice: 2026 provides an unbiased estimate of out-of-sample performance. In practice, the evaluable window contains 22 weeks (from January 2 to May 29, 2026), its upper bound being set by the availability of the Fama-French factors at execution time rather than by the calendar (a limitation discussed in Section 3.7). The use of data posterior to 2025 makes this hold-out possible without removing any regime from the conditional analysis.

The structure, namely the number of lags $k = 2$, the momentum window $n = 12$, the hyperparameters, and the size $W = 156$, is selected on the first development window and then frozen for the whole study, hold-out included (Table 15). Only the estimated quantities, the ARX coefficients, the XGBoost trees, and the mean and standard deviation of the Z-score, are re-estimated at each step of the rolling window.

3.6 Out-of-Sample Evaluation Protocol and Inference

3.6.1 Metrics

The primary metric is the RMSE, computed on the T out-of-sample forecast errors:

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

where $\hat{y}_t$ is the forecast and y_t the realization.

The out-of-sample R^2 (Campbell & Thompson, 2008) compares the model with a naive benchmark, the historical mean:

$$R^2_{\text{OOS}} = 1 - \frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{T}(y_t - \bar{y}_t)^2} \tag{3}$$

where $\bar{y}_t$ is the benchmark evaluated with the information available at $t-1$ alone: the mean return over the 156-week rolling window (primary benchmark) or the expanding mean since the beginning of the sample (complement). The benchmark therefore varies over time and itself contains no future information. The criterion is to beat the benchmark significantly rather than to reach a high R^2, the irreducible noise being large on weekly crypto returns.

Directional accuracy, the proportion of predictions correct in sign, is reported as secondary. It is derived from the same predictions, never optimized as a separate target, and carries managerial value. All metrics are computed globally and then by regime.

3.6.2 Pairwise Inference: The Corrected Diebold-Mariano Test

Comparing configurations on their RMSE indicates which predicts best, but leaves open whether the gap is real or due to sample chance. This is the role of the Diebold-Mariano test (Diebold & Mariano, 1995): week by week, it takes the differential of squared loss between two configurations,

$$d_t = e^2_{1,t} - e^2_{2,t}, \qquad \text{with } e_{i,t} = y_t - \hat{y}_{i,t} \tag{4}$$

and tests the null of equal predictive accuracy against a two-sided alternative,

$$H_0 : \mathbb{E}[d_t] = 0 \qquad \text{against} \qquad H_1 : \mathbb{E}[d_t] \neq 0 \tag{5}$$

where the sidedness is fixed ex ante with the rest of the protocol, since either configuration may in principle predict better. By convention, $\bar{d} = (1/T)\sum_{t=1}^{T} d_t > 0$ means that configuration 2 predicts better, its loss being smaller. The statistic is written

$$\text{DM} = \frac{\bar{d}}{\sqrt{\hat{V}(\bar{d})}} \tag{6}$$

where, at horizon $h = 1$, the variance is estimated as $\hat{V}(\bar{d}) = \hat{\gamma}_0/T$, with $\hat{\gamma}_0$ the empirical variance of d_t, the standard implementation at this horizon; its calibration on data of this type is checked by simulation under the null (Section 4.3). Since the test rests on the same squared loss as the RMSE, the sign of $\bar{d}$ always coincides with the RMSE ranking computed on the same set of weeks: inference only adds a layer of significance to it.

Because the asymptotic test is biased in small samples, the Harvey, Leybourne, and Newbold correction is applied uniformly (Harvey et al., 1997). It multiplies the statistic by an adjustment

factor which, at horizon $h = 1$, reduces to

$$\mathrm{DM}^{*} = \mathrm{DM} \cdot \sqrt{\frac{T-1}{T}} \tag{7}$$

and compares DM^{*} with a Student distribution with $T-1$ degrees of freedom rather than with the normal approximation. The test is run globally and then by regime, on the central comparison (Spec 2 $\rightarrow$ Spec 4) as well as on the secondary comparisons defined in Section 3.3.2. The binary verdict is complemented by the magnitude of the loss difference and its confidence interval, together with a minimum detectable effect, in order to distinguish an absence of effect from a real but undetectable one.

One property of the comparison must be stated, because it defines what the test establishes. When specifications are nested, the larger one estimates additional coefficients; if their true value is zero, these estimates fluctuate around zero from window to window and inject pure noise into the forecast. The richer specification is therefore expected to lose even in a world without signal. This thesis deliberately treats that cost as part of the object measured rather than as a bias to correct: the null tested is equality of finite-sample, real-time predictive accuracy, estimation cost included, since a practitioner exploiting the block would pay that cost too. The verdict consequently bears on the exploitability of the macro-financial block, not on the existence of a population-level relationship behind it, a distinction taken up in Section 5.3.

3.6.3 Global Comparison: The Model Confidence Set

Where the Diebold-Mariano answers a targeted pairwise question, the Model Confidence Set puts the eight configurations in simultaneous competition (Hansen et al., 2011). Let $\mathscr{M}$ denote the set of surviving configurations and $d_{ij,t}$ the loss differential between configurations i and j. At each step, the procedure tests

$$H_{0,\mathscr{M}} : \mathbb{E}[d_{ij,t}] = 0 \quad \forall i, j \in \mathscr{M} \qquad \text{against} \qquad H_{1,\mathscr{M}} : \mathbb{E}[d_{ij,t}] \neq 0 \quad \text{for some } i, j \in \mathscr{M} \tag{8}$$

When $H_{0,\mathscr{M}}$ is rejected, the worst configuration is eliminated and the test is repeated on the reduced set. The procedure stops at the first non-rejection, and the surviving set contains the best configuration with probability at least $1-\alpha$, here $\alpha = 0.10$. Inference proceeds by block bootstrap (implemented through the `arch` package (Sheppard et al., 2025), with a block length $L = 8$ fixed ex ante as $\lceil n^{1/3} \rceil$, the conventional default for block-bootstrap applications (Hall et al., 1995)).

The MCS addresses directly the risk of a chance effect when several configurations are compared, that is, multiple testing bias. A global set retaining a given configuration does not contradict a conditional result by regime: the two operate at different levels of aggregation.

3.6.4 Exclusion of In-Sample Inference

In-sample inference, an F-type test of nested restrictions on the coefficients, is excluded from the protocol for three reasons. The question posed is predictive: it is judged out of sample rather than on in-sample fit. Inference must moreover apply uniformly to both families, and a restriction test has no equivalent for XGBoost and would cover only half the design. Finally, in-sample significance does not imply out-of-sample value.

This exclusion concerns the test of the research question. The coefficients reported for diagnostic purposes in Section 5 are consequently descriptive and carry no significance statement.

3.6.5 Robustness Checks and Hold-Out

The robustness checks are carried out here rather than deferred to future work. Two are executed: sensitivity to the window size $W \in \{130, 156, 208\}$ for both families (Section 3.5.3) and a shift of the regime boundaries by ± 4 weeks. Every check executed is reported. An automated no-leakage test completes the setup: perturbing any data posterior to the prediction date leaves the prediction unchanged to the bit.

In addition, the CSSED, the cumulative sum of squared error differences, separates a steady advantage from one carried by a few extreme weeks, which a favorable Diebold-Mariano verdict could mask.

The whole protocol relies on `statsmodels` (ARX), `xgboost`, and `arch` (Diebold-Mariano, MCS); seeds, grids, and code are provided in Table 15.

3.7 Methodological Limitations

Predictive rather than causal scope. The protocol establishes an incremental predictive value rather than a causal effect: an association may reflect common determinants or indirect links. The results are therefore interpreted in terms of predictability, a causal reading requiring an identification strategy that the design does not provide.

The Peduzzi convention transposed to machine learning. The floor of ten observations per variable is a regression heuristic. Applied to XGBoost, it constitutes a prudential safeguard rather than an established optimum, since the effective complexity of a tree ensemble does not reduce to a parameter count. The convention is itself transposed for the ARX too, its original domain being logistic regression; it serves as an order-of-magnitude safeguard in both families rather than a validated bound in either.

Test power on the short regimes. On 95 to 104 points and with faint effects, the Diebold-Mariano may fail to detect a real effect. The small-sample correction, the magnitude with its confidence

interval, the minimum detectable effect, and the CSSED attenuate this risk without removing it: the conditional conclusions on the short regimes carry more uncertainty than those on R1.

Retrospective scope. The regimes are dated ex post, with hindsight on the events. The study therefore evaluates predictability conditional on known regimes rather than the ability of a system to detect them in real time: the setup is not deployable in real time.

Hold-out size. The hold-out window contains only 22 weeks, bounded by the availability of the Fama-French factors. Its function is to check that the frozen pipeline shows no gross break rather than to settle the central comparison: any strong inferential reading of this slice is ruled out in advance.

4 Results

4.1 Overview

This section tests the protocol of Section 3 against the data: the eight configurations are evaluated on the 510 out-of-sample predictions produced by the rolling window, and then grouped by regime. The presentation follows the order of the inference: overall performance, the verdict of the central comparison and its temporal robustness, the secondary comparisons, the global competition, robustness to the protocol, interpretability, and the hold-out.

4.2 Overall and Conditional Performance

Table 3 reports the overall performance of the eight configurations, grouped by family.

Table 3: Overall performance of the eight configurations

Specification	Family	RMSE	R^2 OOS	DA	MCS 90%
Baseline	ARX	0.09459	−0.0090	0.535	yes
Crypto	ARX	0.09572	−0.0331	0.553	yes
Macro	ARX	0.09734	−0.0685	0.518	yes
Combined	ARX	0.09905	−0.1064	0.520	**no**
Baseline	XGBoost	0.09425	−0.0018	0.551	yes
Crypto	XGBoost	0.09532	−0.0246	0.524	yes
Macro	XGBoost	0.09440	−0.0049	0.514	yes
Combined	XGBoost	0.09464	−0.0101	0.520	yes

Note: RMSE and out-of-sample R^2 on the 510 predictions, against the rolling mean benchmark, whose own RMSE is 0.09417. The R^2 values against the expanding benchmark, all more negative, appear in Table 16. DA: directional accuracy. Membership of the Model Confidence Set is established in Section 4.5.

The out-of-sample R^2 is negative for all eight configurations, against both benchmarks. None

beats the rolling historical mean: the closest, the XGBoost Baseline at −0.0018, reproduces the naive benchmark almost exactly without surpassing it. Since the rolling mean is the more lenient of the two benchmarks, and since that choice was fixed ex ante (Section 3.6.1), a negative R^2 against it is a fortiori negative against the other.

The best overall RMSE belongs to a Baseline specification, 0.09425 for XGBoost and 0.09459 for the ARX, and the highest to the Combined ARX (0.09905), which combines the worst RMSE with the worst R^2. Enriching the information set fails to reduce the error and tends to increase it. The finding here is descriptive; its significance is established in Section 4.3.

The RMSE falls steadily from R1 (0.109 to 0.113 depending on the configuration) to R4 (0.061 to 0.065). This decline measures the fall in Bitcoin's realized volatility rather than any growth in predictive skill, the RMSE following the scale of returns mechanically. RMSEs cannot be compared across regimes: any conditional reading proceeds through the R^2, normalized by the local variance, or through the tests of Section 4.3.

Table 4 breaks this R^2 down by regime and carries the conditional answer to the research question.

Table 4: Out-of-sample R^2 by regime (rolling benchmark)

Configuration	Global	R1	R2	R3	R4
Baseline ARX	−0.0090	0.0011	−0.0299	−0.0071	−0.0226
Crypto ARX	−0.0331	−0.0372	−0.0257	−0.0160	−0.0551
Macro ARX	−0.0685	−0.0265	−0.1417	−0.1080	−0.0792
Combined ARX	−0.1064	−0.0797	−0.1555	−0.1254	−0.1169
Baseline XGBoost	−0.0018	0.0009	0.0015	−0.0105	−0.0130
Crypto XGBoost	−0.0246	−0.0131	−0.0445	−0.0305	−0.0363
Macro XGBoost	−0.0049	−0.0004	−0.0173	−0.0128	0.0121
Combined XGBoost	−0.0101	0.0003	−0.0159	−0.0445	−0.0028

Note: out-of-sample R^2 against the rolling mean benchmark, on the 510 predictions grouped by regime (R1: 208; R2: 104; R3: 95; R4: 103). The same breakdown against the expanding benchmark appears in Table 16.

The R^2 remains negative in almost every cell. The few positive values never exceed +0.0121 and arise only for Baseline specifications in R1 or for the Macro XGBoost in R4; none is backed by a significant test (Section 4.3). Conversely, the Combined ARX is the most negative in each of the four regimes without exception. No regime exists in which macro-financial information raises the forecast.

Directional accuracy, between 0.514 and 0.553, is read against the base rate of 55.29% of positive weeks, the level a constantly bullish rule would reach, rather than against 0.5 (breakdown by regime in Table 23).

The dispersion of the predictions separates the two families sharply. That of XGBoost is two to

three times smaller than that of the ARX with the same specification, with a standard deviation of 0.007 to 0.016 against 0.014 to 0.037: predictions that are almost flat, consistent with the 2 to 6 trees retained by early stopping. Symmetrically, the Combined ARX shows the highest dispersion (0.0366): adding the macro-financial block raises the variance of the linear predictions. The interpretation of this link between enrichment, variance, and degradation is taken up in Section 5.

4.3 Central Verdict: Incremental Value of the Macro-Financial Block

The central comparison sets the Crypto specification (Spec 2) against the Combined one (Spec 4) and isolates the incremental value of the macro-financial block added to an already present crypto-native core (Section 3.3.2). By convention, a positive statistic signals that the Combined predicts better, a negative one that the Crypto predicts better.

Table 5: Corrected Diebold-Mariano, central comparison Spec 2 against Spec 4

Family	Scope	n	DM (HLN)	p	Better spec.
ARX	Global	510	-2.523	**0.012**	Crypto
ARX	R1	208	-1.354	0.177	Crypto
ARX	R2	104	-1.320	0.190	Crypto
ARX	R3	95	-2.113	**0.037**	Crypto
ARX	R4	103	-1.229	0.222	Crypto
XGBoost	Global	510	$+0.918$	0.359	Combined
XGBoost	R1	208	$+0.621$	0.536	Combined
XGBoost	R2	104	$+0.671$	0.504	Combined
XGBoost	R3	95	-0.496	0.621	Crypto
XGBoost	R4	103	$+1.146$	0.254	Combined

Note: a negative statistic favors the Crypto (Spec 2), a positive one the Combined (Spec 4). In bold, p-values below 0.05.

For the linear family, adding the macro-financial block significantly degrades the forecast at the global level: the statistic equals -2.523 ($p = 0.012$) and the confidence interval of the mean loss differential excludes zero. The sign favors the Crypto in all four regimes without exception, and the only one to reach significance is the tightening regime (R3, $p = 0.037$). The linear verdict states more than the absence of help from the macro block: within this family, it harms.

For XGBoost, no scope reaches significance. The global statistic is $+0.918$ ($p = 0.359$), and the sign itself is unstable across regimes, favoring the Combined in R1, R2, and R4 and the Crypto in R3, without any of these oscillations being distinguishable from chance. The nonlinear family, reputed to capture interactions that the linear model misses, finds no macro-financial gain and no clear degradation either.

The test module was validated before application by simulation under the null, with an empirical

rejection rate of 4.40% at $T = 95$ and 5.55% at $T = 510$, close to the nominal level of 5%. Two precautions frame these verdicts. The first concerns multiplicity across regimes. The global verdict, produced by a single test fixed a priori on 510 points, is the primary evidence. The four conditional tests run afterwards within each family provide four opportunities to cross the 5% threshold by chance: under the hypothesis that no regime-specific effect exists, the probability that at least one produces a false positive is $1 - 0.95^4 = 0.186$. The R3 result reads as corroborating the sign of the global verdict rather than as a regime-specific discovery: all four regimes already share the same sign, and R3 is the only one to cross the threshold, and only just.

The second precaution concerns power. For the non-significant verdicts, the observed effect amounts to only 0.18 to 0.48 times the minimum detectable effect: it is too small to be picked up on 95 to 104 points, and their non-significance reflects a lack of power rather than an absence of effect (Section 3.7). The global ARX verdict calls for the opposite caution: significant, but at 0.90 times the MDE, it sits at the edge of what the study can detect. Conviction about the linear degradation rests mainly on its stability across the three window sizes (Section 4.6) and on its regularity over time.

This regularity can be read on the cumulative sum of squared error differences (CSSED), which adds up week after week the loss gap between the two specifications: a steady advantage gives a continuous slope, a one-off advantage an isolated jump. For the linear family, the curve falls continuously over the whole period, with a steepening at the March 2020 boundary; for XGBoost, it stays flat around zero. The linear degradation is not the artifact of a few outliers (Figure 1).

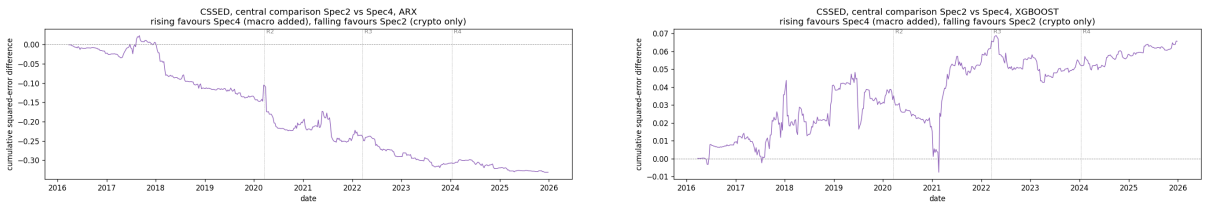


Figure 1: CSSED of the central comparison, Spec 2 against Spec 4. On the left the linear family (ARX), on the right XGBoost.

4.4 Secondary Comparisons

The three secondary comparisons (Section 3.3.2) extend the central verdict. For the linear family, neither the crypto block alone (Spec $1 \to 2$, $p = 0.097$) nor the macro block alone (Spec $1 \to 3$, $p = 0.057$) improves the forecast over the Baseline: both point toward degradation without crossing the threshold. By contrast, adding the crypto block to an already present macro core (Spec $3 \to 4$) significantly degrades the forecast ($p = 0.025$), mirroring the central verdict in which it was the addition of the macro block to a crypto core that harmed. For the ARX,

enriching the information set degrades whichever block is added, and the purely autoregressive Baseline predicts best.

For XGBoost, none of the three comparisons approaches significance (p from 0.126 to 0.713). Neither the crypto block, nor the macro block, nor their combination shifts performance in a way distinguishable from chance, consistent with the neutrality already observed on the central comparison. The details of the six tests appear in Table 17.

4.5 Global Comparison: The Model Confidence Set

The Model Confidence Set (Section 3.6.3) retains seven of the eight configurations. The only one excluded is the Combined ARX, with an MCS p-value of 0.041, below the 0.10 threshold, precisely the configuration that the central Diebold-Mariano identified as significantly the weakest. At the other end, the Baseline XGBoost obtains a p-value of 1.000, the value the best-performing configuration receives by construction, so that no other configuration outperforms it significantly. The two levels of inference, pairwise comparison and global competition, point to the same model as the only one that can be excluded with confidence, the one that adds the macro-financial block to the crypto-native core within the linear family.

The bootstrap block length, fixed ex ante at $L = 8$, is supported ex post by the automatic block-length selection algorithm of Politis and White (2004), applied in its corrected form (Patton et al., 2009). It suggests lengths from 0.85 to 6.33 depending on the configuration, the same order of magnitude as the value retained. The eight p-values appear in Table 18.

4.6 Robustness

Two checks defined ex ante (Section 3.6.5) test the central verdict: the size of the rolling window and the placement of the regime boundaries.

The verdict holds across the three windows: for the ARX, the degradation remains significant throughout, with p from 0.012 to 0.023, always in favor of the Crypto; for XGBoost, the effect stays neutral throughout. The conclusion does not depend on the choice $W = 156$, which makes it stronger than the p-value of the reference window alone would suggest.

One caveat remains: the XGBoost hyperparameters were tuned for $W = 156$ and not re-optimized for 130 and 208, so XGBoost runs at these two windows with settings no longer matched to them. This penalizes its absolute performance, but it applies to Spec 2 and Spec 4 alike and leaves their difference, which is the object of the central comparison, largely unaffected. The penalty bears on the comparison against the naive mean, which has no hyperparameters to mis-tune.

The second check shifts the regime boundaries by ± 4 weeks without retraining any model:

only the labels of the predictions change. The tightening regime (R3) is the most stable, the degradation remaining significant at all three placements (p from 0.035 to 0.037): this is the most robust conditional verdict of the study. R2 and R4 stay non-significant at all three shifts.

The maturation regime (R1) is fragile by contrast: non-significant at the reference boundary ($p = 0.177$), it becomes significant at a four-week backward shift ($p = 0.024$) and stays above the threshold with the forward shift ($p = 0.083$). This behavior is reported as a result rather than resolved by retaining the most favorable shift: choosing the boundary that makes R1 significant would amount to tuning the protocol on the results. R1 is therefore not requalified; its sensitivity to placement follows from the overlap between the 156-week window and the regime boundaries, a prediction just after a boundary being trained partly on the preceding regime.

One point bears directly on the research question: at no shift, in no regime, does a boundary placement produce a significant advantage for the macro-financial block. Where the sign points to a macro gain, it never reaches significance; where an effect is significant, it points to a degradation. The details of the two checks appear in Tables 19, 20, and 21.

4.7 Interpretability

A question distinct from predictive value remains: which variables do the models draw on most when they predict? It is answered through the standardized ARX coefficients and the XGBoost SHAP values, on the full specification (Spec 4), as a diagnostic only (Section 3.4.3).

Table 6: Global interpretability (Spec 4): ARX coefficients and SHAP importance

Variable	Mean ARX coef.	Windows with positive coef.	SHAP importance
CPI (year-over-year)	−0.0140	13%	0.00152
Hash rate	−0.0065	20%	0.00137
Fed rate (gap)	+0.0059	78%	0.00126
$R_m - R_f$	+0.0070	89%	0.00113
Momentum	−0.0021	37%	0.00113
DXY	+0.0066	59%	0.00109
SMB	−0.0038	27%	0.00089
Active addresses	+0.0015	49%	0.00085
y (lag 1)	+0.0080	81%	0.00076
HML	+0.0033	66%	0.00053
Volume	−0.0030	42%	0.00050

Note: standardized coefficients, variables standardized within each window. The column of windows with a positive coefficient measures sign stability: a value close to 0 or to 100% signals a stable sign, a value near 50% an unstable one. SHAP importance is the mean absolute contribution to the forecast. Rows sorted by decreasing SHAP importance.

An apparent convergence emerges: the CPI comes first in both rankings, with the largest coefficient magnitude in the ARX and the highest SHAP contribution in XGBoost, together with a

stable sign, positive in only 13% of the windows. The crypto-native variables, hash rate aside, show far more unstable signs, 49% for active addresses and 42% for volume.

This convergence dissolves at once, and that is what answers the research question. The primacy of the CPI in attribution translates into no predictive value: the block that contains it never predicts better, and predicts significantly worse for the ARX (Section 4.3). The two measures answer different questions: the coefficient and the SHAP value say what the model relies on when it predicts, while the Diebold-Mariano test says whether that reliance improves the forecast. The second is the one that settles the matter, and its answer stays negative.

The breakdown by regime sharpens the finding (Figures 20 and 21). The CPI coefficient is weak and unstable in R1 (-0.005), a period of low inflation, then deepens and stabilizes where inflation varies, reaching -0.029 in R3, where it is negative in every window; the SHAP importance follows the same profile, peaking in R2 and R3. The CPI dominates attribution precisely in R3, the only regime where the macro-financial block significantly degrades the forecast. Maximum attribution and minimum predictive value coincide at the same point.

A second variable shows a distinctive profile. The equity market premium carries a positive coefficient in all four regimes and one of the most stable signs in the table, 89% overall and every window in R3. Its SHAP importance, weak in the early regimes, becomes the highest of all variables in R4 (0.0022). Its economic interpretation, like that of the sign of the CPI, is left to Section 5.

Two caveats frame this reading. The SHAP magnitudes are tiny, from 0.0005 to 0.0015 in log return, consistent with the almost flat predictions of XGBoost and its two trees in Spec 4: a low importance mainly signals a variable rarely selected for a split rather than a variable of no economic content. And the weekly CPI is a step-like series, non-stationary (Section 3.2.5), which inflates its apparent persistence and hence its weight in attribution.

4.8 2026 Hold-Out

The hold-out provides an unbiased evaluation of the frozen pipeline on data never consulted during development (Section 3.5.4). Three precautions must frame the reading of its results, since omitting them would lead to erroneous conclusions.

The hold-out RMSEs, from 0.054 to 0.059, fall well below those of the development phase, around 0.095, but this signals no improvement: the realized volatility of Bitcoin declined over these 22 weeks, with a return standard deviation of 0.054 against 0.093 in development, and the RMSE follows the scale of returns mechanically. Any direct comparison between the two phases would be misleading.

Statistical power is minimal. On 22 points, the minimum detectable effect of the central comparison is six to eleven times the observed effect, with observed-to-MDE ratios of 0.16 for the

ARX and 0.09 for XGBoost. The hold-out could not settle the central comparison, a limitation anticipated in the design (Section 3.7).

Two positive out-of-sample R^2 values appear, $+0.0695$ and $+0.0538$ for the Crypto and Combined XGBoost specifications. On 22 points and with no significant test, they carry no inferential weight: they constitute neither a confirmation nor a refutation of the development verdict, both readings being unfounded for the same reason. Presenting them as a sign that macro or crypto finally works would be an over-interpretation.

With these precautions in place, the only robust finding is one of consistency. The central comparison stays non-significant in both families ($p = 0.656$ and $p = 0.803$) and its sign remains in favor of the Crypto, as in development. The frozen pipeline behaves on unseen data as it did during development, revealing no incremental macro value that the main analysis would have missed. The details appear in Tables 25 and 26.

4.9 Section Summary

The out-of-sample evaluation yields a negative result, consistent across its levels of analysis. None of the eight configurations beats the historical mean, the out-of-sample R^2 being negative everywhere and against both benchmarks. Adding the macro-financial block to the crypto-native core raises the forecast in no regime: it degrades it significantly in the linear family, at the global level ($p = 0.012$) as in the tightening regime ($p = 0.037$), and stays neutral throughout for XGBoost. The two levels of inference agree, pairwise comparison and the Model Confidence Set pointing to the same configuration, the Combined ARX, as the only one that can be excluded. The verdict holds across the three window sizes; under the boundary shift it is stable in R3 and fragile in R1, and the 2026 hold-out, too short to settle anything, only confirms the absence of a break.

These facts call for an interpretation that the present section has refrained from providing. The next section supplies it, draws the implications for practice, and sets out the limitations that bound the scope of the verdict.

5 Discussion

5.1 Answer to the Research Question and Positioning

This research asked whether the macro-financial block improves the weekly forecast of Bitcoin returns beyond the crypto-native block already present. On the 510 out-of-sample predictions, the answer is negative in all four regimes (Sections 4.3, 4.5, and 4.6). This negative result is a robust answer to a well-posed question rather than a flaw in the design: the structure frozen before any access to results, and the hold-out consumed once, rule out that it stems from a

specification chosen ex post.

The verdict extends the literature documenting a macro-crypto link rather than contradicting it, and three arguments articulate this reconciliation.

First, the claims do not bear on the same inferential object. The studies that establish a link, through Granger causality or in-sample regression, estimate their coefficients over the whole period, with hindsight on the full sample. The present thesis evaluates the real-time exploitation of a relationship whose coefficients are re-estimated on a rolling window. Estimating a coefficient has a cost, developed in Section 5.3: a signal that is real but weak can remain detectable with hindsight on the full sample and become unexploitable once that cost is paid. In-sample detection is a necessary but insufficient condition for out-of-sample forecasting.

Second, part of this literature targets volatility, price levels, or the daily horizon, whereas the claim tested here is narrow: weekly returns, out of sample, over 2013–2025. The negative verdict contradicts head-on only a small subset of existing results, those bearing on exactly this object.

Third, the gap between in-sample detection and out-of-sample forecasting is no crypto-specific singularity. Welch and Goyal (2008) show, for the equity risk premium, that almost all predictors significant in sample fail to beat the historical mean in real time; Campbell and Thompson (2008) qualify this finding by restoring weak predictability under economic constraints, without overturning it. The gap is already documented on crypto-assets themselves, where daily predictors well established in sample, trading volume among them, do not survive an out-of-sample comparison (Yae & Tian, 2022); the present result extends it to the macro-financial block and the weekly horizon.

Relative to Palazzi et al. (2026), the closest empirical reference, this thesis builds on their work rather than overturning it. Their in-sample causality called precisely for the test conducted here; the answer is that it does not carry over to out-of-sample forecasting. The move mirrors that of Welch and Goyal (2008): bounding the scope of an established link rather than refuting it.

5.2 What a Comparison Between Dominated Configurations Establishes

The most immediate objection to the central verdict targets its scope: comparing two specifications both dominated by the naive mean (Section 4.2) would amount to ranking two losers, an exercise of no interest. It does not hold, for three reasons.

First, the Diebold-Mariano test compares two series of forecast losses week by week; nowhere does it require that one of the models beat a benchmark. The question it settles, whether the macro-financial block changes forecast accuracy, is well posed whatever the absolute level of the two specifications compared.

Second, the nesting is what licenses the attribution. The Combined specification is exactly the

Crypto specification augmented by the macro-financial block, everything else identical: same lags, same window, same re-estimation, same model family. Any loss difference between the two has only one possible source, the added block, whether through its information content or through the cost of estimating its coefficients. This is an experimental ceteris paribus: if the two specifications also differed in the number of lags or in the family, the gap would no longer be attributable to the macro block alone and the inference would collapse.

Third, a concession bounds what this comparison licenses, and that is what makes it credible. The out-of-sample R^2, negative everywhere, rules out any claim of practical value, for the crypto block as much as for the macro block: on the absolute axis, no configuration recommends itself. Only incremental inference, the object of this thesis, is legitimate here; the choice of benchmark that sets this absolute axis is discussed in Section 5.6.

5.3 Three Explanations for the In-Sample / Out-of-Sample Gap

Section 4 establishes a fact, not its cause: a relationship that the literature detects in sample fails to reappear in out-of-sample forecasting. Three competing explanations can account for it. The first is the absence of signal: the in-sample causality documented by the literature would be an artifact of short samples, repeated specification search, and publication bias. The second is a real signal that is swamped: the information exists, but it falls below the cost of estimation, each added coefficient raising the cost of the forecast by roughly $1/156$ of out-of-sample R^2, so that a signal detectable with hindsight on the full sample becomes unexploitable in real time. The third is instability: the link changes sign or intensity across regimes, and a 156-week window, longer than each of the post-2020 regimes, estimates a blend of them, attenuated toward zero.

The first two are observationally indistinguishable at the available sample size. No result in Section 4 would differ in a world strictly without signal. The linear degradation is expected in both worlds: adding the six coefficients of the macro-financial block, each estimated on noise, injects variance into the forecast without reducing any bias. The neutrality of XGBoost is equally expected: its regularization refrains from estimating useless variables rather than fitting a coefficient to them, so that it does not degrade; this neutrality nevertheless remains a low-power finding, the frozen structure retaining only two to six trees (Sections 4.2 and 5.6). The primacy of the CPI in attribution, finally, would form even on pure noise, where a ranking of contributions always exists (Section 5.4). The conclusion to retain is epistemic and constitutes the interpretive contribution of this section: the data allow at most a weak signal, indistinguishable from zero at this sample size, and do not permit rejecting the pure absence of signal.

The third explanation is testable neither in this design nor in any design feasible on thirteen years of weekly data. The grid of windows tested contains no window fitting inside a regime, 130 weeks already exceeding the 104 of the longest recent regime: the stability of the verdict across the three windows (Section 4.6) is weak evidence against instability rather than a test of

it. And the alternative that would isolate it, a model estimated regime by regime, is structurally infeasible (Section 3.5.2), with only 95 to 104 points in total, below the prudential floor adopted in Section 3. With this history, no estimator is both regime-pure and adequately supplied with data.

A single indication points to a localized macro-financial effect: the fragility of the maturation regime to the boundary shift (Section 4.6). Three placements were tested there, enough multiplicity to produce a false positive. This indication is treated as a documented fragility (Section 5.6) rather than as a result.

5.4 Attribution and the Absence of Predictive Value: The Case of the CPI

The convergence noted in Section 4.7, with the CPI first on both attribution measures, would suggest, taken at face value, that inflation matters for forecasting Bitcoin. It does not show this.

Attribution measures which variables a trained model relies on to produce its outputs: it is an anatomy of the model, not a measure of the quality of its forecasts. An attribution ranking always exists, even on pure noise, since attribution distributes the output across the variables present. “The CPI comes first” therefore carries no predictive content in itself.

The convergence between the two families changes nothing. The ARX coefficients and the SHAP values are constructed independently, but they examine models fitted on the same windows and the same data: an in-sample correlation devoid of predictive value is picked up by both, and both attributions converge on it. Their agreement measures a common property of the data rather than constituting a mutual confirmation. The arbiter of usefulness remains the out-of-sample test of the block, negative everywhere and significantly unfavorable for the ARX (Section 4.3).

A mechanism makes this account coherent. The year-over-year CPI is highly persistent, to the point of being the documented exception to stationarity (Section 3.2.5): within a 156-week window, it varies so slowly that it behaves almost as a constant. Within a window, a near-constant regressor is almost collinear with the intercept, so the regression assigns the CPI a coefficient that partly absorbs the mean return of the window, in the manner of a second intercept, which improves the in-sample fit. At the following window, that mean has changed while the CPI has stayed almost identical, and the inherited coefficient misfires. This mechanism is distinct from, though rooted in the same persistence as, the Stambaugh exposure flagged in Section 3.2.2. The coincidence noted in Section 4.7 between its maximum attribution weight and the significant degradation of the block, both in R3, is the trace of this. The CPI is not the most useful variable; it is the most used one.

This finding leaves room for economic reading; it bounds its object. Two claims must be distinguished: that the macro-financial block is exploitable in forecasting, which this thesis refutes,

and that a lagged association exists in sample between macro-financial variables and Bitcoin returns, which the coefficients document. The second is a legitimate economic object, provided it rests on the statistic that carries it, the sign stability of the coefficients across re-estimations, rather than on the SHAP magnitudes, too weak and too dependent on the two retained trees to support a narrative. Sign stability is a descriptive property of the estimated coefficients, not a significance test: no standard error is attached to them here, and establishing that they differ from zero would require in-sample inference on a single fixed sample, which the current design does not provide. The economic readings that follow are therefore stated as readings of an association whose statistical significance remains untested.

Read this way, the sign of the CPI carries an economic content that deserves to be stated. Negative in 87% of the windows and deepening in R3, a period of strong variation in inflation, it indicates that within the sample high inflation precedes lower Bitcoin returns. This profile is incompatible with the inflation-hedge narrative examined in Section 2 and compatible with the opposite reading, that of a risky asset whose price would react to anticipated tightening of monetary conditions. Neither reading is established here: the design identifies a lagged association, not a mechanism (Section 3.7). A variable can co-vary with returns without that co-variation being stable or strong enough to survive the cost of its estimation (Section 5.3).

The equity market premium offers the symmetric profile. Its coefficient is stable and positive in all four regimes, and its attribution weight becomes the highest of all variables in R4 (Section 4.7). This shift after 2024 is consistent with the hypothesis of growing integration of Bitcoin into equity markets in the institutionalization regime, the hypothesis that motivated the conditional inclusion of the Fama-French block (Section 2). One caveat bounds its scope, and it is a design caveat: integration is ordinarily measured through a contemporaneous exposure, whereas the coefficients estimated here bear on the market premium lagged by one week. The profile is therefore consistent with integration without confirming it, and a dedicated test of it belongs to future work. Two variables, two economic readings, one rule: attribution documents an association, never a predictive value.

5.5 Implications for Practice

Framed from the standpoint of a manager exposed to Bitcoin, an asset manager or a hedge fund allocating at a weekly horizon, the problem is one of resource allocation: should time and resources be devoted to monitoring U.S. macro-financial indicators in order to adjust a position from one week to the next? On the basis of 510 out-of-sample forecasts, the answer is no, the macro-financial block bringing no incremental value to a setup that already contains crypto-native variables and even degrading the forecast in the linear family (Section 4.3). The implication is broader still: since none of the eight configurations beats the historical mean, crypto-native monitoring provides no more justification for departing from a passive position

at this horizon. Stated positively, this finding is consistent with weak-form efficiency of the Bitcoin market at the weekly horizon.

One safeguard bounds this implication, and it matters as much as the implication does. The claim is restricted to point forecasting of weekly returns: it says nothing about volatility, longer horizons, or risk management, for which macro-financial variables may retain a usefulness that this thesis has not examined. The category error charged in Section 2 against certain studies, which transpose to return forecasting results obtained on other objects, applies first of all to the conclusions of the present work.

A complementary economic evaluation sharpens the finding, within limits that must be stated first. Translating forecasts into positions was left unspecified in the frozen protocol, so the exercise is reported as descriptive post-processing: it converts the forecasts of Section 4 into weekly decisions without re-estimating any model, carries no test, and leaves the verdict of the main analysis untouched. Adding a metric liable to contradict a negative result, after having observed that result, would otherwise amount to choosing the measure for the story it tells, against the principle that a backtest is no research tool (López de Prado, 2018). Each week, the forecast is translated into a decision: to be invested if the predicted return is positive and to stay out otherwise, with a second rule taking a short position rather than a flat one.

Table 7: Economic evaluation of the eight configurations, global period

			Long or flat		Long or short	
Configuration	Positive fc.	DA	Cum. return	Max. drawdown	Cum. return	Max. drawdown
Buy-and-hold	100.0%	55.3%	212.4	−81.9%		
Baseline, ARX	85.3%	53.5%	28.2	−84.4%	1.2	−92.5%
Crypto, ARX	69.0%	55.3%	86.3	−76.3%	9.7	−77.6%
Macro, ARX	70.6%	51.8%	91.1	−83.0%	8.8	−88.1%
Combined, ARX	65.3%	52.0%	71.6	−85.4%	5.0	−91.5%
Baseline, XGBoost	94.3%	55.1%	120.4	−81.9%	57.7	−81.9%
Crypto, XGBoost	77.8%	52.4%	33.5	−79.5%	1.8	−90.3%
Macro, XGBoost	86.3%	51.4%	64.0	−81.9%	10.8	−86.7%
Combined, XGBoost	84.5%	52.0%	22.1	−81.8%	0.2	−95.2%

Note: 510 weeks, from March 25, 2016 to December 26, 2025. Transaction costs of 10 basis points. Positive fc.: share of weeks with a positive predicted return. DA: directional accuracy. Cumulative return expressed as a multiple of the initial capital.

None of the eight configurations comes close to buy-and-hold, whose cumulative return reaches 212.4 against 120.4 for the best strategy, and the finding holds at zero transaction cost as at twenty-five basis points; five strategies even incur a maximum drawdown equal to or greater than that of buy-and-hold while delivering a fraction of its return. On the incremental axis, the one that carries the research question, the Combined specification exceeds the Crypto in none of the twelve comparisons conducted.

Directional accuracy gives the most direct formulation of this. The point of comparison is not fifty percent but the base rate of the period, 55.29% of weeks with a rise, since that is the accuracy a permanently long position would obtain. No configuration exceeds it: the best matches it, and the other seven fall short of it by 0.2 to 3.9 points. The two evaluations nevertheless rank the configurations differently, the linear Macro standing seventh on out-of-sample R^2 and second on return, and the rank correlation, weakly negative, is indistinguishable from zero across eight configurations. A strategy depends on the sign of the forecasts and the timing of large moves rather than on mean squared error: a model ranked higher statistically is not thereby the one that best serves an investment decision.

A second implication addresses the practitioner who builds models rather than the one who uses them, and it holds in two rules. First, parsimony: every block of variables added is paid for with a systematic estimation cost while its benefit remains uncertain (Section 5.3), and the linear verdict of this thesis gives a clear illustration of it; the same hierarchy is observed elsewhere, where robust single-predictor regressions outperform more flexible machine learning methods (Yae & Tian, 2022). Second, the prohibition on using attribution as a selection argument: neither a SHAP ranking nor the magnitude of a coefficient establishes that a variable improves the forecast, and only a nested out-of-sample comparison settles it (Section 5.4). These two rules extend beyond crypto-assets; they apply to any machine learning project on financial data with a low signal-to-noise ratio.

5.6 Limitations and Threats to Validity

The limitations that follow fall into two categories. Some are design properties settled before any access to results, others are findings from robustness checks that were themselves pre-specified. None is a rationalization framed after the fact, and each is bounded in what it threatens.

A fragility of the maturation regime at the boundaries. The conditional verdict for R1 flips with boundary placement, non-significant at the reference and significant at the four-week backward shift (Section 4.6). Three placements were tested, enough multiplicity to produce chance significance: the sensitivity is reported as a documented fragility rather than as a localized macro-financial effect.

The choice of comparison benchmark. The rolling historical mean, estimated on 156 observations, is a noisy forecast and thus a relatively easy benchmark, retained for information parity with the models (Section 3.6.1). The out-of-sample R^2 nonetheless stays negative against both benchmarks, rolling and expanding (Section 4.2): losing even against the more lenient opponent is the strongest form this negative verdict can take.

A minimal complexity grid for XGBoost. Early stopping, a procedure frozen before execution, retains only two to six trees depending on the configuration (Section 4.2). This smallness is

informative in itself, since it indicates that the weekly signal supports no more complexity, but it has a counterpart: the neutrality of XGBoost is a low-power finding and cannot count as proof of an absence of effect.

Taken together, these limitations bound the scope of the verdict without eroding it: none supplies a mechanism by which the macro-financial block would have brought incremental value that the design missed. They do map out a research program, which the next section sets out.

5.7 Directions for Future Research

Extensions of this work fall into two families: those that lift a limitation of the design, and those that widen the question to objects it has not examined.

The first addresses the diagnosis stated in Section 5.3, that of a signal possibly real but below the cost of its estimation. The economic constraints of Campbell and Thompson (2008), with the sign of the coefficient fixed ex ante, the forecast bounded, and the coefficients shrunk toward zero, reduce precisely that cost. Transposing them to the design of this thesis would allow a partial separation of the absence of signal from the noise of its estimation, a distinction the present section leaves open. In the same spirit, a declining time weighting of observations, with a speed fixed before execution, would provide a continuous compromise between responsiveness and power where regime-by-regime estimation remains infeasible (Section 3.5.2).

Extension to other crypto-assets, Ether first and foremost, opens the second family: it tests whether the verdict is specific to Bitcoin or characterizes the asset class, and it loosens the power constraint through a panel, to an extent limited by the strong correlation between crypto-assets. Ether further presents an economic reason to be more sensitive to macro-financial variables: since its move to proof of stake, it carries a staking yield that makes it a cash-flow asset, exposed to real rates through a channel that Bitcoin lacks.

A second widening concerns the blocks of variables. The nesting protocol could be applied to a global macro-financial block, including a measure of world economic activity, in order to establish whether the verdict stems from the U.S. perimeter rather than from the macro-financial nature of the variables. Investor sentiment and attention, excluded from the design for scope consistency, rank among the best documented in-sample predictors of the crypto literature: on that account they constitute the most severe test of the stylized fact established here. If variables this favourably documented also fail out of sample, the gap between detection and exploitation generalizes beyond the macro-financial block alone; if they withstand it, the scope of the verdict is bounded to the block actually tested.

The out-of-sample nesting protocol could finally be applied to volatility rather than to returns. Part of the macro-crypto literature targets that variable, and nothing in the results obtained here prejudges what a test conducted on it would establish.

6 Conclusion

This thesis asked whether U.S. macro-financial indicators improve the forecast of weekly Bitcoin returns beyond crypto-native variables, and whether that incremental value depends on the market regime. The design sets four nested specifications against one another, applied identically to a linear family and to a gradient-boosted tree family, on 678 weekly observations spanning 2013 to 2025. Estimation proceeds by a 156-week rolling window whose structure is frozen before any access to results, and produces 510 out-of-sample forecasts evaluated through the small-sample-corrected Diebold-Mariano test and a Model Confidence Set.

The answer is negative, and it is negative in all four regimes. Adding the macro-financial block improves the forecast in no configuration: it degrades it significantly in the linear family and remains without effect for the trees. The two levels of inference agree on this point, and the finding holds across the three estimation windows tested. None of the eight configurations beats the historical mean of returns. These two findings answer distinct questions, the incremental value of the macro-financial block for the first, the predictive claim of the design as a whole for the second, and they converge.

The contribution of this work operates at two levels. The first is empirical. The literature documents a link between macro-financial variables and crypto-assets in abundance, but establishes it in sample, through causality tests or regressions estimated over the complete period. This thesis submits that link to the test it had not undergone, that of its real-time exploitation, and bounds its scope without refuting it. The gap that Welch and Goyal (2008) documented on the equity risk premium is found intact on weekly Bitcoin returns. The second level is methodological, and it does not depend on the result obtained. Four requirements form its core: nested specifications, a structure fixed before seeing any result, a final validation sample opened once, and two levels of inference required to agree. None is specific to Bitcoin, and the same design holds for any question of incrementality posed on noisy financial data. This is where the scope of the verdict is decided: under these constraints, finding nothing is information; without them, finding nothing says no more than that the design was weak.

This verdict is narrow, and it must remain so. It bears on one thing, the return of the week ahead. Volatility, longer horizons, and the use of these same variables to measure risk fall outside what has been tested here. For the manager exposed to Bitcoin, the finding shifts the difficulty: macro-financial information is public, abundant, and already in prices; what is missing is not access to it, but the extraction of a decision that holds when only the past is known. One question remains unsettled by this work, whether the signal is absent or merely too faint to be estimated on this sample. That question is the starting point of the extensions set out in the preceding section.

References

Abid, I. (2026). Bitcoin's short- and long-run interactions with equity markets, oil prices and US macroeconomics: a global perspective. *Studies in Economics and Finance*, *43*(4), 865–894. https://doi.org/10.1108/SEF-08-2025-0618

Ardia, D., Bluteau, K., & Rüede, M. (2019). Regime changes in bitcoin GARCH volatility dynamics. *Finance Research Letters*, *29*, 266–271. https://doi.org/10.1016/j.frl.2018.08.009

Aryan, Y., Soleimani, S., & Shojaee, A. (2025). Investigating the time-varying and conditional causality network among bitcoin, oil, gold and economic uncertainty. *Expert Systems with Applications*, *283*, 127898. https://doi.org/10.1016/j.eswa.2025.127898

Balcilar, M., Bouri, E., Gupta, R., & Roubaud, D. (2017). Can volume predict bitcoin returns and volatility? a quantiles-based approach. *Economic Modelling*, *64*, 74–81. https://doi.org/10.1016/j.econmod.2017.03.019

Baur, D. G., Dimpfl, T., & Kuck, K. (2018). Bitcoin, gold and the US dollar – a replication and extension. *Finance Research Letters*, *25*, 103–110. https://doi.org/10.1016/j.frl.2017.10.012

Baur, D. G., Hong, K., & Lee, A. D. (2018). Bitcoin: medium of exchange or speculative assets? *Journal of International Financial Markets, Institutions and Money*, *54*, 177–189. https://doi.org/10.1016/j.intfin.2017.12.004

Baur, D. G., & McDermott, T. K. (2010). Is gold a safe haven? international evidence. *Journal of Banking & Finance*, *34*(8), 1886–1898. https://doi.org/10.1016/j.jbankfin.2009.12.008

Berger, T., & Koubová, J. (2024). Forecasting bitcoin returns: econometric time series analysis vs. machine learning. *Journal of Forecasting*, *43*(7), 2904–2916. https://doi.org/10.1002/for.3165

Biais, B., Bisière, C., Bouvard, M., Casamatta, C., & Menkveld, A. J. (2023). Equilibrium bitcoin pricing. *The Journal of Finance*, *78*(2), 967–1014. https://doi.org/10.1111/jofi.13206

Bianchi, D., Guidolin, M., & Pedio, M. (2023). The dynamics of returns predictability in cryptocurrency markets. *The European Journal of Finance*, *29*(6), 583–611. https://doi.org/10.1080/1351847X.2022.2084343

Bouri, E., Molnár, P., Azzi, G., Roubaud, D., & Hagfors, L. I. (2017). On the hedge and safe haven properties of bitcoin: is it really more than a diversifier? *Finance Research Letters*, *20*, 192–198. https://doi.org/10.1016/j.frl.2016.09.025

Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: can anything beat the historical average? *Review of Financial Studies*, *21*(4), 1509–1531. https://doi.org/10.1093/rfs/hhm055

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. https://doi.org/10.1145/2939672.2939785

Chen, W., Xu, H., Jia, L., & Gao, Y. (2021). Machine learning model for bitcoin exchange rate prediction using economic and technology determinants. *International Journal of Forecasting*, *37*(1), 28–43. https://doi.org/10.1016/j.ijforecast.2020.02.008

Coin metrics. (2026). Retrieved July 17, 2026, from https://coinmetrics.io

Conlon, T., & McGee, R. (2020). Safe haven or risky hazard? bitcoin during the covid-19 bear market. *Finance Research Letters*, *35*, 101607. https://doi.org/10.1016/j.frl.2020.101607

Corbet, S., Larkin, C., Lucey, B. M., Meegan, A., & Yarovaya, L. (2020). The impact of macroeconomic news on bitcoin returns. *The European Journal of Finance*, *26*(14), 1396–1416. https://doi.org/10.1080/1351847X.2020.1737168

Corbet, S., McHugh, G., & Meegan, A. (2017). The influence of central bank monetary policy announcements on cryptocurrency return volatility. *Investment Management and Financial Innovations*, *14*(4), 60–72. https://doi.org/10.21511/imfi.14(4).2017.07

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

Dyhrberg, A. H. (2016). Bitcoin, gold and the dollar – a GARCH volatility analysis. *Finance Research Letters*, *16*, 85–92. https://doi.org/10.1016/j.frl.2015.10.008

Erfanian, S., Zhou, Y., Razzaq, A., Abbas, A., Safeer, A. A., & Li, T. (2022). Predicting bitcoin (BTC) price in the context of economic theories: a machine learning approach. *Entropy*, *24*(10), 1487. https://doi.org/10.3390/e24101487

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3–56. https://doi.org/10.1016/0304-405X(93)90023-5

Hall, P., Horowitz, J. L., & Jing, B.-Y. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika*, *82*(3), 561–574. https://doi.org/10.1093/biomet/82.3.561

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497. https://doi.org/10.3982/ECTA5771

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, *13*(2), 281–291. https://doi.org/10.1016/S0169-2070(96)00719-4

Hayes, A. S. (2019). Bitcoin price and its marginal cost of production: support for a fundamental value. *Applied Economics Letters*, *26*(7), 554–560. https://doi.org/10.1080/13504851.2018.1488040

Iyer, T. (2022). Cryptic connections: spillovers between crypto and equity markets. *Global Financial Stability Notes*, *2022*(1), 1. https://doi.org/10.5089/9781616358068.065

Liu, Y., & Tsyvinski, A. (2021). Risks and returns of cryptocurrency (I. Goldstein, Ed.). *The Review of Financial Studies*, *34*(6), 2689–2727. https://doi.org/10.1093/rfs/hhaa113

Liu, Y., Tsyvinski, A., & Wu, X. (2022). Common risk factors in cryptocurrency. *The Journal of Finance*, *77*(2), 1133–1177. https://doi.org/10.1111/jofi.13119

López de Prado, M. M. (2018). *Advances in financial machine learning*. Wiley.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions [Version Number: 2]. https://doi.org/10.48550/ARXIV.1705.07874

Nagula, P. K., & Alexakis, C. (2022). A new hybrid machine learning model for predicting the bitcoin (BTC-USD) price. *Journal of Behavioral and Experimental Finance*, *36*, 100741. https://doi.org/10.1016/j.jbef.2022.100741

Nakagawa, K., & Sakemoto, R. (2023). MACRO FACTORS IN THE RETURNS ON CRYPTOCURRENCIES. *Applied Finance Letters*, *11*, 146–158. https://doi.org/10.24135/afl.v11i.540

Nakamoto, S. (2008). Bitcoin: a peer-to-peer electronic cash system. *Cryptography Mailing list at https://metzdowd.com*.

Palazzi, R., de Souza Raimundo Júnior, G., & Klotzle, M. C. (2026, February 8). From network fundamentals to macro-financial integration: the evolving predictability of bitcoin returns. https://doi.org/10.2139/ssrn.6199098

Patton, A., Politis, D. N., & White, H. (2009). Correction to "automatic block-length selection for the dependent bootstrap" by d. politis and h. white. *Econometric Reviews*, *28*(4), 372–375. https://doi.org/10.1080/07474930802459016

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, *49*(12), 1373–1379. https://doi.org/10.1016/S0895-4356(96)00236-3

Politis, D. N., & White, H. (2004). Automatic block-length selection for the dependent bootstrap. *Econometric Reviews*, *23*(1), 53–70. https://doi.org/10.1081/ETC-120028836

Schilling, L., & Uhlig, H. (2019). Some simple bitcoin economics. *Journal of Monetary Economics*, *106*, 16–26. https://doi.org/10.1016/j.jmoneco.2019.07.002

Seabold, S., & Perktold, J. (2010). *Statsmodels: econometric and statistical modeling with python.* (Version 92–96).

Shaikh, I. (2020). Policy uncertainty and bitcoin returns. *Borsa Istanbul Review*, *20*(3), 257–268. https://doi.org/10.1016/j.bir.2020.02.003

Sheppard, K., Khrapov, S., Lipták, G., bot, S., van Hattem, R., mikedeltalima, Hammudoglu, J., Capellini, R., alejandro-cermeno, Hugle, esvhd, Fortin, A., JPN, jake, Judell, M., Russell, R., Li, W., partev, 645775992, … RENE-CORAIL, X. (2025, October 21). *Bashtage/arch: release v8.0.0* (Version v8.0.0). Zenodo. https://doi.org/10.5281/ZENODO.593254

Sihombing, S., Sahputri, R. A. M., Ali, H., Njoman, M. G., Lumbantoruan, R. F., & Syafitri, E. (2025). Determinants of bitcoin returns: an analysis of bitcoin information, macroeconomics, and other cryptocurrency markets. *Jurnal Ilmu Keuangan dan Perbankan (JIKA)*, *15*(1). https://doi.org/10.34010/jika.v15i1.15753

Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics*, *54*(3), 375–421. https://doi.org/10.1016/S0304-405X(99)00041-0

Valadkhani, A. (2024). Asymmetric causality between bitcoin and tech stocks in the US market using mixed frequency data. *Journal of Economic Studies*, *51*(3), 569–586. https://doi.org/10.1108/JES-05-2023-0231

Walther, T., Klein, T., & Bouri, E. (2019). Exogenous drivers of bitcoin and cryptocurrency volatility – a mixed data sampling approach to forecasting. *Journal of International Financial Markets, Institutions and Money*, *63*, 101133. https://doi.org/10.1016/j.intfin.2019.101133

Wang, J., Ma, F., Bouri, E., & Guo, Y. (2023). Which factors drive bitcoin volatility: macroeconomic, technical, or both? *Journal of Forecasting*, *42*(4), 970–988. https://doi.org/10.1002/for.2930

Wang, L., Sarker, P. K., & Bouri, E. (2023). Short- and long-term interactions between bitcoin and economic variables: evidence from the US. *Computational Economics*, *61*(4), 1305–1330. https://doi.org/10.1007/s10614-022-10247-5

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, *21*(4), 1455–1508. https://doi.org/10.1093/rfs/hhm014

Yae, J., & Tian, G. Z. (2022). Out-of-sample forecasting of cryptocurrency returns: a comprehensive comparison of predictors and algorithms. *Physica A: Statistical Mechanics and its Applications*, *598*, 127379. https://doi.org/10.1016/j.physa.2022.127379

A Methodological Appendix

A.1 Target Variable

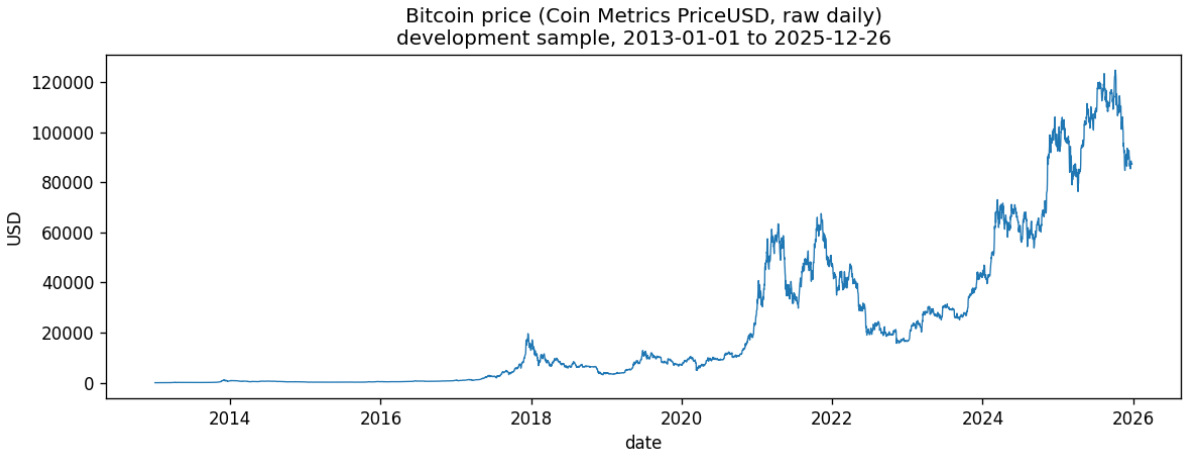


Figure 2: Bitcoin price, weekly closing, 2013–2025.

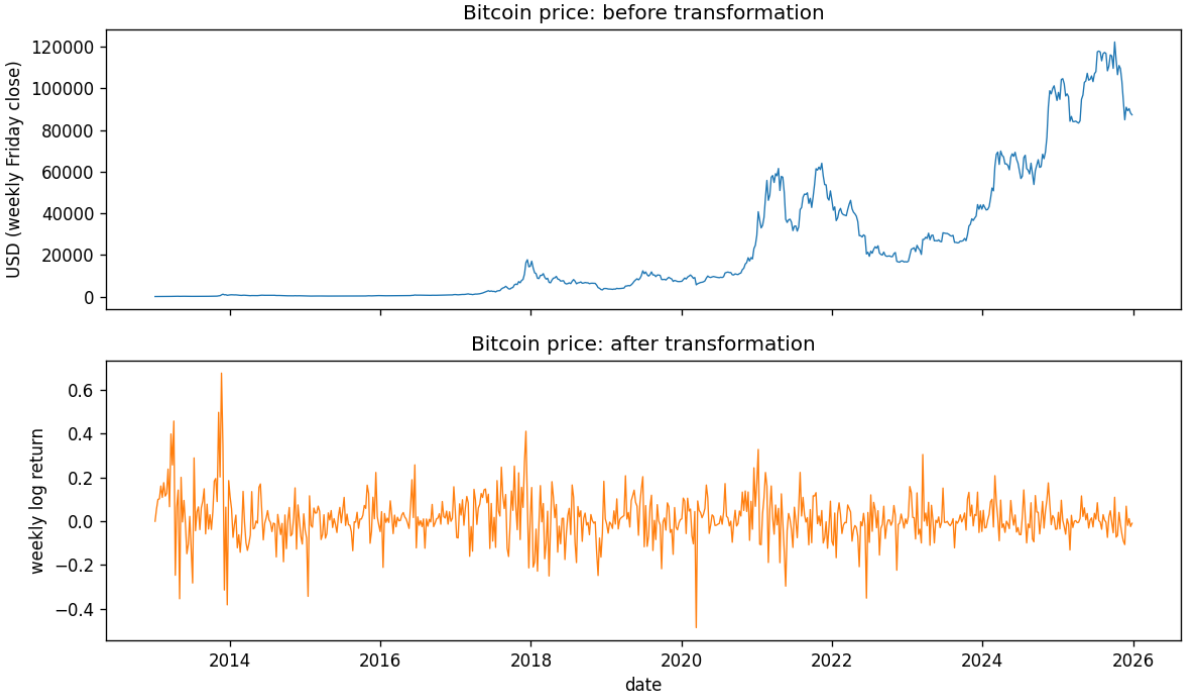


Figure 3: Target variable, weekly log-return $y_t = \ln(P_t/P_{t-1})$.

A.2 Macro-Financial Block

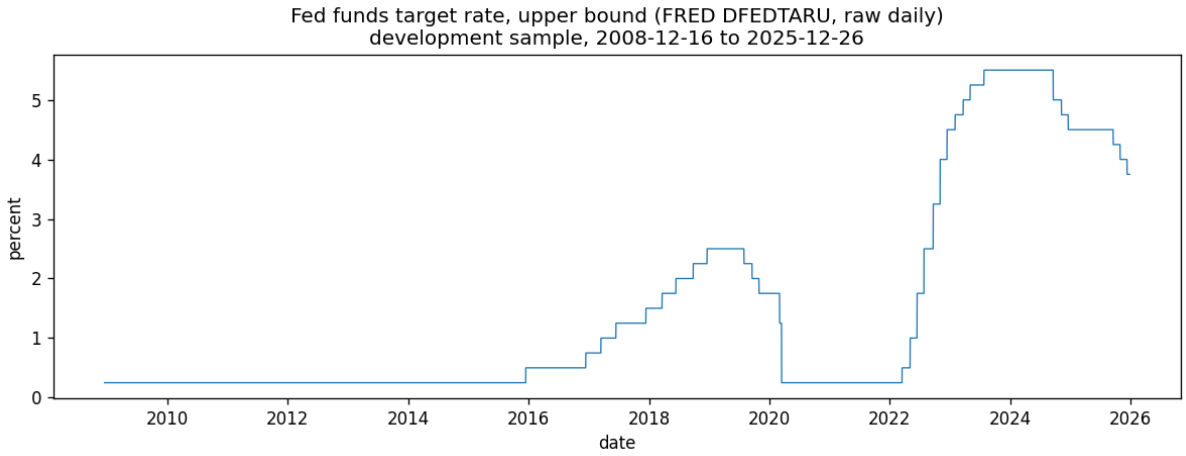


Figure 4: Federal funds target rate, upper bound, weekly level.

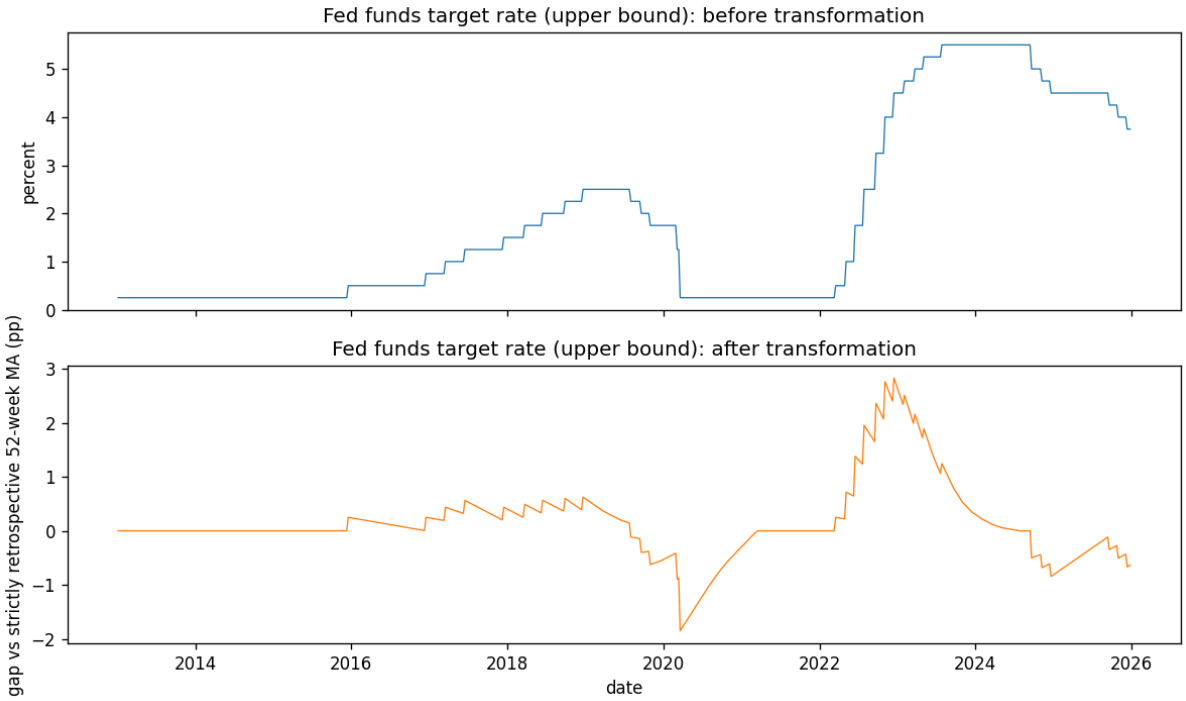


Figure 5: Federal funds rate gap, deviation from its 52-week moving average.

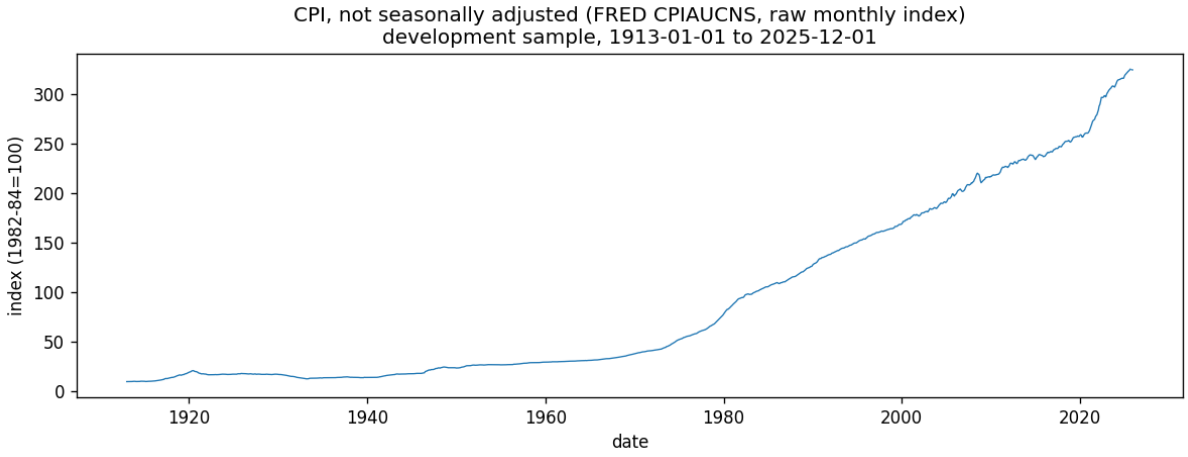


Figure 6: US CPI, not seasonally adjusted, monthly level.

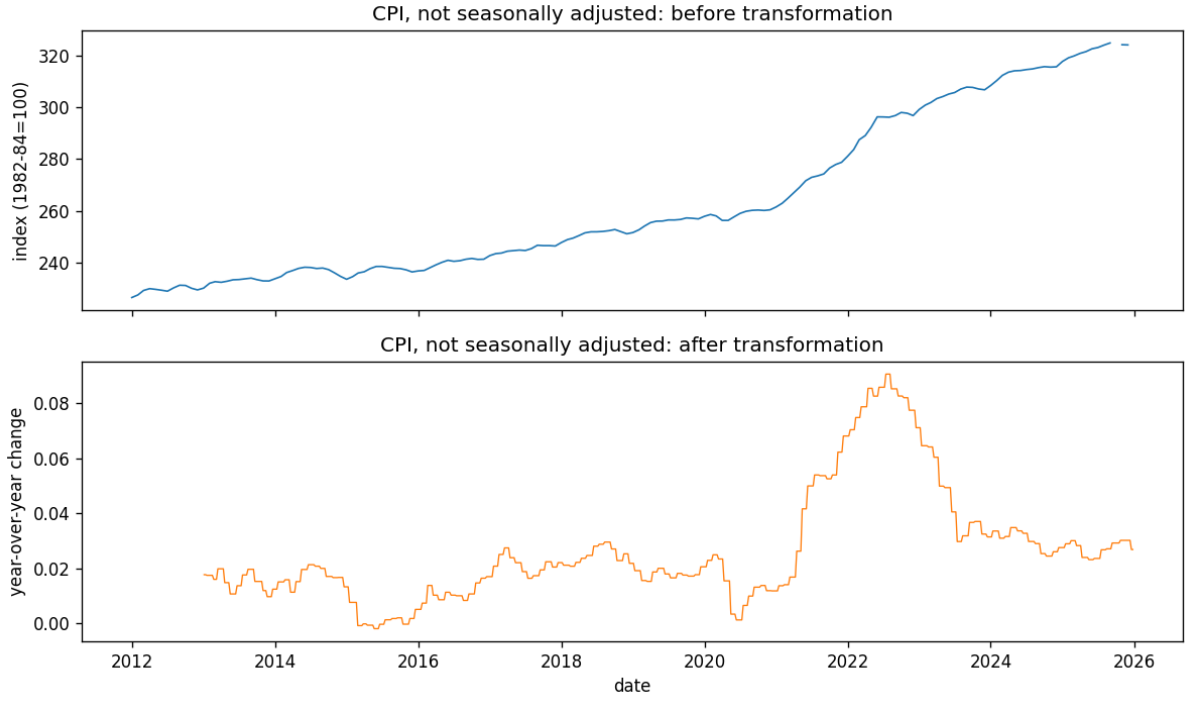


Figure 7: CPI, year-over-year change.

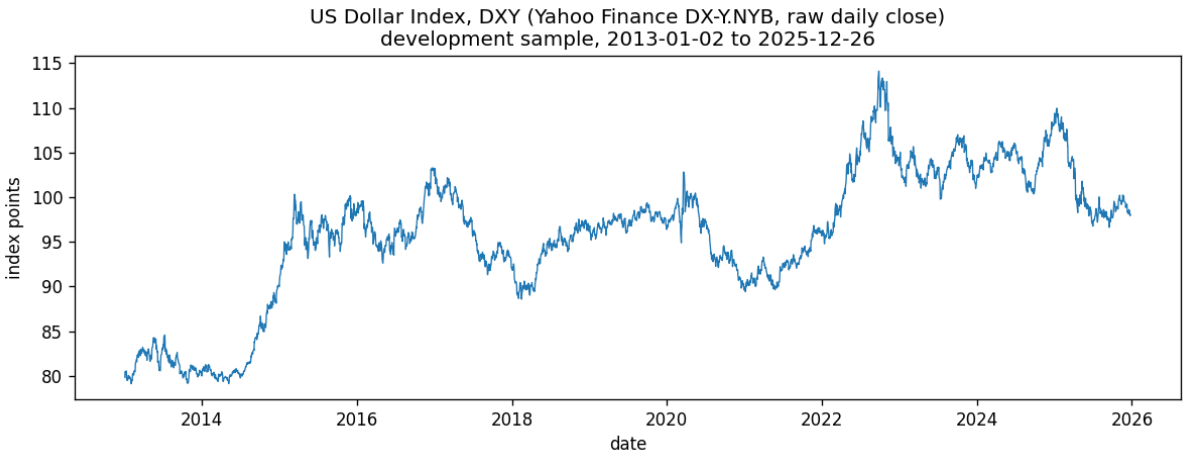


Figure 8: US Dollar Index (DXY), weekly level.

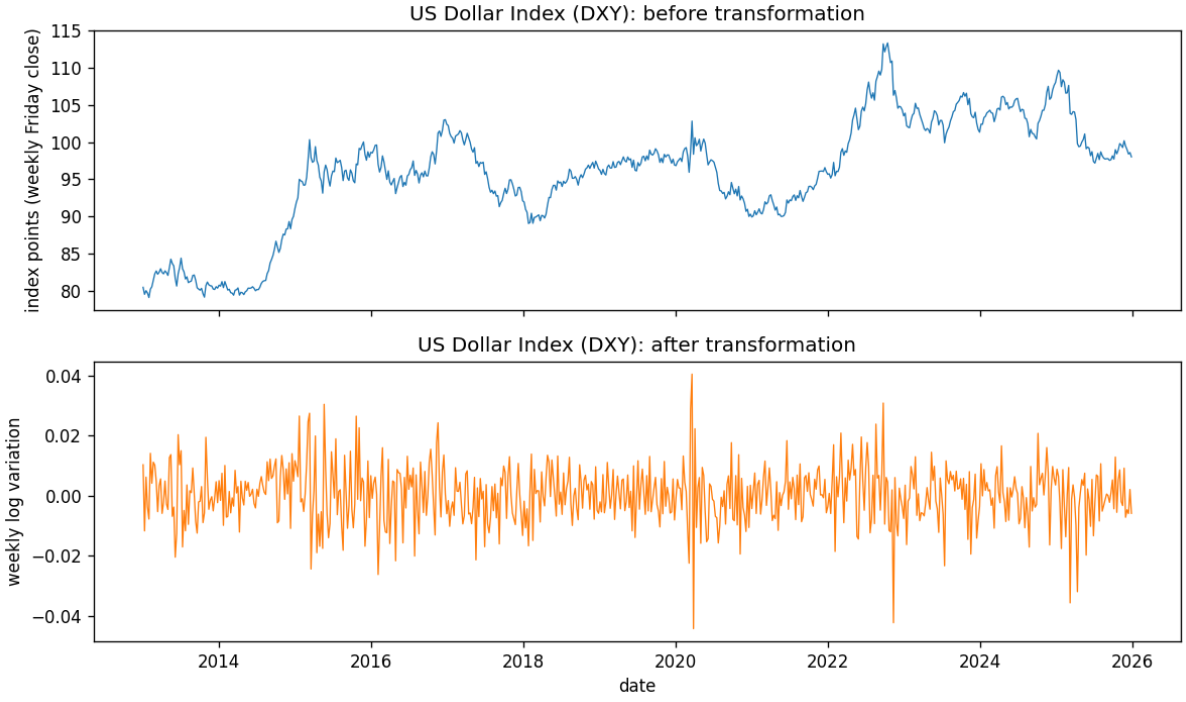


Figure 9: US Dollar Index (DXY), weekly log-variation.

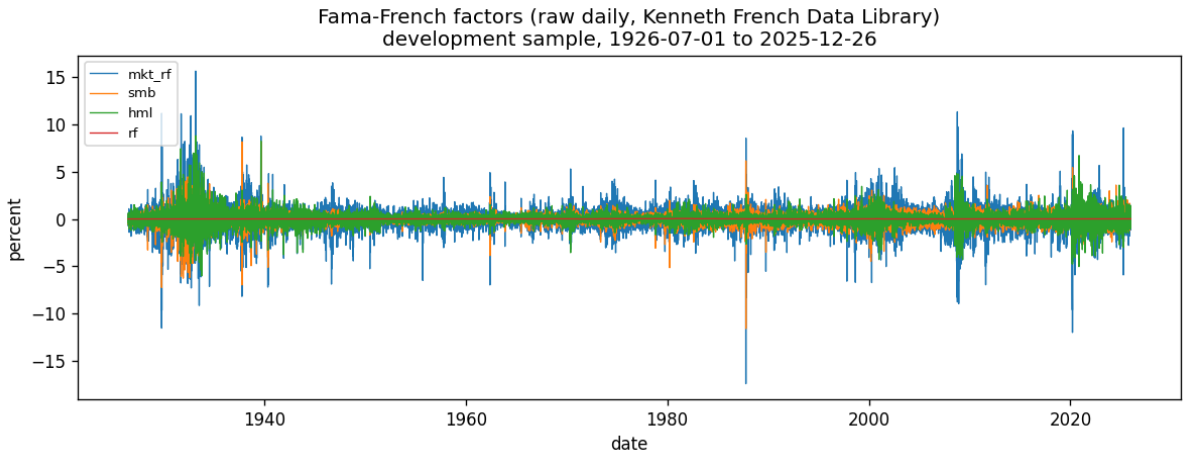


Figure 10: Fama-French three factors ($R_m - R_f$, SMB, HML), weekly returns.

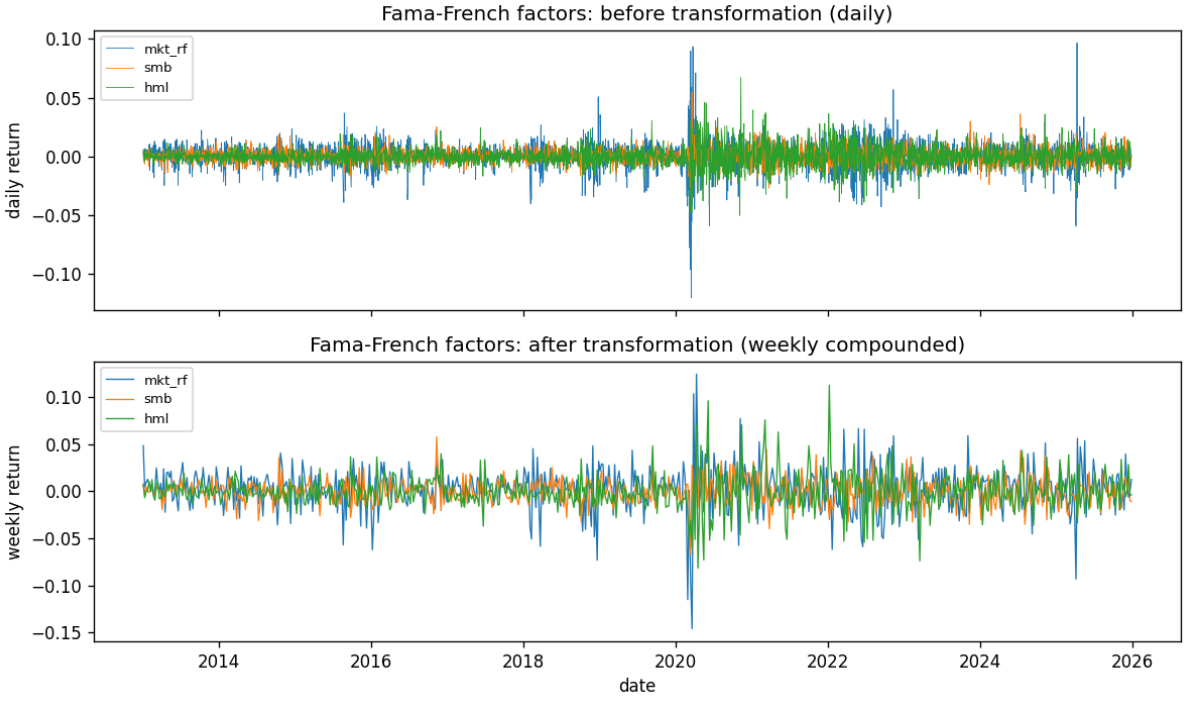


Figure 11: Fama-French three factors ($R_m - R_f$, SMB, HML), as entered in the specification.

A.3 Crypto-Native Block

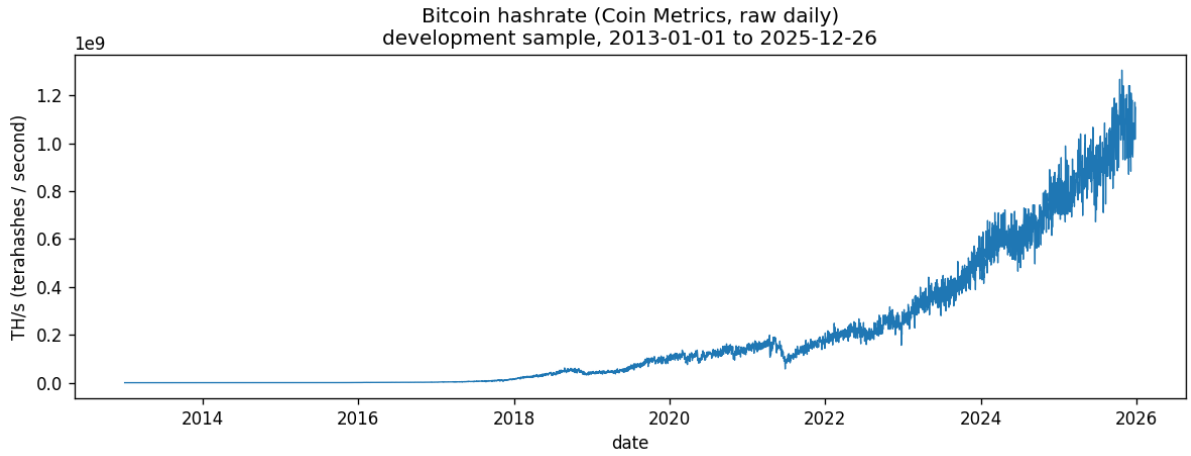


Figure 12: Bitcoin network hash rate, weekly average.

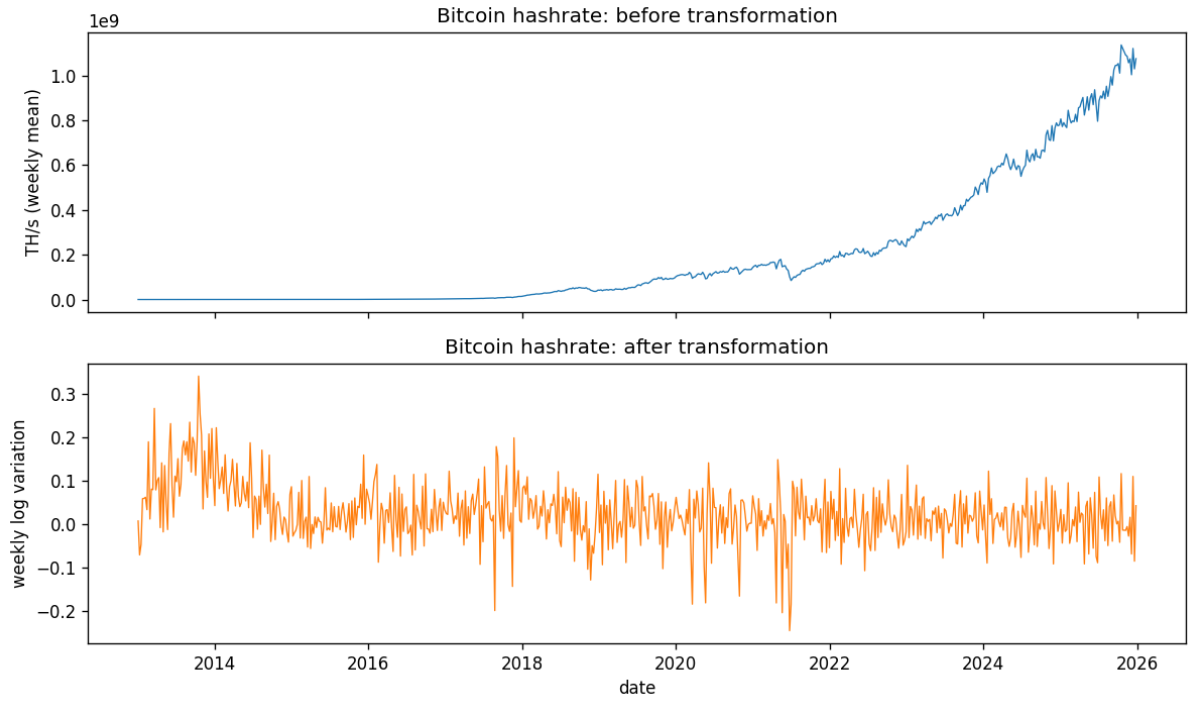


Figure 13: Bitcoin network hash rate, weekly log-variation.

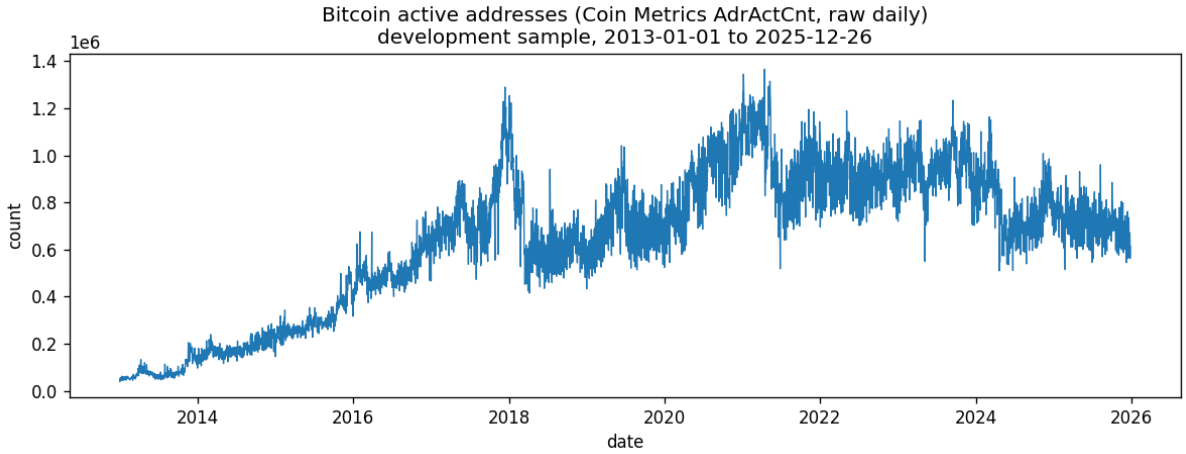


Figure 14: Bitcoin active addresses, weekly average.

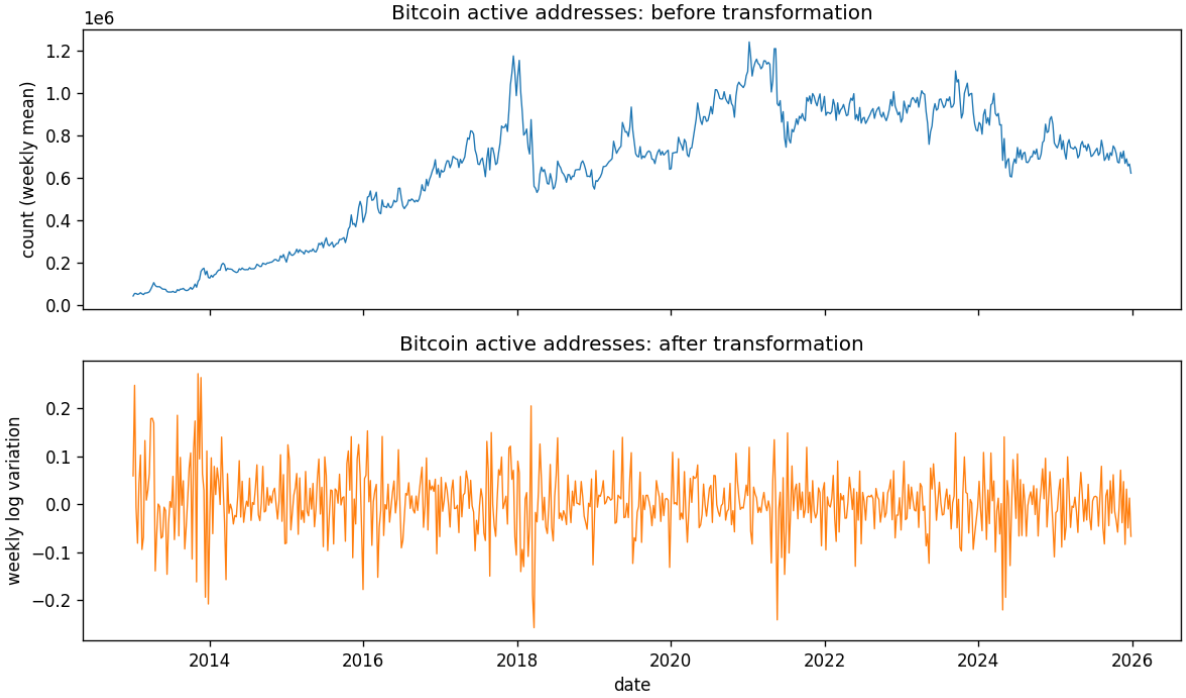


Figure 15: Bitcoin active addresses, weekly log-variation.

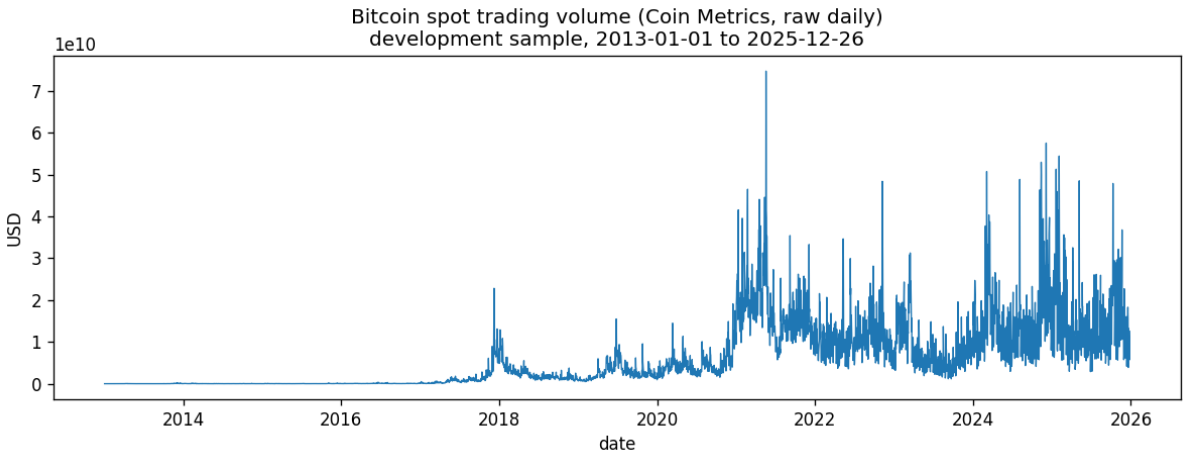


Figure 16: Bitcoin trading volume, weekly sum.

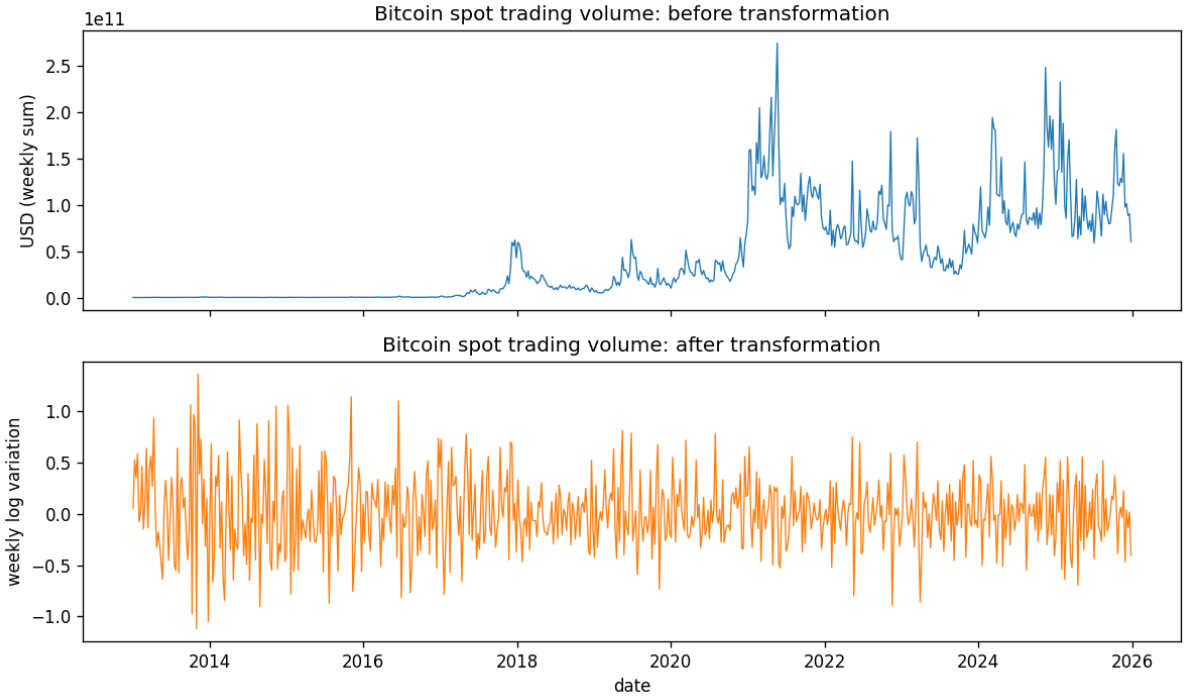


Figure 17: Bitcoin trading volume, weekly log-variation.

A.4 Descriptive Statistics

Table 8: Descriptive statistics of the ten model variables

Variable	n	Mean	Std. dev.	Skew	Kurtosis	Min	Max
y	678	0.0129	0.1054	0.3783	5.0083	−0.4857	0.6760
FedGap	678	0.1572	0.7075	1.2295	3.4195	−1.8510	2.8317
CPI (YoY)	678	0.0266	0.0207	1.4606	1.6224	−0.0020	0.0906
DXY (logvar)	678	0.0003	0.0095	−0.1892	2.0110	−0.0442	0.0404
hash rate (logvar)	678	0.0260	0.0669	0.2742	2.0729	−0.2452	0.3421
Addresses (logvar)	678	0.0040	0.0649	−0.0131	1.9876	−0.2572	0.2725
Volume (logvar)	678	0.0150	0.3561	0.2108	0.6293	−1.1186	1.3634
$R_m - R_f$	678	0.0026	0.0230	−0.5741	5.8321	−0.1457	0.1240
SMB	678	−0.0003	0.0134	0.0772	1.5631	−0.0671	0.0574
HML	678	−0.0002	0.0186	0.3997	5.1129	−0.0815	0.1122

A.5 Correlation Matrix of the Predictors

Table 9: Correlation matrix of the nine predictors (momentum excluded, validated separately)

	FedGap	CPI	DXY	Hashr.	Addr.	Vol.	$R_m - R_f$	SMB	HML
FedGap	1.00	0.58	0.00	−0.02	−0.02	−0.01	−0.04	−0.03	0.00
CPI	0.58	1.00	0.04	−0.14	−0.06	−0.03	−0.07	−0.05	0.03
DXY	0.00	0.04	1.00	−0.07	−0.01	−0.03	−0.24	−0.10	−0.01
hash rate	−0.02	−0.14	−0.07	1.00	0.21	0.00	0.10	−0.02	0.08
Addresses	−0.02	−0.06	−0.01	0.21	1.00	0.41	0.08	−0.01	−0.01
Volume	−0.01	−0.03	−0.03	0.00	0.41	1.00	−0.03	−0.02	−0.05
$R_m - R_f$	−0.04	−0.07	−0.24	0.10	0.08	−0.03	1.00	0.25	−0.02
SMB	−0.03	−0.05	−0.10	−0.02	−0.01	−0.02	0.25	1.00	−0.01
HML	0.00	0.03	−0.01	0.08	−0.01	−0.05	−0.02	−0.01	1.00

Note: highest off-diagonal value, FedGap/CPI (0.58), expected given both derive from monetary and price conditions. No other pair exceeds 0.41 (Addresses/Volume).

A.6 Stationarity Tests (ADF/KPSS), Full Detail

Table 10: ADF and KPSS stationarity tests, all variables

Variable	n	ADF stat.	ADF p	KPSS stat.	KPSS p	Stationary (both)
y	678	-15.2176	< 0.0001	0.2619	≥ 0.10	yes
FedGap	678	-3.1135	0.0256	0.2066	≥ 0.10	yes
CPI (YoY)	678	-1.8999	0.3321	1.5133	≤ 0.01	**no**
DXY (logvar)	678	-27.6150	< 0.0001	0.1641	≥ 0.10	yes
hash rate (logvar)	678	-3.1722	0.0216	1.7833	≤ 0.01	**no**
Addresses (logvar)	678	-9.5379	< 0.0001	0.8364	≤ 0.01	**no**
Volume (logvar)	678	-16.7497	< 0.0001	0.3273	≥ 0.10	yes
$R_m - R_f$	678	-28.5671	< 0.0001	0.0406	≥ 0.10	yes
SMB	678	-27.3753	< 0.0001	0.0802	≥ 0.10	yes
HML	678	-8.0035	< 0.0001	0.0969	≥ 0.10	yes
CPI (YoY, monthly diagnostic)	156	-2.0413	0.2688	0.7393	≤ 0.01	**no**

Note: a series is treated as stationary only when both tests agree. Three exceptions are documented rather than further transformed: CPI (YoY), hash rate and active addresses (log-variation). The last row applies the same tests to the raw monthly CPI series (about 156 points, no weekly forward-fill); its KPSS statistic falls from 1.5133 (weekly) to 0.7393 (monthly), indicating the weekly forward-fill inflates apparent persistence, but the monthly series still fails both tests: the non-stationarity is largely intrinsic to the 2021–2023 inflation episode, not purely a forward-fill artifact.

A.7 Predictive Validation of Constructed Variables

Table 11: Predictive versus contemporaneous correlation with the target

Variable	n	Contemporaneous corr.	Predictive corr. $(t-1)$	Expected sign
Momentum (8 wk)	670	0.4057	0.1259	positive, confirmed
Momentum (12 wk)	666	0.3260	0.1003	positive, confirmed
hash rate (logvar)	677	0.1337	0.0282	positive, confirmed
Addresses (logvar)	677	0.2800	0.0695	positive, confirmed
Volume (logvar)	677	0.1751	0.0480	positive, confirmed

Note: predictive correlation uses the variable at $t-1$ against the return at t; contemporaneous correlation uses both at t. All five variables carry the expected sign, and the sharp drop from contemporaneous to predictive correlation illustrates the modest weekly predictability the study operates under.

A.8 Lag Order and Momentum Window Selection (BIC)

Table 12: BIC selection of the autoregressive lag order k

k	Parameters (incl. constant)	BIC
1	2	−176.645
2	3	−177.655
3	4	−173.231
4	5	−170.891

Note: selected on the Baseline specification, first training window.

Table 13: BIC selection of the momentum window n

n	Parameters (incl. constant)	BIC
8	7	−162.394
12	7	−162.474

Note: selected on Spec 2, k frozen. Margin over $n = 8$ is 0.08 BIC points, documented as a limitation rather than a clean-cut decision (Section 3.3.3).

A.9 XGBoost Hyperparameter Grid

Table 14: XGBoost validation grid, all 36 combinations, four specifications

Spec	Max depth	Learning rate	Trees (early stop)	Validation RMSE
Baseline	2	0.01	14	0.076780
Baseline	2	0.05	2	0.076325
Baseline	**2**	**0.10**	**2**	**0.075784**
Baseline	3	0.01	6	0.076931
Baseline	3	0.05	1	0.076766
Baseline	3	0.10	1	0.076658
Baseline	4	0.01	6	0.076952
Baseline	4	0.05	1	0.076969
Baseline	4	0.10	1	0.077105
Crypto	2	0.01	15	0.076961
Crypto	2	0.05	2	0.077432
Crypto	2	0.10	7	0.077796
Crypto	3	0.01	8	0.076931
Crypto	3	0.05	2	0.076869
Crypto	3	0.10	2	0.076847
Crypto	4	0.01	20	0.076724
Crypto	4	0.05	4	0.076822
Crypto	**4**	**0.10**	**6**	**0.075496**
Macro	2	0.01	28	0.076275
Macro	2	0.05	1	0.076232
Macro	**2**	**0.10**	**4**	**0.075605**
Macro	3	0.01	27	0.076244
Macro	3	0.05	1	0.076232
Macro	3	0.10	1	0.075700
Macro	4	0.01	27	0.076385
Macro	4	0.05	4	0.076546
Macro	4	0.10	1	0.076444
Combined	2	0.01	2	0.076882
Combined	2	0.05	2	0.076365
Combined	2	0.10	2	0.076004
Combined	3	0.01	2	0.076870
Combined	3	0.05	2	0.076421
Combined	3	0.10	2	0.076160
Combined	4	0.01	6	0.076775
Combined	4	0.05	2	0.076083
Combined	**4**	**0.10**	**2**	**0.075670**

Note: subsample and column subsample fixed at 0.8 for all combinations. Selected configuration per specification in bold, lowest validation RMSE, all at learning rate 0.10.

A.10 Frozen Structure, Seeds, and Reproducibility

Table 15: Frozen structure and reproducibility parameters

Element	Frozen value
Lag order k	2
Momentum window n	12
Rolling window W	156
ARX order	$d = 0$, $q = 0$, AR order 2
XGBoost grid	max_depth $\in \{2,3,4\}$, learning_rate $\in \{0.01, 0.05, 0.10\}$, subsample/colsample 0.8
Early stopping	fit weeks 13–132 (120 obs.), validation 133–168 (36 obs.), 50 rounds, cap 2000 trees
Seed scheme	$20260718 + 1000 \times \text{window} + 10 \times \text{spec} + \text{family code}$ (ARX = 1, XGBoost = 2)
Frozen config hash (SHA-256)	`5abad0db74c692a1ab6455a722d33d3e6d8daaec8e6cd7028e8cb892d291cf92`

B Results Appendix

B.1 Out-of-Sample Predictions

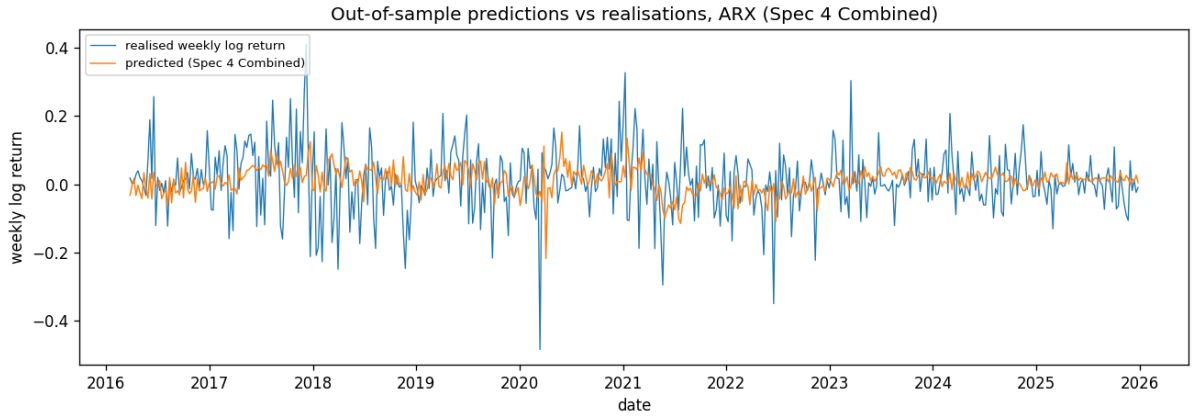


Figure 18: Out-of-sample predictions versus realized weekly log-returns, ARX, Spec 4 (Combined).

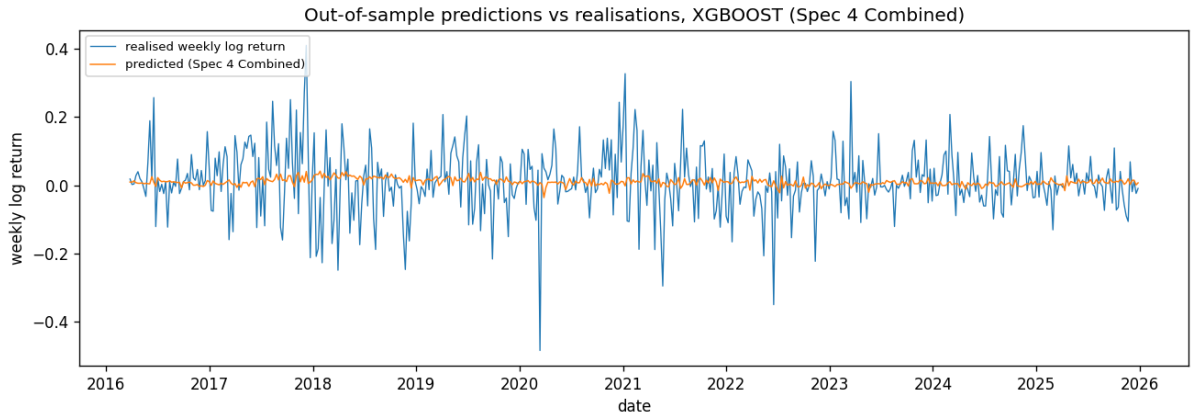


Figure 19: Out-of-sample predictions versus realized weekly log-returns, XGBoost, Spec 4 (Combined).

B.2 Interpretability by Model, Full Regime Breakdown

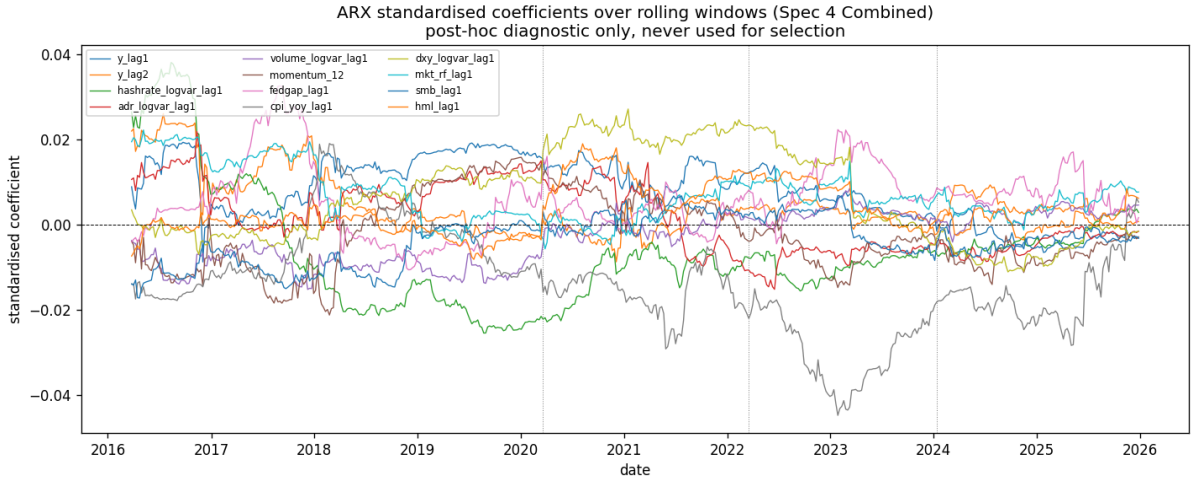


Figure 20: ARX standardized coefficients by regime, Spec 4.

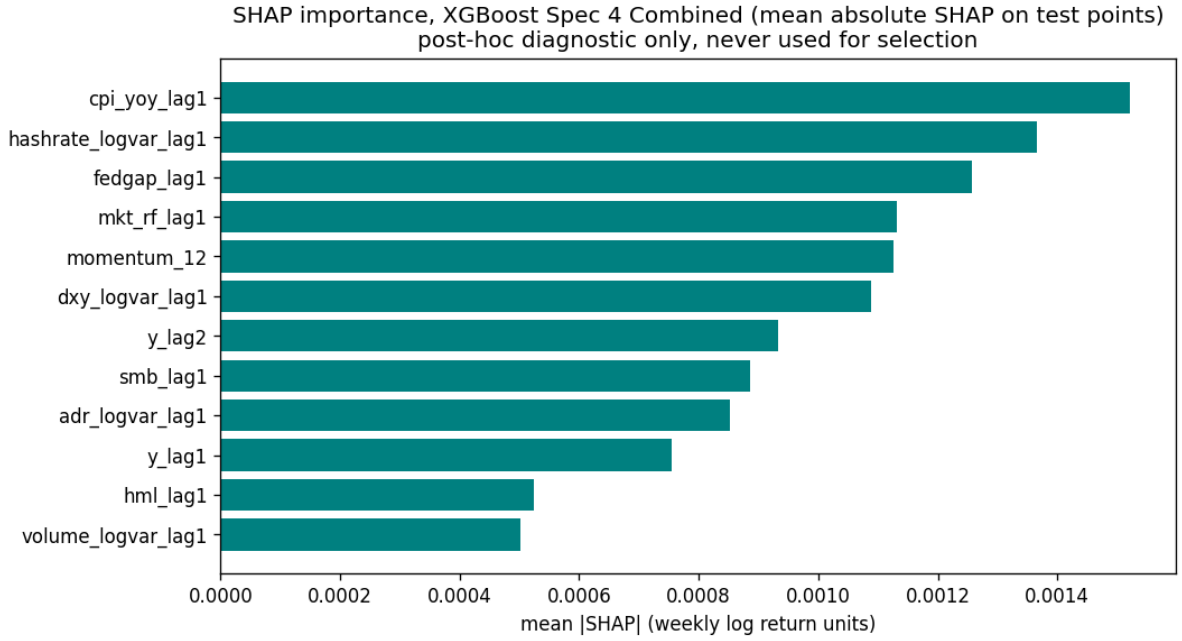


Figure 21: SHAP importance (mean absolute contribution) by regime, XGBoost Spec 4.

B.3 R^2 OOS Against the Expanding Benchmark, by Regime

Table 16: R^2 out-of-sample against the expanding benchmark, by regime

Configuration	Global	R1	R2	R3	R4
Baseline ARX	−0.0244	−0.0160	−0.0632	+0.0102	−0.0402
Crypto ARX	−0.0489	−0.0549	−0.0589	+0.0014	−0.0732
Macro ARX	−0.0848	−0.0441	−0.1786	−0.0890	−0.0978
Combined ARX	−0.1233	−0.0981	−0.1929	−0.1061	−0.1360
Baseline XGBoost	−0.0171	−0.0161	−0.0308	+0.0069	−0.0304
Crypto XGBoost	−0.0402	−0.0304	−0.0783	−0.0128	−0.0541
Macro XGBoost	−0.0202	−0.0175	−0.0502	+0.0046	−0.0049
Combined XGBoost	−0.0255	−0.0168	−0.0487	−0.0265	−0.0200

Note: complements Table 4 in the body, which reports the same breakdown against the rolling benchmark.

B.4 Secondary Comparisons, Full Detail

Table 17: Secondary comparisons (corrected Diebold-Mariano, global)

Comparison	Family	DM (HLN)	p	Signif. 5%	Better spec.
Spec 1 → 2	ARX	−1.665	0.097	no	Baseline
Spec 1 → 2	XGBoost	−1.531	0.126	no	Baseline
Spec 1 → 3	ARX	−1.910	0.057	no	Baseline
Spec 1 → 3	XGBoost	−0.368	0.713	no	Baseline
Spec 3 → 4	ARX	−2.255	0.025	yes	Macro
Spec 3 → 4	XGBoost	−0.502	0.616	no	Macro

Note: Spec 1 → 2 measures the value of the crypto-native block alone; Spec 1 → 3, the macro-financial block alone; Spec 3 → 4, the incremental value of crypto-native beyond macro-financial (mirror of the central comparison).

B.5 Model Confidence Set, Full Results

Table 18: Model Confidence Set at 90%, all eight configurations

Configuration	MCS p-value	Retained (90%)
Baseline XGBoost	1.000	yes
Baseline ARX	0.800	yes
Macro XGBoost	0.800	yes
Combined XGBoost	0.696	yes
Crypto ARX	0.335	yes
Crypto XGBoost	0.319	yes
Macro ARX	0.251	yes
Combined ARX	0.041	**no**

Note: block bootstrap, block length $L = 8$ fixed ex ante (Section 3.6.3). Politis-White rule suggests block lengths from 0.85 to 6.33 depending on configuration, the same order of magnitude as $L = 8$, a mildly conservative but non-alarming choice.

B.6 Robustness to Rolling-Window Size

Table 19: Sensitivity to rolling-window size (central comparison Spec 2 $\rightarrow$ Spec 4)

W	Family	n	DM (HLN)	p	Signif. 5%	Better spec.
130	ARX	536	-2.510	0.012	yes	Crypto
156	ARX	510	-2.523	0.012	yes	Crypto
208	ARX	458	-2.283	0.023	yes	Crypto
130	XGBoost	536	$+0.729$	0.467	no	Combined
156	XGBoost	510	$+0.918$	0.359	no	Combined
208	XGBoost	458	$+0.393$	0.695	no	Combined

Note: XGBoost hyperparameters were tuned for $W = 156$ and not re-optimized for 130 or 208. Regime composition shifts with W: at $W = 130$, R1 absorbs the difference (234 points) while R2–R4 stay at 104, 95, 103; at $W = 208$, R1 falls to 156 points, R2–R4 unchanged.

B.7 Robustness to Regime Boundary Placement, Full Detail

Table 20: Regime boundary shift, ARX family (central comparison Spec 2 → Spec 4)

Shift	Regime	n	DM (HLN)	p	Better spec.
−4 wk	R1	204	−2.281	0.024	Crypto
−4 wk	R2	104	−0.830	0.409	Crypto
−4 wk	R3	95	−2.143	0.035	Crypto
−4 wk	R4	107	−1.094	0.277	Crypto
reference	R1	208	−1.354	0.177	Crypto
reference	R2	104	−1.320	0.190	Crypto
reference	R3	95	−2.113	0.037	Crypto
reference	R4	103	−1.229	0.222	Crypto
+4 wk	R1	212	−1.744	0.083	Crypto
+4 wk	R2	104	−0.861	0.391	Crypto
+4 wk	R3	95	−2.117	0.037	Crypto
+4 wk	R4	99	−1.241	0.218	Crypto

Table 21: Regime boundary shift, XGBoost family (central comparison Spec 2 → Spec 4)

Shift	Regime	n	DM (HLN)	p	Better spec.
−4 wk	R1	204	+0.692	0.490	Combined
−4 wk	R2	104	+0.488	0.626	Combined
−4 wk	R3	95	−0.227	0.821	Crypto
−4 wk	R4	107	+1.044	0.299	Combined
reference	R1	208	+0.621	0.536	Combined
reference	R2	104	+0.671	0.504	Combined
reference	R3	95	−0.496	0.621	Crypto
reference	R4	103	+1.146	0.254	Combined
+4 wk	R1	212	+0.565	0.573	Combined
+4 wk	R2	104	+0.886	0.378	Combined
+4 wk	R3	95	−0.880	0.381	Crypto
+4 wk	R4	99	+1.160	0.249	Combined

Note: no XGBoost cell reaches significance at 5%, at any shift.

B.8 RMSE by Configuration and Regime

Table 22: RMSE by configuration and regime

Configuration	Global	R1	R2	R3	R4
Baseline ARX	0.09459	0.10895	0.09964	0.08348	0.06231
Crypto ARX	0.09572	0.11102	0.09944	0.08385	0.06329
Macro ARX	0.09734	0.11045	0.10491	0.08756	0.06401
Combined ARX	0.09905	0.11327	0.10555	0.08825	0.06511
Baseline XGBoost	0.09425	0.10896	0.09811	0.08362	0.06201
Crypto XGBoost	0.09532	0.10972	0.10035	0.08444	0.06272
Macro XGBoost	0.09440	0.10903	0.09903	0.08371	0.06124
Combined XGBoost	0.09464	0.10899	0.09896	0.08501	0.06170

Note: supports the claim in Section 4.2 that RMSE decreases regularly from R1 to R4, reflecting the decline in realized Bitcoin volatility rather than growing predictive skill.

B.9 Directional Accuracy by Regime

Table 23: Directional accuracy by configuration and regime

Configuration	Global	R1	R2	R3	R4
Baseline ARX	0.535	0.567	0.510	0.547	0.485
Crypto ARX	0.553	0.558	0.558	0.642	0.456
Macro ARX	0.518	0.538	0.481	0.537	0.495
Combined ARX	0.520	0.538	0.500	0.526	0.495
Baseline XGBoost	0.551	0.582	0.596	0.484	0.505
Crypto XGBoost	0.524	0.538	0.538	0.505	0.495
Macro XGBoost	0.514	0.548	0.490	0.474	0.505
Combined XGBoost	0.520	0.572	0.529	0.421	0.495

B.10 Prediction Dispersion by Configuration

Table 24: Standard deviation of out-of-sample predictions, by configuration

Spec	Family	Std. dev. of predictions
Baseline	ARX	0.01442
Crypto	ARX	0.02224
Macro	ARX	0.03233
Combined	ARX	0.03660
Baseline	XGBoost	0.00726
Crypto	XGBoost	0.01554
Macro	XGBoost	0.01107
Combined	XGBoost	0.01035

B.11 2026 Hold-Out, Full Detail

Table 25: Hold-out 2026: descriptive performance (22 weeks)

Spec	Family	RMSE	R^2 OOS (rolling)	DA
Baseline	ARX	0.05701	−0.0292	0.545
Crypto	ARX	0.05713	−0.0335	0.364
Macro	ARX	0.05785	−0.0598	0.591
Combined	ARX	0.05858	−0.0865	0.500
Baseline	XGBoost	0.05671	−0.0182	0.500
Crypto	XGBoost	0.05421	+0.0695	0.636
Macro	XGBoost	0.05750	−0.0469	0.545
Combined	XGBoost	0.05466	+0.0538	0.545

Table 26: Hold-out 2026: central comparison Spec 2 (Crypto) versus Spec 4 (Combined)

Family	n	Mean loss diff.	95% CI	DM (HLN)	p	Better spec.
ARX	22	−0.0001675	[−0.0009398, 0.0006047]	−0.4512	0.656	Crypto
XGBoost	22	−0.0000497	[−0.0004594, 0.0003600]	−0.2523	0.803	Crypto

Note: single execution, 2026-07-19. Both non-significant, sign favoring Crypto, consistent with development (Section 4.3).

B.12 Minimum Detectable Effect, Central Comparison

Table 27: Minimum detectable effect (80% power), central comparison Spec 2 → Spec 4

Family	Scope	n	Observed diff.	MDE (80% power)	Ratio	Reaches 80% power
ARX	Global	510	−0.0006495	0.0007209	0.901	no
ARX	R1	208	−0.0005052	0.0010448	0.484	no
ARX	R2	104	−0.0012512	0.0026539	0.471	no
ARX	R3	95	−0.0007566	0.0010027	0.755	no
ARX	R4	103	−0.0002346	0.0005344	0.439	no
XGBoost	Global	510	+0.0001287	0.0003928	0.328	no
XGBoost	R1	208	+0.0001587	0.0007163	0.222	no
XGBoost	R2	104	+0.0002761	0.0011525	0.240	no
XGBoost	R3	95	−0.0000966	0.0005456	0.177	no
XGBoost	R4	103	+0.0001272	0.0003107	0.409	no

Note: MDE approximated as $2.8 \times sd(d)/\sqrt{T}$. A ratio near 1 with a significant verdict (ARX global, 0.901) sits at the edge of the study's resolving power; a ratio well below 1 with a non-significant verdict is uninformative about the absence of an effect, it reflects limited power.

Code and Data Availability

The full codebase producing every result, table, and figure of this thesis, is available at:

https://github.com/hbelguenani/macro_crypto_btc_forecast_thesis

Raw data are not redistributed; the repository documents the sources (Coin Metrics, FRED, Kenneth French Data Library, Yahoo Finance) and the retrieval scripts.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management
Chaussée de Binche 151, 7000 Mons, Belgique | **www.uclouvain.be/lsm**