Model Aggregation for Improving Forecasts of Canada's Core Inflation Rate

Alexander Murray[1], **Sandi Alakhras**[1], **Josh Brault**[1], **Stenio Fernandes**[1], **Pierre Siklos**[2]

[1] Bank of Canada
[2] Wilfrid Laurier University and Balsillie School of International Affairs

Abstract

We present a model aggregation approach for forecasting Canadian core inflation that is based on a collection of heterogenous forecasting sub-models from a wide variety of sources. These models range in type from structural, and semi-structural, to time series models (single and multiple equations). Each sub-model was chosen for its predictive power and uniqueness relative to the others in the aggregate model. The sub-models' forecasts are combined linearly with time-varying weights that are estimated using ridge regression. Canadian data are used to train and evaluate both the sub-models and the aggregate model. It is found that the aggregate model has greater predictive power than any of the individual models over the evaluation data. One policy implication is that central banks can improve inflation performance by combining models as opposed to combining individual point forecasts.

1. Introduction

Macroeconomic models serve as a key decision-making input for policy makers in conducting forecasting and economic analysis exercises. As former Bank of Canada Deputy-Governor, John Murray, pointed out Murray (2012), the first of five stages in the central bank's preparations for setting the policy rate at its Governing Council meetings is the presentation of staff projections. Models that serve as the basis for forecasts used by central banks have, over the years, become more complex and varied, spurred on by a succession of financial and economic crises experienced by the world economy since 2008. The last two decades have encouraged modelers to relax more assumptions embedded in their models, generated by the widening gap between the realities of what drive inflation and economic activity more generally, and the capacity of existing models to accommodate behaviour unaccounted for by these models. Moreover, older models made it even more difficult to accurately forecast the outlook in part because the forces that drive inflation (e.g., aggregate demand versus aggregate supply) have also changed over time. Unsurprisingly, much of the focus since 2021 has been on how the most recent surge in inflation (both core and headline) did not show up in the outlook provided by most projection models in the quarters leading up

Alexander Murray: amurray@bank-banque-canada.ca; Sandi Alakhras: SAlakhras@bank-banque-canada.ca; Josh Brault: JBrault@bank-banque-canada.ca; Stenio Fernandes: SFernandes@bank-banque-canada.ca; Pierre Siklos: psiklos@wlu.ca

We thank Paul Corrigan, Tolga Ozden, and Tony Chernis for their many helpful comments and feedback, and Karsten Gerdrup (Norges Bank), Reuben Jacob (Reserve Bank of New Zealand), May Rostom (Bank of England), Ingvar Strid (Riksbank), Shunichi Yoneyama (Bank of Japan), Mariana Garcia-Schmidt (Bank of Chile), and Guillermo Escudé (Central Bank of Argentina) for research assistance.

to the global rise inflation. Kryvtsov et al. (2023), relying on Canadian data, conclude that the shocks experienced after 2019 were exceptional and likely triggered by the pandemic. Blanco et al. (2022) focus on inflation surges in history in a data set covering over 50 countries and find these episodes were almost wholly unanticipated. Moreover, its effects dissipate slowly over time (i.e., 3 to 4 years). Eggertsson and Kohn (2023) consider the US experience and also argue that pandemic related and geopolitical shocks were critical ingredients in the sharp rise in inflation beginning in 2021. However, they also add that certain features of the Fed's modified monetary policy strategy delayed the eventual tightening of its stance thereby exacerbating the inflation control problem. Finally, Lane (2022) exhaustively documents the sources of the inflation surge in the eurozone and, as many others have reported, finds multiple culprits that translated into large relative price shocks (e.g., war and pandemic related).

While the Bank of Canada's remit aims for 2 percent in headline CPI inflation central banks are especially interested in core inflation because, as former Governor of the US Federal Reserve, Frederic Mishkin pointed out "...focusing on core inflation can help a central bank from responding too strongly to transitory movements in inflation." Mishkin (2007). This does not detract from the Bank of Canada's responsibility to discuss or emphasize headline inflation in its public communication.

Despite shifts away from treating the financial sector as a veil in models, allowing for the existence of heterogeneous economic agents, relaxing assumptions related to flexibility in price setting behaviour, or treating expectations as being formed rationally Gauss and Gibbs (2018), the complex and interconnected nature of most economies renders prediction, even under normal conditions, an exceptionally challenging task. The inherent difficulty of attempting to predict human decision-making necessitates that assumptions and simplifications are made to suit a particular context or purpose for which the model is designed. For example, in economies where the export of commodities plays a large role, there will typically be more focus in the model on the dynamics surrounding the extraction and value of those commodities. Those dynamics may be completely absent from a model designed for an economy where, for example, the financial sector plays a much larger role.

Globalization adds another layer of model complexity. In response to this complexity one approach is for models to become even more realistic so that as many of the potential sources of shocks are accounted for. However, decision-makers understandably also prize tractability. Generally, they also prefer models that are clear in how their results are derived. This clarity aids not only in policy discussions, but also helps in the communication of policy decisions. While voluminous research finds that judgmental forecasts are hard to beat Ang et al. (2007); Faust and Wright (2013), models offer decision-makers clearer views about the economic forces that drive the outlook. Ideally, the profession should strive to develop models that can provide both superior forecasts and assist with the narrative that guides how the stance of monetary policy is set. Unfortunately, the resulting model is likely to be "...unwieldy and not cost effective." (op.cit., p.23) Faust and Wright (2013). Instead, this paper proposes to harness the predictive power of a large collection of models as a means of generating superior forecasts via model aggregation, in a manner to be defined below, and it provides empirical support for this strategy in the Canadian context. Our focus is on core inflation. A by-product of our approach is that it can offer some insights into links between model elements and forecast performance over time. Moreover, since the aggregate model consists of several heterogeneous models, this presents central banks with an opportunity to communicate to the wider public its awareness that improved forecasts of inflation requires richness in its modelling strategy. However, we leave these extensions for future research.

While model averaging is not unheard of within economics, it is used infrequently Bajari et al. (2015). When it is used, a relatively homogenous set of models is often chosen. The main contribution of this paper is a novel approach to forecasting Canadian inflation, which is based on Dynamic Model Averaging (DMA) using a heterogenous set of sub-models. It is also demonstrated that the aggregate model consistently produces better out-of-sample (OOS) inflation forecasts for Canada and that model heterogeneity is critical to obtaining this conclusion.

The rest of the paper is organized as follows. The next section explains the role and potential for model aggregation and a brief overview of the relevant literature. It attempts to draw succinctly from different approaches to obtaining forecasts. Section 3 briefly describes the universe of models, both structural, semi-structural, and of the time series variety. Their sources, properties and a few illustrations are provided in this section, as well as details about the clustering approach used to arrive at the representative set of models used by the aggregate models. The section concludes by briefly describing the final set of models used in the model aggregation exercise. Section 4 describes model aggregation methodologies. Section 5 discusses the main findings while Section 6 concludes with a summary, policy implications, and suggests avenues for future research.

2. Model Aggregation and Model Forecasts

2.1. Model Aggregation: Key Considerations

Models have a role as a communication device.[1] Hence, there is an advantage in relying on simplicity as a modelling strategy. Of course, there is a trade-off with realism that usually translates into greater complexity. The latter feature might be useful in improving the quality of forecasts. However, model aggregation does not typically seek to find the optimal level along this trade-off. Instead, a large set of existing models of various sizes and complexity is taken as a given and combined to generate improved inflation forecasts. Model aggregation, therefore, should be seen as a form of hedging, that is, as a device to check against benchmark forecasts or individual models that are consulted routinely.

One issue that can be raised is that some models are intended for purely forecasting purposes while others provide projections.[2] Generally, models are constructed to generate predictions, that is, some indication of the economic outlook. The assumptions and the modalities underlying the construction of models intended for a variety of uses are, of course, different. However, all have in common the need to be forward-looking, an essential ingredient in the conduct of monetary policy. Some research has also examined whether central bank projections can influence private sector (viz., professional) forecasts.[3] Population, energy consumption, and climate change effects, in addition to standard economic variables such as inflation and real economic growth, have all been the subject of this line of research.

Although there have been many recent advances in our understanding of the patterns inherent in deriving aggregate economic performance Slobodyan and Wouters (2012), much work remains to be done. Indeed, most forecasting models used by central banks begin with a theoretical rationale inspired by two sets of theoretical developments over the past few decades. The simplest expression of the New Keynesian (NK) approach is built around three equations. Clarida et al. (1999) provides an early survey, while Woodford (2003) provides the seminal theoretical foundation. Paralleling this development are real business cycle models which add micro-foundations and optimizing behaviour assuming rational agents (e.g., see Goodfriend and King (1997)). These advances would inspire the specification of Dynamic Stochastic General Equilibrium (DSGE) models that represent the workhorse models used by central banks for simulation and policy analysis. The DSGE methodology has evolved greatly to incorporate a variety of economic frictions and loosen some of the strong assumptions made in early vintages of these models.[4]

Given the wide variety of econometric forecasting methodologies, some central banks have

[1] Indeed, (García-Schmidt et al., 2020) advocates that the conduct of monetary policy should consist of a model of the central bank's likely future behaviour. Burgess et al. (2013) makes the case for models constrained by rules-like behaviour and such that, despite the many challenges surrounding them, they are easier to communicate.

[2] Projections here refer to scenarios that are viewed as being "central" or "baseline", but often supplemented with judgment. Forecasts, in turn, are meant to provide of what is expected to happen in future. It should be noted that there is a long tradition in the literature in evaluating projections by treating them as if they are forecasts. Inter alia, see Morris and Shin (2005); Smith (1987); Doan et al. (1984); Jain and Sutherland (2018) for illustrations where the two concepts related to prediction are often blended.

[3] This idea stems from Morris and Shin (2005) seminal piece about how greater central bank transparency about the outlook impacts the private sector's desire to generate its own independent judgment about the future.

[4] A challenge with DSGE models is validating them by comparing its impulse response functions with those of a vector autoregression (VAR). While the details, and other criticisms levelled at DSGE models are beyond the scope of this paper (see, however, Morris and Shin (2005) and Smith (1987)) our approach is agnostic on the subject since both kinds of models are included in the universe of models considered.

begun investigating and implementing model aggregation.[5] in order to achieve superior forecasts Araujo and Gaglianone (2023); Chakraborty and Joseph (2017b). Model aggregation is a modelling approach dating back to at least the 1970's (Leamer, 1978) and can take a number of different forms.[6] In it's most basic form, an aggregate model consists of a linear combination of the output of each sub-model with some weight on each sub-model output (Barutçuoglu and Alpaydin, 2003; Prudencio and Ludermir, 2006; Bajari et al., 2015).[7] The principal benefit of model aggregation is that the advantages of varying modelling approaches can be harnessed jointly such that better predictive power is achieved relative to any individual model. An obvious analogy is the oft reported finding that forecast averaging usually outperforms individual forecasts over time (Timmermann, 2006). Of course, the simplest solution is to give each model equal weight. However, as explained in Section 5, this strategy does not necessarily generate the best forecast as the optimal weight allocation for each model may vary over time as different models could provide better value for decision making depending on the state of the economy at the time the forecasts are generated.

While each sub-model serves as an abstraction of the economy, when deployed in combination, they may create a new meta-model that is closer to reality than any of them individually. The primary disadvantage of model aggregation is its complexity. Each sub-model must be individually estimated, and thus there is potentially a much higher computational burden in comparison with a single large model. However, it can be the case that the design or estimation of a single large model is simply intractable, or that there exists a model forecast which relies on confidential or qualitative data[8] In such cases, model aggregation can serve improve upon each of the individual forecasts.

In addition to improving predictive power, some aggregate models can help policy makers to observe which set of models performs best over time. This can be done by examining the weights placed on the individual forecasts made in an aggregated model in order to observe which models are most important in the aggregation. However, any analysis of the aggregated model output essentially requires analysis of each of the included models that contribute most to improving forecasts. This element of aggregation provides an additional benefit since multiple perspectives on the same phenomenon can be provided. For example, one sub-model may imply, say, that variable A is the primary driver of inflation, while another sub-model finds variable B to be the most important factor. The weights in the aggregate model would provide a notion of relative importance between the two perspectives. Therefore, instead of running a horserace to determine which model generates the best forecast, the relative contribution of all available models can be exploited to achieve the desired objective, namely obtaining a superior inflation forecast. This strategy is also motivated by the knowledge that the 'best' model will change over time.

The ability of model aggregation techniques to balance the biases of each of the sub-models to provide a less-biased forecast is well-documented in a variety of contexts (Barutçuoglu and Alpaydin, 2003; Prudencio and Ludermir, 2006; Bajari et al., 2015; Timmermann, 2006). However, in economic modelling, the scale and heterogeneity of the aggregation exercise is typically rather limited. For example, Check and Piger (2021) employs the Bayesian model averaging (BMA) method on a collection of auto-regressive models to form their forecast model. In contrast, the approach presented in this paper initially draws upon a set of 300 models. These include the current models available at several central banks (Corrigan et al., 2021; Gervais and Gosselin, 2014; Hirakata et al., 2019; García-Schmidt et al., 2020; Burgess et al., 2013), and the Macro Model Database (Wieland, 2023). Of this set of 300 models, a smaller set of representative models is selected using a clustering-based approach. Each of the models is algorithmically labelled according to the presence (or lack thereof) of the following features, namely linearity; occasionally binding constraints (e.g., zero or effective lower bound; ELB); existence of a steady state; and the assumption of rational expectations.

As discussed below, this results in 16 possible model classifications. Other model features could

[5] In this paper, model aggregation refers to the construction of a meta-model. That is, a model of models, which is the structure defining the composition of multiple other models.

[6] For example, the use of Bagging in ensemble learning can be seen as a type of model aggregation

[7] This approach is sometimes referred to as akin to "thick modelling" (Granger and Jeon, 2004).

[8] Such as the forecast given by a professional forecaster, which may take concepts like politics, climate, and geography into account which may be difficult to incorporate into an economic model.

have been chosen (e.g., homogeneous or heterogeneous agents, incorporating a finance or housing sector, and so on). However, the number of possible model classifications increases rapidly as the number of classifications rises. Of the 16 classifications, only 6 are non-empty. One representative model is selected from each of the non-empty classifications to construct the set of models to aggregate. Once the models are selected, a dynamic model averaging approach is applied as described in detail in Section 4. DMA constructs the aggregate model as a linear combination of the forecasts of each of the sub-models.[9] The weights on the sub-model forecasts at each time step are allowed to vary, albeit with a penalty parameter on changes between time steps. This is done primarily to allow for regime changes, such as introducing inflation targeting, or other institutional changes, for which one model may be better suited over another. Once forecasts from each sub-model and the aggregate model have been generated, a variety of metrics can be deployed to evaluate predictive power.

While it is shown in Section 5 that model aggregation can improve inflation forecasts, the use of time-varying weights in particular is found to produce the best forecasts. This result stands in contrast with a frequently cited result from the point forecast combination literature (e.g. Timmermann (2006)), where the simple average of point forecasts is often found to outperform individual forecasts. Another advantage of time-varying weights is that it can serve as a new monitoring tool. Indeed, there is the potential for this approach to provide a signal to consumers of forecasts about when to rely more heavily on one type or set of models over others when judging the most plausible forecast. Finally, it should be noted that many models currently employed by central banks and academics to derive inferences from shocks to the economy, or generate forecasts, share a high degree of similarity (e.g., see Binder et al. (2019)). The deliberate reliance on model heterogeneity may also be a useful device to mitigate the impact of groupthink in generating estimates for the economic outlook and especially inflation. That said, in the present context, there are significant data related challenges, as we shall see, when aiming for model heterogeneity when useful models used elsewhere (i.e., outside Canada) are also adapted to the Canadian environment in the interests of increasing the heterogeneity of the universe of available models.

2.2. Forecast Evaluation and Properties

Two strategies are often adopted in forecast evaluation.[10] A researcher can consider how well the model performs in-sample. A more stringent test, as far as policy makers, financial markets, and the public are concerned, is how well a model performs out-of-sample (OOS). Typically, only part of the available data is used for model estimation so that some can be reserved for forecast evaluation.[11] Although both types of tests are provided, OOS forecast evaluation is the most demanding metric that dictates the net benefits of the proposed model aggregation exercise employed in our study.

Forecasting exercises must also confront methodological issues. While purely judgment or sentiment-based predictions sometimes play a role, central bank forecasts are typically primarily based on structural models that are consistent with some underlying economic theory. The current standard at most central banks relies on one (or more) Dynamic Stochastic General Equilibrium

[9] Hauzenberger et al. (2023) also adopt this strategy as well as considering non-linear shrinkage techniques which can, under certain conditions, out-perform linear approaches for US inflation data. However, their application relies on a flexible state space models that, while able to encompass a wide range of models, cannot match the heterogeneity we exploit with our methodology. Model aggregation is relatively less common than the aggregation of time series via techniques such as principal components or factor models have proved very popular recently. Indeed, for the US, Low and Meghir (2017) have developed and regularly update a large data set (https://research.stlouisfed.org/econ/mccracken/fred-databases/) used to extract common factors.

[10] The focus here is on the forecasting process alone. There are separate questions related to the plausible economic interpretation of, say, coefficient estimates of a model. In what follows we are not concerned with this step in the modelling strategy or, rather, we assume that all models considered pass such a test.

[11] A complication that arises is that some of the series in any model can be revised over time. Ideally, one would prefer to use real time data to carry out forecast evaluation. Given the size and complexity of many candidate models in this study, this is not feasible. It may be noted that inflation data does not exhibit significant revisions over time. The same is not true, however, of output or output gaps which are also present in many, but not all, forecasting models considered. As a result, a parallel development is the nowcasting exercise that utilize data available before official data are released. There may be a subjective element in such approaches, but this is also outside the scope of this study.

(DSGE) models built from micro-foundations of individual utility maximizing behaviour. Policy makers are assumed to be credible and subject to a policy constraint (e.g., an inflation objective). A simple example of theory guiding model structure is given by the example where a tightening of monetary policy, via higher policy rates, is expected to reduce inflation over time, usually within 2 years. Expectations are typically also assumed to be anchored which facilitates the reduction in inflation following the introduction of a more hawkish monetary policy (however, see below).

In general, DSGE models are derived from the so-called New Keynesian paradigm which also incorporates elements from the neoclassical synthesis. There is insufficient space to go into the details about the inspiration for such models (however, see e.g.Goodfriend and King (1997); Woodford (2003); Romer (1993); Del Negro and Schorfheide (2013); Christiano et al. (2018)). Suffice it to say that since the seminal works of Smets and Wouters (2007) and Christiano et al. (2018) (also see Dotsey et al. (2011)) the level of sophistication and complexity of these models have grown considerably. The Global Financial Crisis (GFC) led to models that incorporate financial shocks, often via the inclusion of a banking sector, and other financial frictions. Later vintages of DSGE models admit the possibility that agents are heterogeneous instead of relying on a representative agent (e.g., some consumers are more patient than others) and there is growing acceptance of less than rational consumers, at least relative to the rational expectations benchmark.[12] The DSGE approach continues to evolve as modellers seek to incorporate more frictions or forms of behaviour that appear to have significant macroeconomic consequences but were previously not accounted for. Finally, in the case of small open economies such as Canada's, a further distinction is between DSGE models of a closed economy versus open economy models that admit the effects of external shocks.[13]

A potential risk, however, is that parallel developments in DSGE modeling risk reducing model heterogeneity. As Dorich et al. (2017) made clear, model heterogeneity is essential in generating net benefits from model aggregation exercises aimed at improving economic forecasts. While there is no consensus definition of what constitutes sufficient heterogeneity, examples of heterogeneous models are ones that incorporate a housing sector versus ones that do not, models that incorporate a significant role for an important driver of economic activity such as an energy or resource sectors, or ones that admit behaviourally motivated expectations. Other variants can also be imagined.

Next to structural models that are frequently used by central banks and in academic assessments of forecasting performance are semi-structural models. As the name implies, these models retain some basic theoretical precepts (e.g., higher policy rates will eventually reduce inflation, there exists persistence in inflation movements) but relax some of the restrictions incorporated into structural models. As a result, many such models end up being reduced forms where cause and effect is more difficult to establish than in, say, DSGE models. Indeed, identification assumptions are required to estimate the structural parameters (e.g., see Ramey (2016) for a survey). Several identification methods exist, and new ones are always being proposed. However, since our concern in this study is to determine whether any model type, when combined, can improve forecasts, these concerns are secondary. That said, semi-structural models have the limitation that, as Low and Meghir (2017) point out, they are unable to "...define how outcomes relate to preferences and to relevant factors in the economic environment, identifying mechanisms that determine outcomes." (op.cit. p. 33).

A large class of models can be placed under the semi-structural umbrella. Generally, they range from the univariate variety, where the variable of interest, such as inflation, is forecasted exclusively from its own history to the multivariate kind where economic theory guides the forecaster's choice of series that interact with each other econometrically. Models of this variety include autoregressive integrated moving average (ARIMA) models, or some variant (e.g., see Enders (2015)), or vector autoregressive (VAR) and vector error-correction (VECM) models.[14]

The profession has been aware for some time of the challenges to forecasting created when

[12] Often referred to as Full Information Rational Expectations (FIRE).

[13] Many DSGE models are semi-structural so that their attributes are derived exclusively from economic theory combined with features that are obtained from a time series analysis of the data.

[14] In the case of VARs, it should be noted that many variations that have been proposed since Sims (1980) introduced this methodology. Previous surveys of VARs include Hyndman and Khandakar (2008); Chen et al. (2012); Akiba et al. (2019).

shocks are historically large.[15] For example, Del Negro and Schorfheide (2013) evaluates how DSGE type models perform when subjected to a shock such as the financial crisis in 2008-9 since referred to as the Global or Great Financial Crisis (GFC).

In addition to DSGE and statistical time series forecasting models, machine learning (ML) models have established themselves as strong competitors to the classical methods in the domain of timeseries forecasting Ahmed et al. (2010). ML models leverage their ability to learn complex patterns and relationships in the data without relying on a predetermined equation. In the case of time series forecasting, whether in the univariate or multivariate setting, the forecasting problem is typically approached as a supervised regression problem where the data are re-arranged in an input-output form. In order to transform the data in this manner, most implementations follow a time-delay embedding technique that encodes the autoregressive structure of the time series by utilising the lagged values of the target variable (or with the addition of other variables in the case of multivariate series) along with other possible temporal feature engineered embeddings Joseph (2022).

At least two other issues are worth noting when considering the OOS forecasting properties of models. First, whereas it is often the case that a single metric, typically RMSE, is used to determine the superior forecast (e.g., as in Binder et al. (2019)), and we follow this approach for the most part, other metrics are also available.[16]

A second issue worth highlighting is that forecasters are expected to generate forecasts over different horizons. For example, the frequently used convention is that a change in the stance of monetary policy usually is expected to take anywhere from 4 to 6 quarters to complete.[17] Nevertheless, most observers are also keenly interested in forecasts at shorter horizons. This may be particularly true when significant revisions to forecasts are likely. Alternatively, longer horizons might be of interest when multiple shocks imply that the time span to a return to some equilibrium or desired level of inflation is longer than the 8 quarters ahead convention. There exists a forecast horizon beyond which predictions are not more informative than some mean value for the series of interest (e.g., see Galbraith and Tkacz (2007)). In the empirical results reported below we consider several horizons ranging from the one quarter ahead up to 16 quarters ahead horizons.

If model aggregation is one solution to the problem of improving inflation forecasts one must still confront how the set of models being considered are aggregated. A few strategies are available. Granger and Jeon (2004) propose the thick modelling strategy wherein all estimated specifications are retained and the one chosen relies on a metric (e.g., fit or RMSE). Indeed, the theory of forecast combination clearly demonstrates the net benefits of utilizing many forecasts in part because it is highly unlikely that the same forecaster will produce the best forecast over time and for all forecast horizons and time periods. While the use of equally weighted and carefully selected models has been shown to work well Granger and Jeon (2004),[18] it is shown in Section 5 this is not the case when the models are not carefully selected and that time-varying weights ultimately has the best overall RMSE OOS regardless of the set of models included in the aggregate model.[19] It is also debatable how much heterogeneity is admitted when relying on thick modelling. That said, bootstrapping methods can be employed to generate confidence intervals for thick models. McAdam and McNelis (2005) is one application to US, euro area and Japanese data that uses thick modelling to deliver relatively superior inflation forecasts.

[15] There is no formal definition of the distinction between a large and a small shock. Nevertheless, one can classify shocks according to their size and use percentiles or the number of standard deviations from the mean to define the size of a shock.

[16] These include: mean absolute prediction error (MAPE), mean absolute error (MAE), and the maximum informative forecast horizon (MIFH), to name the better-known ones. Since forecasts are evaluated the record over a considerable span of time, relying on multiple metrics potentially alerts us to the various sources of forecast errors while, for example, large forecast errors tend to favour using the RMSE. Some metrics (e.g., MAE) are more sensitive to outliers while others (viz. namely MAPE) are less informative when forecasts and outturns are small numbers. Finally, forecast evaluations can also account for the possibility that consumers of forecasts can assign different weights to positive versus negative forecast errors.

[17] See, for example, https://www.bankofcanada.ca/2021/04/understanding-how-monetary-policy-works/.

[18] It is possible that this simple strategy succeeds due to its inherent hedging of the weight allocated to each sub-model OOS.

[19] Although it should be added that Timmermann (2006) also consider alternatives such as weights obtained, for example, via regression methods.

One method of constructing a set of "best" models is through the use of model confidence sets. Model confidence sets (MCS) are reminiscent of the usual statistical problem of constructing a confidence interval. The investigator does not need to know the true model but the approach of Hansen et al. (2011) allows one to leave out models that are uninformative. Bootstrapping is used to reduce the dimensionality of the problem when the number of candidate models is too large. One of the applications considered by Hansen et al. (2011) is the problem of forecasting inflation and an advantage of the technique is that it does not require a benchmark model which aids in reducing the so-called 'data-snooping' problem (e.g., see White (2000)). Hence, MCS is a model selection technique. However, the MCS method potentially shares the same drawback as thick modelling because it limits model heterogeneity. On the other hand, the MCS approach can deal with, for example, models specified in both levels and first differences. Hansen et al. (2011) use MCS to investigate the performance of the Phillips curve but finds that factor models or standard forecast combination methods can generate better forecasts.

The approach adopted in this paper is an unsupervised machine learning technique. A clustering method is also employed to select the smallest set of models that provides a certain degree of heterogeneity. Chakraborty and Joseph (2017a), and references therein, offers a more complete description of clustering techniques while additional details are also provided below. Perricone (2018) is an application to inflation, although the focus is on AR-type models only.

3. Model Heterogeneity: A Summary of Model Types

The models considered below capture the essence of model heterogeneity used to illustrate the potential for model aggregation to improve inflation forecasts. Table 1 provides a summary of model types that are candidates for aggregation. The classification is a simplified one to fix ideas. Since the universe of models considered in this study consists of approximately 300 models it is impractical to list them all here.[20] In the case of structural (i.e., DSGE type) models shown the aim is to highlight that models considered are generally ones used by central banks in Canada and elsewhere. In the case of non-Canadian models, we replace the original data with Canadian data that best matches the variables developed by central banks outside of Canada. The remaining models are time series models that are specified and estimated using a wide variety of methodologies and assumptions about the number of variables considered and the degree of endogeneity and exogeneity of the series included.

A practical issue arises because different models can potentially rely on different information sets. However, this does not invalidate the aggregation methodology.[21] For example, two VARs that are estimated for different lag lengths can be compared even if the effective degrees of freedom are different. Similarly, two VARs with different dimensions (i.e., the number of variables included) can be compared because, a priori, we cannot know whether a selection or all the variables in both models significantly improve forecast performance. Finally, even if a model contains more variables or equations does not imply that every one of its components significantly contribute to generating better forecasts at all horizons. The statistical and machine learning models are listed according to whether they are univariate or multivariate specifications.[22]

In what follows we provide a few essential details about the class of models beginning with DSGE type models which are representative of the class of structural models. This is then followed by brief summaries of select semi-structural and time series models. This class of modelling is well known. Hence, the narratives are very brief.

[20] Provision will be made to provide all necessary data related details and other materials.

[21] Indeed, using diverse information sets among the individual forecasts typically leads to improved aggregate model forecasts due to higher forecast heterogeneity. Furthermore, forecast aggregation may be the only option when it is infeasible to construct a single model that incorporates all of the information sets Timmermann (2006).

[22] Note that both statistical and machine learning models are sometimes referred to as ”time series models” when time series data is given as input.

Table 1. Examples of Model Types Used to Generate Projections

Structural (source)	**Semi-Structural**	**Univariate Statistical/ML**	**Multivariate Statistical/ML**
ToTEM (Canada)	FRB-US (USA)	ARIMA	VAR
COMPASS (UK)	LENS (Canada)	UCSV	SVAR
TNT (Chile)	ECB-BASE (Europe)	Regression	TVP-VAR
NZSIM (New Zealand)	Q-JEM (Japan)	Markov Switching	VARMA
MAJA (Sweden)	MARTIN (Australia)	Moving Average	Factor Model
World3 (Club of Rome)		N-BEATS	XGBoost
SAMBA (Brazil)		Auto-regressive	
ARGEM (Argentina)			

Note: ToTEM is Terms of Trade Economic Model; COMPASS is Central Organizing Model for Projection Analysis and Scenario Simulation; TNT is Tradeable and Non-Tradeable Sectors; NZSIM is New Zealand Structural Inflation Model; MAJA is Modell för Allmän Jämvikts Analys (Model for General equilibrium Analysis); ARGEM is Argentina is Argentina Economic Model; SAMBA (Stochastic Analytical Models with a Bayesian Approach); LENS is Large Empirical and Semi-Structural Model; ECB-BASE is the European Central Bank Multi-country Semi-Structural Model; FRB-US is Federal Reserve Board-United States; Q-JEM is quarterly Japan Economic Model; Martin is Macroeconomic Relationships for Targeting Inflation; IMPACT is International Model for Projecting activity. Source refers to the country whose central bank developed the model in question.

3.1. Selected Examples of Structural Models

To conserve space, we briefly describe the individual forecasting performance of a selection of models covering the model types discussed above. OOS forecast performance is evaluated for horizons that range from one to 16 quarters ahead. We begin with a brief graphical description of forecast performance followed by more formal performance metrics below. It is also worth reminding readers that we are interested in forecasts of core inflation.

ToTEM The terms of trade economic model (ToTEM) is a large-scale open economy DSGE model of the Canadian economy. One of its distinctive properties is that it features significantly more firm- and household-level disaggregation than well-known DSGE models, such as those of Christiano et al. (2005) and Smets and Wouters (2007). On the firm side, the model features five distinct sectors producing final goods for consumption, residential investment, business investment, government spending and non- commodity exports. The model also includes a separate commodity-producing sector. This structure helps the model capture the composition of Canadian gross domestic product (GDP), which is important to accurately evaluate monetary policy frameworks that target the level or growth rate of nominal GDP or incorporate some role for the output gap. The firms responsible for producing final goods face nominal rigidities when setting their prices. More specifically, in a given final-good-producing sector, some of the firms re-optimize their prices in a forward-looking but staggered fashion, as in the literature following Smith (2017), while the other firms set their prices using a rule of thumb (RoT) similar to that in Hyndman and Khandakar (2008). Estimations of the sector-specific shares for each of these two pricing types find that the share of RoT price setters is relatively high in some sectors during the time since the introduction of inflation-targeting regime in Canada.

The household block of ToTEM features three prominent household types differing in terms of the financial markets they have access to and their status as savers or borrowers in those markets. On the saver side, the model follows Chen and Guestrin (2016) and Akiba et al. (2019) in assuming that some savers are ”restricted” (they can access only long- term debt markets) while others are ”unrestricted” (they have access to short- and long-term debt markets). As a result of these two saver types, ToTEM allows short- and long-term interest rates to influence aggregate household spending in distinct ways. Taken together, the two saver types—restricted and unrestricted—account for roughly half of all households in the economy. A single borrower type accounts for most of the remaining households in the economy. Borrowers in ToTEM finance part of their spending using long-term loans secured from saver households. When doing this, borrowers are assumed to face a collateral constraint under which new loans must be backed by some combination of new housing investment and home equity.

ToTEM follows most of the DSGE literature in assuming that workers enjoy some degree of wage-setting power but are subject to nominal rigidities similar to those faced by price setters. Lastly, monetary policy in ToTEM follows a simple rule under which the interest rate is set as a linear function of the previous period's interest rate, the output gap and the deviation of expected inflation over the next four quarters from the central bank's inflation target. ToTEM is estimated by Bayesian methods in a multi-step process. First, the variables are de-trended. Next, the rest-of-world block of the model is estimated. The estimated parameters of the international part of the model are then fed into the domestic part of the model wherein the rest of the parameters are estimated. Like many such models, ToTEM has evolved over time and continues to do so.

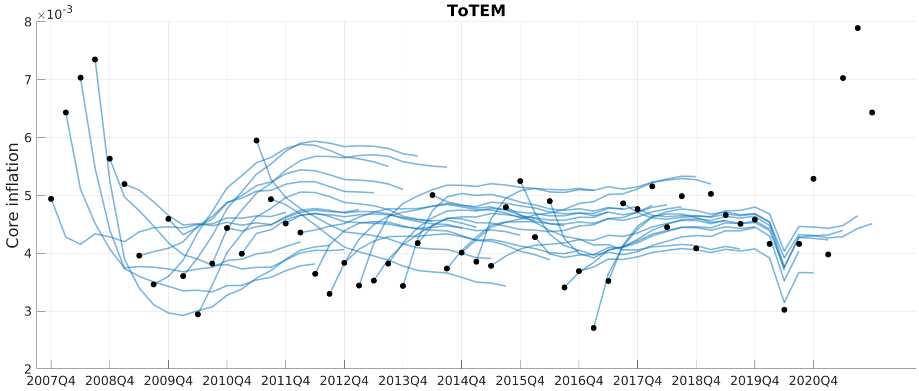


Figure 1. The 16 step ahead forecasts for ToTEM at each quarter from 2007Q4 to 2017Q4.

Figure 1 displays OOS forecasts for ToTEM included in the model aggregation exercise carried out below. The dots are the observed core inflation and each line plots 16 step ahead forecasts using data up to and including the point where it begins. Overall, the forecasts generated by ToTEM appear reasonable, although they display a tendency to overshoot. The wide undershooting of the model at the end of the sample is also clearly visible.

LENS The large empirical and semi-structural (LENS) model is as its name suggests, a semi-structural model driven more by the empirical properties of the data than economic theory. Like ToTEM, it is built to model the Canadian economy. However, while ToTEM is entirely inspired by economic theory, LENS is instead composed of reduced-form equations that are partly informed by theory but not fully micro-founded. As a result, the coefficients in LENS are less likely to be invariant to policy, making it less suitable for counterfactual analyses. An advantage of LENS is that its reduced-form structure makes the model easier to update as it can be estimated by blocks of equations rather than as a full system.

Long-run trends are also part of LENS and respond endogenously to shocks, meaning no ex-ante data detrending is required. LENS has a rich model structure which can be used for forecasting at a disaggregated level. For example, the model provides an explicit decomposition of the exports forecast into non-commodity, energy, and non-energy commodity exports, each of which depends on different foreign demand variables and relevant exchange rate measures.

The LENS forecasts frequently predict a sharp increase at short horizons followed by a gradual decrease, despite the fact that this behaviour seems to be rarely observed in the data.

COMPASS is the main organizing device used by the Monetary Policy Committee of the Bank of England for analyzing, forecasting, and conducting counterfactual experiments. Compared to many other DSGE models employed by central banks, COMPASS is relatively small in terms of

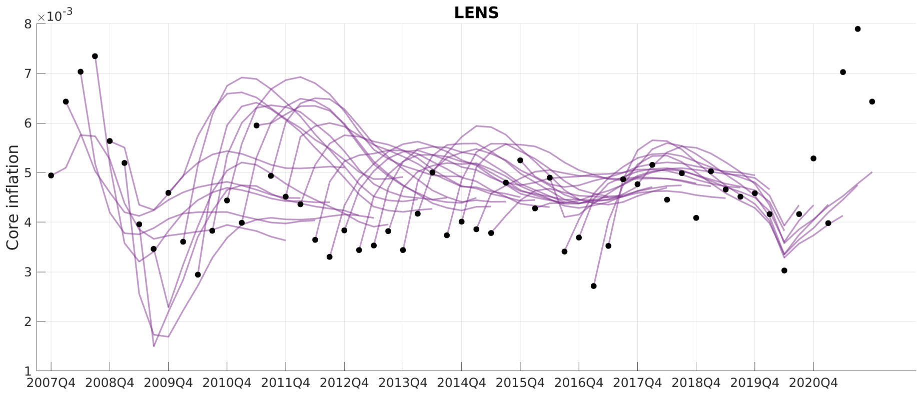


Figure 2. The 16 step ahead forecasts for LENS at each quarter from 2007Q4 to 2017Q4.

number of variables. However, its limited size is by design and COMPASS is complemented by a suite of over 50 other models which are used to assess its outputs and provide valuable input on missing channels. Examples of missing channels include: financial disruptions, energy price shocks, unconventional monetary policy actions, and fiscal policy that is more realistic than the model's highly simplified treatment.

COMPASS is characterized by the interactions among households, firms, the fiscal and monetary authorities, and a rest-of-world block, subject to 18 different types of exogenous shocks. Fiscal policy follows an exogenous AR(1) process and monetary policy is conducted according to a Taylor rule. There are two types of households referred to as unconstrained and RoT consumers. Unconstrained households operate in a utility maximizing way by choosing consumption (both current and future) and savings, subject to their respective budget constraints and external shocks. RoT consumers do not have access to savings and consume their entire income each period. Production in COMPASS does not distinguish between different sectors. Many of the steady state features of the model are calibrated according to national accounts data, while the remaining parameters are estimated using Bayesian methods. The estimation procedure uses 15 observables, primarily consisting of the aggregate expenditure components, prices and wages, and proxies for activity and prices in the rest of the world. According to COMPASS, domestic markup and monetary policy shocks have accounted for significant inflationary pressures after 2010, with imported price shocks playing a more prominent role prior to 2010.

The fact that COMPASS resides at the center of a larger network of models provides several benefits. First, because of its limited size, the model can be re-estimated frequently as new data arrives or are revised. Second, the flexibility of the approach allows for judgement and external forecasts to be included in the model. Third, the network of models can challenge and inform COMPASS outputs. However, the flexibility of this approach can also present some challenges, especially when large gaps between COMPASS's forecasts and policy prescriptions and ones generated by its extended network of models.

Keeping in mind that the forecasts shown in Figure 3 are based on adapting COMPASS to Canadian data, the resulting forecasts in the early portion of the sample, that is, until 2011, show a decline in core inflation at first before rising again. This pattern is reversed during the remaining years of the sample. This may explain why one of its sister models may be needed to produce more accurate forecasts. Note, however, that forecasts are overestimates the year before the pandemic strikes and then, in parallel with the preceding two models, the model is unable to capture the surge in inflation.

TNT Like many DSGE models, the TNT model of the Central Bank of Chile is characterized by an economy with households, firms, monetary and fiscal policy, and a rest-of-world block. However, TNT features much more heterogeneity on the firm side, distinguishing between firms producing in commodities, imports, exports, and non-tradeables. The difference in approaches comes from the

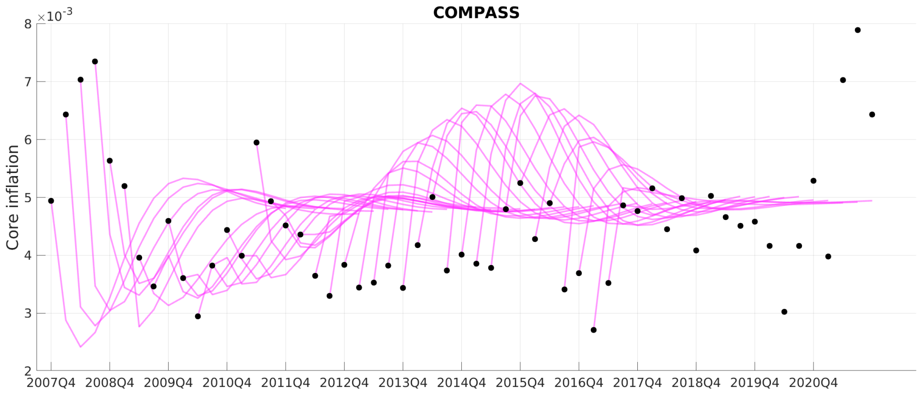


Figure 3. The 16 step ahead forecasts for COMPASS at each quarter from 2007Q4 to 2017Q4.

fact that explaining inflation in Chile requires a much more detailed specification for tradable/non-tradable sectors and the role of exchange rate passthrough to domestic production. This is evident by the fact that 29% and 37% of consumption and investment production in Chile is composed of importable intermediate goods. Fiscal policy in the model follows an exogenous AR(1) process and monetary policy is conducted according to a Taylor rule responding to both total and core inflation.

TNT features 20 exogenous shocks, with shocks originating in both the domestic and foreign blocks of the model. Similar to COMPASS, a number of steady state ratios are calibrated according to national accounts data. The remaining parameters are estimated with Bayesian methods, using 23 observables. The observables include breakdowns of value added according to tradable and non-tradable sectors, prices and wages, and rest-of-world measures such as the world interest rate, output, and inflation (commodity, imports, and consumption goods). According to the unconditional variance decomposition provided by TNT, the major shocks affecting inflation in Chile are world interest rates, uncovered interest rate parity, and world prices.

TNT does not confront the same dilemma as COMPASS regarding differing forecasts/policy prescriptions since it is a singular model. However, the model is subject to a criticism that can be levelled at all DSGE models namely that, if the model is mis-specified, this can lead to forecasts and policy prescriptions which are at odds with optimal policy.

Both COMPASS and TNT build on the medium-scale New Keynesian (NK) DSGE literature, with model economies characterized by sticky prices and wages (e.g. Christiano et al. (2005) and Smets and Wouters (2007)). In this framework the expectations of households and firms play an important role in the evolution of real and nominal activity, and monetary policy can play a stabilizing role through its adjustment of the short-term policy rate. However, monetary policy can only impact real activity in the short-run, and in the long-run real economic activity is determined by supply side factors such as total factor productivity growth. While COMPASS and TNT are broadly characterized as NK-DSGE models, at a more granular level they emphasize very different channels and propagation mechanisms due to the inherent differences between the UK and Chilean economies.

Figure 4 displays forecasts of core Canadian inflation once the model has been adapted to the Canadian environment. The forecasts generated by TNT seem to consistently predict a downturn in core inflation at short horizons followed by a steady increase, despite this behaviour being rarely observed in observed Canadian data used for estimation. Indeed, the model predicts deflation around the time of the GFC. In line with all of the previous examples, the 2021 inflation surge is missed.

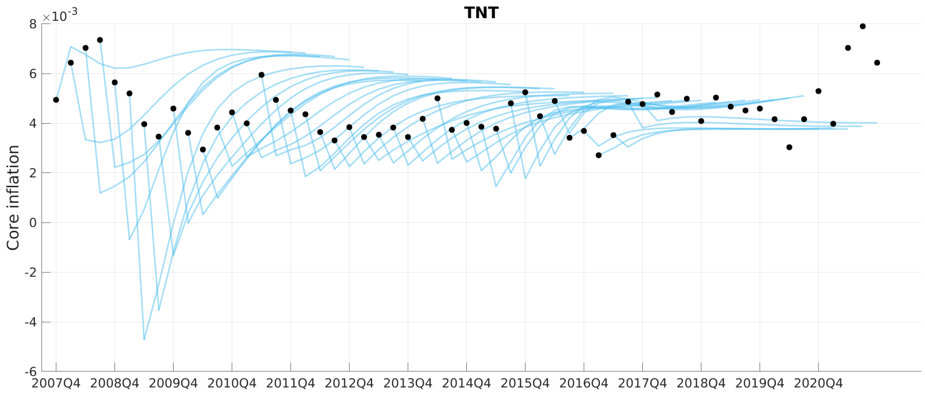


Figure 4. The 16 step ahead forecasts for TNT at each quarter from 2007Q4 to 2017Q4.

3.2. Selected Examples of Semi-Structural and Time Series Models

Naive A set of "naïve" forecasts are also included in the ensemble since it often serves as a baseline for the other forecasting models and is easy to implement. The naïve forecasts simply predict that, whatever the current state, it will not change over the entire forecast horizon. For example, if inflation is at 2.5% at t=0 then the naïve model will predict that inflation will remain at 2.5% from h=1 until h=16 where h is the forecast horizon.

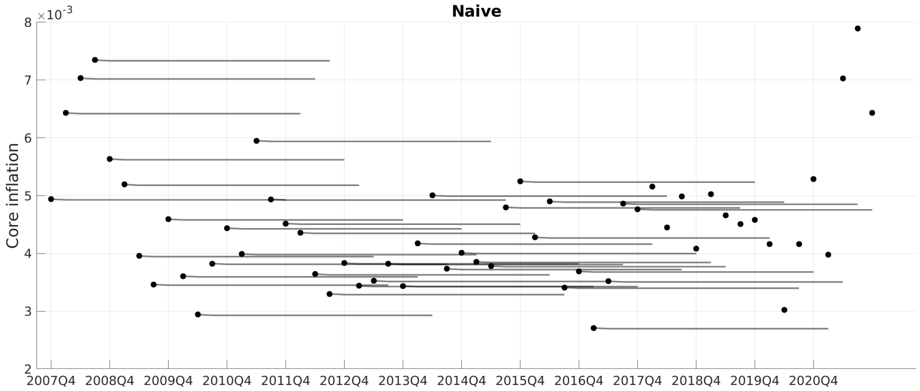


Figure 5. The 16 step ahead forecasts for the naïve model at each quarter from 2007Q4 to 2017Q4.

Figure 5 displays the naïve core inflation forecasts. Unsurprisingly, this forecast does poorly in the COVID era. However, it is also worth noting that the naïve forecast does well when inflation fluctuates in a narrow band as it did between approximately 2012-2015.

ARIMA The autoregressive integrated moving average (ARIMA) model is one of the most widely used traditional statistical time series forecasting models Calvo (1983). ARIMA directly assumes a linear relationship between the future value of a stationary univariate time series' variable and its past observations and errors as per the following equation:

$$y_t = \theta_0 + \psi_1 y_{t-1} + \psi_2 y_{t-2} + \ldots + \psi_p y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \ldots - \theta_q \epsilon_{t-q}$$

where y_t is the prediction of the variable y at time t, ϵ_t is the random error at time t, ψ_j $(j \in 1, 2, 3, \ldots, p)$ are the autoregressive parameters, and θ_i $(i \in 0, 1, 2, \ldots, q)$ are moving average parameters. Due to its relative simplicity in interpretation and ease of implementation, ARIMA has been a cornerstone in time series forecasting research for the past few decades Calvo (1983).

However, its simplicity can prevent it from being able to capture more complex non-linear relationships that often characterize real world data.

Autocorrelation and partial autocorrelation functions are used to estimate an interval of the possible model orders p and q.[23] The Darts (Data Analysis and Regression Time Series) time series forecasting library is chosen as the main software library for the implementation of the ARIMA model. In the case of ARIMA, Darts implements a wrapper model around the pmdarima AutoARIMA model (Gali and Gertler, 1999). In order to find the best model, the pmdarima AUTOARIMA model uses the step-wise selection algorithm proposed by Hyndman and Khandakar in 2008 (see Andres et al. (2004)) to select and identify the optimal model parameters from a given interval. The models are then evaluated and optimized per the user's choice of validation criteria[24] and the model with the best Akaike Information Criterion is returned Gali and Gertler (1999).

Shown in Figure 6 are the 16 step ahead forecasts generated by ARIMA at each quarter from 2007Q4 to 2017Q4. The forecasts are generated using input data ranging from 1987Q1 up to the quarter before the forecast begins. In common with other individual models considered the models perform most poorly during the COVID and recovery periods (i.e., 2020 and 2021).

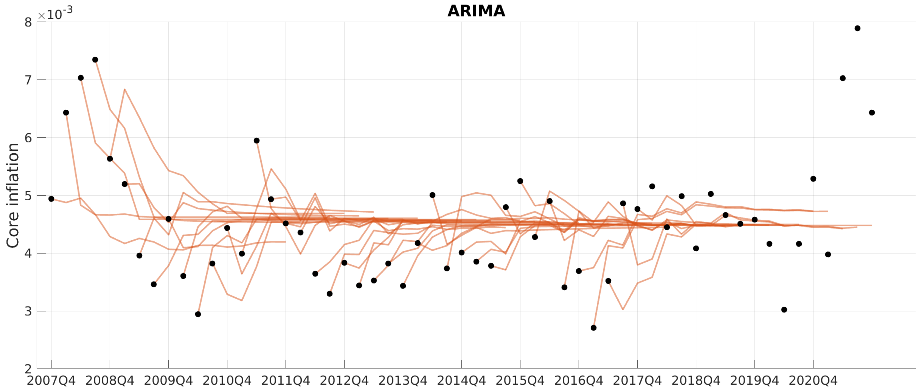


Figure 6. The 16 step ahead forecasts for ARIMA at each quarter from 2007Q4 to 2017Q4.

XGBoost Extreme Gradient Boosting (XGBoost) was initially proposed in 2016 by Chen and Guestrin and is based on the gradient boosting framework (Chen et al., 2012). XGBoost uses a boosting strategy that combines numerous weak prediction models, often decision trees, into a strong ensemble model. It iteratively constructs trees based on the residual data to minimise a loss function, allowing it to capture complicated relationships and patterns in the given data Chen et al. (2012). XGBoost has various advantages, including scalability, missing data handling, and the ability to handle both classification and regression tasks. Nevertheless, if not adequately regularised, XGBoost, like similar boosting algorithms, is prone to overfitting the training data leading to poor predictive performance on out-of-sample data.

The XGBoost model used in this paper is estimated in a multi-step process. First, the model is trained on 8 timeseries which were predetermined to hold particular economic significance.[25] Next, new time series are added one at a time to the estimation dataset. This iterative process continues until a plateau in the model's performance is observed. At each stage, the time series that provided the greatest improvement in OOS RMSE is chosen and incorporated into the final estimation dataset. It should be noted that the hyper parameter tuning library OPTUNA (Zhang, 2003) is used to tune the model in terms of lags, feature lags and maximum tree length. XGBoost-h denotes a version of XGBoost targeting h steps ahead.

The forecasts shown in Figure 7 obtained XGBoost are erratic, which could be due to the

[23] In accordance with the Box-Jenkis methodology Box and Jenkins (1970).

[24] Such as the Akaike Information Criterion (AIC), Corrected Akaike Information Criterion (AICC), Bayesian Information Criterion (BIC), Hannan-Quinn Information Criterion (HQIC) or out-of-bag (OOB)

[25] The time series in question are inflation, foreign exchange rate, output gap, short-term interest rest-of-world (ROW) inflation, ROW commodity prices, ROW output gap, and ROW short-term interest rate.

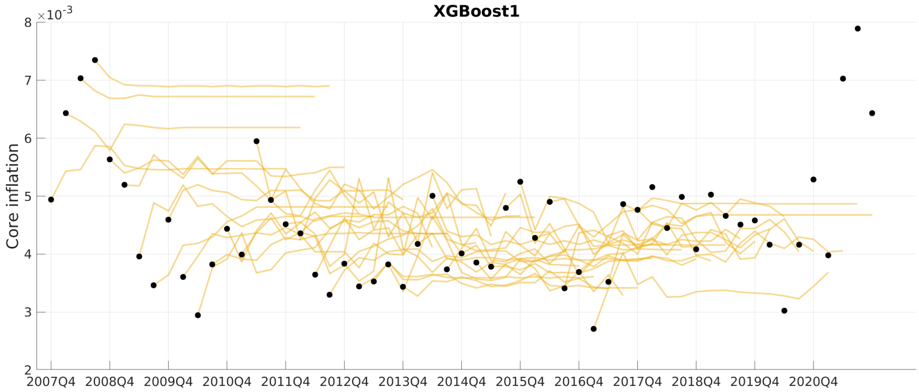


Figure 7. The 16 step ahead forecasts for XGBoost at each quarter from 2007Q4 to 2017Q4.

relatively small amount of input data (51 timeseries over a sample of up to 124 quarters). Improved forecasts are likely possible with increased tuning. In common with the other models examined so far, the post-COVID inflation surge is missed while, once again, model performance is improved when inflation is low and stable.

Vector Autoregressions (VAR) Vector Autoregressive (VAR) models are a class of time series models used in econometrics and statistics to analyze and forecast multiple time series variables simultaneously. Like other time series models, VAR models assume that the variables are stationary, meaning that their statistical properties, such as mean and variance, remain constant over time. VAR models use lagged values of the variables as predictors. The number of lags p is determined by the modeler and can vary depending on the data and the desired model complexity, as per the following equation:

$$y_t = \theta_0 + \psi_1 y_{t-1} + \psi_2 y_{t-2} + \ldots + \psi_p y_{t-p} + \epsilon_t$$

Where p is the number of lags, $y_t \in \mathbb{R}^N$ is the data to be modelled and the parameters to be estimated are $\theta_0 \in \mathbb{R}^N$ and $\psi_i, \forall i \in 1, \ldots, p$. Like ARIMA, this structure requires stationarity and assumes some degree of linearity in the temporal relationships between the variables.

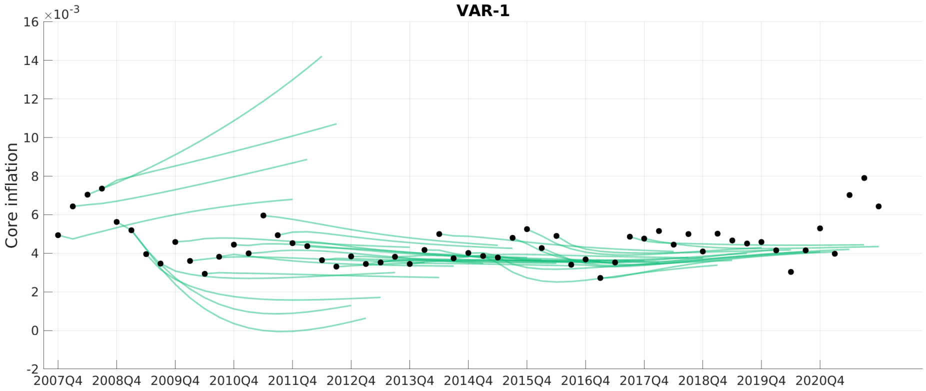


Figure 8. The 16 step ahead forecasts for VAR-1 at each quarter from 2007Q4 to 2017Q4.

The VAR models in this paper are estimated using Ordinary Least Squares (OLS), which assumes Gaussian error terms and the forecasts are shown in Figure 8. VAR-p denotes a VAR model with p lags. While the forecasts which come out of the VAR-1 model are very poor early on, the majority of its forecasts are similar to those of ARIMA and trend towards the sample mean.[26] To estimate the model eight times eries are given as input: quarterly inflation, exchange rate, output gap, interest rate, rest of world (ROW) quarterly inflation, ROW output gap, ROW interest rate, and ROW commodity prices.

4. Model Aggregation

The aggregate model is constructed by selecting weights $w \in \mathbb{R}^2$ for each forecast f_m in the set of M models at each timestep t. This is done by minimizing the squared difference of the weighted sum of forecasts to the univariate timeseries data d_t over the period $t = 1, \ldots, T$. Additionally, changes in the weights between timesteps are penalized with penalty parameter λ. The resulting optimization problem can be represented as follows:

$$\begin{aligned} &\min_{w} \sum_{t=1}^{T} \left(\left(\sum_{m=1}^{M} w_{t,m} f_m(t) \right) - d_t \right)^2 + \lambda \sum_{t=1}^{T} \sum_{m=1}^{M} (w_{t+1,m} - w_{t,m})^2 \\ &s.t. \quad \sum_{m=1}^{M} w_{t,m} = 1, \qquad \forall t \in 1, \ldots, T \end{aligned} \tag{1}$$

In practice, it is found that a penalty parameter of about 0.1 works best for the given data.[27] The penalty term in the optimization problem serves to reduce fluctuations in the weights while still allowing for some regime shifting, overall improving OOS fit. Similarly, the constraint on the weights also improves OOS fit by forcing the aggregate model to generate forecasts between $[min(f_1(t), \ldots, f_M(t)), max(f_1(t), \ldots, f_M(t))]$ at each timestep t.

While expression 1 can be employed to construct an ensemble model, it can also be solved iteratively using sets of $1, \ldots, N$ step-ahead forecasts. The result is a new ensemble model for each step ahead, with weights which may differ greatly from each other. The ensemble of these ensemble models can then be used to generate a forecast for the period $t = T + 1, \ldots, T + N$. As explained below, different models are weighted differently depending on the number of steps ahead being forecast.

5. Empirical Results

The models are estimated over the 1987Q1-2017Q3 sample and then OOS are generated beginning with 2017Q4 for horizons of up to 16 quarters ahead. The forecasts and outturns discussed below are measured on a quarter-to-quarter basis. It is straightforward to convert the data to annual or annualized rates. Core inflation is defined here as the quarterly rate of change in an average of the trim, median, and common inflation rates published by the Bank of Canada.[28]

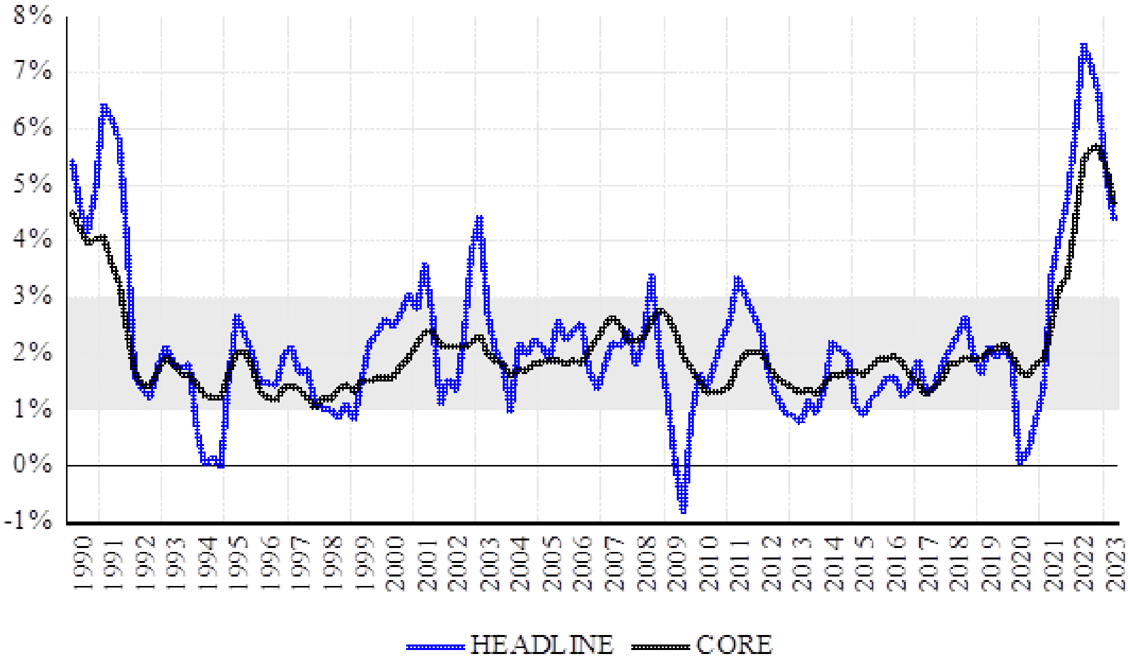
The top portion of Figure 9 plots headline inflation against core inflation while the bottom portion displays the three individual inflation series that make up core inflation. As is well known headline inflation is considerably noisier than core inflation. However, deviations between core and headline inflation are not very persistent. Hence, differences between the two series are stationary.[29] More generally, core inflation is stationary. The bottom figure illustrates that there are different ways of extracting core measures of inflation depending on how one removes the noisiest components on inflation. Since, a priori, there is little theoretical guidance about the ideal measure of core inflation the average of the three proxies shown in the bottom portion of the figure is used.[30]

Shown below in Figure 10 are the weights obtained by solving expression 1 using the $1, 2, \ldots, 16$ step-ahead forecasts generated by the models described in Section 3 given inflation data from

[26] One explanation for this may be because using too few lags can result in autocorrelated errors. Likewise, using too many lags results in over-fitting, and thus reduced out-of-sample predictive power. Selection of an appropriate lag is critical to inference in VARs Lutkepohl (2017).

[27] This corresponds to a penalty parameter that is 2 orders of magnitude larger than the given data.

[28] The CPI data are part of the “key monetary policy variables” used by the Bank of Canada. See https://www.bankofcanada.ca/rates/indicators/key-variables/. In principle, our results can accommodate headline



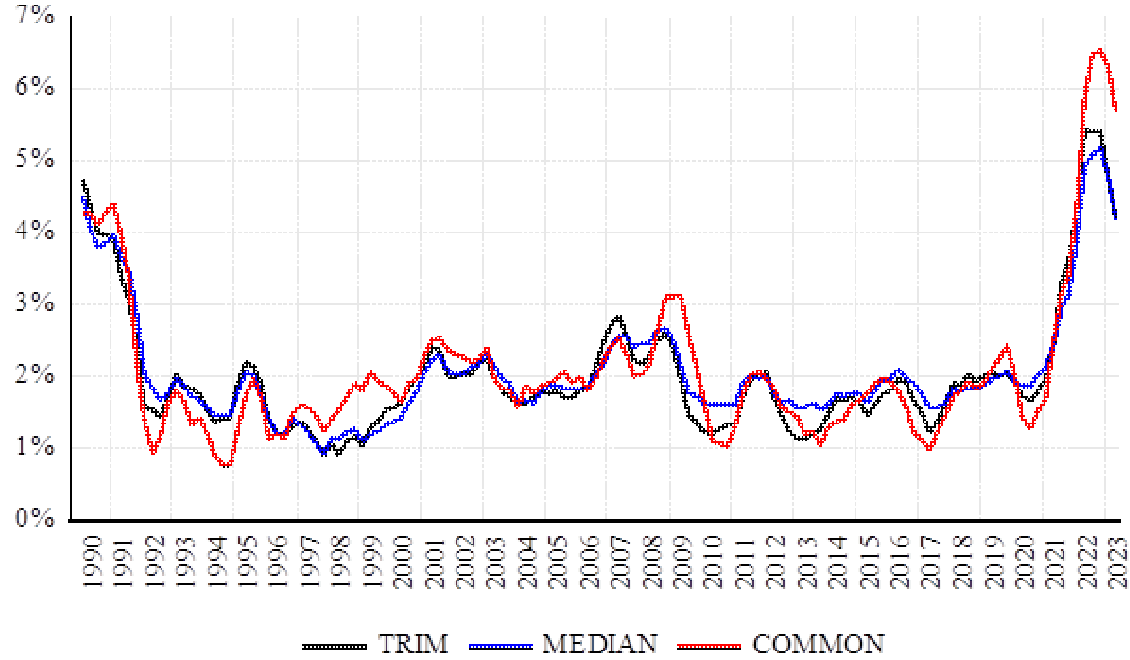


Figure 9. (Top) Headline and core inflation. The latter is the series being forecast. (Bottom) The three core inflation measures averaged to produce core inflation. Details about the definitions of TRIM, MEDIAN, and COMMON inflation can be found in boc (2023). Essentially, all three measures seek to remove the noisy components of headline inflation. Note also that the figures show inflation as the year over year change in CPI. The variable being forecast is the quarterly rate of change which, as seen below, is a noisier series. We show these versions to facilitate the description of a few stylized facts.

inflation. We leave this extension for future research.

[29] Tests (not shown) confirm this to be the case.

[30] A variety of so-called unit root tests, with and without allowance for structural breaks, confirm that inflation is

2007Q4 until 2017Q4. Interestingly, there is a high degree of variation across the forecast horizons, which indicates that model usefulness changes in the short, medium, and longer term.

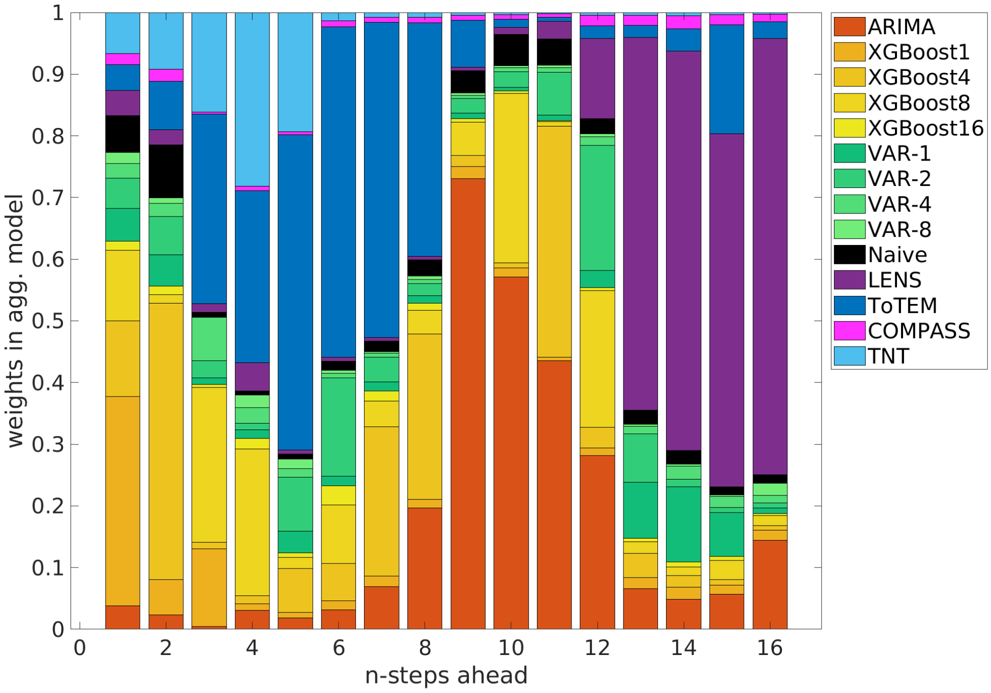


Figure 10. Weights assigned to each forecast in each n-step ahead aggregate model given forecasts from 2007Q4 to 2017Q4.

Unsurprisingly, the naïve forecast is most useful in the very short term since, generally speaking, inflation does not change dramatically each quarter. What is more surprising is that the naïve forecasts retain some weight around 3 years out, which could be an indication of a slight 3-year cycle in Canadian inflation from 2007 to 2017. The somewhat cyclical behaviour of the inflation data in this period seems to have been captured by ARIMA and the XGBoost models, which is why they are weighted so highly overall. Interestingly, despite the VAR models having access to more timeseries during model estimation, they are generally given less weight than ARIMA.

Another interesting observation that can be made from Figure 10 is that the structural models are assigned little weight in both the short-term (1 and 2 steps ahead) and longer-term (9-12 steps ahead). The reason is that their RMSE over the sample period is worse than the other models and the objective function of the aggregate model in expression 1 is effectively seeking to minimize RMSE. While LENS is weighted highly for the very long-term forecasts (13 to 16 steps ahead) in Figure 10, it should be noted that most of the models considered in Section 3 have forecasts which tend towards the sample mean. When the estimation period is changed the other weights stay more, or less, the same while those for 14 to 16 steps ahead can change considerably.

Given the weights shown in Figure 10, the aggregate model's OOS forecast is shown below in Figure 11 along with the individual forecasts and observed values. The forecasts generated by the aggregate model at each point within the sample period are shown in Figure 12. Note that the

stationary over the full sample. A structural break is identified for core inflation in 2021Q1 when inflation begins to show signs of surging, but the assumption of stationarity in inflation is not easily be rejected. For example, the so-called Dickey-Fuller test, with or without a break, rejects the null of a unit root (i.e., non-stationarity) at least at the 10% significance level for headline inflation. The same conclusion holds for core inflation (p-value of 1%). A break is detected in headline inflation very early in the inflation targeting era (1992Q1). Of course, more than one break is possible. More detailed test results are available on request. A variety of unit root tests also support the foregoing conclusions.

aggregate model's forecasts are less volatile than many of the component models.

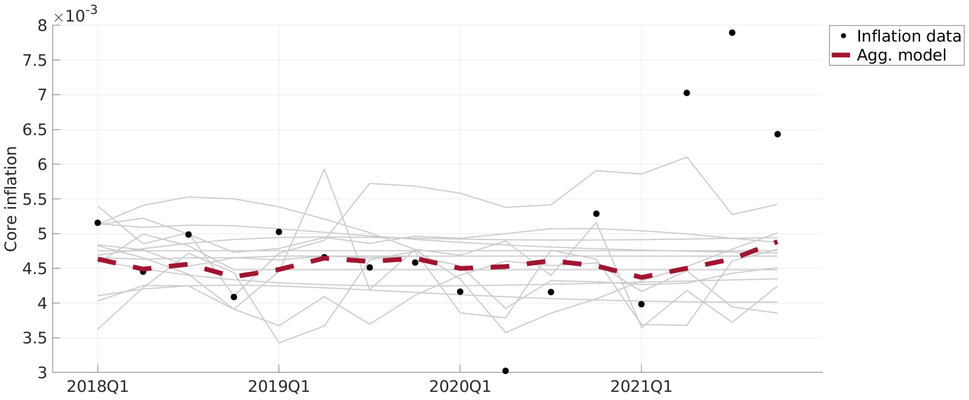


Figure 11. Out-of-sample forecasts for each model given estimation data up to 2017Q4. The aggregate model (dashed red line) uses the weights shown in Figure 8. The black dots represent the observed values for quarterly inflation.

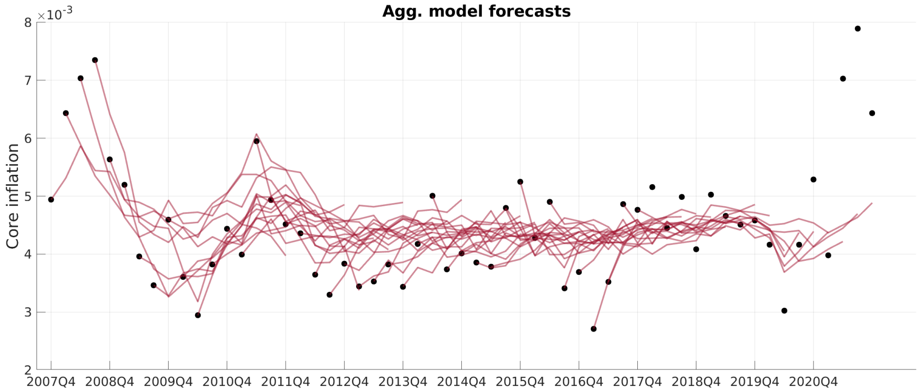


Figure 12. The 16 step ahead forecasts for the aggregate model at each quarter from 2007Q4 to 2017Q4.

Shown below in Table 2 are the average RMSE, MAPE, and MAE values for each model. The average error metrics are calculated by first estimating expression (1) over the period from 2007Q4 to 2010Q1, $\ldots$, 2017Q4 and then calculating the error metrics in the 16 step ahead OOS period starting just after the end point of the estimation sample period.

Table 2. Average OOS error for each model (10^{-3})

Model Name	16 step ahead RMSE	16 step ahead MAPE	16 step ahead MAE	8 step ahead RMSE	8 step ahead MAPE	8 step ahead MAE
Agg. model	0.8224	174.47	0.687	0.7448	168.93	0.6526
ARIMA	0.8068	170.72	0.667	0.7749	173.58	0.6702
XGBoost1	0.8368	164.49	0.681	0.7900	162.48	0.6528
XGBoost4	0.8760	173.61	0.710	0.8268	172.69	0.6870
XGBoost8	0.8791	167.74	0.691	0.7777	157.56	0.6300
XGBoost16	1.1388	235.01	0.936	0.9367	200.09	0.7829
Naive	0.9807	193.62	0.817	0.8928	175.16	0.7549
ToTEM	0.8624	181.95	0.713	0.9187	182.95	0.7739
LENS	1.1468	246.33	0.947	0.9616	207.88	0.8285
COMPASS	1.2229	259.73	0.989	1.4392	317.57	1.2486
TNT	1.8287	405.88	1.597	0.9514	198.39	0.8182
VAR-1	0.9380	173.07	0.762	1.3440	319.11	1.2088
VAR-2	0.9444	176.77	0.763	0.7679	174.73	0.6767
VAR-4	1.0121	215.08	0.857	1.3786	317.02	1.2065
VAR-8	1.8609	402.27	1.597	1.5892	347.92	1.4085

Overall, the error metrics in Table 3 correlate quite closely with one another, indicating that the results are robust to the choice of metric used in evaluating predictive power. While the aggregate model achieves the best predictive power in the 8 step ahead forecasting period, its predictive power is slightly less than that of ARIMA in the 16-step ahead OOS forecast period. This contradicts the hypothesis that the aggregate model would achieve greater predictive power by incorporating more and more diverse perspectives. This then begs the question: ”does a subset of forecasts exist that yields the best predictive power for the aggregate model?”

If only ARIMA, XGBoost8, ToTEM, VAR1 and VAR2 are included in the aggregate model then it can achieve an average RMSE of 0.00072895 in the 16 step ahead OOS forecasting period. Of course, averaging over a 16 quarters horizon is not likely to be the objective policy makers have in mind when deliberating the ideal course to set for monetary policy.

As mentioned in Section 2, there are alternative model aggregation approaches to expression (1). One could for example use equal weights, or non-equal but non-time-varying weights. The former is equivalent to the special case of $w_{t,m} = 1/M$ for all t and m, while the latter is the case where $\lambda = \infty$ in expression (1). For the sake of comparison, the average OOS errors of each approach are shown below in Table 3.

Table 3. Average 16 step-ahead OOS error for alternative model aggregation methods (10^{-3})

	RMSE:	MAPE:	MAE:
Expression (1) – all models	0.8224	174.47	0.6871
Const. weights – all models	0.8383	176.61	0.6952
Equal weights – all models	0.8534	180.82	0.7124
Expression (1) – best RMSE	0.7289	142.23	0.5907
Const. weights – best RMSE	0.7309	145.36	0.5958
Equal weights – best RMSE	0.7585	155.06	0.6197

A striking result from Table 3 is that the average OOS predictive power can be drastically increased by limiting the selection of models available to the aggregate model. Specifically, the set of models that leads to the best outcome for the aggregate model is quite heterogenous. It includes a variety of statistical models, an ML model, and a structural model. This seems to indicate that model heterogeneity is key in model aggregation. By breaking down the average OOS RMSE by predictive horizon, as shown in Figure 13, we see that the improvement is primarily coming from improved predictive power in the first five steps in the OOS forecast. This does seem to come at the cost of some predictive power in the longer term time horizons, along with some increased bias (see Figure 14).

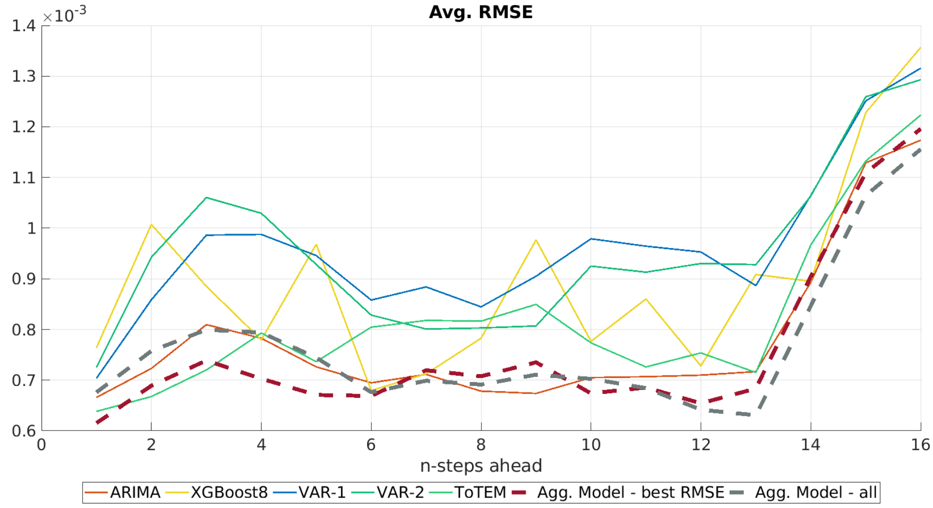


Figure 13. Average OOS RMSE of model forecasts from 2012Q1 to 2017Q4

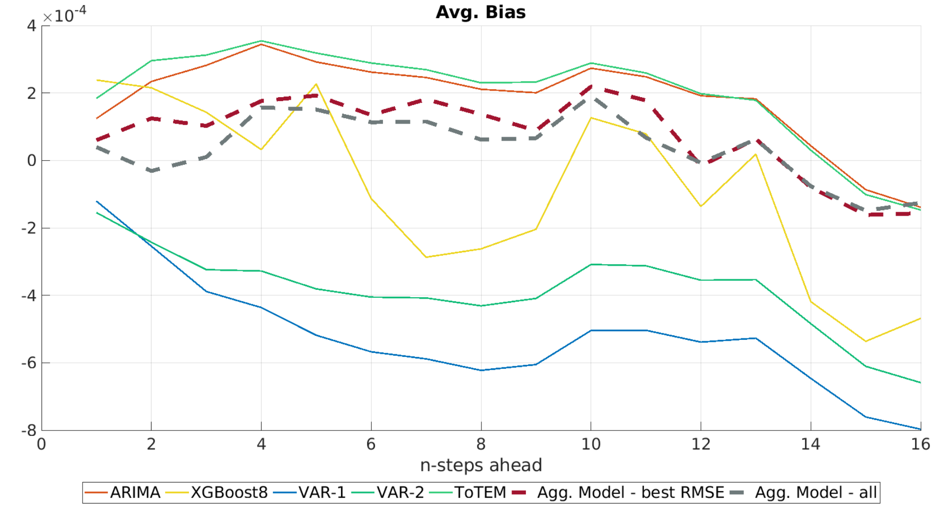


Figure 14. Average OOS bias of model forecasts from 2012Q1 to 2017Q4

6. Conclusions

Skepticism about inflation forecasts has received a lot of attention recently. While completely unexpected shocks are unlikely to ever be captured in existing models there exist strategies that have yet to be exploited that offer the promise of improved forecasts. This paper considers aggregating forecasts from a variety of model types ranging from time series to structural models. That said, a restriction that needs to be imposed is to select a set of models with diverse forecast biases (i.e., errors) and which may be better or worse at a variety of forecast horizons. The aim is not to replace any specific model. Rather, the methodology employed in this paper exploits the underlying heterogeneity of existing models. After all, experience has shown that some models outperform others at various times for a variety of reasons including the underlying information set, the complexity of the model, or the dynamics imposed on the specification employed.

The main finding is that model aggregation can improve forecasts of inflation. However, significant forecast improvement requires that models are weighted in a time-varying manner. This result stands in contrast with the related literature that investigates how best to combine point forecasts and often concludes that unweighted averaging does just as well as more complex forecast combinations. Moreover, there are significant differences in forecast performance at different forecast horizons. Second, model heterogeneity is found to be an important device to mitigate the impact of groupthink in generating estimates for the outlook on inflation. This is also a critical finding given the similarity of many models that are currently employed by central banks and academics to derive inferences from shocks to the economy or in producing forecasts.

Nevertheless, several extensions and improvements remain to be implemented. While the results presented in this study examine core inflation forecast performance, policy makers also require headline inflation forecasts. After all, the public is more likely to focus on overall CPI behaviour than on its counterparts stripped of its most volatile elements. Furthermore, a particular methodology is applied to aggregate the models. An additional robustness test calls for repeating the estimates of this study relying on other ways of aggregating models. The aggregation process focuses on certain characteristics that are considered key in influencing forecasts. However, other characteristics such as how far backward-looking or forward-looking the various models are, the vintage of the model and the data, the nature of price stickiness in the structural and semi-structural models, which sectors of the economy are excluded (e.g., commodities, finance, housing), or model scale, could be considered. A Bayesian approach to aggregation is another possibility that should be attempted wherein confidence intervals around aggregated forecasts could be constructed. Additionally, the model aggregation approach could yield interesting results when applied to other variables on interest, such as real output or unemployment. Finally, a positive side-effect from this exercise is that the set of models included in the aggregated model may, someday, provide clues about model elements that contribute to ensuring forecast errors are minimized over time. These extensions are left for future research.

References

Consumer price index, 2023. URL https://www.bankofcanada.ca/rates/price-indexes/cpi/.

N. Ahmed, A. Atiya, N. Gayar, and H. El-Shishiny. An Empirical Comparison of Machine Learning Models for TIme Series Forecasting. *Economic Reviews*, 29:594–621, 2010.

T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A Next-Generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

J. Andres, J. D. López-Salido, and E. Nelson. Tobin's imperfect asset substitution in optimizing general equilibrium. *Journal of Money, Credit and Banking*, pages 665–690, 2004.

A. Ang, G. Bekaert, and M. Wei. Do macro variables, asset markets, or surveys forecast inflation better? *Journal of Monetary Economics*, 54(4):1163–1212, 2007.

G. S. Araujo and W. P. Gaglianone. Machine learning methods for inflation forecasting in brazil: New contenders versus classical models. *Latin American Journal of Central Banking*, 4(2): 100087, 2023.

P. Bajari, D. Nekipelov, S. P. Ryan, and M. Yang. Demand estimation with machine learning and model combination. *National Bureau of Economic Research*, 2015.

Z. Barutçuoglu and E. Alpaydin. A comparison of model aggregation methods for regression. *Artificial Neural Networks and Neural Information Processing—ICANN/ICONIP 2003*, pages 76–83, 2003.

M. Binder, P. Lieberknecht, J. Quintana, and V. Wieland. Model Uncertainty in Macroeconomics: On the Implications of Financial Frictions. *[Online]. Available: https://macromodelbase.com. [Accessed 05 01 2023].*, pages 679–778, 2019.

A. Blanco, P. Ottonello, and T. Ranosova. The dynamics of large inflation surges. Technical report, National Bureau of Economic Research, 2022.

G. Box and G. Jenkins. *Time Series Analysis, Forecasting and Control.* Holden-Day, 1970.

S. Burgess, E. Fernandez-Corugedo, C. Groth, R. Harrison, F. Monti, K. Theodoridis, M. Waldron, et al. The Bank of England's forecasting platform: COMPASS, MAPS, EASE and the suite of models. *working document*, (471):13, 2013.

G. A. Calvo. Staggered prices in a utility-maximizing framework. *Journal of monetary Economics*, 12(3):383–398, 1983.

C. Chakraborty and A. Joseph. Machine Learning at Central Banks. *Bank of England working paper*, (674), 2017a.

C. Chakraborty and A. Joseph. Machine learning at central banks. 2017b.

A. Check and J. Piger. Structural breaks in us macroeconomic time series: A bayesian model averaging approach. *Journal of Money, Credit and Banking*, 53(8):1999–2036, 2021.

H. Chen, V. Curdia, and F. A. The macroeconomic effects of large-scale asset purchase programmes. *The economic journal*, 122(564):289–315, 2012.

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

L. Christiano, M. Eichenbaum, and M. Trabandt. On DSGE Models. *Journal of Economic Perspectives*, 32:113–140, 2018.

L. J. Christiano, M. Eichenbaum, and C. L. Evans. Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of political Economy*, 113(1):1–45, 2005.

R. Clarida, J. Gali, and M. Gertler. The science of monetary policy: a new keynesian perspective. *Journal of economic literature*, 37(4):1661–1707, 1999.

P. Corrigan, H. Desgagnés, J. Dorich, V. Lepetyuk, W. Miyamoto, and Y. Zhang. Totem iii: The bank of canada's main dsge model for projection and policy analysis. Technical report, Bank of Canada, 2021.

M. Del Negro and F. Schorfheide. DSGE Model-Based Forecasting. *Handbook of Economic Forecasting*, 2:57–140, 2013.

T. Doan, R. Litterman, and C. Sims. Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, 3(1):131–144, 1984.

J. Dorich, O. Kryvtsov, V. Traclet, and J. Witmer. Workshop Summary: Central Bank Models: The Next Generation. *30 January 2017. [Online]. Available: https://www.bankofcanada.ca/2016/11/central-bank-models-next-generation/*, 2017.

M. Dotsey, M. Del Negro, A. Sbordone, and K. Sill. System DSGE Project Documentation. *Federal Reserve Board memo, 9 June 2011. [Online]. Available: https://www.federalreserve.gov/monetarypolicy/files/FOMC20110609memo02.pdf*, 2011.

G. B. Eggertsson and D. Kohn. The inflation surge of the 2020s: The role of monetary policy. *Presentation at Hutchins Center, Brookings Institution*, 23, 2023.

W. Enders. *Applied Econometric Time Series, 4th Edition.* John Wiley and Sons, 2015.

J. Faust and J. Wright. Forecasting inflation. *Handbook of Economic Forecasting*, 2(A):4–51, 2013.

G. Galbraith and G. Tkacz. How Far Can Forecasting Models Forecast? *Bank of Canada working paper*, 2007.

J. Gali and M. Gertler. Inflation dynamics: A structural econometric analysis. *Journal of monetary Economics*, 44(2):195–222, 1999.

M. García-Schmidt, J. García-Cicco, et al. *A TNT DSGE model for chile: explaining the ERPT*. Banco Central de Chile, 2020.

E. Gauss and C. Gibbs. Expectations and the empirical fit of dsge models. *Forth-coming in Studies in Nonlinear Dynamics and Econometrics*, 2018. URL http://https://christopherggibbs.weebly.com/uploads/3/8/2/6/38260553/exp_and_dsge_2018.pdf.

O. Gervais and M.-A. Gosselin. Analyzing and forecasting the canadian economy through the lens model. Technical report, Bank of Canada, 2014.

M. Goodfriend and R. G. King. The new neoclassical synthesis and the role of monetary policy. *NBER macroeconomics annual*, 12:231–283, 1997.

C. W. Granger and Y. Jeon. Thick modeling. *Economic Modelling*, 21(2):323–343, 2004.

P. Hansen, A. Lunde, and J. Nason. The Model Confidence Set. *Econometrica*, 79:453–497, 2011.

N. Hauzenberger, F. Huber, and K. Klieber. Real-time inflation forecasting using non-linear dimension reduction techniques. *International Journal of Forecasting*, 39:901–921, 2023.

N. Hirakata, K. Kan, A. Kanafuji, Y. Kido, Y. Kishaba, T. Murakoshi, T. Shinohara, et al. The quarterly japanese economic model (Q-JEM): 2019 version. 2019.

R. Hyndman and Y. Khandakar. Automatic Time Series Forecasting: The forecast package for R. *Journal of Statistical Software*, 27:1–22, 2008.

M. Jain and C. Sutherland. How do central bank projections and forward guidance influence provate-sector forecasts, 2018.

M. Joseph. *Modern time series forecasting with python explore inustry-ready time series forecasting using modern machine learning and deep learning.* 2022.

O. Kryvtsov, J. J. MacGee, and L. Uzeda. The 2021–22 surge in inflation. Technical report, Bank of Canada, 2023.

P. Lane. Inflation diagnostics, 2022. URL https://www.ecb.europa.eu/press/blog/date/2022/html/ecb.blog221125~d34babdf3e.en.html.

E. E. Leamer. *Specification searches: Ad hoc inference with nonexperimental data*, volume 53. John Wiley & Sons Incorporated, 1978.

H. Low and C. Meghir. The use of structural models in econometrics. *Journal of Economic Perspectives*, 31(2):33–58, 2017.

H. Lutkepohl. Estimation of Structural Vector Autoregressive Models. *Communications for Statistical Applications and Methods*, 24:421–441, 2017.

P. McAdam and P. McNelis. Forecasting Inflation with Thick Models and Neural Networks. *Economic Modelling*, 22:848–867, 2005.

F. Mishkin. Headline versus core inflation in the conduct of monetary policy, October 2007. URL https://www.federalreserve.gov/newsevents/speech/mishkin20071020a.htm. speech given at the Business Cycles, International Transmission and Macroeconomic Policies Conference, Montreal, Canada.

S. Morris and H. Shin. Central bank transparency and the signal value of prices. *Brookings Papers on Economic Activity*, 36(2):1–66, 2005.

J. Murray. Monetary policy decision-making at the bank of canada. *Remarks at the Mortgage Brokers Association of B.C.*, 2012.

C. Perricone. Clustering Macroeconomic Variables. *Structural Change and Economic Dynamics*, 44:23–33, 2018.

R. Prudencio and T. Ludermir. A machine learning approach to define weights for linear combination of forecasts. *International Conference on Artificial Neural Networks*, pages 274–283, 2006.

V. Ramey. Macroeconomic shocks and their propagation. *J.B. Taylor and H. Uhlig (Eds.), Handbook of Macroeconomics*, 2A:71–162, 2016.

D. Romer. The New Keynesian Synthesis. *Journal of Economic Perspectives*, 7:5–22, 1993.

C. Sims. Macroeconomics and Reality. *Econometrica*, 48:1–48, 1980.

S. Slobodyan and R. Wouters. Learning in a medium-scale dsge model with expectations based on small forecasting models. *American Economic Journal: Macroeconomics*, 4(2):65–101, 2012.

F. Smets and R. Wouters. Shocks and frictions in us business cycles: A bayesian dsge approach. *American economic review*, 97(3):586–606, 2007.

S. Smith. Tests of forecast accuracy and bias for country population projections. *Journal of the American Statistical Association*, 82:991–1003, 1987.

T. Smith. pmdarima: ARIMA estimators for Python. *[Online]. Available: http://www.alkaline-ml.com/pmdarima*, 2017.

A. Timmermann. Forecast combinations. *Handbook of economic forecasting*, pages 135–196, 2006.

H. White. A Reality Check for Data Snooping. *Econometrica*, 68:1097–1126, 2000.

V. Wieland. Macroeconomic model database. *[Online]. Available: https://macromodelbase.com. [Accessed 05 01 2023].*, 2023.

M. Woodford. Interest and prices. 2003.

G. Zhang. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, pages 159–175, 2003.