# Inflation: High Frequency Estimation and Forecasting

Sonan Memon*

This version: August 2022

**Abstract**

I begin by motivating the utility of high frequency inflation estimation and review recent work done at the State Bank of Pakistan for inflation forecasting and now-casting GDP using machine learning (ML) tools. I also present stylized facts about the structure of historical and especially recent inflation trends in Pakistan. However, since the available data and already used methods cannot achieve high frequency forecasting, I discuss 3 novel techniques from recent literature including *web scrapping, scanner data and synthetic data*. Due to lack of access to scanner and web scrapped data for Pakistan, I generate synthetic data using *generative* ML models (Gaussian Copula and PAR models) and numerical analysis (cubic spline interpolation) methods. I use cubic splines to estimate monthly inflation rate from quarterly data and unknown high frequency, weekly inflation rate from actual monthly data. Meanwhile, I use a probabilistic auto-regressive ML model to forecast future short-run inflation for Pakistan from 2020 to 2023. I evaluate the accuracy of ML forecasts by comparing them with forecast error variances and predictions from conventional reduced form vector auto-regressive models (VAR)[1].

**Keywords:** High Frequency Inflation Estimation and Forecasting. Forecast Accuracy. Synthetic Data. Machine Learning. Hyperinflation. Forecasts of Inflation in Pakistan. VAR Models. Web Scrapping and Scanner Data.

**JEL Classification:** E30, E31, E32, E37, E47, E52, E58, C53.

*Research Economist, PIDE, Islamabad. smemon@pide.org.pk

[1] The replication code of this paper, using Python, R and Julia is available on my github page: https://github.com/sonanmemon/High-Frequency-Inflation-Forecasting

# Contents

# 1. Motivation

Accurate forecasting of inflation is a concern for market players, central banks, and also governments. The market participants want to update their inflation expectations in line with new information revelation so that their investment strategies are optimal. Meanwhile, central banks typically have mandates for price stability (see Cukierman et al. (1992)) and they routinely collect data on inflation expectations and forecast inflation. Hyperinflation episodes dramatically hurt hand to mouth households and this extreme economic turmoil has political consequences for governments, especially when the election period is a few quarters away (see for instance Binder (2021)).

Governments have an incentive to make inflation control a priority and interfere with central bank independence in these election periods. In fact, there is a well known classic literature on the so called political business cycle, initiated by Nordhaus (1975). For instance, Abrams and Butkiewicz (2012) revealed evidence from the Nixon tapes that President Nixon manipulated Arthur Burns[2] and the Federal Reserve into creating a political business cycle which helped ensure his reelection victory in 1972. While Nixon understood the risks that his monetary policy imposed, he chose to trade longer-term economic costs to the economy for his own short-term political profits. At the most fundamental level, hyperinflation episodes are humanitarian and social crises which can be addressed to some extent if we develop better inflation forecasting methods and independent central banks with mandates for price stability.

While central banks do collect data on consumer price indices, the frequency of collection does not allow accounting for sudden swings in inflation *and* inflation expectations. Some examples of standard measures include the HICP (Harmonized Consumer Price Index) data used in the Euro area and the CPI (consumer price index) data from USA. Such data typically tends to be quarterly in worst cases or in best cases monthly, but results are revealed in the next month after collection. However, when for instance, in a matter of few days and weeks, news about the Ukraine and Russian crisis changed the inflation landscape for many products, conventional price indices had little forecasting potential for the following inflation crisis. Similarly, inflation shocks can result from sudden change of central bank's governors or government change, terrorism episodes or political turmoil, especially in developing economies, where inflation tends to more volatile and central banks are less independent (see Vuletin and Zhu (2011)).

[2] Former Head of the Federal Reserve Bank in USA.

# 2. RESEARCH AT SBP

The State Bank of Pakistan's (SBP) research department has done some work on inflation forecasting by using machine learning methods (e.g Neural Networks) and monthly year on year (YoY) inflation rates of Pakistan from Jan 1958 to Dec 2017 Hanif et al. (2018). The *Thick ANN* (Artificial Neural Networks) model developed in this paper is found to outperform all the 15 econometric models of Pakistan economy previously developed in forecasting 24 months ahead headline inflation.

Similarly, the SBP's research team has worked on *nowcasting* GDP using data on large scale manufacturing growth (LSM) in Pakistan (see Hussain et al. (2018)) and LASSO type[3] ML methods. The models are used to extract the unique *information* from a range of variables having close association with LSM in Pakistan. The results displayed in Figure 1 below from Hussain et al. (2018) reveal that the predicted LSM series closely tracks the actual LSM series. Since LSM is available at relatively higher frequency (monthly) relative to the actual GDP (annual), it is a predictor for determinants of economic activity such as key sectors, prices, credit, interest rates and tax collection, external trade and inflows. This is in line with emerging methodologies among central banks worldwide, which are all moving toward big data and machine learning methods (see Doerr et al. (2021)).

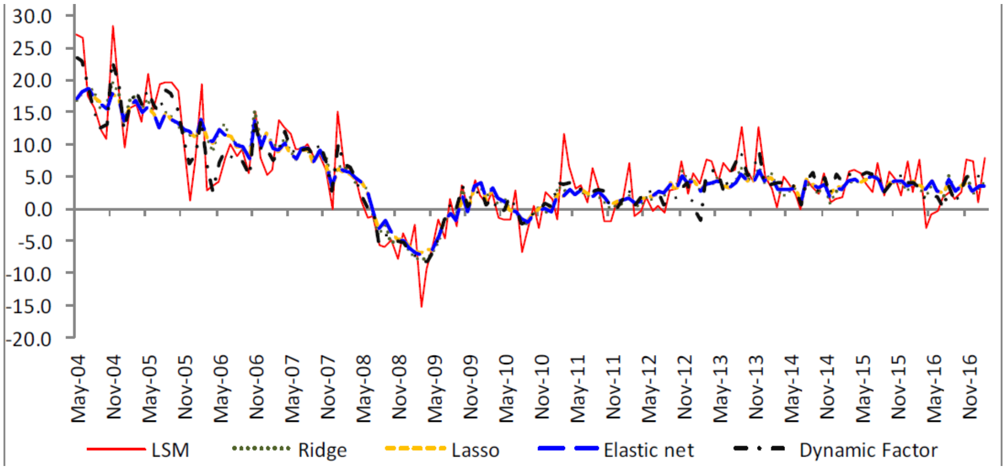

Figure 1: Nowcasting LSM For Pakistan (Source is Hussain et al. (2018))

However, lack of availability of high frequency data on the order of weeks and months, poses a limitation on forecasting inflation. Hence, I argue that we need more *granular* data for enhancing forecasting. Next, I discuss methods for collecting such high frequency

[3] Least Absolute Shrinkage Operator, Ridge Regressions and Elastic Nets.

data, currently used at the current frontier of research on inflation in economics.

# 3. Review of Modern Forecasting

In this section, I will briefly review three methods including *web scrapping* inflation data, using *scanner data* from supermarkets and *synthetic* or artificial data for high frequency inflation forecasting.

## 3.1. Web Scrapping

In recent literature, the daily consumer price index (CPI) produced by the Billion Prices Project (BPP CPI) of Cavallo and Rigobon (2016) offers a glimpse of the direction taken by consumer price inflation in *real time*. For instance, Figure 2 is based on web scrapping online inflation data for Argentina (see Cavallo and Rigobon (2016)). It shows that the official CPI significantly under-stated actual inflation, when measured by web scrapping. An added benefit of such data is that it reveals the partisan measurement and particularly disclosure of CPI data in developing economies such as Argentina, where central bank independence is low.

Should we expect a similar lack of correspondence between official inflation data of the SBP (State Bank of Pakistan) and non-partisan research measures? Not much is known about the political business cycle in Pakistan and I believe that independent research, not originating from SBP is needed to address the question. Given the low levels of central bank independence (henceforth CBI) in Pakistan and existing literature on CBI (see Vuletin and Zhu (2011)), we should expect that higher levels of price stability can be achieved if governor appointments and turnovers are not manipulated by political power.

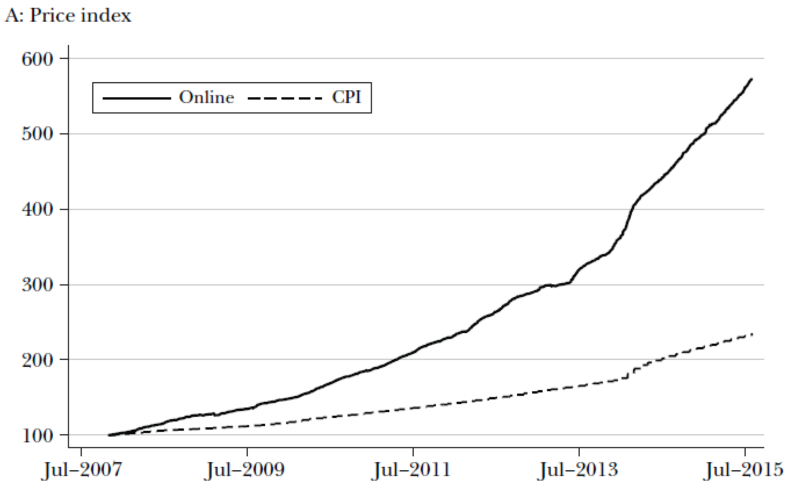

Figure 2: Inflation in Argentina (Source is Cavallo and Rigobon (2016))

With increasing scope of online transactions in Pakistan, *web scrapping* can also be informative, despite absence of Amazon type large scale online transaction services in the country.

## 3.2. Scanner Data

Meanwhile, another branch of emerging literature uses scanner-based data (see for instance Beck et al. (2020)) on prices rather than web-scrapping them. In Figure 3 below, recent scanner-based price indices for Germany are disclosed from the work of Beck et al. (2022). The data compares trajectories in 2022 (red and orange solid lines below) with their historical averages from 2019 to 2021 (blue and purple solid lines) along with historic minimum and maximum values (shaded areas). The data indicates a very strong increase in prices for sunflower oil and flour in light of the Ukraine conflict, accompanied by temporarily higher sales. The price increase of sunflower oil was rather gradual and already started as of early February. In contrast, prices for flour increased very sharply, but only more than two months after the invasion. However, in both cases, sales went far beyond their average levels, suggesting increased demand and possibly stockpiling behavior from pessimistic consumers (see Cavallo and Kryvtsov (2021)). Concerning the more recent period up to June 2022, prices for both products seemed to have stabilized at a very high level, whereas quantities have converged back to their average levels.

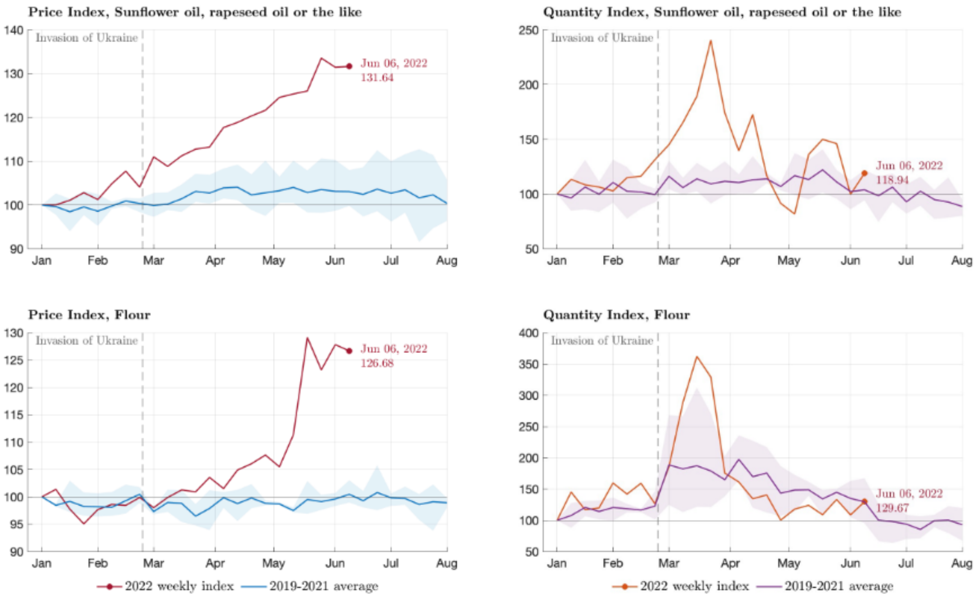

Figure 3: Source is Beck et al. (2022)

I propose that high frequency *scanner data* from super markets in Pakistan can improve inflation forecasting. In Karachi, Carrefour, Metro, Chase, Chase Up, Imtiaz supermarket and Bin Hashim are some major super markets. Similarly, Al-Fatah, Carrefour and Imtiaz supermarket are some major super market players in Lahore. However, lack of availability for super market scanner data is a constraint which must be overcome.

## 3.3. Synthetic Data

Synthetic data is artificial data (see Nikolenko (2021)) which is generated to mimic key information of the actual data and provide the ability to draw valid statistical inferences. It allows widespread access to data for analysis while overcoming privacy, confidentiality and cost of data collection concerns (also see Raghunathan (2021)).

For instance, Patki et al. (2016) has developed the SDV (Synthetic Data Vault) which uses multivariate *Gaussian Copula* (see Chapter 5.1.5 of Stachurski (2016)) to calculate covariances across input columns. The distributions and covariances are sampled from the copula to produce synthetic data. As proof of pudding, relational data sets were synthetically generated and used by freelance data scientists to develop predictive models. The researchers found no significant difference between the results produced using the synthetic versus true data Patki et al. (2016). For review of other generative models using synthetic data and advanced methods such as generative adversarial networks for economists, refer to Koenecke and Varian (2020).

Synthetic data is particularly useful for me since high frequency inflation data is not available for Pakistan. It is also essential to state that even at a private level, State Bank of Pakistan does not have high frequency data, so when I use the term *synthetic*, I mean artificially constructed high frequency data from the available low frequency data. This is in contrast with synthetic methods which are solving an information revelation problem by generating synthetic data from actual data of same frequency (see for instance Patki et al. (2016)). The accuracy of forecasting using *synthetic* data is certainly questionable and using high frequency data through scanner data and web-scrapping is part of my long term research agenda. This synthetic data exercise can motivate policy makers and State Bank of Pakistan to initiate collection of high frequency data by providing a glimpse of the utility of this data.

## 4. Stylized Facts For Pakistan's Inflation

In this section, I describe and visualize some historical, stylized facts about inflation in Pakistan. I present data on quarterly and monthly inflation rates during the period of 1958 to 2022 or subsets of this maximum range. Moreover, I describe recent trends in inflation after 2018 in more detail, breaking them down across headline inflation, food inflation, core measures (non-food and non-energy), clothing, health, transport and education sectors.

In Figure 4 below, I have plotted the quarterly inflation series for Pakistan from the first quarter of 1958 to the third quarter of 2020, based on IMF data. During 1980 to 2008, average annualized quarterly inflation rate was around 8% (representated by the horizontal dotted line). Pakistan had a severe hyperinflation crisis during the 1970's and other major inflation periods were during 2007-2009 (great recession period). Relative to the disinflation observed during 2011 to 2015, there was an inflationary period after 2015, which accelerated particularly in 2018 and 2019. Thus, the upward trend of inflation had already began before the COVID shock in 2020 and the hyperinflation crisis of 2022.

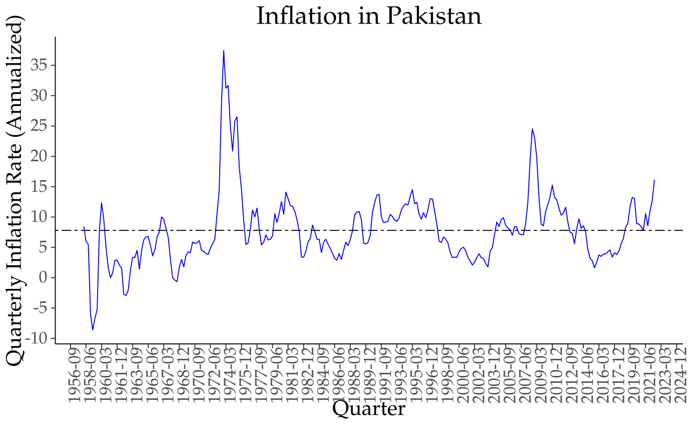

Figure 4: Data is From IMF

The latest available consumer price index (CPI Inflation with Base Year of $2015 - 16 = 100$) data from the State Bank of Pakistan recorded inflation at 8 percent on year-on-year basis in December 2020 (see Table 1 below) and 12.3% in December 2021. Moreover, core measure of inflation which excludes food and energy (Non-food, Non-Energy (NFNE)) inflation was recorded at 6.4 % in December 2020 and 8.5% in December 2021. Meanwhile, across various sectors, food (12.9 %) had among the highest inflation rates and education sector had among the lowest rates at 1.3% in December 2020. In the case of transport sector, we actually observed deflation of -3.5% in December 2020, which reflects the demand shock due to lower mobility and the COVID crisis. On the other hand, in particularly the transport sector, followed by clothing had among the highest inflation rates in December 2021, which reflects a massive transformation in the transport sector within one year. Education remained among the sectors with lowest inflation rates in December 2021.

Table 1: Annual National Inflation (December 2017 to December 2021)

| **Year on Year (%)** | **Dec 2017** | **Dec 2018** | **Dec 2019** | **Dec 2020** | **Dec 2021** |
|---|---|---|---|---|---|
| **Categories** | | | | | |
| Headline Inflation | 4.6 | 5.4 | 12.6 | 8 | 12.3 |
| Food Inflation | 3.8 | 0.6 | 17.9 | 12.9 | 10.6 |
| Core Measure (NFNE) | 5.5 | 7.64 | 7.7 | 6.4 | 8.5 |
| Clothing | 3.6 | 6.3 | 9.8 | 9.7 | 11.2 |
| Health | 10.9 | 7.1 | 11.3 | 8.1 | 9.4 |
| Transport | 4.5 | 18.4 | 14.7 | -3.5 | 24.1 |
| Education | 12.4 | 9.8 | 6 | 1.3 | 2.8 |

Note: Data is from State Bank of Pakistan. Base year is 2015-2016 for all columns, apart from Dec 2017 column for which it is 2007-2008.

In Figure 5, we have monthly data for inflation in Pakistan from the IMF. These trends are similar to the quarterly data but we can see more granular and monthly fluctuations within each year from January 1958 to sixth month (June) of 2022. The horizontal dotted line depicts mean inflation which is close to 8% for the whole sample. With this data, we can also measure the effect of COVID shock and recent hyperinflation crisis of 2022. As I am writing, the current inflation crisis is evolving and Pakistan's exchange rate with respect to USA is finally appreciating since August 2022 after a period of sharp depreciation.

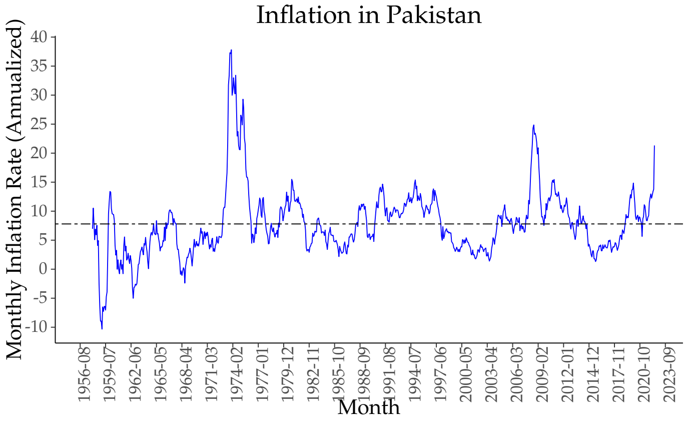

Figure 5: Data is From IMF

Lastly, Figure 6 plots only the most recent trends in monthly, year on year inflation (headline inflation rate) for Pakistan. The data is from State Bank of Pakistan and covers the months from January 2020 to June 2022. The graph indicates that during 2020, year on year inflation was actually falling despite the COVID shock. In fact, monthly and year on year inflation was close to 5% in the beginning of 2021. However, in 2022 and especially after the debt crisis of 2022, inflation rates have sky rocketed to more than 20%. In section 1 of the appendix, I present some additional data on weighted price index, annual consumer price index and annual inflation, based on linked GDP deflator for Pakistan.

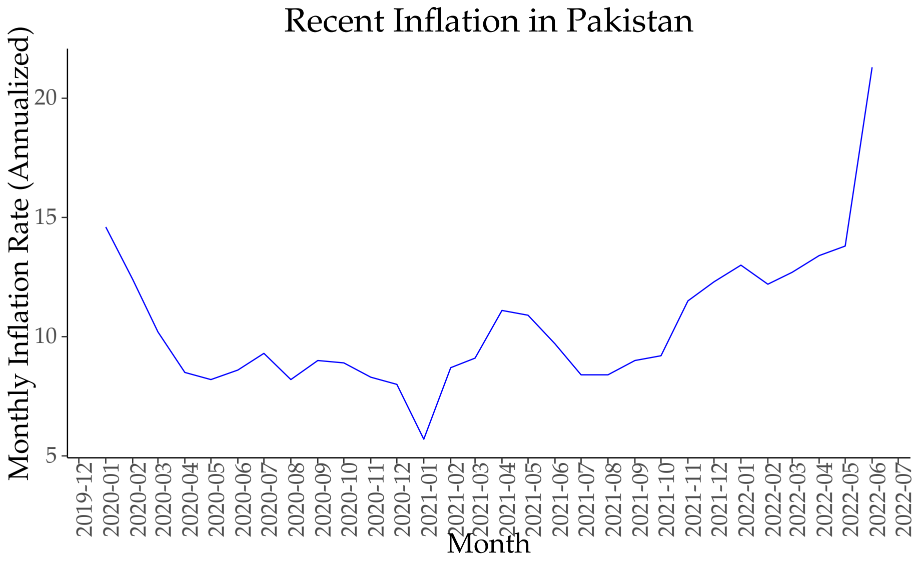

Figure 6: Data is From State Bank of Pakistan

# 5. Methodology

I will generate unknown, higher order synthetic series using quarterly and monthly inflation data for Pakistan from the IMF and forecast short-run, future inflation. I use machine learning methods such as multivariate *gaussian copula* (see Patki et al. (2016)) and *probability autoregressive models* for forecasting inflation. I also use *cubic spline interpolation* (numerical analysis) to estimate higher order inflation series such as weekly inflation series.

In order to evaluate these models, I use a reduced form VAR[4] models to forecast future inflation in Pakistan and compare the accuracy of these forecasts with ML model forecasts. Furthermore, for evaluation of the accuracy of higher order interpolation from cubic splines, I first use only quarterly series to interpolate monthly inflation and then, compare these forecasts with actual monthly inflation.

## 5.1. Synthetic Data From Copula

A copula $C$ in space $\mathbb{R}^n$ is a multivariate CDF (cumulative density function) supported by the unit hyperplane $[0,1]^N$ with the property that all of its marginals are uniformly distributed on $[0,1]$ (see Stachurski (2016)). Formally, $C$ is the function of the form below,

[4] Vector Autoregressive Model

where $0 \leq s_n \leq 1$ and $u_n \sim U[0,1], \forall n$.

$$C(s_1, s_2, s_3, ..., s_N) = \mathbb{P}\{u_1 \leq s_1, ..., u_N \leq s_N\} \tag{1}$$

While each $u_n$ has its marginal distribution pinned down, there can be infinitely many ways to specify the joint distribution. For instance, the independence copula, gumbel copulas and clayton copulas are some different types of joint distributions (see Stachurski (2016)). Figure 7 represents the general structure of a generative model which uses Guassian Copula so that $F(s_1, s_2, ..., s_N) = C(F_1(s_1), ...F_N(s_N))$ and $F_1, F_2, ..., F_N$ are univariate normal distributions.

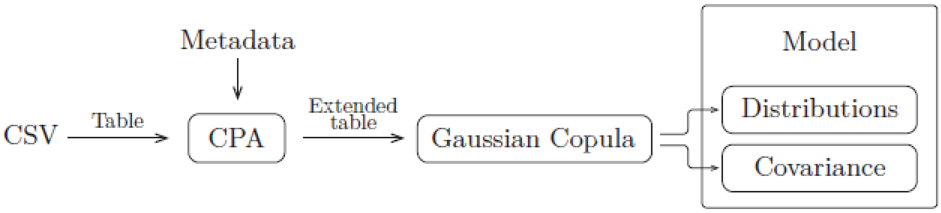

Figure 7: Source is Patki et al. (2016)

In Figure 8 below, I use Gaussian Copula to generate synthetic time series for inflation in Pakistan for 250 quarters from quarter 1 which is 1958Q2 and ending at 2020Q3 [5]. A quick comparison with previous Figure 4 above can reveal that the series roughly estimates the actual, quarterly inflation. I have generated simulations for 500 draws but the results are robust to other simulation sizes. The average inflation in simulation closely approximately the actual, average quarterly inflation of around 8%. However, there is a lot of variation across quarters due to the noise introduced by the gaussian copula.

[5] In order to apply the gaussian copula model on my data, I use the synthetic data vault (SDV) package developed by Patki et al. (2016): https://sdv.dev/

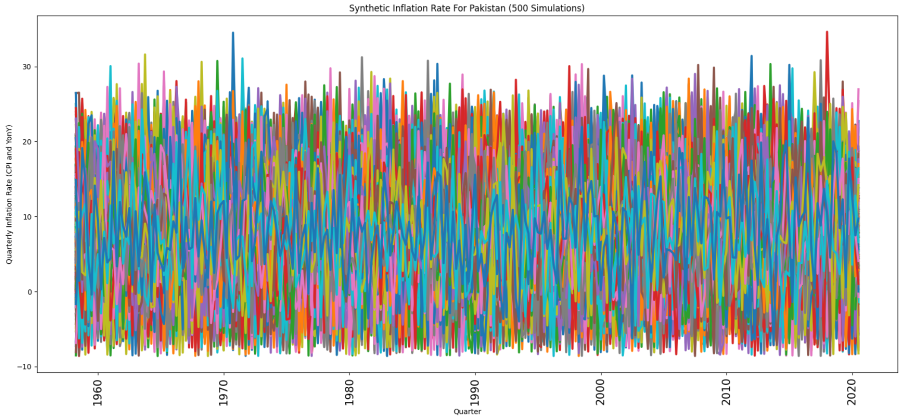

Figure 8: Author's Simulations

## 5.2. Synthetic Data From PAR

Probability Autoregessive Model (PAR) is a synthetic data creation methodology which is well suited for time series models and accounts for the auto-correlation structure of time series data. The PAR class allows learning multiple types, multivariate time series data and generating new synthetic data that has the same format and properties as the learned one. Salinas et al. (2020) have done path-breaking work at the frontier by developing probabilistic forecasting models with autoregressive, recurrent *neural networks*.

Assume that we are given access to a data set $\mathcal{D}$, consisting of $n$-dimensional data points $x$. For simplicity, let us assume that the data points are binary, i.e. $x \in (0,1)^N$. By the chain rule of probability, we can factorize the joint distribution over the $n$-dimensions as:

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(x_i | x_1, x_2, x_3, ..., x_{i-1}) \tag{2}$$

The chain rule factorization can be thought of as a Bayesian network. Such a Bayesian network that makes no conditional independence assumptions is said to obey the *autoregressive property*. We fix an ordering of the variables $(x_i|x_1, x_2, x_3, ..., x_{i-1})$ and the distribution for the $i^{th}$ random variable depends on the values of all preceding random variables in the chosen ordering: $(x_1, x_2, x_3, ..., x_{n-1})$. In an autoregressive generative model, the conditionals are specified as parameterized functions with a fixed number of parameters (see Salakhutdinov (2015). That is, we assume the conditional distributions to correspond to a Bernoulli (*Bern* in equation 3 below) random variable and learn a function that maps the preceding random variables $(x_1, x_2, x_3, ..., x_{i-1})$ to the mean of this distribution. Hence, we get the following equation, where the number of parameters of an autoregressive generative model are given by $\sum_{i=1}^{N} |\theta_i|$ (see Grover et al. (2018))[6].

$$p_{\theta_i}(x_i|x_1, x_2, x_3, ..., x_{i-1}) = Bern(f_i(x_1, x_2, x_3, .., x_{i-1})) \tag{3}$$

In order to apply the PAR model on the data, I use synthetic data vault package, developed by Patki et al. (2016) (see https://sdv.dev/). In Figure 9 below, I reveal my results from applying the PAR model which includes only the inflation time series from 1958Q2 to 2020Q3 for Pakistan. A combination of 100 simulations from the fitted PAR model indicates that this model has roughly the same mean inflation of around 8% as the output from gaussian copula. Occasionally, the PAR model draws values above 30% inflation and below $-5\%$.

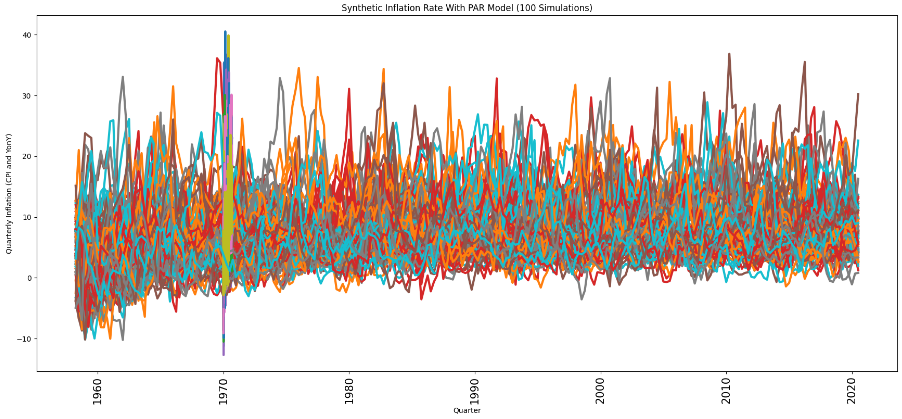

Figure 9: Author's Simulations

## 5.3. Cubic Spline Interpolation

Cubic spline interpolation is an interpolation method used in numerical analysis. It uses *cubic polynomials* to connect the existing data nodes, which allows estimation of unknown and high frequency intermediate data points. For a mathematical and formal review of

[6] For further review of PAR models, refer to https://deepgenerativemodels.github.io/notes/autoregressive/.

cubic spline interpolation, you can refer to Burden et al. (2015) and the appendix of this paper.

Consider the following data points $(x_i, y_i)$ in equation 4:

$$(x_0, y_0), (x_1, y_1), ...., (x_n, y_n) \tag{4}$$

where $x_0 < x_1 < ... < x_n$. In equation 5 below, the cubic polynomial's interpolating pairs of data are labeled as $S_0, ..., S_{n-1}$. The polynomial $S_i$ interpolates the nodes $(x_i, y_i)$ and $(x_{i+1}, y_{i+1})$. Let:

$$S_i(x) = a_i + b_i x + c_i x^2 + d_i x^3, \forall i = 0, 1, 2, ..., n-1 \tag{5}$$

Based on quarterly inflation series for Pakistan, I carry out cubic spline interpolation exercises, displayed in figures below. Despite having access to only quarterly data, visible in the dots of Figure 10, the interpolation allows me access to a higher order approximation for monthly inflation rates in this period. Similarly, I can use monthly inflation data to approximate the unknown weekly inflation series (see Figure 11). The quarterly data is for period from 1958 Q2 to 2020 Q3, whereas monthly data incorporates 774 months from January 1958 to June 2022.

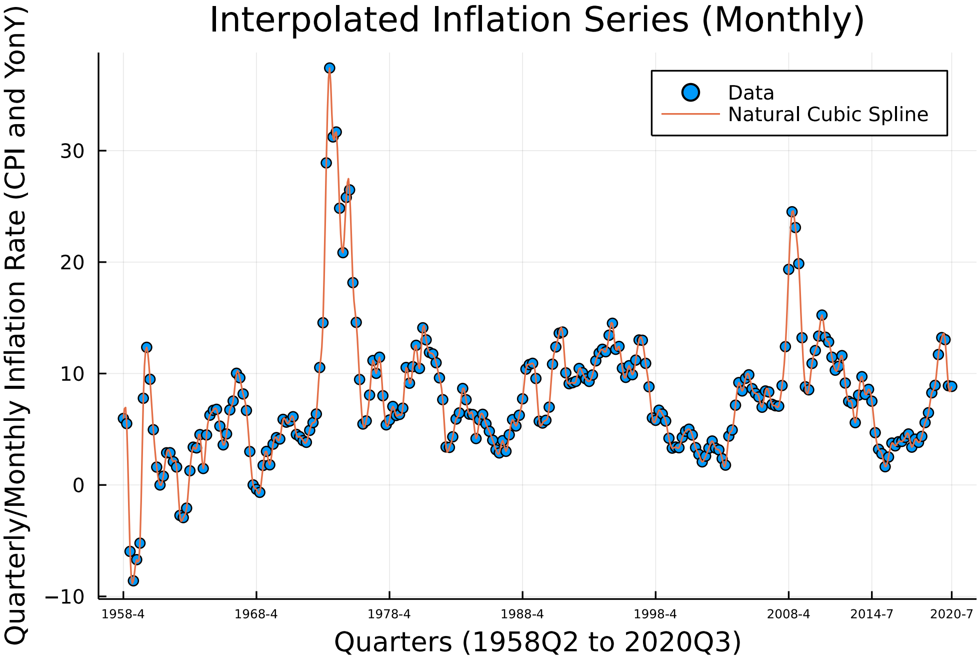

Figure 10: Using Quarterly Inflation to Interpolate Monthly Inflation

A comparison between predicted inflation from cubic spline interpolation at monthly frequency, displayed in Figure 10 and *actual monthly inflation* data indicates that the predictions are fairly accurate. In Table 2, I provide data on formal deviation measures between actual monthly data and interpolated/estimated monthly data from quarterly data, where the total number of observations is 748. For instance, absolute deviation measure displayed below computes $|a_i - b_i|$ for each pair of data points, before taking averages across these absolute differences between actual monthly and interpolated monthly inflation i.e $\frac{1}{N}\sum_{i=1}^{n}|a_i - b_i|$. Similarly, RMSQ (root mean squared error), takes the average of squared deviations between actual and interpolated data before taking a root in the end i.e $\sqrt{\frac{\sum_{i=1}^{N}(a_i-b_i)^2}{N}}$. Whereas, maximum and minimum AD's are merely maximum and minimum deviations observed. Squared $L_2$ distance is $\sum_{i=1}^{n}|a_i - b_i|^2$, $L_2$ distance is $\sqrt{\sum_{i=1}^{n}|a_i - b_i|^2}$ and finally $L_1$ distance is $\sum_{i=1}^{n}|a_i - b_i|$.

Table 2: Error Stats for Interpolated Monthly Inf Relative to Actual Monthly Inflation

| **Summary Stats** | **Mean** |
|---|---|
| **Measures** | |
| Absolute Deviation (AD) | 2.12 |
| Maximum AD | 19.91 |
| Minimum AD | 0.01 |
| Root Mean Squared Error | 3.25 |
| Squared $L_2$ Distance | 7914.3 |
| $L_2$ Distance | 88.96 |
| $L_1$ Distance | 1586.28 |

Table 2 reveals that even though the maximum deviation is above 19%, average deviation in absolute terms, as well as root mean squared error terms is fairly low and hence, the interpolated series closely approximates the actual monthly inflation. Based on this estimation accuracy, I can extrapolate that the accuracy of unknown, weekly inflation forecasts from monthly inflation series, displayed in Figure 11 is likely to be quite accurate. Hence, in the next graph (Figure 11), I demonstrate my estimation of unknown weekly inflation series of Pakistan from 1958 to 2022 using real monthly data for Pakistan (the dots in graph).

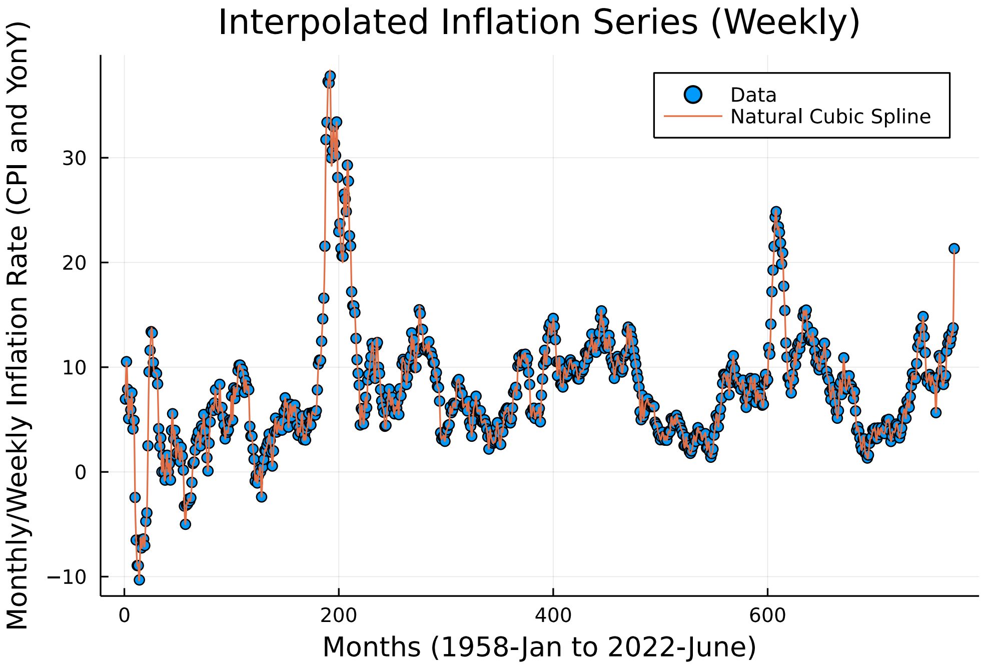

Figure 11: Using Monthly Inflation to Interpolate Weekly Inflation

## 6. Forecasts and Forecast Evaluation

The latest and most recognized work on forecasting in Pakistan is done by Ahmad et al. (2019) and Syed and Lee (2021). The former use quarterly data (1980Q4-2017Q2) on real GDP, CPI, USD/PKR exchange rate and Call Money Rate. Since actual quarterly data of GDP is not available in Pakistan, they approximate GDP data from the work of Hanif et al. (2018), where this series is available till 2012Q2. For 2012Q2-2017Q2, they interpolate annual real GDP series to generate quarterly series. For foreign variables, they use USA's GDP, CPI and 3-months T-Bill (Treasury Bill) rate. Whereas Syed and Lee (2021) use machine learning methodology and monthly inflation data from July 2007 to July 2017 to forecast the CPI inflation, GDP growth and the weighted average overnight repurchase rate in Pakistan. They use naive mean model and the autoregressive model as benchmarks and compare their forecasting performance against the dynamic factor model (DFM) and basic machine learning methods such as the ridge regressions, LASSO, elastic nets and bagging etc.

Even though my forecasting exercise is closely related to the recent work of Syed and Lee (2021), who also forecasts inflation based on ML techniques but the main contribution of my paper is to use cutting-edge methods using ML *and* synthetic data as opposed to ridge regressions (see Syed and Lee (2021)). To the best of my knowledge, this combination of methodologies has not been applied in Pakistan to forecast and estimate high frequency inflation. For evaluation of ML predictions, I use reduced form VAR models and discuss forecast error variance decomposition results in addition to fan chart outcomes and compare them with ML forecasts.

## 6.1. Fan Charts and FEVD

The following fan charts of Figure 12 are based on forecasts from a simple reduced form VAR model for Pakistan. Vector Autoregressions are standard tools of empirical macroeconomics and for a review of these methods refer to Walsh (2017). Note that my model is not a structural VAR (SVAR) but only a reduced form VAR, since I am only interested in forecasting rather than causal inference. I estimate two versions of The VAR model; the first one is the Big VAR which includes variables such as CPI (Quarterly, Year on Year Inflation), short term external debt measure, M2 (measure of money), tax revenue, imports and SR rate (short term interest rate) variables for Pakistan. As a robustness check, I also estimate a simpler VAR with the variables CPI and imports only.

The models are estimated for the period from 2006Q2 to 2020Q2 due to restrictions on availability of data. Standard information criteria such as AIC[7] are used to select lag order which are 7 quarter lags for the Big VAR and 6 quarter lags for small VAR with only imports. Having stationary variables is ideal in our VAR even though not required property. Hence, I use the Phillips Perron test, which has a null hypothesis of non-stationarity and results show that all variables in the VAR model are non-stationary, apart from imports and tax revenues. The forecasts are evaluated for 12 quarters after 2020Q2, starting from 2020 Q3 and ending at 2023 Q3. After the estimation of VAR, I also analyze standard diagnostic characteristics for model evaluation. All the diagnostic tests are reported in the appendix (section 8.2), since these are standard, run-of-the-mill tests in the VAR literature.

We know from our stylized facts that actual inflation during 2020 Q3 was around 10% and reached up to 15% by first quarter of 2022 and rose to more than 20% by the 2nd quarter of 2022. Meanwhile, both the VAR using only imports and the Big VAR using 5 variables predicts a deflation of up to negative 13% and 20% respectively for 2020 to 2022 (see Figure 12). For the last 4 quarters from 2022Q3 to 2023Q3, we get predictions of massive inflation of around 30% using Big VAR and around 5% using imports only

[7] Akaik Information Criteria.

VAR. Hence, we get grossly inaccurate forecasts using VAR methodology for 2020 to 2022 period since the actual inflation was increasing in this period and reached above 20% during 2022. On other hand, the predictions for 2022 to 2023 are yet to be examined but appear to be highly unlikely in the Big VAR case which predicts 30% inflation by 2023 Q3. Lastly, the imports' (small) VAR predicts a rise in inflation from negative 13% to almost 10% during 2022 to 2023. Even if the 2023 forecasts turn out to be accurate, the overall pattern of prediction from small VAR is completely inconsistent with actual data.

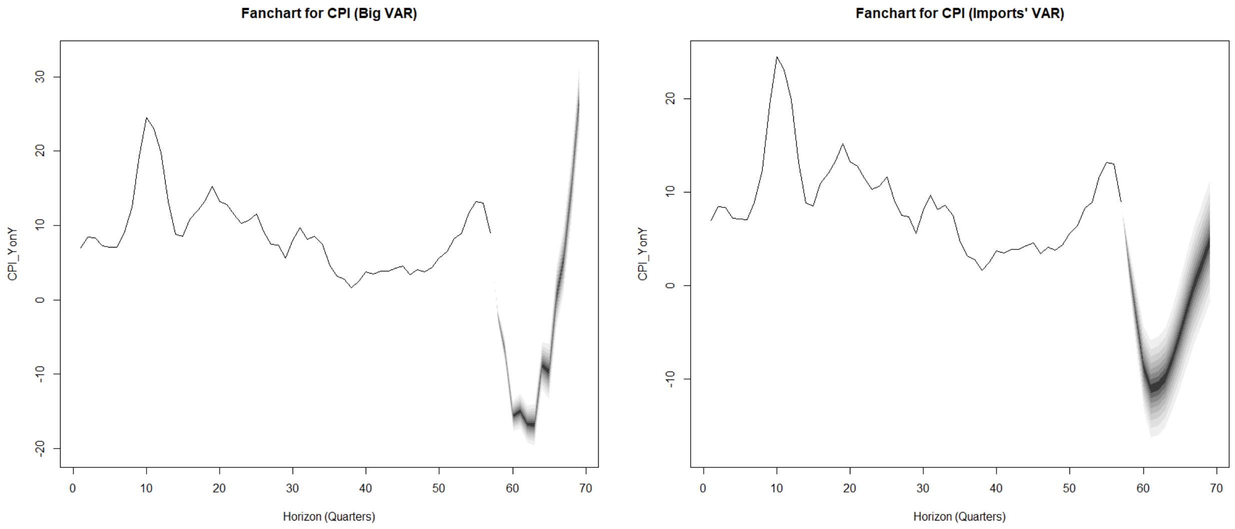

Figure 12: Fan Charts Using VAR Models (Big and Imports Only)

I also present FEVD (forecast error variance decomposition) results[8] for the Big VAR model in Figure 13. The horizon for FEVD is 12 quarter ahead and it demonstrates that non-CPI variables drive a larger share of the forecast error at longer horizons. For instance, non-CPI variables explain more than 70% of the forecast error variance at the 8th quarter ahead forecasts and almost 80% of the error variance at 12th quarter ahead. Meanwhile, for the first 3 quarters, CPI explains more than 60% of the forecast errors.

[8] For an introduction to FEVD refer to Lütkepohl (2010).

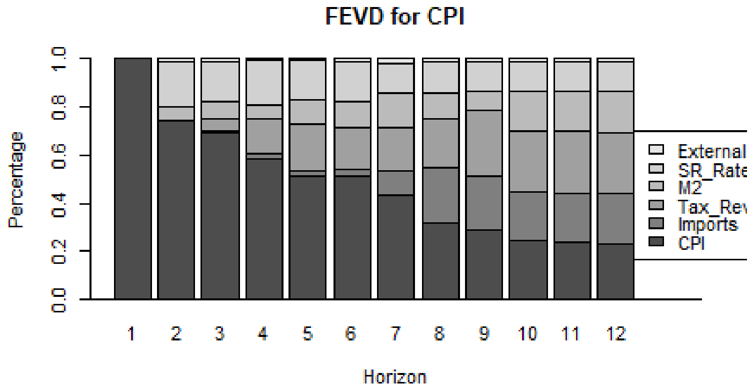

Figure 13: Forecast Error Variance Decomposition For VAR Big

## 6.2. FORECASTS FROM PAR MODELS

In order to generate forecasts based on PAR models, I use new and currently evolving estimation techniques developed by Alexandrov et al. (2019)[9]. To maintain consistency with VAR model from last section, I apply the probabilistic autoregressive model (PAR) on inflation series of Pakistan, starting from 2006Q2 and ending at 2020Q3. For monthly series, I use data from April 2006 to September 2020, which corresponds with the start and end points of the frequencies for quarterly data, which ends in quarter 3 and therefore roughly includes 9th month as well.

The top graph in Figure 14 presents 12 quarter ahead forecasts from the PAR model, starting from 2020Q4 to 2023 Q4. Similarly, the bottom graph of Figure 14 presents 36 month ahead forecasts from October 2020 to October 2023.

[9] I use Python code and packages such as *gluonts* and *pytorchts*. All code is available on my github page for replication.

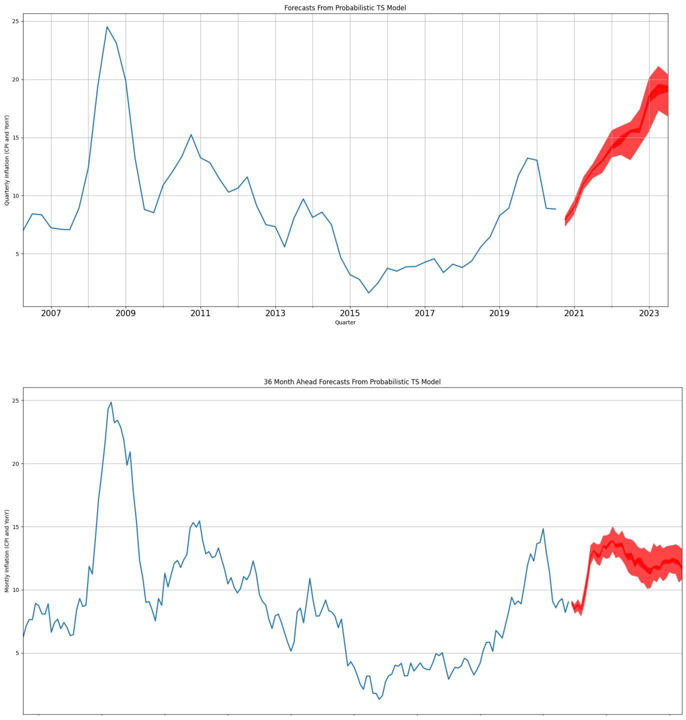

Figure 14: 12 Quarter/36 Month Ahead Forecasts From PAR Models

A quick eye-ball comparison with previous Figure 6 reveals that the PAR model is more accurate than the VAR fan charts of last section for both the quarterly and monthly specifications of PAR. However, the forecasts based on quarterly inflation series are more accurate predictors of the hyperinflation crisis of 2021 and 2022. For 2021 and 2022, the monthly PAR model can predict a rise in inflation which is close to 15% but the quarterly data predicts close to 20% inflation, which is closer to actual inflation.

Nevertheless, another interesting difference is that for 2022 to 2023, quarterly data continues to predict an year on year inflation of above 15%. However, monthly PAR model predicts a stabilization and downward adjustment of inflation, which is close to 10% by October 2023. Given the current trend toward mild downward stabilization of exchange rate and inflation in Pakistan, the monthly PAR model may be more accurate after 1 year. Meanwhile, for comparable data range, the VAR model's forecast for inflation was completely off the charts and predicted a dis-inflationary and even deflationary period during 2020Q4 to 2022Q4.

My results are robust to estimation based on entire time series period from 1958 to 2020 for both quarterly and monthly data in addition to variations in size of testing data cuts. My default cut for training data is 11 quarters in quarterly estimation and 69 months for monthly data. For robustness, I evaluate the forecasts for training data with 33 (11), 69 (23) and 105 (35) monthly (quarterly) cuts for monthly (quarterly) data respectively[10]. In all these cases, the forecast accuracy is superior to the VAR models.

# 7. CONCLUSION

In this working paper, I review cutting edge methodologies for inflation forecasting, while being motivated by the current inflation crisis of Pakistan. Making use of high frequency scanner, web-scrapped and synthetic data can make inflation forecasting and measurement more accurate, which can make policy interventions more well-informed about the ground realities However, data from scanners and web-scrapping are currently not available, which leads me to use synthetic data and numerical techniques to estimate the unknown high frequency inflation series for Pakistan in the period from 1958 to 2022.

More specifically, I mainly use *probability autoregressive models* and *cubic spline* interpolation. I find that we can approximate monthly and other low order (weekly) inflation series for Pakistan using cubic splines. I evaluate the forecast accuracy through measures such as absolute deviation, RMSQE[11] etc, applied to a comparison between cubic spline interpolation and actual monthly inflation. In addition, I use standard, reduced form, vector autoregressive models to forecast inflation and compare the forecasting potential of VAR versus my ML model forecasts, which are based on *probability autoregressive models*. Thus, both of my methods demonstrate a lot of promise for estimation of historically unknown high frequency inflation as well as short-run inflation forecasting.

It may be fruitful to expand the data availability by collecting web-scrapped and scanner data on high frequencies for Pakistan. Moreover, another extension could be to use a combination of ML based methods and VAR's (Vector Autoregressive), which would be an extension of the machine learning model used in this paper, which is autoregressive but does not include other variables. Nevertheless, it is quite informative that even a probability autoregressive model can achieve more accurate inflation forecasts relative to VAR with multiple variables.

[10] I can share these additional robustness checks if requested by any reader.

[11] Root Mean Squared Error.

# 8. Appendix

## 8.1. VAR Model Diagnostics

Standard model diagnostics for VAR's are applied, including ARCH property, residual auto-correlation, normality of residual distributions and stability tests for structural breaks.

The ARCH property is strongly rejected, since I cannot reject the null hypothesis of no degree of heteroscedasticity at even 20% significance level. Hence, there are is significant evidence for no ARCH (Autoregressive Conditional Heteroskedasticity) effects or evidence for constant volatility in the model. Another assumption is that the residuals should be non-autocorrelated. In other words, residuals should be white noise and thus uncorrelated with the previous periods. Unfortunately, the null hypothesis is rejected with a $p$ value of less than 1% and model residuals are autocorrelated.

A soft requirement is also that the normality of the distribution of the residuals. To test for the normality of the residuals, I use the standard jarque-bera test, the kurtosis test, and the skewness test. The model residuals fail to reject the null hypothesis of normality, based on all these three measures used, with typical $p$ values, far above 0.7. This is an encouraging result and the requirement is satisfied.

The stability test is a test for the presence of structural breaks. We know that if we are unable to test for structural breaks and if there happened to be one, the whole model becomes invalidated. Fortunately, we have a simple test for this which uses a plot of the sum of recursive residuals. It is reassuring that the stability test is satisfied since I was not able to detect presence of structural breaks.

In Table 3, I summarize the model diagnostics for the Big VAR model[12] There is no evidence for heteroskedasticity and structural breaks which is encouraging; on the other hand, there is evidence for normality of residual distributions, which is also positive. The only disappointing result is that there is evidence for residual autocorrelation.

[12] Results are similar for small, imports only VAR.

Table 3: Model Diagnostics for Big VAR Model

| Presence of Property | Yes/No |
|---|---|
| **Properties** | |
| Heteroskedasticity | NO |
| Residual Autocorrelation | YES |
| Normality of Residual Distributions | YES |
| Structural Breaks | NO |

## 8.2. Review of Cubic Spline Interpolation

Cubic spline interpolation is the most common spline interpolation method. It uses cubic polynomials to connect the nodes. Consider the data in equation 6:

$$(x_0, y_0), (x_1, y_1), ...., (x_n, y_n) \tag{6}$$

where $x_0 < x_1 < ... < x_n$. In equation 7 below, the cubic polynomial's interpolating pairs of data are labeled as $S_0, ..., S_{n-1}$. The polynomial $S_i$ interpolates the nodes $(x_i, y_i)$ and $(x_{i+1}, y_{i+1})$.

$$S_i(x) = a_i + b_i x + c_i x^2 + d_i x^3, \forall i = 0, 1, 2, ..., n-1 \tag{7}$$

Under the above formulation, there are $4n$ unknowns to be determined and the following four set of equations (8 to 11) must be satisfied by the interpolating function. The first set of two conditions below are merely consistency conditions with the peer of data inputs. The last two equations are referred to as smoothness or boundary conditions. We have to choose boundary conditions with two possible choices, a free or natural boundary (equation 10) or a clamped boundary (equation 11). In sum, there are $4n$ equations since the first set of equations give us $2n$ equations, the second set of conditions gives us $2n-2$, and last set (10 or 11) gives us 2 equations ($2n + 2n - 2 + 2 = 4n$). It turns out that this system of equations has one unique solution (for a proof see Burden et al. (2015)).

$$S_i(x_i) = y_i, S_i(x_{i+1}) = y_{i+1} \tag{8}$$

$$S'_{i-1}(x_i) = S'_i(x_i), S''_{i-1}(x_i) = S''_i(x_i) \tag{9}$$

Free or Natural Boundary:

$$S_0^{''}(x_0) = S_{N-1}^{''}(x_n) = 0 \tag{10}$$

Clamped Boundary:

$$S_0^{'}(x_0) = f'(x_0), S_{N-1}^{'}(x_n) = f'(x_n) \tag{11}$$

# References

**Abrams, Burton A and James L Butkiewicz**, "The Political Business Cycle: New Evidence From The Nixon Tapes," *Journal of Money, Credit and Banking*, 2012, *44* (2-3), 385–399.

**Ahmad, Shahzad, Adnan Haider et al.**, "An Evaluation of the Forecast Performance of DSGE and VAR Models: The Case of A Developing Country," *Business Review*, 2019, *14* (1), 28–52.

**Alexandrov, A., K. Benidis, M. Bohlke-Schneider, V. Flunkert, J. Gasthaus, T. Januschowski, D. C. Maddix, S. Rangapuram, D. Salinas, J. Schulz, L. Stella, A. C. Turkmen, and Y. Wang**, "GluonTS: Probabilistic Time Series Modeling in Python," *arXiv preprint arXiv:1906.05264*, 2019.

**Beck, Guenter W, Hans-Helmut Kotz, and Natalia Zabelina**, "Price Gaps at the Border: Evidence from Multi-Country Household Scanner Data," *Journal of International Economics*, 2020, *127*, 103368.

**Beck, Guenter W., Kai Carstensen, and Jan-Oliver Menz**, "Real-time Food Price Inflation in Germany in Light of the Russian Invasion of Ukraine," *VOXEU*, 2022.

**Binder, Carola Conces**, "Political Pressure on Central Banks," *Journal of Money, Credit and Banking*, 2021, *53* (4), 715–744.

**Burden, Richard L, J Douglas Faires, and Annette M Burden**, *Numerical Analysis*, Cengage Learning, 2015.

**Cavallo, Alberto and Oleksiy Kryvtsov**, "What can Stockouts tell us about Inflation? Evidence from Online Micro Data," Technical Report, National Bureau of Economic Research 2021.

_ **and Roberto Rigobon**, "The Billion Prices Project: Using Online Prices for Measurement and Research," *Journal of Economic Perspectives*, 2016, *30* (2), 151–78.

**Cukierman, Alex, Steven B Web, and Bilin Neyapti**, "Measuring the Independence of Central Banks and its Effect on Policy Outcomes," *World Bank Economic Review*, 1992, *6* (3), 353–398.

**Doerr, Sebastian, Leonardo Gambacorta, José María Serena Garralda et al.**, "Big Data and Machine Learning in Central Banking," *BIS Working Papers*, 2021, (930).

**Grover, Aditya et al.**, "Course on Deep Generative Models," *Autogregessive Models, Github*, 2018.

**Hanif, Muhammad Nadim, Khurrum S Mughal, and Javed Iqbal**, "A Thick ANN Model for Forecasting Inflation," Technical Report, State Bank of Pakistan, Research Department 2018.

**Hussain, Fida, Kalim Hyder, and Muhammad Rehman**, "Nowcasting LSM growth in Pakistan," *State Bank of Pakistan*, 2018, p. 1.

**Koenecke, Allison and Hal Varian**, "Synthetic Data Generation for Economists," *arXiv. arXiv:2011.01374*, 2020.

**Lütkepohl, Helmut**, "Variance Decomposition," in "Macroeconometrics and Time Series Analysis," Springer, 2010, pp. 369–371.

**Nikolenko, Sergey I**, *Synthetic Data For Deep Learning*, Vol. 174, Springer, 2021.

**Nordhaus, William D**, "The Political Business Cycle," *Review of Economic Studies*, 1975, *42* (2), 169–190.

**Patki, Neha, Roy Wedge, and Kalyan Veeramachaneni**, "The Synthetic Data Vault," in "2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)" IEEE 2016, pp. 399–410.

**Raghunathan, Trivellore E**, "Synthetic Data," *Annual Review of Statistics and Its Application*, 2021, *8*, 129–140.

**Salakhutdinov, Ruslan**, "Learning Deep Generative Models," *Annual Review of Statistics and Its Application*, 2015, *2*, 361–385.

**Salinas, David, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski**, "DeepAR: Probabilistic Forecasting With Autoregressive Recurrent Networks," *International Journal of Forecasting*, 2020, *36* (3), 1181–1191.

**Stachurski, John**, *A Primer in Econometric Theory*, MIT Press, 2016.

**Syed, Ateeb Akhter Shah and Kevin Haeseung Lee**, "Macroeconomic Forecasting for Pakistan in a Data-Rich Environment," *Applied Economics*, 2021, *53* (9), 1077–1091.

**Vuletin, Guillermo and Ling Zhu**, "Replacing a "Disobedient" Central Bank Governor with a "Docile" one: A Novel Measure of Central Bank Independence and its Effect on Inflation," *Journal of Money, Credit and Banking*, 2011, *43* (6), 1185–1215.

**Walsh, Carl E**, *Monetary Theory and Policy*, MIT press, 2017.