Empirical Economics (2023) 65:149–188
https://doi.org/10.1007/s00181-022-02329-y

Big data forecasting of South African inflation

Byron Botha[1] · Rulof Burger[2,3] · Kevin Kotzé[3,4] · Neil Rankin[3] ·
Daan Steenkamp[1,2]

Received: 19 February 2022 / Accepted: 18 October 2022 / Published online: 8 November 2022

Abstract

We investigate whether the use of statistical learning techniques and big data can enhance the accuracy of inflation forecasts. We make use of a large dataset for the disaggregated prices of consumption goods and services, which we partially reconstruct, and a large suite of different statistical learning and traditional time-series models. The results suggest that the statistical learning models are able to compete with most benchmarks over medium to longer horizons, despite the fact that we only have a relatively small sample of available data. This may imply that the ability of statistical learning models to explain nonlinear relationships, or as an alternative, restrict the set of predictors to relevant information, is of importance. These characteristics of the statistical learning models may be particularly useful during periods of crisis, when deviations from the steady state are more persistent. We find that the accuracy of the central bank's near-term inflation forecasts compares favourably with those of

The authors would like to thank Patrick Kelly and Marietjie Bennett from Statistics South Africa for their assistance with the data. The insightful comments of the anonymous referees have also lead to a number of improvements, for which the authors are grateful. The remaining errors are those of the authors.

✉ Kevin Kotzé
kevin.kotze@uct.ac.za

Byron Botha
byron@codera.co.za

Rulof Burger
rulof@sun.ac.za

Neil Rankin
neil@predictiveinsights.co.za

Daan Steenkamp
daan@codera.co.za

[1] Codera Analytics, 42 Ennis Road, Parkview, Gauteng 2193, South Africa

[2] Department of Economics, University of Stellenbosch, Stellenbosch 7601, South Africa

[3] Predictive Insights, 3 Meson Street, Techno Park, Stellenbosch 7600, South Africa

[4] School of Economics, University of Cape Town, Rondebosch 7701, South Africa

other models, while the inclusion of off-model information, such as electricity tariff adjustments and other sources of within-month data, provides these models with a competitive advantage. Lastly, we also investigate the relative performance of the different models as we experienced the effects of the recent pandemic and identify the most important contributors to future inflationary pressure.

Keywords Micro-data · Inflation · High-dimensional regression · Penalised likelihood · Bayesian methods · Statistical learning

JEL Classification C10 · C11 · C52 · C55 · E31

1 Introduction

Accurate near-term inflation forecasts are important to most central banks as they contribute towards a more precise assessment of the economic outlook and an appropriate policy stance. This variable would also usually have a significant influence over the evolution of the short-term interest rate, which would potentially affect the activity of a diverse set of economic agents. In addition, such forecasts anchor inflation expectations, which may improve policy efficacy and economic stability, to provide an improved foundation for higher levels of economic growth. For this reason, it is important that central banks continue to assess the relative performance of different forecasting approaches and the use of different information sets. More recently, such investigations have considered the use of statistical learning methodologies that utilise different types of big data to inform policy decisions.[1] In addition, studies that make use of these methodologies have also influenced those policy decisions that consider the impact of the COVID-19 pandemic on economic activity.[2]

[1] Agrawal et al. (2019), Athey (2017, 2018), Athey and Imbens (2019), Mullainathan and Spiess (2017) and Varian (2014) contain overviews of selected research and discussions relating to the potential use of statistical learning methods within the field of economics. In a recent survey, Doerr et al. (2021) note that ± 80% of central banks discuss the topic of big data formally, where 70% of these monetary authorities use it for economic research, while 40% use it to inform policy decisions. Around two-thirds of respondents indicated that they wanted to start new big data projects in 2020/2021. The results from the study also suggest that the number of central bank speeches that mention the use of big data has increased significantly over recent periods of time and that most do so in a positive light. For an earlier discussion on the use of statistical learning methods within central banks and other policymaking institutions, see Wibisono et al. (2019), Tissot (2019), Mehrhoff (2017), Hammer et al. (2017), Baldacci et al. (2016), Florescu et al. (2014) and World Bank (2014) and United Nations Global Pulse (2012).

[2] For example, Blumenstock (2020) describes a few practical cases where the use of these techniques may be applied in developing countries, while the Organisation for Economic Co-operation and Development include mention of ways in which these techniques may be used to identify potential responses that may ease the effects of the pandemic (OECD 2020). In addition, Buckman et al. (2020) make use of the method that is employed in Shapiro et al. (2017) to report on changes in consumer sentiment following the onset of the pandemic, while Chetty et al. (2020) construct daily indices on consumer spending and other indicators, disaggregated by zip code, industry and income to show that high-income households reduced spending by more than low-income households, which has contributed towards job losses among the low-income households that provide services to high-income households. Similar changes in consumption behaviour have been noted in Baker et al. (2020), who make use of transaction-level financial data to explore how household consumption responded to the onset of the pandemic, while Baker et al. (2020) report on the

The relative accuracy of inflation forecasts has been considered in a number of important studies, where in an early investigation, Stock and Watson (2007) suggest that relatively simple univariate models provide good comparative forecasts of inflation in the USA, where a model that incorporates both unobserved components and stochastic volatility provides a reasonably good forecast of inflation.[3] Similarly, Faust and Wright (2013) also advocate for the use of a relatively straightforward approach and suggest that judgement forecasts, such as those from the Federal Reserve or inflation expectation surveys, tend to be more accurate than various forecasting model predictions. These findings, which largely relate to a period of stable economic activity for a developed economy, should not be too surprising as the conditional mean of inflation was highly persistent (Fuhrer 2010; Wolters and Tillmann 2015). However, over periods where economic activity is not particularly stable, a forecast that is close to the previously observed mean may not be terribly accurate and the models may need to allow for sustained departures from steady-state values. Hence, it may be necessary to make a number of amendments to traditional forecasting models during a period of economic crisis, or where the rate of inflation is relatively variable, as in the case of several low- and middle-income countries.

To address some of the challenges that may arise when looking to forecast macroeconomic variables, following a significant structural change or large departure from steady-state values, Galvao (2021) summarises a number of developments from the international literature, while Castle et al. (2021) and Coulombe et al. (2021) note that statistical learning models that are able to adapt to various changes may perform better than well-specified structural models.[4] In addition, the use of statistical learning models that incorporate nonparametric nonlinear features have gained significant attention over recent periods of time, partially due to the fact that they may be applied to large datasets to yield impressive results. For example, Medeiros et al. (2021) make use of nonlinear statistical learning models that are able to learn complex unknown functional forms, which may be useful when there are potential structural changes in both the mean and trend, to forecast inflation. Their results suggest that these techniques may provide superior forecasts over medium to longer horizons, when making use of the large macroeconomic dataset for the USA (the construction of which is described

Footnote 2 continued
changes in uncertainty relating to consumption spending. Other research by Chakrabarti et al. (2020a, b) make use of large datasets to investigate changes in consumer spending and business revenue in response to state re-openings, while Carvalho et al. (2020) note that in Spain, the consumption baskets converge towards the goods basket of low-income households. Similar changes in the consumption basket over this period of time have also been observed in Cavallo (2020) for a number of countries.

[3] In addition, for the period that incorporates the financial crisis, Stock and Watson (2010) suggest that this model should incorporate a stochastic trend that reacts to the unemployment recession gap, where the short-term response of inflation is consistent with an increase in this gap, while the long-term response is dependent upon the persistence in trend inflation. As is the case with most low- and middle-income countries, South Africa does not have a reliable measure for the unemployment recession gap that could be applied in an investigation that makes use of monthly observations of time-series variables.

[4] In particular, Coulombe et al. (2021) advocate for the use of nonlinear statistical learning models, while in a similar investigation, Koop et al. (2021) suggest that modelling specifications that accommodate time variation in forecasting uncertainty may also provide improved results.

in McCracken and Ng (2016)).[5] Coulombe et al. (2022) confirm these results and suggest that the statistical learning models that are able to incorporate nonparametric nonlinear features are responsible for the most significant performance gains, when comparing the predictions of a large suit of different forecasting approaches.

In addition to the above, there are also a number of similar studies that consider the relative merits of employing statistical learning approaches that seek to summarise all the available information (which is contained in the set of potential predictors), as opposed to only selecting those variables from a set of potential predictors that provide useful predictive power (i.e. the *density* versus *sparsity* debate). Giannone et al. (2021) have suggested that when making use of various macroeconomic and financial data sets for the USA, the forecasts of dense models are more accurate than the sparse counterparts. Similar arguments are made in Coulombe et al. (2022), who note that the use of sparse techniques would usually contribute towards to a substantial decline in forecasting accuracy. However, the results that are contained in Joseph et al. (2021) note that when restricting the subset of potential predictors, which incorporate disaggregated consumption price indices for the UK, the sparse models provide more impressive results.[6]

In this paper, we make use of four broad categories of models to predict future measures of inflation. The first of these relate to the *benchmark models*, which include traditional random walk, autoregressive and Bayesian vector autoregressive (BVAR) specifications. In addition, we also include the forecasts from the South African Reserve Bank (SARB) disaggregated inflation model (DIM), which is largely responsible for influencing the monthly near-term inflation forecasts, along with the actual monthly inflation forecasts that are presented to the monthly Monetary Policy Committee (MPC) meetings. The second group of models make use of *dimensionality reduction* techniques that seek to summarise all of the data from the potential predictors and would include those frameworks based upon principal component analysis. The third group of models make use of *variable selection* techniques and would include methods that make use of shrinkage estimators, penalised likelihood functions or Bayesian model selection techniques. And then finally, the fourth group of models include the use of *nonlinear statistical learning* forecasting models, such as the random forest and neural network, which may also incorporate nonparametric features.

These models are applied to data that is measured at different levels of aggregation for consumer prices, which is collected by Statistics South Africa (StatsSA) to con-

[5] There are several important differences between the setup that has been used in this paper and the one that was used in Medeiros et al. (2021). For example, their sample period is much longer and extends back to January 1960. It also does not include any data that arose over the period of the COVID-19 pandemic, as the final observation pertains to December 2015. Furthermore, their set of predictors incorporates a number of different measures of economic activity and is not limited to information about prices. The maximum forecasting horizon in their paper is also different, as it extends over twelve months, and they also make use of a rolling-window forecasting scheme, which is usually preferable when the sample extends over a lengthy period, as is the case in their study.

[6] Joseph et al. (2021) also find that after incorporating additional measures of macroeconomic activity, the dense models then provide more accurate forecasts, which support the findings of Giannone et al. (2021).

struct the South African Consumer Price Index (CPI).[7] The raw dataset incorporates prices for 34,075 unique goods and services, which were collected by fieldworkers that are dispersed across the country. Through various methods of aggregation and with the assistance of StatsSA, we were then able to reconstruct the set of 216 disaggregated predictors for the period between January 2009 and March 2021. In addition, we also make use of the publically available dataset for CPI that is measured at a slightly higher level of aggregation (and includes data for 46 different items). Hence, these datasets also incorporate a period of 12 months, over which various lockdown measures were imposed. When measured at higher levels of disaggregation, it has been suggested that such a dataset contains information that relates to the idiosyncratic behaviour of consumer prices, where the frequency and dispersion of price adjustments can vary across items and over time (Chu et al. 2018; Petrella et al. 2019; Stock and Watson 2020; Chetty et al. 2020; Carvalho et al. 2020; Cavallo 2020). Given these characteristics of the data, we could conceive that when the price indices are subjected to various forms of aggregation, their predictive power may decline. For example, if the disaggregated price index for brown bread has impressive predictive power, while the other products in the category for breads and cereals are poor predictors, then the signal that is provided by brown bread may be obscured if we were to restrict the analysis to use the aggregate data for the category rather than the individual goods. Previous findings in Hubrich and Hendry (2005) suggest that the use of disaggregated CPI components for the USA does not result in a meaningful improvement in forecasting accuracy, while studies that were conducted for Mexico and Portugal suggest that the use of disaggregated components could provide notable improvements (Ibarra 2012; Duarte and Rua 2007).

Our results suggest that despite the limitations of the data, which largely pertain to the number of available observations, the combined use of big data and statistical learning methods provide results that are potentially able to compete with most benchmarks over medium to longer horizons. However, many of the traditional benchmarks are superior over shorter horizons. In addition, after making use of data that is measured at different levels of aggregation, we note that the use of more disaggregated data results in an improved forecasting performance over all horizons. The results also suggest that the forecasts of several sparse models are superior to those of dense models, when making use of more disaggregated data. For example, both the least absolute shrinkage and selection operator (LASSO) and the ridge regression provide results that are superior to the dynamic factor models, over most horizons, when using data for headline inflation. Hence, there would appear to be advantages to identifying those variables that contribute towards the underlying predictive signal in the data, by restricting the information that is used in the construction of the forecast to those variables that have substantive predictive power. Furthermore, we also note that the relative performance of the statistical learning methods is more impressive when the

[7] Previous studies that have used the disaggregated consumer price survey data for assessing pricing behaviour in South Africa include Creamer and Rankin (2008) ,Creamer et al. (2012), Ruch et al. (2016), and Ruch et al. (2016). Restricting this analysis to the use of consumer prices is of importance to the SARB, as the forecasts from the current DIM model make use of CPI data that is measured at a relatively high level of aggregation.

rate of inflation deviates from its steady state during the period that incorporated a number of economic lockdowns.[8]

The remaining sections of this paper are organised as follows: Section 2 contains a review of the inflation forecasting models that have been applied to South Africa data, while Sect. 3 describes the methodology of the various models that have been specified in this study. Details relating to the data are discussed in Sect. 4 and the results from the different forecasting models are presented in Sect. 5. Then finally, Sect. 6 concludes.

2 Review of inflation forecasting in South Africa

A number of studies have considered the relative performance of inflation forecasting models in South Africa. These include those that emphasise the structural features of an economy, where in an early study, Woglom (2005) notes that inflation forecasts that are generated from a simple Phillips curve are not particularly accurate. However, when making use of a more expansive variant of a structural model, Smal et al. (2007) suggest that such models are capable of producing quarterly forecasts for CPIX[9] inflation that are more accurate than either the DIM or autoregressive integrated moving average model. In addition, these forecasts were also shown to be more accurate than the Reuters consensus forecast over their particular sample. Subsequent structural models, which include Liu et al. (2009), suggest that the forecasts from a small closed-economy New Keynesian dynamic stochastic general equilibrium (NKDSGE) model outperform those that are generated by classical vector autoregressive (VAR) and BVAR models for the South African GDP deflator. However, these authors also note that the difference in root-mean-squared error (RMSE) was not significant in most cases. Thereafter, Steinbach et al. (2009) extended the NKDSGE model to incorporate small open-economy features and found that the model's forecasts for CPIX inflation provided a lower RMSE, when compared to the Reuters consensus forecast, over a horizon that extends between four- and seven-quarters ahead. Similarly, Alpanda et al. (2011) built upon the small open-economy NKDSGE model that is discussed in Alpanda et al. (2010) and Alpanda et al. (2010), to show that their model provides better forecasts for consumer price inflation over shorter horizons. Furthermore, they also show that the difference in performance relative to classical VAR, BVAR and random-walk models is significantly different from zero.

This literature has subsequently been extended to consider the performance of a small open-economy NKDSGE-VAR model in Gupta and Steinbach (2013), which generates CPIX inflation forecasts that are superior to classical VAR and most BVAR models (with the exception of a BVAR model that incorporates a stochastic search variable selection prior) over a one-quarter ahead horizon. Other researchers have considered the role of nonlinearities within structural models, where Balcilar et al. (2015) make use of a nonlinear NKDSGE model, which employs the second-order solution

[8] In addition, we find that the accuracy of SARB's near-term inflation forecasts compare favourably to those of the other models that we have utilised in this study, reflecting the importance of the inclusion of off-model information, such as electricity tariff adjustments and the availability of within-month data.

[9] A measure of consumer price inflation that excludes the effects of interest rates on mortgage bonds.

method of Schmitt-Grohé and Uribe (2004) and a particle filter to evaluate the likelihood function, to provide forecasts for consumer inflation that have a lower RMSE, when compared to a large variety of BVAR models (including those that employ variable selection priors). Furthermore they found that the difference in forecasting performance is usually statistically significant, when compared to a random-walk and linear NKDSGE model (particularly over longer horizons). However, when considering the use of regime-switching nonlinearities, Balcilar et al. (2017) note that the out-of-sample forecasts for South African inflation that are generated by different forms of Markov-switching NKDSGE models are largely inferior to the single regime counterpart.

There are also a number of papers that focus on the application of different nonstructural statistical techniques to forecast South African inflation, to which this paper contributes. For example, in an attempt to reduce the potential effects of an omitted variable bias, Gupta and Kabundi (2011) make use of the Stock and Watson (2002b) and Forni et al. (2000) large factor models to forecast the percentage change in the implicit GDP deflator, along with the percentage change in real per capita GDP and the 91-day Treasury Bill rate in South Africa, over a one- to four-quarter ahead period from 2001Q1 to 2006Q4. They make use of 267 quarterly macroeconomic series to show that the factor models tend to outperform the unrestricted VAR, BVAR and small closed-economy NKDSGE models. Similar results are provided in Gupta and Kabundi (2010), where it is noted that large-scale data-rich models are better suited to forecasting key macroeconomic variables, relative to small-scale models. As an alternative, Kanda et al. (2016) is one of the few studies that make use of monthly data to focus on evaluating the performance of a suite of univariate nonlinear models, which include a locally linear model tree, neuro-fuzzy, multilayered perceptron, artificial neural network, nonlinear autoregressive, and genetic algorithm-based forecasting model. Their findings suggest that the locally linear model tree provides forecasts that can compete with the linear autoregressive model and is generally superior over longer horizons. In addition, Ruch et al. (2020) derive forecasts for quarterly measures of core inflation in South Africa with the aid of time-varying parameter vector autoregressive models (TVP-VARs), factor-augmented VARs, and structural break models to show that small TVP-VARs outperform all their other models, where additional information on the growth rate of the economy and the interest rate is sufficient to forecast core inflation accurately.

3 Methodology

To describe the methodology that has been employed by the various models, it is necessary to introduce some notation. In all that follows, we assume that $\boldsymbol{y} = \{y_1, \ldots, y_n\}$ is a vector of data for the measure of inflation, where the observations that arise over time are denoted, $i \in \{1, \ldots, n\}$. The matrix for the set of predictors that include the price indices for the different products or categories that are sampled to construct the CPI are contained in $\boldsymbol{X} = \{x_{1,1}, \ldots, x_{n,p}\}'$, which is of dimension $(n \times p)$, while $j \in \{1, \ldots, p\}$ is used to denote each of the different predictors in the matrix. We made use of four lags for the predictors in each of the models. To consider the relative forecasting accuracy of the different models, we employ a recursive out-of-sample scheme that extends over a horizon of between one and twenty-four months ahead, where the data that is used to test the predictions extends over a four-year period.

The motivation for making use of a recursive forecasting scheme, as opposed to a rolling-window scheme, is that the number of available observations that have been measured over time is relatively small and the forecasts over more recent periods of time may have benefited from the use of the maximum available number of observations. For example, if we were to make use of a rolling-window scheme, then we would have been limited to making use of a constant in-sample period for the predictors of just over five years to generate a twenty-four month-ahead forecast, when using most of the statistical learning models. Since we have a large number of potential predictors, we have assumed that by making use of a slightly larger in-sample dataset, we could possibly generate more accurate forecasts for the observations that arise over more recent periods of time. Furthermore, as the structural change that is attributed to the pandemic arose relatively suddenly, and very close towards the end of the sample, there are probably few (if any) gains that could be made by making use of a rolling-window scheme for the forecasts over this period.

The statistics that are used to evaluate the out-of-sample performance of the respective models include the root-mean-squared error (RMSE), the mean absolute percentage error (MAPE), and the Diebold and Mariano (1995) statistics.[10] When reporting on the results, we consider the year-on-year forecasts of headline and core inflation.

3.1 Benchmarks models

To evaluate the relative forecasting performance of the statistical learning models, we consider the use of a number of benchmarks, which are provided by autoregressive, large-scale Bayesian vector autoregressive, stochastic volatility, and random-walk models. Additional benchmarks include the model that is currently used by the central bank in South Africa to generate short-term monthly inflation forecasts, and the actual forecasts that are presented to the MPC. The latter incorporate off-model information such as electricity tariff adjustments and within-month data releases. Where there are relatively few predictors, we also include the results from a linear regression model. Further details relating to the specification of the benchmark models are included in section A of the online appendix.

3.2 Dynamic factor models

To compare the results of competing models against various dense models, that make use of principal components to summarise linear combinations of the original predic-

[10] Although there will be cases where these models will be nested, which would imply that the use of the statistics that are discussed in Clark and West (2007) and McCracken (2007) would be preferred to the Diebold and Mariano (1995) statistic, these models are not always nested within one another. Therefore, for the sake of consistency, we make use of the Diebold and Mariano (1995) statistic to evaluate all of the models. This would suggest that the results may favour the more parsimonious model in those cases where the models are nested.

tors, we utilise two different variants of the dynamic factor model (DFM). The first of these builds upon the framework of the traditional DFM, which largely follows the seminal work of Forni et al. (2000), Stock and Watson (2002a, b) and Bai (2003). In addition, it also makes use of the target factor approach that follows the work of Bai and Ng (2008), while the second approach utilises the three-pass regression filter of Kelly and Pruitt (2013, 2015). Section B of the online appendix contains additional details that pertain to the specification of these models, which seek to summarise the information that is contained in a large set of predictors, or explanatory variables.

3.3 Variable selection models

The literature on the development of statistical learning models that employ different variable selection methods, which are particularly useful when working with a large set of sparse predictors, is extensive. In this paper, we make use of a number of alternative methods that utilise a penalised likelihood function, where the parameters are estimated with frequentist techniques, as well as some of the Bayesian model selection counterparts. In particular, we employ the least absolute shrinkage and selection operator (LASSO), which was initially proposed by Tibshirani (1996), where the size of the penalty is determined by cross-validation. Furthermore, we also make use of the adaptive LASSO of Zou (2006), which may reduce the potential over-selection problem that has been encountered with the traditional LASSO. As an alternative to making use of the adaptive LASSO, we also make use of post-selection inference, to exclude those predictors that may not be able to make a significant contribution towards the explanation of future inflationary pressure. This exercise involves the application of methods that are discussed in Lee et al. (2016). The econometrics literature also makes extensive reference to the Post-LASSO estimation methods that are discussed in Belloni et al. (2011, 2013, 2014, 2017), which, in this particular setting, would motivate for the use of the methods in Belloni et al. (2013), to reduce the set of predictors to those that may be relevant.

As an alternative to imposing L_1 penalties, we also make use of methods that seek to implement L_0 penalties, which in general have improved properties, but require the use of methods that are not as efficient from a computational perspective. To implement these models, we follow the work of Rossell (2021). In addition, we also make use of models that impose L_2 penalties, such as the case of ridge regressions that were first discussed in Hoerl and Kennard (1970a, b), which seek to adjust the coefficient estimates to values of zero when they are deemed to be insignificantly different from zero. Models that make use of combinations of both L_1 and L_2 penalties are also implemented, which include the elastic net and smoothly clipped absolute deviation model of Fan and Li (2001).

As an alternative to making use of frequentist techniques, we also employ Bayesian model selection methods that consider the use of models that contain different sets of regressors, following Johnson and Rossell (2010, 2012) and Rossell and Telesca (2017). The results for the single specification that is most likely to contain the most useful predictors are reported along with the specifications that are summarised with

Bayesian model averaging techniques. Additional details relating to the use of each of the variable selection methods have been included in section C of the online appendix.

3.4 Nonlinear statistical learning models

The suite of nonlinear statistical learning models, which may also contain nonparametric features, include ensemble methods, random forests, gradient boosting and neural networks. Further details relating to each of these methods are contained in section D of the online appendix.

3.4.1 Ensemble methods

For comparative purposes, we have also made use of an ensemble method, which takes the form of the complete subset regression (CSR) framework of Elliott et al. (2013, 2015). This procedure makes use of the results from independent models that are then combined with a deterministic calculation. In many respects, it is similar to the bagging procedure of Breiman (1996) and provides an intuitive method for generating forecasts from many variables. To apply this methodology, we fit a linear regression model that seeks to explain y_i using each of the individual regressors in x_{i-h}. To identify the *best* predictors, we would then rank the absolute value of the t-statistics from the initial coefficient estimates. These predictors are then used to generate a number of individual forecasts, which are combined to provide the CSR forecast.

3.4.2 Random forests

The random forest model of Breiman (2001) reduces the variance of regression trees, which are nonparametric models that approximate an unknown nonlinear function with local predictions using recursive partitioning of the parameter space. They are based on bootstrap aggregation (bagging) methods for randomly constructed regression trees that take the form of nonparametric models that approximate an unknown nonlinear function with local predictions, using recursive partitioning of the parameter space that pertains to the covariates. Hence, to implement these methods, the parameter space is split successively to minimise the sum of squared errors in the regression.

3.4.3 Gradient boosting

As an alternative to random forests, gradient boosting seeks to build a model by repeatedly fitting a regression tree to the residuals. After each tree has grown to model the residuals, it is shrunk down by a factor before it is added to the current model. This would allow us to explain certain elements (including nonlinear relationships) that may have been discarded in the residual. A general gradient descent *boosting* paradigm has been developed for additive expansions based on any fitting criterion. It utilises developments discussed in Friedman et al. (2000) and Friedman (2001), where special enhancements are derived for the particular case where the individual additive components are regression trees. In general, it has been suggested that gradient boosting of regression trees produces competitive, highly robust results.

3.4.4 Deep learning (neural networks)

Neural network models usually take the form of highly parameterised nonparametric specifications that are able to potentially explain any nonlinear function. These models would often make use of a large number of parameter weights that transform the data that is contained in the set of predictors to fit the target variable. These parameter weights are learnt through repeated exposure to different subsets of the data. Deep learning methods make use of layered representations of neural network models that are stacked on top of each other to provide a mathematical framework for learning the rules that would allow for the mapping of the characteristics of the predictors to the target variable. Such models could potentially explain behaviour that is extremely complex, although there is also a significant possibility that the model may be prone to over-fitting errors. In our case, we have utilised a relatively simple model structure that will hopefully circumvent such concerns, where we have incorporated three hidden layers and a relatively parsimonious combination of 32, 16, and 8 nodes.

4 Data

The South African Consumer Price Index (CPI) measures changes in the general level of prices of consumer goods and services. It is a fixed-basket price index, in that it represents the cost of purchasing a fixed basket of consumer goods and services of constant quality and similar characteristics (Statistics South Africa 2017a). The items that are included in the basket seek to represent average household expenditure, using information from the Income and Expenditure Survey (IES) and more recently from the Living Conditions Survey (LCS), which was last conducted in 2014/15.[11] Note that the index only incorporates data on those products that contribute at least 0.1% of total household expenditure. Additional data sources such as regulatory reports, excise tax receipts, industry association reports and summarised transaction data from retailers are then used to align the data from the respective surveys with the data that goes into the household final consumption expenditure in the national accounts. The last update to the items that are included in the CPI basket was in January 2017, and the next update is expected to take place during 2021 (Statistics South Africa 2017b).

Since 2006, StatsSA has made use of fieldworkers who are responsible for collecting the relevant prices from the retail outlets directly. Each province has its own basket and every product that appears in at least one provincial basket is included in the national basket. The current CPI contains 412 products, which is slightly more than the previous basket, which included 393 products (Statistics South Africa 2017b), and its composition follows the United Nations Statistical Division (UNSD) standard for classifying household expenditure on goods and services. This standard is termed the Classification of Individual Consumption by Purpose (COICOP) and it currently incorporates 14 high-level (or 2-digit) categories (e.g. 01-Food and non-alcoholic beverages). Table 1, which is taken from Statistics South Africa (2017a), shows how

[11] Household expenditure in the LCS is surveyed in the same way as in the IES. However, the LCS also includes measures on a range of additional poverty indicators.

Table 1 Convention for COICOP classification

COICOP	Level	Name	Example
01	2-digit	Category	Food and non-alcoholic beverages
011	3-digit	Class	Food
0111	4-digit	Group	Bread and cereals
01112	5-digit	Product	Bread
01112001	8-digit	Commodity	Loaf of white bread
01112001$wxyz$	12-digit	Sampled product	Loaf of white bread for specific Brand, size, outlet (within an area)

the naming convention of the COICOP has been applied to the different levels of products and categories in South Africa.

In the subsequent analysis, we make use of the monthly four-digit data on consumer prices between January 2008 to March 2021, since a slightly different methodology was used to collect and classify the data for prior periods of time.[12] This dataset includes a total of 46 different predictors. In addition, we have also made use of a new dataset that contains more disaggregated data on the prices of goods in the consumption basket. This dataset includes information on 216 products or categories, where food products are measured at the 8-digit level and all other goods and services are measured at the 5-digit level. Unfortunately, the first observation in this dataset relates to January 2017, which would make for an extremely small in-sample training period in our case. Therefore, with the help of StatsSA we have now extended this dataset, going back to January 2009, by making use of the fieldworker data, which has been collected for 34,075 different products, across 5,505 outlets, that arise in 85 different areas.

To obtain a measure for the changes in prices over time, we calculate the price relative indices for the available fieldworker data, utilising the method that is used in the compilation of the respective CPI indices. This procedure involves the construction of a Jevon's index, which is defined as the unweighted geometric mean of the price ratios that utilise data for the current and previous survey periods for a particular commodity (i.e. at the 8-digit level). Such a Jevon's index may be constructed as follows:

$$\mathbb{I}_i^J = \prod_{i=1}^{n} \left(\frac{P_{\theta,i}}{P_{\theta,i-1}} \right)^{1/\xi} \tag{1}$$

where $\mathbb{I}_t^J$ denotes the Jevon's index, while $P_{\theta,t}$ is the price of commodity θ in period i, and ξ refers to the total number of items that are included in this calculation. In this study, we calculate a number of different variants of price relative indices to obtain information about price movements. This is then used to construct individual indices for each of the components at different levels of aggregation. After completing this

[12] These changes are discussed in Statistics South Africa (2007).

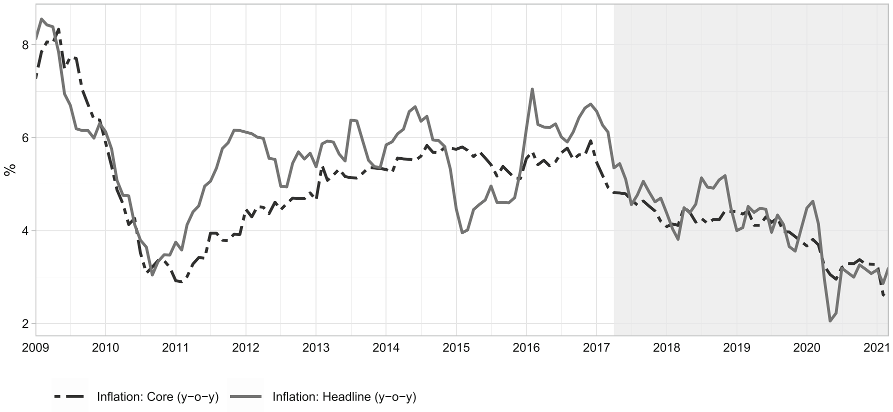

Fig. 1 Inflation over initial in-sample and out-of-sample period (year-on-year)

process, we are left with 216 predictors for the eight/five digit data, over the sample period from January 2009 to March 2021.

Figure 1 displays the measures of headline inflation and core inflation over the entire sample, where the shaded area relates to the entire out-of-sample period, where we assume that we do not have future information relating to the outcome variable and predictors, when estimating the parameters in the different models. The initial observation in the out-of-sample period is April 2017. Note that the trend in both measures of inflation has declined over the out-of-sample period, which would suggest that most mean reverting models will produce a negative forecast bias. In addition, as would be expected, headline inflation is certainly much more volatile than core inflation, where over the out-of-sample period, core inflation has a variance of 0.61%, while the variance of headline inflation is 1.04%.

5 Results

To evaluate the relative performance of the different models, we make use of a recursive out-of-sample forecasting exercise and a forecasting horizon of between one and twenty-four months ahead. The statistics that are used to evaluate the out-of-sample performance of the respective models include the root-mean-squared error (RMSE), the mean absolute percentage error (MAPE) and the Diebold and Mariano (1995) statistics.[13] When reporting on the results, we compare the year-on-year inflation forecasts against the official CPI release for headline and core inflation.[14] To gener-

[13] Although there will be cases where these models will be nested, which would imply that the use of the statistics that are discussed in Clark and West (2007) and McCracken (2007) would be preferred to the Diebold and Mariano (1995) statistic, these models are not always nested within one another. Therefore, for the sake of consistency, we make use of the Diebold and Mariano (1995) statistic to evaluate all of the models. This would suggest that the results may favour the more parsimonious model in those cases where the models are nested.

[14] Over the sample period, core inflation is derived from the prices of goods and services in the consumption basket, excluding food and non-alcoholic beverages, fuel and energy.

Table 2 Root-mean-squared error

	1-step	2-step	3-step	4-step	6-step	12-step	18-step	24-step
Headline inflation								
SARB	**0.15**	**0.25**	**0.37**	**0.57**	**0.84**	1.42	1.59	1.63
4 XG-BOOST	0.45	0.64	0.78	0.95	1.1	**1.28**	**1.33**	**1.48**
8/5 NEURAL	0.5	0.66	0.8	0.99	1.06	**1.48**	**1.5**	**1.45**
Core inflation								
SARB	**0.13**	**0.24**	**0.35**	0.46	0.63	1.19	1.25	1.32
4 RAND-FOREST	0.27	0.41	0.52	0.63	0.8	**0.71**	**0.85**	**1.17**
8/5 NEURAL	0.25	0.39	0.39	**0.42**	**0.55**	**0.86**	**0.91**	**1.09**

Boldface indicates that the RMSE is lower, when measured relative to the SARB forecast
Model acronyms: "SARB"—official SARB forecast reported to MPC, "4"—Four-digit data, "8/5"—Eight/five-digit data, "NEURAL"—neural network, "RAND-FOREST"—random forest, "XG-BOOST"—boosting

ate forecasts, we mostly make use of the direct forecasting approach, where the only exceptions pertain to the random walk, DIM, autoregressive, and vector autoregressive models.[15] To apply the direct forecast approach over the forecasting horizon of $h = \{1, \ldots, 24\}$, we estimate the model $y_i = \sum_{j=1}^{p} x_{i-h,j}\beta_j + \epsilon_i$ and use the coefficients to find $\mathbb{E}_i\left[y_{i+h}|x_{i,j}\right] = \sum_{j=1}^{p} x_{i,j}\hat{\beta}_j$, where the predictors may include lagged values of the target variable.

Table 2 summarises the main results, where we compare the RMSE for the official SARB forecasts, which were reported to the MPC, to the best-performing statistical learning model. We note that over shorter horizons, the official SARB forecasts, which have benefited from the use of off-model information and within-month data updates, are generally superior. However, over longer horizons, the nonlinear statistical learning models provide more impressive results. When considering the results for headline inflation, the boosting model that is applied to the four-digit data provides the most attractive results, when the horizon is twelve months or greater. Similarly, when applied to the eight-/five-digit data, the neural network model also appears to be responsible for lower RMSE statistics over longer horizons, when compared to the official SARB forecasts. However, they are not superior to the results of the boosting model that is applied to four-digit data.

For core inflation, the results are similar, as the official SARB forecasts are superior over a horizon of between one and three months. However, from four-steps-ahead to longer horizons, the neural network model is able to generate a lower RMSE, when applied to eight/five-digit data. Furthermore, the random forest model, which is applied to four-digit data, is also responsible for a smaller forecasting error (compared to the other predictions in the table), when the horizon is greater than a year and less than two years ahead.

In addition to these results, we also report on the Shapley values for a selected statistical learning model that appears to provide attractive out-of-sample results, to

[15] Explicit forecast functions for these models have been included in the above description of the respective models.

identify the important drivers of future inflationary pressure. This work follows Lundberg and Lee (2017) and Joseph et al. (2020), and the results are contained in section H of the online appendix.

5.1 Four-digit data: headline inflation

Headline inflation is made up of forty-six different price indices that are measured at the four-digit level of aggregation. These indices are used to generate the DIM forecast, which has a significant influence over the official central bank near-term monthly inflation forecast. Given the relatively small number of predictors, there are sufficient degrees-of-freedom to be able to include the forecasts from a linear regression model in this case.

Table 3 contains the out-of-sample RMSE statistics. When comparing the relative performance of all the benchmark models, we note that with the exception of the linear regression model, the errors are all fairly similar, where the DIM and official SARB forecasts are superior over the short term, while the random walk and stochastic volatility models are superior over longer horizons. Note also that over the first three months, the official SARB forecasts provide a RMSE that is about half the size of the DIM, which suggests that the use of off-model information has reduced the forecasting error by a relatively large amount over these horizons.

Turning our attention to the relative forecasting performance of the linear regression model, we note that it provides results that are indicative of a model that is prone to the over-fitting problem, since the models that make use of variable selection techniques often provide more impressive results. This would also suggest that the matrix that contains the predictors may be sparse, although we do not make use of a specific definition for statistical sparsity, as in McCullagh and Polson (2018). Further support for this finding is included in section E of the online appendix, where the in-sample estimation results for the full sample suggest that a model that makes use of twelve explanatory variables is able to provide a near-perfect explanation of the behaviour that is measured by headline inflation.

When we compare the accuracy of the forecasts from the *dense* models relative to the *sparse* models, we note that the results are somewhat mixed, where although there are a number of cases where the variable selection techniques provide more impressive results, the DFMs are at least competitive in all cases and superior over longer horizons. In the case of the twenty-four-month-ahead forecast, this would imply that the observed values of the predictors from twenty-four months ago, which provide the best explanation of current headline inflation, are not necessarily going to be the same, as the ones that provide the best twenty-four-month-ahead forecast, from the current point in time. Furthermore, we also note that in this case, the variable selection techniques would in most cases appear to be inferior to the benchmarks, which include the autoregressive, stochastic volatility and random walk models. Note also that the results for the nonlinear statistical learning models are in most cases similar to those of the DFMs models, however, over horizons that are longer than six months, the model that makes use of boosting methods provides forecasts that are more accurate than both *dense* and *sparse* models.

Table 3 Root-mean-squared error (four-digit headline inflation)

Model	1-step	2-step	3-step	4-step	6-step	12-step	18-step	24-step
AR(1)	0.41	0.68	0.82	0.93	1.09	1.68	1.75	1.68
AR-DIRECT	0.42	0.7	0.85	0.97	1.12	1.55	1.66	1.69
STOCH-VOL	0.42	0.64	0.76	0.84	0.91	1.21	**1.31**	**1.38**
RAND-WALK	0.41	0.64	0.76	0.84	0.9	**1.19**	1.33	1.45
BVAR	0.6	0.79	0.93	1.04	1.2	1.41	1.45	1.47
DIM	0.36	0.59	0.72	0.82	0.97	1.49	1.61	1.69
SARB	**0.15**	**0.25**	**0.37**	**0.57**	**0.84**	1.42	1.59	1.63
LINEAR	0.5	0.8	1.15	1.11	1.56	2.42	3.06	3.05
DFM-TF	0.43	0.68	0.84	0.97	1.2	1.85	1.98	1.89
DFM-3PRF	0.45	0.7	0.85	0.97	1.23	1.8	1.77	2.03
LASSO	0.39	0.73	0.94	1.16	1.56	2.11	2.91	2.28
LASSO-PSI	0.76	1.87	1.49	1.25	1.82	3.4	3.25	2.53
ADAP-LASSO	0.4	0.74	1	1.18	1.6	2.14	2.74	2.16

Table 3 continued

Model	1-step	2-step	3-step	4-step	6-step	12-step	18-step	24-step
POST-LASSO	0.44	0.68	0.66	0.76	1.09	1.77	1.78	2.53
LASSO-ZERO	0.42	0.7	0.87	1.04	1.2	2.46	2.19	3.47
RIDGE	0.48	0.9	1.13	1.29	1.68	2.25	3.07	2.46
ELASTIC	0.39	0.79	1.02	1.18	1.68	2.11	3.04	2.37
ADAP-ELASTIC	0.39	0.81	1.07	1.23	1.7	2.19	2.97	2.31
SCAD	0.44	0.83	0.88	1.02	1.43	2.2	2.12	3.33
BMS	0.44	0.71	0.9	0.95	1.13	2.43	1.5	2.88
BMA	0.44	0.72	0.78	0.87	0.94	1.34	1.4	1.68
CSR	0.41	0.65	0.75	0.82	0.92	1.77	1.66	2.11
NEURAL-NET	0.45	0.74	0.96	1.01	1.21	2	1.83	1.85
RANDOM-FOREST	0.51	0.72	0.9	1.03	1.25	1.43	1.57	1.64
XG-BOOST	0.45	0.64	0.78	0.95	1.1	1.28	1.33	1.48

Boldface indicates the lowest RMSE at a particular horizon

Model acronyms: “AR(1)"—first-order autoregressive, “AR-DIRECT"—first-order autoregressive with direct forecasting function, “STOCH-VOL"—stochastic volatility, “RAND-WALK"—random-walk, “BVAR"—large Bayesian vector autoregressive model, “DIM"—disaggregated inflation model, “SARB"—official SARB forecast reported to MPC, “LINEAR"—lienar regression model, “DFM-TF"—dynamic factor model with target factors, “DFM-3PRF"—dynamic factor model with three pas filter, “LASSO"—least absolute shrinkage and selection operator, “LASSO-PSI"—LASSO with post-selection inference, “ADAP-LASSO"—adaptive LASSO, “POST-LASSO"—post-OLS LASSO, “LASSO-ZERO"—LASSO with L_0 penalty, “RIDGE"—ridge regression, “ELASTIC"—elastic net, “ADAP-ELASTIC"—adaptive elastic net, “SCAD"—smoothly clipped absolute deviation, “BMS"—Bayesian model selection, “BMA"—Bayesian model averaging, “CSR"—complete subset regression, “NEURAL-NET"—neural network, “RANDOM-FOREST"—random forest, “XG-BOOST"—boosting

Table 4 Diebold–Mariano statistics (four-digit head)

Model	1-step	2-step	3-step	4-step	6-step	12-step	18-step	24-step
AR(1)	0.06	− 0.84	− 0.72	− 0.82	− 0.88	− 0.68	− 0.45	− 0.32
AR-DIRECT	− 0.2	− 1.65	− 1.21	− 1.26	− 1.41	− 0.79	− 0.5	− 0.38
STOCH-VOL	− 0.26	0.04	0.1	0.07	− 0.12	− 0.07	0.04	0.17
BVAR	− **2.21**	− 1.77	− 1.94	− 1.6	− 1.16	− 0.46	− 0.19	− 0.03
DIM	*2.19*	1.46	0.91	0.46	− 0.61	− 0.78	− 0.5	− 0.43
SARB	*3.67*	*2.3*	1.84	1.88	0.35	− 0.61	− 0.56	− 0.37
LINEAR	− 1.46	− 1.94	− **2.3**	− 1.84	− **1.97**	− 1.26	− 1.33	− **4.22**
DFM-TF	− 0.77	− 0.71	− 1.05	− 1.21	− 1.19	− 1.09	− 0.62	− 0.44
DFM-3PRF	− 1.36	− 1.04	− 1.44	− 1.39	− 1.44	− 0.89	− 0.48	− 1.21
LASSO	0.51	− 0.89	− 1.41	− **2.16**	− **3.22**	− 1.66	− 1.04	− **3.54**

Table 4 continued

Model	1-step	2-step	3-step	4-step	6-step	12-step	18-step	24-step
LASSO-PSI	– **2.58**	– **2.4**	– **2.94**	– **2.72**	– **3.96**	– **3.04**	– 1.26	– 1.81
ADAP-LASSO	0.43	– 0.94	– **2.08**	– **2.3**	– **2.77**	– 1.63	– 0.86	– **3.09**
POST-LASSO	– 0.97	– 0.68	1.05	0.81	– 0.78	– 1.15	– 0.75	– **2.19**
LASSO-ZERO	– 0.32	– 1.68	– 1.01	– 1.76	– 0.89	– **7.11**	– **2.73**	– 0.97
RIDGE	– 0.99	– **2.01**	– **2.26**	– **2.46**	– **4.28**	– **2.93**	– **5.03**	– **4.61**
ELASTIC	0.5	– 1.34	– **2.31**	– **2.23**	– **3.75**	– 1.38	– 1.15	– **4.17**
ADAP-ELASTIC	0.53	– 1.45	– **2.52**	– **2.62**	– **3.32**	– 1.27	– 1.13	– **4.07**
SCAD	– 1.01	– **2.46**	– 1.41	– **2.54**	– 1.62	– **2**	– 1	– **3.48**
BMS	– 1.5	– 1.62	– 1.41	– 0.97	– 0.72	– **5.15**	– 0.48	– **2.85**
BMA	– 0.95	– 1.33	– 0.15	– 0.31	– 0.33	– 0.82	– 0.55	– **3.77**
CSR	0.43	– 0.17	0.21	0.4	– 0.11	– 0.9	– 0.38	– **4.61**
NEURAL-NET	– 1.05	– 1.39	– **2.13**	– **2.63**	– 1.59	– 0.97	– 0.45	– 0.82
RANDOM-FOREST	– 1.24	– 1.05	– 1.53	– 1.65	– 1.21	– 0.5	– 0.42	– 0.4
XG-BOOST	– 1.37	0.03	– 0.28	– 1.78	– 1.07	– 0.21	0	– 0.24

Boldface indicates that the difference in forecasting accuracy is significantly different from zero in favour of the random-walk forecast, while *italics* indicate that the difference is significant, but in favour of the competing model

See Table 3 for model acronyms

Table 4 contains the Diebold and Mariano (1995) statistics for the forecasts of the different models, relative to what is produced by the random-walk model. In this case, we note that the only forecasts that are significantly superior to the random-walk forecast are provided by the DIM over a one-month horizon and by the official SARB forecast over a one- and two-month horizon. Furthermore, the forecasting performance of the DFMs, relative to the random-walk forecast, is not significantly different from zero over all horizons, while the random-walk forecast provides a significant improvement in forecasting performance over several horizons, relative to most of the models that employ variable selection techniques.

5.2 Eight/five-digit data: headline inflation

When using data that is measured at the eight-digit level of aggregation for food items and at the five-digit level for most other goods, we have a total of two-hundred-and-ten different price indices for headline inflation. Given the relatively large number of predictors, we do not have sufficient degrees of freedom to estimate a linear regression model. Table 5 contains the out-of-sample RMSEs for the different models, where most of the results that pertain to the benchmarks are similar to what was provided when using four-digit data, with the exception of the BVAR, which has experienced a slight deterioration in performance.

Note that the RMSEs for *sparse* models are lower when using the more disaggregated data, which would suggest that the combined use of more disaggregated data and variable selection techniques allows for an improved forecasting performance, as it would discard some of the noise that may be included in the variables when they are subject to greater degrees of aggregation. This is in contrast to the results of the *dense* models, which provide forecasts that are slightly more inaccurate than those that were derived from the four-digit data. And then finally, the results for the nonlinear statistical learning models are in most cases comparable to those that make use of the four-digit data.

Table 6 contains the Diebold and Mariano (1995) statistics, which are measured relative to the forecasts from the random-walk model. Once again, the only forecasts that are significantly superior to the random walk are provide by the DIM over a one-month horizon and by the official SARB forecast over a one- and two-month horizon. Furthermore, the forecasting performance of the LASSO at the three-month horizon is significantly more accurate than what is provided by the random-walk model, while most of the other variable selection forecasts are either positive (which is due to their lower RMSE) or not significantly different from zero.

5.3 Four-digit data: core inflation

In what follows, we repeat the above analysis, but in this case, the target variable is core inflation, which is derived from the measure of CPI that excludes the effects of changes in the prices of food, non-alcoholic beverages, fuel and energy. When using the four-digit data, we are able to make use of thirty-three different predictors for core inflation. After applying this data to the respective models, we calculate the RMSEs,

Table 5 Root-mean-squared error (8/5-digit headline)

Model	1-step	2-step	3-step	4-step	6-step	12-step	18-step	24-step
AR(1)	0.41	0.68	0.83	0.95	1.13	1.72	1.76	1.66
AR-DIRECT	0.41	0.69	0.83	0.94	1.08	1.41	1.47	1.49
STOCH-VOL	0.42	0.65	0.76	0.84	0.92	1.23	1.34	**1.41**
RAND-WALK	0.41	0.64	0.76	0.84	0.9	**1.19**	**1.33**	1.45
BVAR	0.51	0.68	0.79	0.9	1.09	1.45	1.61	1.72
DIM	0.36	0.59	0.72	0.82	0.97	1.49	1.61	1.69
SARB	**0.15**	**0.25**	**0.37**	**0.57**	**0.84**	1.42	1.59	1.63
DFM-TF	0.43	0.66	0.82	0.98	1.28	1.89	1.69	1.74
DFM-3PRF	0.46	0.7	0.82	0.95	1.3	2.03	1.97	2.38
LASSO	0.43	0.62	0.65	0.8	1.18	1.85	1.79	1.85
LASSO-PSI	1.59	1.18	2.02	1.98	1.66	2.51	2.33	2.08

Table 5 continued

Model	1-step	2-step	3-step	4-step	6-step	12-step	18-step	24-step
ADAP-LASSO	0.43	0.62	0.66	0.81	1.18	1.86	1.79	1.9
POST-LASSO	0.46	0.8	0.91	1.07	1.35	1.69	1.94	1.99
LASSO-ZERO	0.53	1.51	2.21	2.62	5.75	3.03	2.24	2.02
RIDGE	0.37	0.57	0.72	0.81	1.2	2.24	1.8	2.55
ELASTIC	0.42	0.6	0.66	0.78	1.13	1.72	1.84	2
ADAP-ELASTIC	0.43	0.58	0.67	0.77	1.15	1.72	1.91	2
SCAD	0.41	0.78	0.93	1.01	1.44	2.09	1.86	2.33
BMS	1.57	2.26	2.18	2.41	2.31	3.65	3.06	2.39
BMA	2.24	1.85	2.45	2.77	2.53	4.74	2.98	2.8
CSR	0.41	0.66	0.84	1.02	1.21	1.63	1.91	2.22
NEURAL-NET	0.5	0.66	0.8	0.99	1.06	1.48	1.5	1.45
RANDOM-FOREST	0.58	0.8	1.02	1.2	1.32	1.52	1.7	1.7
XG-BOOST	0.49	0.74	0.99	1.15	1.39	1.57	1.71	1.72

Boldface indicates the lowest RMSE at a particular horizon
See Table 3 for model acronyms

Table 6 Diebold–Mariano statistics (8/5-digit head)

Model	1-step	2-step	3-step	4-step	6-step	12-step	18-step	24-step
AR(1)	0.11	− 0.84	− 0.8	− 0.98	− 1.09	− 0.8	− 0.55	− 0.31
AR-DIRECT	0.02	− 1.06	− 0.86	− 0.9	− 0.9	− 0.46	− 0.21	− 0.06
STOCH-VOL	− 0.37	− 0.1	− 0.03	− 0.13	− 0.27	− 0.15	− 0.03	0.1
BVAR	− 1.42	− 0.53	− 0.54	− 0.69	− 1.04	− 0.58	− 0.39	− 0.22
DIM	*2.19*	1.46	0.91	0.46	− 0.61	− 0.78	− 0.5	− 0.43
SARB	*3.67*	*2.3*	1.84	1.88	0.35	− 0.61	− 0.56	− 0.37
DFM-TF	− 0.69	− 0.29	− 0.7	− 1.33	− 1.42	− 1.31	− 0.47	− 0.58
DFM-3PRF	− 1.32	− 0.85	− 0.67	− 0.78	− 1.62	− 0.87	− 0.81	− **2.84**
LASSO	− 0.37	0.51	*2.09*	0.55	− 1.03	− 1.34	− 1.05	− 1
ADAP-LASSO	− 0.37	0.56	1.8	0.42	− 1.03	− 1.35	− 1.05	− 1.5
POST-LASSO	− 1.38	− 1.55	− **1.98**	− **2.64**	− 1.29	− 0.75	− 0.77	− 1.7

Table 6 continued

Model	1-step	2-step	3-step	4-step	6-step	12-step	18-step	24-step
LASSO-ZERO	− **2.94**	− **3.04**	− **3.95**	− **2.81**	− 1.59	− **4.79**	− 1.83	− 0.56
RIDGE	0.92	1.08	0.53	0.29	− 0.83	− 0.92	− 0.73	− **3.42**
ELASTIC	− 0.19	0.68	1.54	0.81	− 0.83	− 0.94	− 1.12	− 1.8
ADAP-ELASTIC	− 0.28	1.06	1.4	0.91	− 0.89	− 0.93	− 1.37	− **2.14**
SCAD	0.25	− 1.67	− 1.81	− **2.47**	− 1.67	− 0.97	− 0.94	− 1.1
BMS	− **3.71**	− **4.83**	− **4.69**	− **2.93**	− **3.3**	− **3.69**	− **2.24**	− 0.79
BMA	− **4.09**	− **4.14**	− **5.68**	− **3.06**	− **3.78**	− **5.77**	− **2**	− 0.99
CSR	0.15	− 0.38	− 1.36	− **2.03**	− 1.42	− 0.7	− 0.56	− 1.41
NEURAL-NET	− 1.95	− 0.49	− 0.72	− 1.61	− 1.5	− 0.64	− 0.5	0.01
RANDOM-FOREST	− 1.93	− **2.26**	− **3.25**	− **2.82**	− 1.66	− 0.6	− 0.44	− 0.38
XG-BOOST	− **2.17**	− 1.43	− **3.16**	− **2.18**	− 1.86	− 0.64	− 0.47	− 0.47

Boldface indicates that the difference in forecasting accuracy is significantly different from zero in favour of the random-walk forecast, while *italics* indicate that the difference is significant, but in favour of the competing model

See Table 3 for model acronyms

which are displayed in Table 7. Note that in this case, the SARB forecasts are superior over a one- and two-month horizon, while the random-walk model provides superior forecasts for between three and six months ahead. Thereafter, the lowest RMSEs are provided by the ridge and random forest models. Once again, some of the worst results are provided by the linear regression model and after generating the in-sample summary statistics for the models that make use of variable selection techniques, we observe that the matrix that contains the predictors displays sparse characteristics.

Table 8 contains the Diebold and Mariano (1995) statistics, which suggest that there is no occasion where the difference in forecasting performance, relative to the random-walk, is significantly different from zero, in favour of the competing model (even in the case of the short-term SARB forecasts). In addition, there are also a number of occasions where the random-walk model provides results that are significantly more accurate than any of the competing models.

5.4 Eight/five-digit data: core inflation

After excluding those items that are not included in the definition of core inflation, we are left with eighty-three price indices, which are measured at a five-digit level, since this measure does not include any food items. Table 9 contains the out-of-sample RMSEs for the different models, where we note that the results are fairly similar to the case where we made use of less disaggregated data. In this case, there is only one occasion where the random-walk model does not generate the lowest RMSE over the medium- to long-term horizon.

The Diebold and Mariano (1995) statistics that are contained in Table 10 suggest that there is no occasion where there is a significant difference in forecasting performance, in favour of the models that are competing with the random-walk.

5.5 Change in the inflationary level or trend

Following the onset of the COVID-19 pandemic, South Africa initially went into lockdown on 27 March 2020. The use of these regulations resulted in what could be described as a level-shift in the rate of inflation, where between April 2019 and March 2020, year-on-year headline inflation averaged 4.2%, while between April 2020 and March 2021, it only averaged 2.9% (which is below the lower bound of the inflation target). In what follows, we discuss the relative performance of the different models following this change in the data-generating process, given the limitation that we only have twelve observations that arise after the onset of the pandemic.

In what is similar to Table 2, we compare the RMSE results for the official SARB forecasts to the best-performing dynamic factor, variable selection and nonlinear statistical learning models for the out-of-sample period that extends between April 2020 and March 2021. The full results for all the models over this out-of-sample period have been included in section G of the online appendix. Note that for the twelve-step-ahead forecast, we are only able to calculate the RMSE for a single realisation, and as such

Table 7 Root-mean-squared error (four-digit core)

Model	1-step	2-step	3-step	4-step	6-step	12-step	18-step	24-step
AR(1)	0.19	0.3	0.39	0.48	0.58	0.97	1.17	1.26
AR-DIRECT	0.19	0.3	0.38	0.48	0.59	0.99	1.2	1.29
STOCH-VOL	0.18	0.28	0.34	0.41	0.51	0.83	1.01	1.11
RAND-WALK	0.18	0.26	**0.32**	**0.39**	**0.47**	0.76	0.94	1.04
BVAR	0.33	0.47	0.57	0.65	0.78	0.99	1.04	1.05
DIM	0.2	0.33	0.42	0.52	0.68	1.22	1.36	1.45
SARB	**0.13**	**0.24**	0.35	0.46	0.63	1.19	1.25	1.32
LINEAR	0.24	0.4	0.5	0.66	0.69	2.36	1.76	1.65
DFM-TF	0.23	0.36	0.48	0.62	0.82	1.44	1.6	1.8
DFM-3PRF	0.24	0.35	0.46	0.63	0.95	1.55	1.44	1.43
LASSO	0.28	0.37	0.43	0.49	0.67	1.38	1.28	1.27
LASSO-PSI	0.47	0.56	0.62	1	0.64	1.33	1.18	1.84
ADAP-LASSO	0.27	0.38	0.44	0.5	0.69	1.4	1.27	1.26
POST-LASSO	0.25	0.33	0.44	0.54	0.76	1.34	1.04	1.53

Table 7 continued

Model	1-step	2-step	3-step	4-step	6-step	12-step	18-step	24-step
LASSO-ZERO	0.18	0.29	0.39	0.54	0.65	1.52	1.46	6.18
RIDGE	0.26	0.41	0.61	0.75	0.71	1.76	1.34	**0.96**
ELASTIC	0.25	0.37	0.45	0.49	0.61	1.39	1.23	1.22
ADAP-ELASTIC	0.26	0.39	0.46	0.49	0.63	1.4	1.2	1.25
SCAD	0.18	0.29	0.33	0.39	0.68	1.02	1.22	1.51
BMS	0.23	0.38	0.48	0.59	0.64	1.93	1.41	5.28
BMA	3.11	1.81	2.27	2.84	1.96	2.8	2.62	2.69
CSR	0.2	0.31	0.41	0.51	0.71	1.31	1.37	1.38
NEURAL-NET	0.3	0.44	0.47	0.56	0.74	1.2	1.31	1.25
RANDOM-FOREST	0.27	0.41	0.52	0.63	0.8	**0.71**	**0.85**	1.17
XG-BOOST	0.23	0.37	0.46	0.58	0.73	0.84	0.88	1.28

Boldface indicates the lowest RMSE at a particular horizon
See Table 3 for model acronyms

Table 8 Diebold–Mariano statistics (four-digit core)

Model	1-step	2-step	3-step	4-step	6-step	12-step	18-step	24-step
AR(1)	− **2.5**	− **2.25**	− **2.29**	− 1.89	− 1.16	− 0.56	− 0.39	− 0.33
AR-DIRECT	− **2.62**	− **2.65**	− **2.55**	− **2.79**	− **2.05**	− 0.98	− 0.67	− 0.49
STOCHVOL	− **2.57**	− **2.44**	− **2.39**	− 1.95	− 1.72	− 0.92	− 0.51	− 0.41
BVAR	− **3.96**	− **3.02**	− **2.77**	− **2.28**	− 1.72	− 0.65	− 0.22	− 0.03
DIM	− 1.78	− **2.6**	− **2.91**	− **3.02**	− **3.33**	− **2.21**	− 1.25	− 0.89
SARB	1.37	0.49	− 0.68	− 1.49	− 1.82	− 1.84	− 0.84	− 0.56
LINEAR	− **2.87**	− **2.55**	− **2.84**	− **2.86**	− **4.2**	− 1.5	− **2.42**	− 0.72
DFM-TF	− **3.59**	− **2.89**	− **2.49**	− **2.31**	− **2.27**	− **2.87**	− 1.54	− 1.77
DFM-3PRF	− **3.15**	− **2.46**	− **2**	− **2.06**	− **1.98**	− **4.22**	− 0.94	− 0.69
LASSO	− **3.64**	− **2.21**	− **2.12**	− **2.08**	− 1.68	− **5.54**	− **2.42**	− 1.35
LASSO-PSI	− **5.24**	− **6.47**	− **4.29**	− **3.67**	− **2.08**	− **2.03**	− 1.45	− 0.79
ADAP-LASSO	− **3.98**	− **2.27**	− **2.47**	− **2.23**	− 1.88	− **5.72**	− **4.57**	− **2**

Table 8 continued

Model	1-step	2-step	3-step	4-step	6-step	12-step	18-step	24-step
POST-LASSO	− **2.9**	− 1.55	− 1.83	− 1.88	− 1.29	− **3.55**	− 0.37	− 1.49
LASSO-ZERO	− 0.96	− 0.89	− 1.23	− **2.2**	− **2.33**	− **11.19**	− 1.26	− 0.67
RIDGE	− **2.84**	− **2.63**	− **2.3**	− 1.44	− **2.89**	− 1.93	− **2.33**	0.56
ELASTIC	− **3.09**	− **2.1**	− **2.58**	− **2.17**	− 1.55	− **3**	− **6.79**	− **4.93**
ADAP-ELASTIC	− **3.81**	− **2.25**	− **2.53**	− **2.09**	− 1.61	− **2.36**	− **4.82**	− **5.11**
SCAD	0	− 1.47	− 0.49	− 0.09	− 1.78	− 1.7	− 1.36	− 0.9
BMS	− **2.74**	− **2.19**	− **2.71**	− **2.86**	− 1.87	− 1.34	− **2.28**	− 0.68
BMA	− **7.41**	− **4.59**	− **5.32**	− **2.27**	− **3.52**	− **3.09**	− 1.58	− 1.88
CSR	− **2.37**	− 1.86	− **2.76**	− **4.31**	− **4.59**	− 1.78	− 0.76	− 0.55
NEURAL-NET	− **3.87**	− **3.07**	− **2.11**	− 1.81	− **2.32**	− 0.89	− 0.49	− 0.31
RANDOM-FOREST	− **3.07**	− **2.67**	− **3.1**	− **3.59**	− **3.24**	0.29	0.36	− 0.32
XG-BOOST	− **2.72**	− **2.16**	− **2.01**	− **2.46**	− **3.42**	− 0.66	0.15	− 0.5

Boldface indicates that the difference in forecasting accuracy is significantly different from zero in favour of the random-walk forecast, while *italics* indicate that the difference is significant, but in favour of the competing model

See Table 3 for model acronyms

Table 9 Root-mean-squared error (8/5-digit core)

Model	1-step	2-step	3-step	4-step	6-step	12-step	18-step	24-step
AR(1)	0.18	0.28	0.36	0.45	0.55	1	1.27	1.43
AR-DIRECT	0.19	0.29	0.36	0.45	0.55	0.89	1.06	1.12
STOCH-VOL	0.18	0.27	0.34	0.41	0.5	0.8	0.98	1.07
RAND-WALK	0.18	0.26	**0.32**	**0.39**	**0.47**	**0.76**	0.94	**1.04**
BVAR	0.3	0.42	0.51	0.59	0.72	1.01	1.13	1.22
DIM	0.2	0.33	0.42	0.52	0.68	1.22	1.36	1.45
SARB	**0.13**	**0.24**	0.35	0.46	0.63	1.19	1.25	1.32
DFM-TF	0.18	0.27	0.33	0.39	0.55	1.09	1.31	1.46
DFM-3PRF	0.27	0.43	0.6	0.77	1.08	1.63	1.47	1.57
LASSO	0.28	0.36	0.38	0.55	0.83	1.49	1.04	1.51

Table 9 continued

Model	1-step	2-step	3-step	4-step	6-step	12-step	18-step	24-step
LASSO-PSI	0.52	0.85	0.96	0.65	1.03	1.09	1.72	1.37
ADAP-LASSO	0.27	0.37	0.37	0.52	0.82	1.52	1.03	1.52
POST-LASSO	0.22	0.36	0.45	0.54	0.72	1.14	1.25	1.46
LASSO-ZERO	0.92	1.13	0.8	1.08	1.35	1.86	2.28	1.44
RIDGE	0.28	0.47	0.56	0.53	0.65	1.09	1.32	1.6
ELASTIC	0.26	0.34	0.36	0.59	0.72	1.51	1.04	1.52
ADAP-ELASTIC	0.27	0.35	0.36	0.57	0.7	1.49	1.03	1.53
SCAD	0.19	0.32	0.4	0.54	0.78	1.13	1.31	1.6
BMS	0.73	1.03	1.13	1.24	1.03	2.05	2.19	1.59
BMA	0.64	0.53	0.95	0.83	0.78	2.15	1.78	1.48
CSR	0.2	0.33	0.43	0.5	0.63	1.02	1.28	1.32
NEURAL-NET	0.25	0.39	0.39	0.42	0.55	0.86	**0.91**	1.09
RANDOM-FOREST	0.24	0.35	0.41	0.51	0.64	0.91	1.24	1.36
XG-BOOST	0.22	0.37	0.38	0.47	0.6	0.89	1.19	1.36

See Table 3 for model acronyms. **Boldface** indicates the lowest RMSE at a particular horizon

Table 10 Diebold–Mariano statistics (8/5-digit core)

Model	1-step	2-step	3-step	4-step	6-step	12-step	18-step	24-step
AR(1)	− **1.99**	− **2.7**	− **2.98**	− **2.77**	− **2.09**	− 1.36	− 1.05	− 0.94
AR-DIRECT	− **2.17**	− **2.05**	− 1.93	− **2.03**	− 1.41	− 0.63	− 0.36	− 0.21
STOCH-VOL	− **2.65**	− **2.36**	− **2.28**	− 1.77	− 1.46	− 0.62	− 0.31	− 0.18
BVAR	− **3.09**	− **2.78**	− **2.7**	− **2.31**	− 1.95	− 0.92	− 0.44	− 0.34
DIM	− 1.78	− **2.6**	− **2.91**	− **3.02**	− **3.33**	− **2.21**	− 1.25	− 0.89
SARB	1.37	0.49	− 0.68	− 1.49	− 1.82	− 1.84	− 0.84	− 0.56
DFM-TF	− 0.79	− 0.16	− 0.4	− 0.16	− 0.66	− 0.86	− **2.38**	− 1.32
DFM-3PRF	− **2.87**	− **1.98**	− 1.89	− 1.71	− 1.85	− **2.26**	− **6.55**	− 1.77
LASSO	− **3.32**	− **2.73**	− 1.51	− **2.21**	− **5.8**	− 1.6	− 0.22	− 0.73
LASSO-PSI	− **4.55**	− **3.1**	− **2.46**	− **2.96**	− 1.11	− 0.75	− **5.89**	− 0.47
ADAP-LASSO	− **3**	− **2.87**	− 1.4	− **2.06**	− **5.64**	− 1.51	− 0.19	− 0.71
POST-LASSO	− **2.36**	− 1.76	− **1.98**	− 1.68	− 1.81	− **6.37**	− 1.6	− 1.21

Table 10 continued

Model	1-step	2-step	3-step	4-step	6-step	12-step	18-step	24-step
LASSO-ZERO	– **9.43**	– **3.49**	– **3.03**	– **2.89**	– **2.82**	– **3.89**	– **2.46**	– **2.18**
RIDGE	– **3.66**	– **4.11**	– **3.75**	– **2.06**	– 1.94	– 0.94	– 0.52	– 0.74
ELASTIC	– **3.15**	– **2.36**	– 1.11	– **2.13**	– **3.7**	– 1.27	– 0.21	– 0.69
ADAP-ELASTIC	– **3.22**	– **2.33**	– 1.17	– **2.14**	– **3.66**	– 1.18	– 0.19	– 0.68
SCAD	– 1.62	– **2.3**	– **2.12**	– **2.24**	– **2.92**	– **2.08**	– 1.3	– 0.92
BMS	– **7.78**	– **4.97**	– **3.82**	– **3.86**	– 1.93	– **2**	– 1.13	– 1.23
BMA	– **3.44**	– **2.75**	– **3.14**	– **3.29**	– **2.11**	– **3.16**	– 0.86	– 1.01
CSR	– **2.38**	– **2.12**	– **2.39**	– **2.09**	– 1.74	– **4.76**	– 1.07	– 0.51
NEURAL-NET	– **2.64**	– **2.13**	– 1.21	– 0.91	– 0.98	– 0.37	0.06	– 0.1
RANDOM-FOREST	– **2.58**	– **2.28**	– **2.42**	– **2.9**	– **2.88**	– 0.95	– 0.59	– 0.63
XG-BOOST	– **2.67**	– **2.12**	– 1.51	– 1.56	– **3.34**	– 0.59	– 0.68	– 0.66

Boldface indicates that the difference in forecasting accuracy is significantly different from zero in favour of the random-walk forecast, while *italics* indicate that the difference is significant, but in favour of the competing model. See Table 3 for model acronyms

Table 11 Root-mean-squared error

	Step-1	Step-2	Step-3	Step-4	Step-6	Step-8	Step-10	Step-12
Headline inflation								
SARB	1.06	1.08	0.86	0.63	0.64	1.13	1.37	1.7
4 DFM-3PRF	0.64	0.97	0.92	0.81	1.16	1.79	1.75	1.7
4 LASSO	**0.47**	**0.72**	**0.58**	**0.51**	1.82	**0.9**	2.8	1.94
4 CSR	0.52	0.8	0.7	0.53	0.7	1.21	1.88	2.32
8/5 DFM-TF	0.6	0.92	0.93	0.96	1.44	2.05	2.56	2.51
8/5 ELASTIC	**0.47**	**0.7**	**0.67**	**0.6**	1.32	**1.02**	**1.29**	2.32
8/5 CSR	0.52	0.81	0.8	0.76	0.79	1.08	1.32	1.88
Core inflation								
SARB	0.47	0.51	0.51	0.51	0.56	0.87	1.16	1.42
4 DFM-TF	**0.28**	**0.42**	0.45	0.45	**0.37**	**0.38**	0.84	1.67
4 LASSO	0.35	0.43	0.38	**0.41**	0.51	0.47	1.26	1.94
4 NEURAL	0.37	0.43	**0.34**	0.49	0.48	0.61	**0.74**	1.74
8/5 DFM-TF	**0.25**	**0.35**	0.39	0.44	0.43	**0.27**	0.45	1.15
8/5 POST-LA	0.27	0.42	0.42	**0.35**	**0.34**	0.62	0.67	0.99
8/5 NEURAL	0.3	0.42	**0.35**	0.39	0.45	0.56	**0.14**	**0.77**

Boldface indicates the lowest RMSE at a particular horizon and dataset
Data acronyms: “4”—Four-digit data, “8/5”—Eight/five-digit data. Model acronyms: “SARB"—official SARB forecast reported to MPC, “DFM-TF"—dynamic factor model with target factors, “DFM-3PRF"—dynamic factor model with three pas filter, “LASSO"—least absolute shrinkage and selection operator, “POST-LA"—post-OLS LASSO, “CSR"—complete subset regression, “NEURAL"—neural network

it is difficult to read too much into this result, while the RMSE for the one-step ahead forecast is computed over the average of twelve successive forecasts.[16]

These results suggest that if we were to impose a limit on the forecasting horizon at eight-steps ahead (or where we have at least 5 successive forecasts to evaluate), then there is always a statistical learning model that provides a lower RMSE, relative to the official SARB forecasts that may benefit from the use of off-model and within-month information. In addition, we also note that in general, for headline inflation, the variable selection models perform reasonably well, which may suggest that the removal of those variables that are unable to make a significant contribution towards the predictive ability of the model provide more accurate forecasts (where one would presume that the variables that are removed are unable to contribute towards the explanation of the level shift). Similarly, for core inflation, where the number of available predictors is somewhat limited, combining all the available information within a DFM would in most cases provide the most desirable forecast.

Figure 2 contains the results of the recursive one-step ahead forecasts that were generated for headline and core inflation, by the LASSO and DFM (with target factors), between April 2020 and March 2021. In both cases, the statistical learning models appear to have done a reasonable job of detecting the relative change in the inflationary

[16] To identify the best model within each class, we take the mean of the RMSE for the one-to-six-step-ahead forecasts.

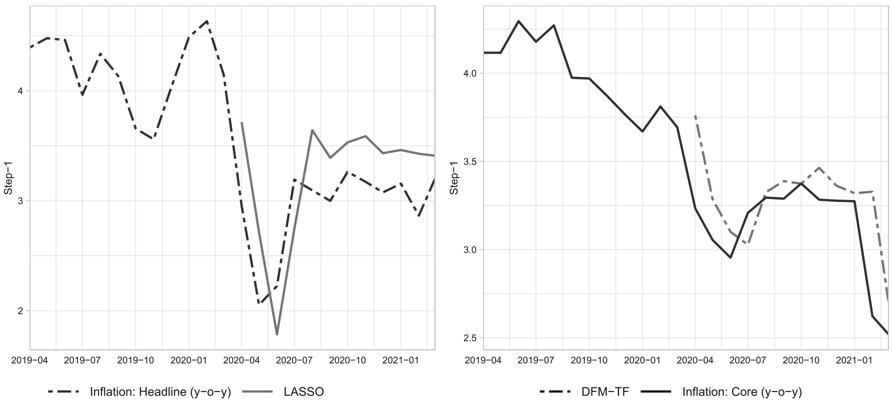

Fig. 2 Forecasts from benchmark and statistical learning models—headline inflation

level or trend.[17] While these results are of interest, particularly to those who are concerned with the relative performance of various models over the pandemic, one should be cautious of reading too much into them as they have been generated from a very small sample.

6 Conclusion

We assess the potential predictive power of a number of different forecasting models that may be applied to large datasets that are used to measure inflation. We find that the models that employ variable selection techniques and nonlinear statistical learning techniques provide impressive results, despite the fact that the number of observations in the dataset is limited. We also note that when comparing the use of models that seek to exploit any potential sparsity in the set of predictors, relative to those that seek to summarise all of the available information, the results are somewhat mixed, over the entire out-of-sample period. Over horizons that are longer than three months, the statistical learning models would also appear to provide results that are even more accurate than the sparse models, where the neural network and boosting models provide the most accurate results. However, the results of simple forecasting models continue to produce results that are in many cases superior to those of the statistical learning models. Hence, one would conclude that from a practical perspective, the use of statistical learning models in this particular setting may not provide forecasts that are consistently superior to what is provided by a simple random-walk model, although they are certainly competitive.

Furthermore, the results suggest that for headline inflation the official central bank forecast that is presented to the MPC, which incorporates various sources of off-model and within-month information, is more accurate than any of the other models, over

[17] Previously, Stock and Watson (2010) suggested that to account for the change in the inflationary trend one could augment the previous specification that was utilised in Stock and Watson (2007), with a stochastic trend that reacts to the unemployment recession gap. However, as is the case with most low- and middle-income countries, South Africa does not have a reliable measure for the unemployment recession gap.

the first three months. Similarly, over a one-month horizon the central bank forecast for core inflation is more accurate than any of the other models. Hence, the use of judgement has systematically improved the SARB forecasts over a short-term horizon. Another important finding relates to the use of more disaggregated data, where the results from the eight/five-digit data are generally more accurate than when we report on the use of the four-digit data, which suggests that the use of more disaggregated data provides more desirable results. In particular, those models that are able to distinguish between information that may or may not be of potential use are able to provide more accurate forecasts when they are applied to more disaggregated data. As has been shown, we can also use the output from the models to generate Shapley values, which provide policymakers with information that pertains to the drivers of future inflationary pressure. In addition, when we consider the relative performance of the benchmark models, which include a number of mean-reverting specifications, for the period that includes the effects of economic lockdowns over the pandemic, we note that the statistical learning models are able to detect the decrease in the trend of the respective measures of inflation reasonably quickly, to provide short-term forecasts that are more accurate than what was provided to the MPC.

Subsequent research into the use of alternative sources of big data, as well as the potential use of alternative statistical learning model specifications, may provide more promising forecasting results in the future. As has been noted, the number of available observations over time for this dataset is relatively limited, and as it is generally acknowledged that to provide impressive results in such a setting, statistical learning models, and in particular the nonlinear variants of these models, would usually require a relatively large number of observations that have been measured over time. Nevertheless, the fact that the forecasts from many of these models are competitive, despite the limitation of the data, may provide encouraging signs for researchers in this field of study.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00181-022-02329-y.

Declarations

Conflict of interest The authors declare that they have no conflicts of interest.

Human and animal rights This article does not contain any studies with animals performed by any of the authors. This article does not contain any studies with human participants or animals performed by any of the authors.

References

Agrawal A, Gans J, Goldfarb A (2019) The economics of artificial intelligence: an agenda. University of Chicago Press, Chicago

Alpanda S, Kotzé K, Woglom G (2010) The role of the exchange rate in a new Keynesian DSGE model for the South African economy. S Afr J Econ 78(2):170–191

Alpanda S, Kotzé K, Woglom G (2010) Should central banks of small open economies respond to exchange rate fluctuations? The case of South Africa. ERSA working paper no. 174, Economic Research Southern Africa

Alpanda S, Kotzé K, Woglom G (2011) Forecasting performance of an estimated DSGE model for the South African economy. S Afr J Econ 79(1):50–67

Athey S (2017) Beyond prediction: using big data for policy problems. Science 355(6324):483–485

Athey S (2018) The impact of machine learning on economics. University of Chicago Press, pp 507–547

Athey S, Imbens GW (2019) Machine learning methods that economists should know about. Annu Rev Econ 11(1):685–725

Bai J (2003) Inferential theory for factor models of large dimensions. Econometrica 71(1):135–171

Bai J, Ng S (2008) Large dimensional factor analysis. Found Trends Econom 3(2):89–163

Baker SR, Bloom N, Davis SJ, Terry SJ (2020) Covid-induced economic uncertainty. Working Paper 26983, National Bureau of Economic Research

Baker SR, Farrokhnia RA, Meyer S, Pagel M, Yannelis C, Pontiff J (2020) How does household spending respond to an epidemic? Consumption during the 2020 COVID-19 pandemic. Rev Asset Pric Stud 10(4):834–862

Balcilar M, Gupta R, Kotzé K (2015) Forecasting macroeconomic data for an emerging market with a nonlinear DSGE model. Econ Model 44:215–228

Balcilar M, Gupta R, Kotzé K (2017) Forecasting South African macroeconomic variables with a Markov-switching small open-economy dynamic stochastic general equilibrium model. Empir Econ 53(1):117–135

Baldacci E, Buono D, Kapetanios G, Krische S, Marcellino M, Mazzi GL, Papailias F (2016) Big data and macroeconomic nowcasting: from data access to modelling. European Union, Eurostat, Luxembourg

Belloni A, Chernozhukov V, Fernández-Val I, Hansen C (2017) Program evaluation and causal inference with high-dimensional data. Econometrica 85(1):233–298

Belloni A, Chernozhukov V, Hansen C (2013) Inference on treatment effects after selection among high-dimensional controls. Rev Econ Stud 81(2):608–650

Belloni A, Chernozhukov V, Hansen C (2014) High-dimensional methods and inference on structural and treatment effects. J Econ Perspect 28(2):29–50

Belloni A, Chernozhukov V, Wang L (2011) Square-root LASSO: pivotal recovery of sparse signals via conic programming. Biometrika 98(4):791–806

Blumenstock J (2020) Machine learning can help get COVID-19 aid to those who need it most. Nature

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Buckman SR, Shapiro AH, Sudhof M, Wilson DJ (2020) News sentiment in the time of COVID-19. FRBSF Economic Letter 2020-08, Federal Reserve Bank of San Francisco

Carvalho VM, Hansen S, Ortiz A, Ramón García J, Rodrigo T, Rodriguez Mora S, Ruiz J (2020) Tracking the COVID-19 crisis with high-resolution transaction data. CEPR Discussion Papers 14642, C.E.P.R. Discussion Papers

Castle JL, Doornik JA, Hendry DF (2021) The value of robust statistical forecasts in the COVID-19 pandemic. Natl Inst Econ Rev 256:19–43

Cavallo A (2020) Inflation with COVID consumption baskets. Working Paper 27352, National Bureau of Economic Research

Chakrabarti R, Heise S, Melcangi D, Pinkovskiy M, Topa G (2020) Did state reopenings increase consumer spending? Liberty street economics, Federal Reserve Bank of New York

Chakrabarti R, Heise S, Melcangi D, Pinkovskiy M, Topa G (2020) How did state reopenings affect small businesses? Liberty street economics, Federal Reserve Bank of New York

Chetty R, Friedman JN, Hendren N, Stepner M, Team TOI (2020) The economic impacts of COVID-19: Evidence from a new public database built using private sector data. NBER Working Papers 27431, National Bureau of Economic Research, Inc

Chu B, Huynh K, Jacho-Chavez D, Kryvtsov O (2018) On the evolution of the United Kingdom price distributions. Ann Appl Stat 12(4):2618–2646

Clark TE, West KD (2007) Approximately normal tests for equal predictive accuracy in nested models. J Econom 138:291–311

Coulombe PG, Leroux M, Stevanović D, Surprenant S (2022) How is machine learning useful for macroeconomic forecasting? J Appl Econom 37:920–964

Coulombe PG, Marcellino M, Stevanović D (2021) Can machine learning catch the COVID-19 recession? Natl Inst Econ Rev 256:71–109

Creamer K, Farrel G, Rankin N (2012) What price-level data can tell us about pricing conduct in South Africa. S Afr J Econ 80(4):490–509

Creamer K, Rankin N (2008) Price setting in South Africa 2001 to 2007 stylised facts using consumer price micro data. J Dev Perspect 4(1):93–118

Diebold FX, Mariano RS (1995) Predictive accuracy. J Bus Econ Stat 13(3):253–263

Doerr S, Gambacorta L, Serena JM (2021) Big data and machine learning in central banking. BIS Working Papers 930, Bank of International Settlements

Duarte C, Rua A (2007) Forecasting inflation through a bottom-up approach: how bottom is bottom? Econ Model 24(6):941–953

Elliott G, Gargano A, Timmermann A (2013) Complete subset regressions. J Econom 177(2):357–373

Elliott G, Gargano A, Timmermann A (2015) Complete subset regressions with large-dimensional sets of predictors. J Econ Dyn Control 54(C):86–110

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96(456):1348–1360

Faust J, Wright JH (2013) Forecasting inflation. In: Elliott G, Granger C, Timmermann A (eds) Handbook of economic forecasting, 2:2–56. Elsevier

Florescu D, Karlberg M, Reis F, Rey Del Castillo P, Skaliotis M, Wirthmann A (2014) Will 'big data' transform official statistics? European Union, Eurostat, Luxembourg

Forni M, Hallin M, Lippi M, Reichlin L (2000) The generalized dynamic-factor model: identification and estimation. Rev Econ Stat 82(4):540–554

Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting. Ann Stat 28(2):337–374

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232

Fuhrer JC (2010) Inflation persistence. In Friedman BM, Woodford M (eds) Handbook of Monetary Economics, 3(9):423–486. Elsevier

Galvao AB (2021) The COVID-19 pandemic and macroeconomic forecasting: An introduction to the spring 2021 special issue. Natl Inst Econ Rev 256:16–18

Giannone D, Lenza M, Primiceri GE (2021) Economic predictions with big data: the illusion of sparsity. Working Paper Series 2542, European Central Bank

Gupta R, Kabundi A (2010) Forecasting macroeconomic variables in a small open economy: a comparison between small- and large-scale models. J Forecast 29(1–2):168–185

Gupta R, Kabundi A (2011) A large factor model for forecasting macroeconomic variables in South Africa. Int J Forecast 27(4):1076–1088

Gupta R, Steinbach R (2013) A DSGE-VAR model for forecasting key South African macroeconomic variables. Econ Model 33:19–33

Hammer C, Kostroch D, Quiros-Romero G (2017) Big data: Potential, challenges and statistical implications. Washington, DC

Hoerl AE, Kennard RW (1970a) Ridge regression: applications to nonorthogonal problems. Technometrics 12(1):69–82

Hoerl AE, Kennard RW (1970b) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12(1):55–67

Hubrich K, Hendry DF (2005) Forecasting aggregates by disaggregates. Computing in Economics and Finance 2005 270, Society for Computational Economics

Ibarra R (2012) Do disaggregated CPI data improve the accuracy of inflation forecasts? Econ Model 29(4):1305–1313

Johnson VE, Rossell D (2010) On the use of non-local prior densities in Bayesian hypothesis tests. J R Stat Soc Ser B (Stat Methodol) 72(2):143–170

Johnson VE, Rossell D (2012) Bayesian model selection in high-dimensional settings. J Am Stat Assoc 107(498):649–660

Joseph A, Kalamara E, Kapetanios G, Potjagailo G (2020) Forecasting UK inflation bottom up. Nontraditional Data and Statistical Learning with Applications to Macroeconomics, Bank of Italy and Federal Reserve Board

Joseph A, Kalamara E, Kapetanios G, Potjagailo G (2021) Forecasting UK inflation bottom up. Staff Working Paper 915, Bank of England

Kanda PT, Balcilar M, Bahramian P, Gupta R (2016) Forecasting South African inflation using non-linear models: a weighted loss-based evaluation. Appl Econ 48(26):2412–2427

Kelly B, Pruitt S (2013) Market expectations in the cross-section of present values. J Finance 68(5):1721–1756

Kelly B, Pruitt S (2015) The three-pass regression filter: a new approach to forecasting using many predictors. J Econom 186(2):294–316

Koop G, McIntyre S, Mitchell J, Poon A (2021) Nowcasting 'true' monthly US GDP during the pandemic. Natl Inst Econ Rev 256:44–70

Lee JD, Sun DL, Sun Y, Taylor JE (2016) Exact post-selection inference with the LASSO. Ann Stat 44(3):907–927

Liu GD, Gupta R, Schaling E (2009) A New-Keynesian DSGE model for forecasting the South African economy. J Forecast 28(5):387–404

Lundberg S, Lee S-I (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30:4765–4774

McCracken MW (2007) Asymptotics for out of sample tests of granger causality. J Econom 140(2):719–752

McCracken MW, Ng S (2016) FRED-MD: a monthly database for macroeconomic research. J Bus Econ Stat 34(4):574–589

McCullagh P, Polson NG (2018) Statistical sparsity. Biometrika 105(4):797–814

Medeiros MC, Vasconcelos GFR, Veiga Á, Zilberman E (2021) Forecasting inflation in a data-rich environment: the benefits of machine learning methods. J Bus Econ Stat 39(1):98–119

Mehrhoff J (2017) Central banks' use of and interest in big data. Bank for International Settlements (ed) Big Data, IFC Bulletins. Bank for International Settlements

Mullainathan S, Spiess J (2017) Machine learning: an applied econometric approach. J Econ Perspect 31(2):87–106

OECD (2020) Using artificial intelligence to help combat COVID-19. Organisation for Economic Co-operation and Development

Petrella I, Santoro E, Simonsen LP (2019) Time-varying price flexibility and inflation dynamics. EMF Research Papers 28, Economic Modelling and Forecasting Group

Rossell D (2021) Concentration of posterior probabilities and normalized L_0 criteria in regression. Bayesian Anal 1(1):1–27

Rossell D, Telesca D (2017) Nonlocal priors for high-dimensional estimation. J Am Stat Assoc 112(517):254–265

Ruch F, Balcilar M, Gupta R, Modise MP (2020) Forecasting core inflation: the case of South Africa. Appl Econ 52(28):3004–3022

Ruch F, Rankin N, du Plessis S (2016) Decomposing inflation using micro-price data: sticky-price inflation. South African Reserve Bank Working Paper Series 7354, South African Reserve Bank

Ruch F, Rankin N, du Plessis S (2016) Decomposing inflation using micro price level data: South Africa's pricing dynamics. Working Papers 7353, South African Reserve Bank

Schmitt-Grohé S, Uribe M (2004) Solving dynamic general equilibrium models using a second order approximation of the policy function. J Econ Dyn Control 28:755–75

Shapiro AH, Sudhof M, Wilson DJ (2017) Measuring news sentiment. Working Paper 2017-01, Federal Reserve Bank of San Francisco

Smal D, Pretorius C, Ehlers N (2007) The core forecasting model of the South African Reserve Bank. Working Paper WP/07/02, South African Reserve Bank

Statistics South Africa (2007) Shopping for two: the CPI new basket parallel survey—results and comparisons with published CPI data. Statistics South Africa

Statistics South Africa (2017a) Consumer price index: the South African CPI sources and methods manual. Statistics South Africa

Statistics South Africa (2017b) Introduction of new weights and basket for the consumer price index. Statistics South Africa

Steinbach R, Mathuloe P, Smit B (2009) An open economy New Keynesian DSGE model of the South African economy. Working Papers 3431, South African Reserve Bank

Stock JH, Watson MW (2002a) Forecasting using principal components from a large number of predictors. J Am Stat Assoc 97(460):1167–1179

Stock JH, Watson MW (2002b) Macroeconomic forecasting using diffusion indexes. J Bus Econ Stat 20(2):147–162

Stock JH, Watson MW (2007) Why has U.S. inilation become harder to forecast? J Money Credit Bank 39:3–33

Stock JH, Watson MW (2010) Modeling inflation after the crisis. Working Paper 16488, National Bureau of Economic Research

Stock JH, Watson MW (2020) Slack and cyclically sensitive inflation. J Money Credit Bank 52(S2):393–428

Tibshirani R (1996) Regression shrinkage and selection via the LASSO. J R Stat Soc Ser B (Stat Methodol) 58(1):267–288

Tissot B (2019) Big data for central banks. The use of big data analytics and artificial intelligence in central banking, IFC Bulletins. Bank for International Settlements

United Nations Global Pulse (UNGP) (2012) Big data for development: Challenges and opportunities

Varian HR (2014) Big data: new tricks for econometrics. J Econ Perspect 28(2):3–28

Wibisono O, Ari HD, Widjanarti A, Zulen AA, Tissot B (2019) The use of big data analytics and artificial intelligence in central banking. IFC Bulletins, Bank for International Settlements

Woglom G (2005) Forecasting South African inflation. S Afr J Econ 73(2):302–320

Wolters MH, Tillmann P (2015) The changing dynamics of US inflation persistence: a quantile regression approach. Stud Nonlinear Dyn Econom 19(2):161–182

World Bank (2014) Central America: Big data in action for development. Washington, DC

Zou H (2006) The adaptive LASSO and its oracle properties. J Am Stat Assoc 101(476):1418–1429

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.