WORKING PAPER

ESTIMATION OF INFLATION PERSISTENCE IN SERBIA: RESULTS OF APPLYING QUANTILE AUTOREGRESSION ANALYSIS

Mirjana Miletić, Savo Jakovljević

Available at www.nbs.rs

Economic Research and Statistics Department

NATIONAL BANK OF SERBIA

Belgrade, 12 Kralja Petra Street

Telephone: (+381 11) 3027 100

Belgrade, 17 Nemanjina Street

Telephone: (+381 11) 333 8000

www.nbs.rs

Estimation of inflation persistence in Serbia: results of applying quantile autoregression analysis

Mirjana Miletić, Savo Jakovljević

Abstract: The aim of this paper is to analyse the persistence of inflation in Serbia and its main components, using a quantile autoregression model as the primary analytical method. The results indicate that, on average, inflation in Serbia is a stationary process, i.e. it fluctuates around its equilibrium level, but that there is asymmetry in inflation's response to positive and negative shocks, with positive shocks exerting a more durable effect on inflation. Looking at the individual inflation components, unprocessed food prices exhibit the lowest degree of persistence, followed by prices of non-food industrial products, while processed food prices and services show a higher degree of persistence.

Key words: inflation, persistence, quantile autoregression model, quantile unit root.
[JEL Code]: E31, E52, C22

Non-technical summary

The term inflation persistence, or inflation inertia, refers to the durability of inflation over time. The higher the degree of inflation persistence, the longer the effects of price shocks last and the more time inflation needs to return to its equilibrium level. A wide range of factors can contribute to greater inflation persistence and its slower return to a stable level or target. These include shocks originating from the international environment (heightened geopolitical tensions, wars, energy shocks, etc.), as well as unanchored inflation expectations, changes in the behaviour of economic agents, and other factors. The main concern associated with persistent inflation is that it reduces business certainty and may adversely affect business and investment confidence. From the perspective of monetary policy, this phenomenon implies higher costs of inflation stabilisation when inflation is elevated, i.e. a stronger monetary policy response is needed to achieve the desired effects.

Traditional studies of inflation persistence have generally assumed that inflation invariably behaves in the same way regardless of whether the economy is experiencing a crisis or not. In that case, the empirical assessment is typically based on conventional unit root tests used to examine the stationarity of inflation series, i.e. the property that inflation always returns to a stable level irrespective of the magnitude of the shock affecting it. In contrast, the quantile unit root approach makes it possible to analyse the stationarity of inflation under different scenarios (quantiles), capturing its asymmetric behaviour depending on its position in the distribution and the nature of the shocks driving its movements. Lower quantiles correspond to periods of very low inflation or deflation, when inflation is affected by negative shocks. Middle quantiles represent periods in which inflation remains stable around the target, while upper quantiles correspond to periods of elevated inflation driven by strong positive shocks. Non-stationarity typically occurs in the upper quantiles, where there is a risk that, once inflation has reached a high level, it tends to remain elevated for a prolonged period. This may result from rising inflation expectations and the emergence of an inflationary spiral.

Inflation persistence has significant implications for monetary policy. When inflation tends to remain elevated as a result of strong upward shocks, a stronger monetary policy response is needed to bring inflation back to a stable level or to the target.

We applied quantile unit root tests to headline inflation in Serbia and its key components: unprocessed and processed food, non-food industrial products, services, and core inflation (inflation excluding food, energy, alcohol and cigarette prices). Unlike conventional unit root tests based on the conditional mean, which confirmed the stationarity of these inflation indicators, i.e. their tendency to revert to a stable level, quantile unit root tests for the upper quantiles indicate that the null hypothesis of a unit root generally cannot be rejected. This is particularly the case for processed food prices, services, and core inflation.

The application of the quantile methodology revealed certain asymmetries in the mechanisms through which inflation adjusts to its equilibrium level, depending on both the magnitude and the direction of shocks. The results indicate that both the sign and the size of shocks have a significant impact on the speed at which inflation converges to its long-run equilibrium. In particular, negative inflation shocks tend to revert to the mean more rapidly, whereas strong positive shocks result in greater inflation persistence.

In the final part of the paper, we applied a methodology that allows us to examine the dynamic relationship between inflation persistence and its components (processed food prices, energy prices, and core inflation) across three dimensions: over time, across different observation horizons (short-term and long-term), and across different quantiles of the distribution. The results indicate that, for processed food prices and core inflation, the correlation with headline inflation is stronger in the upper quantiles, i.e. during periods of stronger positive inflation shocks, and that it is more pronounced over the long term than over the short term. The weakest degree of correlation was recorded during the 2017–2020 period. In the initial phase of the pandemic, namely in 2020, the correlation was even negative, whereas the strongest correlation was observed in the post-pandemic period, particularly during 2022–2024. The relationship between headline inflation and energy prices is less pronounced and exhibits a weaker association in the presence of positive shocks.

Contents

1 Introduction

Inflation persistence, or inflation inertia, commonly defined as the speed at which inflation returns to its equilibrium level (the target, stable level, or long-term average) following an economic shock, is an important concept for understanding inflation dynamics. A higher degree of inflation persistence implies that shocks have more long-lasting effects and that inflation requires a longer period to return to its equilibrium level. From the perspective of monetary policy, this phenomenon entails higher costs of inflation stabilisation when inflation is elevated, i.e. a stronger monetary policy response is needed to achieve the desired effects. This is important to bear in mind because an excessive monetary policy response may simultaneously lead to a slowdown in economic activity. As a result, monetary policy may face an unnecessary trade-off between containing higher inflation and avoiding a deceleration in economic growth.

To examine how the speed of inflation adjustment varies with the magnitude and direction of economic shocks, this paper employs a quantile autoregression analysis, which makes it possible to assess how the rate at which inflation returns to equilibrium changes depending on whether inflation is high, low, or moderate. While standard linear regression models measure average inflation persistence, assuming a uniform speed of reversion to the equilibrium level regardless of whether shocks are positive or negative, quantile regression models allow for the identification of asymmetries in the adjustment process by examining different parts (quantiles) of the inflation distribution.

Given the importance of this concept and its implications for monetary policy, as well as the fact that the NBS pursues an inflation-targeting monetary policy regime, this paper analyses the persistence of monthly inflation in Serbia and its main components using the quantile autoregression model proposed by Koenker and Xiao (2004). The core purpose of the analysis is to test whether the inflation series is characterised by a stationary process, i.e. whether it exhibits reversion to its conditional mean, or whether the assessment of a unit root depends on the particular segment of the inflation distribution being examined. An additional advantage of this approach is that quantile regression is less sensitive to violations of the assumption of normally distributed error terms, and to the presence of outliers (Koenker and Bassett, 1978).

We applied quantile unit root tests to monthly annualised rates of headline inflation and its key components: unprocessed and processed food, non-food industrial products, services, and core inflation (inflation excluding food, energy, alcohol and cigarette prices). The results indicate that, at higher quantiles, inflation tends to revert more slowly to its equilibrium level when exposed to strong upward shocks. A higher degree of persistence is observed for processed food prices, services, and core inflation. The application of the Bai and Perron (2003) test identified two structural breaks in the headline and core inflation series. The first, occurring in late 2012 and early 2013, separates the period of high and volatile inflation from the period of low and stable inflation that lasted until the outbreak of the COVID-19 pandemic. The second structural break occurred in mid-2021, when global inflation began to rise following the reopening of the world economy after the first wave of the pandemic. Once these structural breaks are taken into account, inflation exhibits stationarity across all quantiles of the distribution.

The paper is structured as follows. Section 2 provides an overview of the relevant empirical literature on measuring inflation persistence in other countries using various methodological approaches. Section 3 briefly presents the estimation methodology employed in this paper, while Section 4 presents the results of the analysis. The final section summarises the main findings of the analysis.

2 Overview of the empirical literature

The analysis of inflation persistence has been the subject of numerous empirical studies. Examples of such analyses for the USA include: Taylor (2000), Pivetta and Reis (2007), Stock and Watson (2007), Benati (2008), etc. A large body of research on inflation persistence has also been conducted for industrial economies and the euro area, including studies by: Levin and Piger (2003), Cecchetti and Debelle (2006), Altissimo, Ehrmann and Smets (2006), Meller and Nautz (2012), Lopez and Papell (2012), etc. Inflation persistence has likewise been a frequent topic of empirical research in the countries of Central and South-Eastern Europe [see, for example, Darvas and Varga (2014), Mladenović and Nojković (2012), Franta, Saxa, and Šmídková (2009)].

Several methodologies are used in the analysis of inflation persistence. Most empirical studies are based on autoregression models and standard unit root tests. A second group of approaches comprises reduced-form Phillips curve models or structural macroeconomic models that incorporate a broader set of factors estimated using the Kalman filter. More recent studies have employed quantile autoregressive analyses and fractionally integrated models, commonly referred to as ARFIMA models.

In a large number of empirical studies, both for advanced economies [Tillmann and Wolters (2014), Tsong and Lee (2011), Manzan and Zerom, (2015)] and for emerging economies [Phiri (2018), Çiçek and Akar (2013), Gaglianone, Guillén and Figueiredo (2018) …], quantile-based inflation analysis has been employed to assess inflation persistence. The findings generally suggest that, at higher levels of inflation, shocks tend to exert more persistent effects on inflation.

Tillmann and Wolters (2014) analysed inflation persistence in the United States using annualised monthly inflation rates based on the consumer price index and the personal consumption expenditures price index, as well as annualised quarterly inflation rates measured by the GDP deflator. The analysis covered the period from 1947 to 2013. The results identified a structural break in the early 1980s. Applying a quantile unit root test, the authors found that inflation persistence in the United States declined significantly after 1981, i.e. following Volcker's disinflation policies. In contrast to the preceding period, during which inflation reverted only slowly to its equilibrium level in the presence of large positive shocks – as evidenced by unit root tests for the higher quantiles – the post-1981 period is characterised by the rejection of the unit root hypothesis across all quantiles of the conditional distribution.

Following the methodology proposed by Koenker and Xiao (2004), Tsong and Lee (2011) analysed inflation dynamics in 12 OECD countries over the 1957–2010 period using quarterly data series on year-on-year inflation rates. Their findings point to the existence of asymmetry where large negative shocks tend to induce a return of inflation to its mean level, whereas this is generally not the case for large positive shocks.

Phiri (2018) employed quantile autoregression analysis to assess inflation persistence in the BRICS countries over the 1996–2016 period. The results indicate that, at higher quantiles, inflation series exhibit a unit root, implying that once inflation exceeds a certain threshold, shocks tend to have more persistent effects on inflation.

Çiçek and Akar (2013) examined the dynamics of inflation and its components in Turkey over the 1994–2012 period using quantile autoregression analysis. The results of their study indicate that inflation persistence in Turkey declined significantly after 2002, when an explicit inflation-targeting regime was adopted as the monetary policy strategy.

Gaglianone, Guillén and Figueiredo (2018) demonstrated that Brazilian inflation is generally stationary, though in 28% of the sample period (1995–2017), i.e. at higher quantiles, it exhibits non-stationary behaviour.

Using quantile autoregression, Aguilar, Garibay and Quineche (2026) showed that inflation in Chile, Colombia, and Peru does not exhibit a symmetric response; rather, its persistence and transmission of shocks vary depending on the level of inflation. Inflation persistence declined significantly as the process of macroeconomic stabilisation progressed in all three countries. Although all three countries display similar patterns of asymmetry, the intensity and form of inflation persistence differ across Chile, Colombia, and Peru, reflecting country-specific economic and institutional factors.

Ajit and Ghosh (2025) similarly examined changes in inflation persistence in India across different monetary policy regimes, including monetary aggregate targeting, a multiple-objective regime, and explicit inflation targeting. Their study shows that the adoption of inflation targeting in India significantly reduced inflation persistence before the COVID-19 pandemic. However, food price shocks that emerged during and after the pandemic continue to generate prolonged inflationary pressures that are difficult for monetary policy alone to contain.

Using the quantile autoregression method, Gbolonyo, Boyetey and Yang (2020) examined inflation dynamics in Ghana and showed that inflation in that country has been relatively well controlled in the long run, as it tends to revert to equilibrium following shocks. However, inflation dynamics are not uniform across all conditions, with inflation behaviour differing between periods of low and high inflation, which makes quantile models more suitable than standard econometric approaches for analysing inflation. The authors also emphasise the importance of monitoring food and non-food product prices, as they play a significant role in shaping headline inflation developments.

Gbolonyo, Zhang and Yang (2020) also examined the persistence of inflation in China, namely the extent to which the effects of inflation shocks remain after they occur. The results of traditional unit root tests indicate that, over the full sample period (January 1987–May 2019), inflation tends to revert to its long-run mean, suggesting that shocks are not permanent.

However, quantile unit root tests show that approximately 44.85% of observations exhibit locally non-stationary behaviour, implying that, during certain periods, the effects of shocks may persist for a very long period.

3 Methodological concept of quantile regression

This section will present a brief overview of the quantile autoregression analysis model and of the quantile unit root test proposed by Koenker and Xiao (2004).

We will first consider the autoregression model of order p, AR(p) represented in the following form:

$$y_t = \beta_1 y_{t-1} + \sum_{j=1}^{p} \beta_{j+1} \Delta y_{t-j} + \varepsilon_t \tag{1}$$

where in the specific case $y_t = \pi_t - \widehat{\mu_t}$, where π_t is the inflation rate, and $\widehat{\mu_t}$ its multiyear average or equilibrium level.

Parameter β_1 is the autoregression coefficient; if $|\beta_1| = 1$, the y_t process is non-stationary, whereas if $|\beta_1| < 1$, it is a stationary process.

The quantile representation of the previous model for quantile τ is:

$$Q_{yt}(\tau|\Omega_{t-1}) = Q_e(\tau) + \beta_1(\tau) y_{t-1} + \sum_{j=1}^{p} \beta_{j+1}(\tau) \Delta y_{t-j} \tag{2}$$

where Ω_{t-1} is the set of information available from the preceding period, $Q_e(\tau)$ denotes the τ-th quantile of the model error term, which captures the effects of unanticipated shocks of different magnitudes that lead to fluctuations in series y_t, and where $Q_{yt}(\tau|\Omega_{t-1})$ is the τ-th conditional quantile of series y_t on Ω_{t-1}.

If we make some substitutions: $\beta_0(\tau) = Q_e(\tau)$, $\beta(\tau) = \big(\beta_0(\tau), \beta_1(\tau) \dots . \beta_{p+1}(\tau)\big)$, $x_t = \big(1, y_{t-1}, \Delta y_{t-1}, \dots\ \Delta y_{t-p}\big)'$, the previous equation takes the following form:

$$Q_{yt}(\tau|\Omega_{t-1}) = x_t{}'\beta(\tau) \tag{3}$$

The estimation of the QAR model presented in equation (3) involves solving the following problem:

$$\min_{\beta} \sum_{t=1}^{n} \varphi_\tau |y_t - x_t{}'\beta(\tau)| \tag{4}$$

where $\varphi_\tau = \big(\tau - I(y_t - x_t{}'\beta(\tau) < 0)\big)$ and I is an indicator function that takes the value of one if $y_t - x_t{}'\beta(\tau) < 0$, and zero otherwise.

The estimation of the QAR model represents a variation of the LAD (least absolute deviation) method and can be expressed as follows:

$$\min_{\beta} \sum_{t:y_t \geq x_t{}'\beta(\tau)} \varphi_\tau \big(y_t - x_t{}'\beta(\tau)\big) + \min_{\beta} \sum_{t:y_t < x_t{}'\beta(\tau)} (\varphi_\tau - 1)\big(y_t - x_t{}'\beta(\tau)\big) \tag{5}$$

Koenker and Xiao (2004) proposed the use of a quantile unit root test, which enables testing the stationarity of inflation at each quantile, in contrast to standard unit root tests that assess stationarity only at the conditional mean. Applying this test for the τ-th quantile involves testing the null hypothesis that $\beta_1(\tau) = 1$ and calculating the corresponding test statistic, $t_n(\tau)$, in the following way:

$$t_n(\tau) = \frac{f\left(\widehat{F^{-1}(\tau)}\right)}{\sqrt{(1-\tau)\tau}} (\Pi'_{-1} P_x \Pi_{-1})^{\frac{1}{2}} \left(\widehat{\beta_1}(\tau) - 1\right) \tag{6}$$

where Π_{-1} is the vector of lagged values of the dependent variable y_{t-1}, P_x is the projection matrix onto the space orthogonal to $X = \left(1, \Delta y_{t-1}, \ldots \Delta y_{t-p}\right)$, f and F denote the probability density function and the cumulative distribution function of the error term from equation (1).

The consistent estimate for $f(F^{-1}(\tau))$ is:

$$f\left(\widehat{F^{-1}(\tau)}\right) = \frac{\tau_i - \tau_{i-1}}{x_t'(\widehat{\beta}(\tau_i) - \widehat{\beta}(\tau_{i-1}))} \tag{7}$$

where $\tau_i \in \Gamma = [0{,}1; 0{,}2; \ldots; 0{,}9]$.

Therefore, in order to analyse inflation persistence and differences in its dynamics depending on the occurrence of positive and negative shocks, the Koenker and Xiao (2004) unit root test can be applied over a range of quantiles. Furthermore, to examine inflation persistence over a broader range of quantiles, Koenker and Xiao (2004) and Galvao (2009) propose the use of the quantile Kolmogorov–Smirnov (QKS) test, which is obtained as follows:

$$QKS = sup|t_n(\tau)| \tag{8}$$

The test statistic $t_n(\tau)$ and QKS have non-standard distribution. According to Galvao (2009), the limiting distribution $t_n(\tau)$can be written as:

$$t_n(\tau) \xrightarrow{D} \delta \left(\int_0^1 \underline{W_1^2}(r)dr\right)^{-\frac{1}{2}} \int_0^1 \underline{W_1}\,(r)dW_1(r) + \sqrt{1-\delta^2}N(0,1) \tag{9}$$

where $\xrightarrow{D}$ denotes convergence in distribution, and where $W_1(r)$ is Brownian motion, $\underline{W_1}(r) = W_1(r) - \left[\int_0^1 (4-6s)W_1(s)ds - r\int_0^1 (6-12s)W_1(s)ds\right]$and $N(0,1)$ is standardised normal distribution of random variable independent from limiting distribution.

Let $\psi_\tau(x) = \tau - I(x < 0)$ and let $\varepsilon_{t\tau}$ denote the residuals from the regression equation (2). The weighting parameter δ is then defined as follows:

$$\delta = \frac{\sigma_{\varepsilon\psi}}{\sigma_\varepsilon\sqrt{\tau(1-\tau)}} \tag{10}$$

where σ_ε is the long-term variance of ε_t, and $\sigma_{\varepsilon\psi}$ is the long-term covariance of ε_t and $\psi_\tau(\varepsilon_{t\tau})$.

The approach proposed by Koenker and Xiao (2004) and Galvao (2009) is employed for the estimation of these nuisance parameters.

The limiting distribution for *QKS* may be obtained using the simulation strategy proposed by Galvao (2009) in the following way:

1. $\delta(\tau_i)$ is calculated for every $\tau_i \in \Gamma$;
2. for each $\delta(\tau_i)$, one realisation is independently simulated from the Dickey–Fuller and standard normal distributions, and the expression from equation (9) is calculated;

   $t_n(\tau_i) = \delta(\tau_i)\left(\int_0^1 \underline{W_1^2}(r)dr\right)^{-\frac{1}{2}} \int_0^1 \underline{W_1}(r)dW_1(r) + \sqrt{1-\delta(\tau_i)^2}N(0,1)$ and maximum absolute value for every τ_i is taken;
3. finally, by repeating the previous step a large number of times, the critical value at the 5% significance level can be calculated as the 0.95^{th} quantile of the empirical distribution.

In our specific analysis of inflation persistence, the estimate $\beta(\tau)$ represents the persistence estimate for the τ-th quantile, conditional on past inflation values y_{t-1}, …., y_{t-p}. In this way, the application of quantile analysis enables the assessment of inflation dynamics depending on the magnitude and direction of shocks. If inflation is high relative to its previous values, this indicates that a large positive shock has occurred, causing inflation to move above the conditional mean and into one of the upper conditional quantiles. Similarly, if inflation is lower than in the previous period, a negative shock has occurred, and inflation is located in one of the lower conditional quantiles.

To investigate the dynamic dependence of inflation persistence on its components across three dimensions – over time, across different observation horizons (short-term and long-term), and across different quantiles of the distribution – the Rolling Window Wavelet Quantile Correlation (RWWQC) methodology can be applied. This approach measures the correlation between two time series, X and Y, by combining wavelet transform, rolling-window method, and the quantile distribution structures. Rather than being calculated over the entire sample period, the correlation is estimated for a specific subsample (window), which is then moved forward through time while maintaining a constant window length for each calculation.

This indicator is calculated as follows:

$$RWWQC_t(s_d, w_i[X], s_d, w_i[Y]\ ) = \frac{q\delta_t(s_d, w_i[Y], s_d, w_i[X])}{\sqrt{\sigma^2\left(\phi_\tau\left(s_j, w_i[Y] - Q_{\tau, s_j, w_i[Y]}\right)\right)\sigma^2\left(s_j, w_i[X]\right)}}$$

where $q\delta_t(s_d, w_i[Y], s_d, w_i[X])$ is the covariance for the specific quantile τ of scale s_d and window w_i, $Q_{\tau,s_d,w_i[Y]}$ is the conditional quantile function for each window w_i, ϕ_τ is the quantile-specific transformation function that captures asymmetric dependence and where $\sigma^2(s_d, w_i[X])$ и $\sigma^2\left(\phi_\tau\left(s_d, w_i[Y] - Q_{\tau,s_d,w_i[Y]}\right)\right)$ are the wavelet variances of scale s_d within window w_i. Scale s_d refers to the D-th resolution level of the wavelet transform. Wavelet decomposition is performed using Maximal Overlap Discrete Wavelet Transform (MODWT) with the least asymmetric Daubechies wavelet filter of length L=8, employing four or five resolution levels, D=4, D=5 [for more detail, see Ghosh and Ajit (2025)].

Results of the analysis

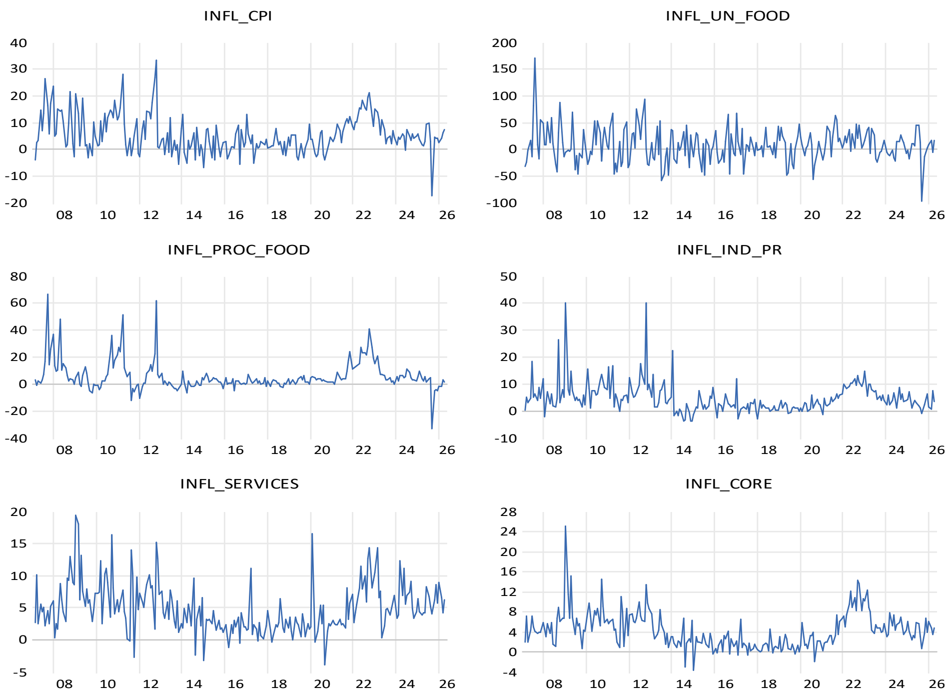
The analysis was conducted for the period from January 2007 (when the SORS began publishing consumer price index data) until and including April 2026. The analysis uses the headline consumer price index (index_cpi) and its most important price components: unprocessed food prices (index_un_food), processed food prices (index_proc_food), industrial non-food product prices (index_ind_pr), services prices (index_services), and the core inflation measure excluding food, energy, alcohol, and cigarettes (index_core). To eliminate the effects of seasonal factors that could obscure the assessment of inflation persistence, all inflation indicator series were first seasonally adjusted.

The analysis focuses on annualised monthly rates of inflation indicators. In the case of the headline consumer price index (infl_cpi), these rates were calculated using the following formula:

$$\pi_t = 1200 * \left(\frac{index\_CPI_t}{index\_CPI_{t-1}} - 1\right)$$

Price changes for the main inflation components – unprocessed food (infl_un_food), processed food (infl_proc_food), industrial non-food products (infl_ind_pr), services (infl_services), and core inflation (infl_core) – were calculated in the same manner.

Chart 1 **Monthly annualised growth rates of inflation indicators (seasonally-adjusted)**



The graphical presentation of the annualised monthly series of inflation indicators (Chart 1) shows that higher inflation rates were recorded at the beginning of the observed period (up to 2013), followed by a period of relatively low and stable inflation that lasted until 2021. The outbreak of the COVID-19 pandemic and the subsequent reopening of the global economy led to a significant increase in global demand which, together with rising world oil prices and disruptions in global supply chains, contributed to a surge in inflation worldwide. This was followed in 2022 by an energy crisis and a sharp increase in global prices of primary agricultural commodities triggered by the outbreak of the conflict in Ukraine. From mid-2023 onward, inflation gradually stabilised. In September 2025, a government regulation was adopted that capped wholesale and retail trade margins on food products and household chemicals at 20% for a period of six months, resulting in a significant monthly decline in food prices and headline inflation in that month. Excluding this effect, both headline and core inflation have exhibited relatively stable movements over the past two years, albeit at somewhat higher levels than those observed prior to the pandemic.

Based on the statistical indicators presented in Table 1, it can be observed that, over the period under review, unprocessed food prices recorded the highest average growth rates, while core inflation recorded the lowest. All the analysed series exhibit right-skewed distributions and higher-than-normal kurtosis, indicating heavier tails. The hypothesis of normality is rejected in all cases, while the results of the standard DF-GLS unit root test indicate that the inflation series is stationary, i.e. it returns to its equilibrium level. However, when an inflation time series departs from the normal distribution, i.e. exhibits a higher degree of skewness to the right or heavier tails, as is the case with the inflation indicators series analysed in this paper, conventional unit root tests may produce biased estimates.

Table 1 **Summary statistics for inflation indicators and results of the standard unit root test**

	Mean	Standard deviation	Skewness	Kurtosis	JB statistic	DF-GLS unit root test
infl_cpi	5.498	6.913	0.818	4.576	49.666***	2.122**
infl_un_food	7.075	31.594	0.636	5.708	86.144***	2.338**
infl_pr_food	5.784	10.85	2.317	12.437	1064.021***	4.152***
infl_core	4.32	3.572	1.361	7.539	269.574***	2.964***
infl_ind_pr	4.896	5.655	2.689	15.721	1836.014***	2.581***
infl_services	4.698	3.66	1.106	4.581	70.046***	2.317**

Note: *** denotes statistical significance at the 1% level; ** denotes statistical significance at the 5% level. JB test denotes the Jarque-Bera test statistic for assessing normality. The optimal lag length in the DF-GLS test was selected based on the Schwarz Information Criterion (SIC).

Table 2 shows, for each inflation indicator, the estimated values of the constants $\beta_0(\tau)$ and the autoregressive coefficient $\beta_1(\tau)$ across different quantiles, together with the results of the QKS test. The estimates of $\beta_0(\tau)$ capture systematic differences in the level of inflation across quantiles, i.e. the magnitude of the inflation shock at quantile τ. These estimates are negative up to the median (the 50th percentile), are close to zero around the median, and then increase with the quantiles, reaching their highest value at the uppermost quantile. This indicates a positive inflation shock at the highest quantile, whereas the negative values at the lowest quantile indicate a negative inflation shock. The estimates of $\beta_1(\tau)$, which measure inflation persistence, also increase across quantiles, indicating that inflation persistence is greater in the presence of larger positive inflation shocks. For the lower quantiles, the null hypothesis of a unit root is generally rejected, suggesting that, at lower inflation rates and under smaller positive and negative shocks, inflation tends to revert to its equilibrium level. This is not the case, however, when inflation is exposed to larger positive shocks.

Table 2 **Results of the quantile unit root test for different inflation indicators**

		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
infl_cpi	$\beta_0(\tau)$	-7.269***	-4.797***	-3.033***	-1,487***	-0.432	0.823*	2.407***	4.694***	7.554***
	$\beta_1(\tau)$	0.278***	0.250***	0.299***	0.499***	0.549***	0.632***	0.71*	0.773	0.692**
	QKS					6.971***				
infl_un_food	$\beta_0(\tau)$	-37.287***	-23.447***	-14.670***	-8.825***	-2.665	4.233*	10.481***	26.124***	37.311***
	$\beta_1(\tau)$	0.017***	0.085***	0.038***	0.101***	0.126***	-0.036***	0.143***	0.245***	0.257***
	QKS					12.28***				
infl_proc_food	$\beta_0(\tau)$	-5.483***	-3.755***	-2.815***	-1.441***	-0.608	0.204	1.812***	3.661***	6.446***
	$\beta_1(\tau)$	0.433***	0.453***	0.532***	0.663***	0.726**	0.772*	0.864	0.961	1.103
	QKS					10.77***				
infl_ind_pr	$\beta_0(\tau)$	-4.233***	-2.763***	-1.874***	-1.030***	-0.458*	0.259	1.188***	2.419***	4.562***
	$\beta_1(\tau)$	0.282***	0.375***	0.531***	0.592***	0.634***	0.672***	0.715***	0.76	0.791
	QKS					7.126***				
infl_services	$\beta_0(\tau)$	-3.500***	-2.509***	-1.772	-1.188***	-0.342	0.215	1.001***	2.309***	3.807***
	$\beta_1(\tau)$	0.269***	0.291***	0.442***	0.574***	0.664***	0.634***	0.666***	0.847	0.931
	QKS					8.75***				
infl_core	$\beta_0(\tau)$	-2.487***	-1.773***	-0.961***	-0.457***	0.020***	0.091***	0.644***	1.381***	3.196***
	$\beta_1(\tau)$	0.626***	0.663***	0.774***	0.752***	0.744***	0.745***	0.829	0.892	1.007
	QKS					4.831***				

Note: *** denotes statistical test significance at the 1% level; ** statistical significance at the 5% level, and * statistical significance at the 10% level. The significance of β0(τ) is tested using the t test, while the significance of β1(τ) is tested using t_n (τ).

Based on the results reported in Table 2, the following conclusions can be drawn:

1. Unprocessed food prices exhibit the lowest degree of persistence. The null hypothesis of a unit root is rejected across all quantiles, indicating that unprocessed food prices revert to their average rate of increase even in the presence of large positive and negative shocks. Even at the highest quantile, the coefficient $\beta_1(\tau)$ is below 0.3. Unprocessed food prices mostly depend on the character of the agricultural season and market supply. When the agricultural season is favourable, prices can decline easily, whereas the opposite occurs when the season is unfavourable;

2. Unlike unprocessed food prices, processed food prices exhibit a higher degree of persistence. At the 5% significance level, the null hypothesis of a unit root can no longer be rejected from the 0.6 quantile onwards;

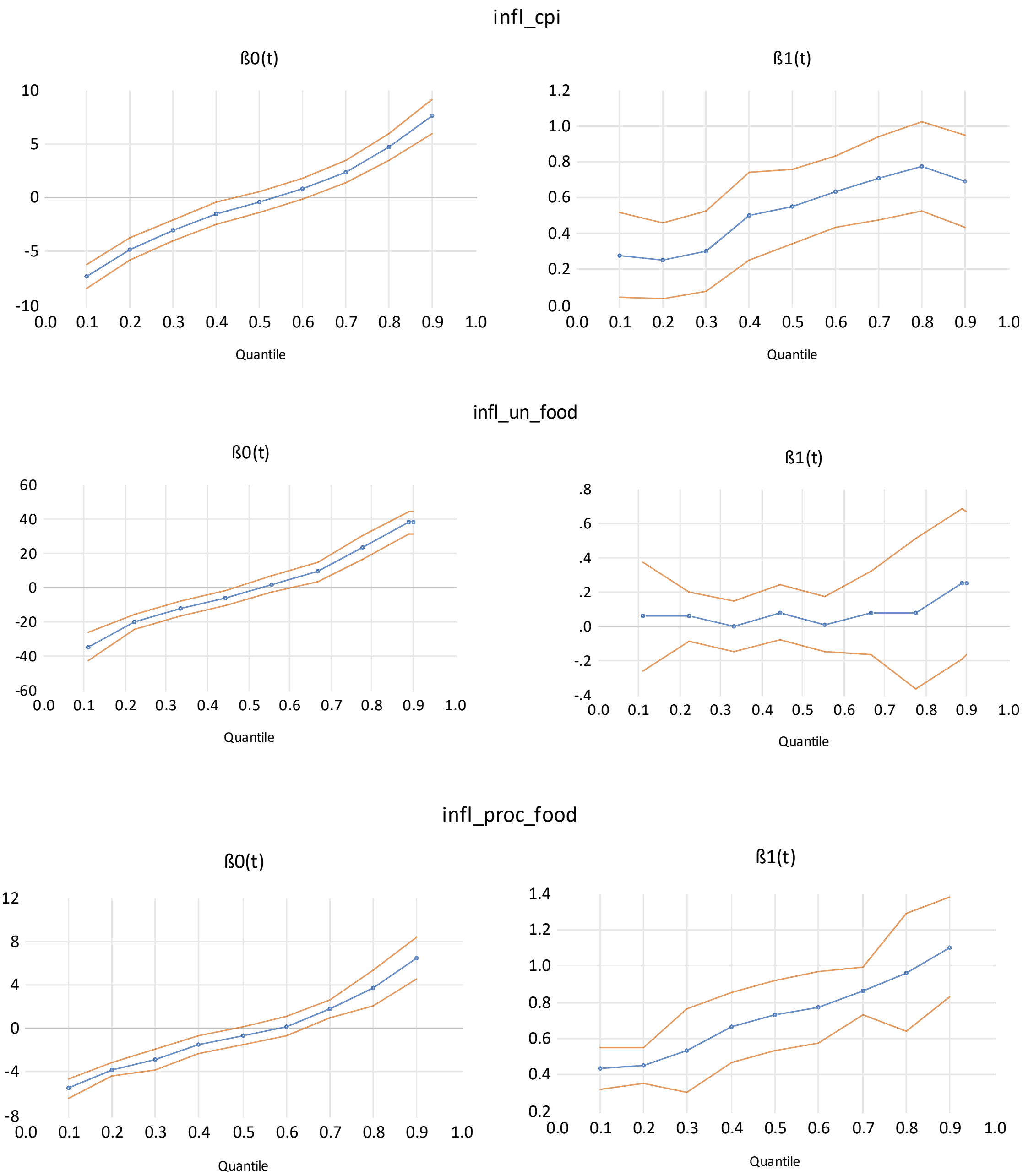
3. Core inflation exhibits a higher degree of persistence than headline inflation, with service prices displaying greater persistence than prices of non-food industrial products;

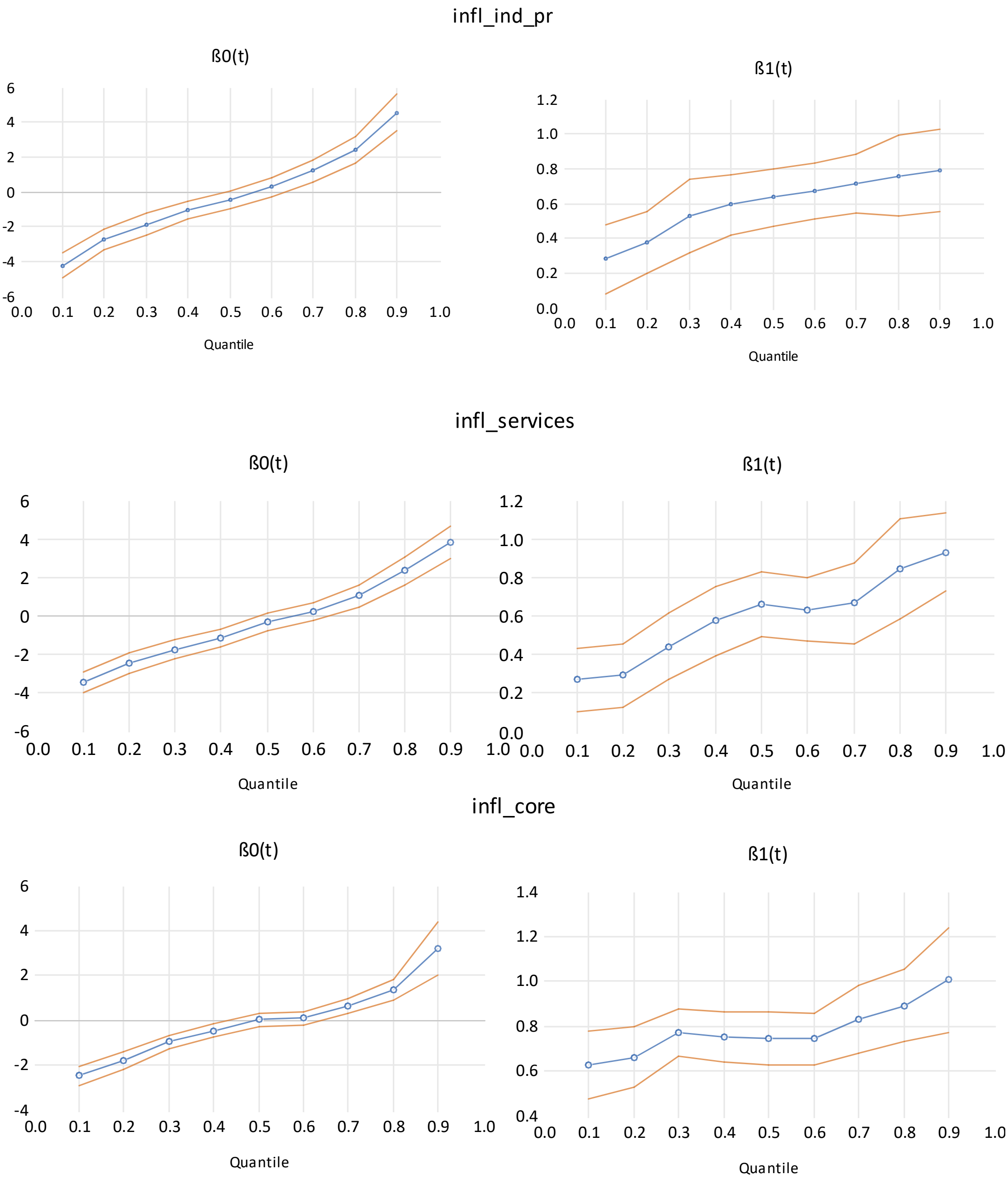
4. As regards headline inflation, as measured by the Consumer Price Index (CPI), the null hypothesis of a unit root cannot be rejected at the 5% significance level for the 0.7 and

0.8 quantiles. However, at very high inflation rates, headline inflation nevertheless reverts more rapidly to its equilibrium level;

5. The QKS test rejects the hypothesis of a unit root at the 1% significance level for all inflation indicators, indicating that all the series are stationary overall. This finding is consistent with the results of the DF-GLS test shown in Table 1.

Chart 2 **Estimated β_0 and β_1 parameters by quantile, with 95% confidence intervals**





We also calculated the speed of reversion of inflation and its components to the conditional mean for each quantile, denoted by HL, which is obtained using the following formula:

$$HL(\tau) = \frac{log(0,5)}{log(\beta_1(\tau))}$$

Table 3 **HL values by quantile for different inflation indicators**

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
infl_cpi	2.34	2.16	2.48	4.31	5.00	6.53	8.75	11.63	8.14
infl_un_food	0.74	1.22	0.92	1.31	1.45	***	1.54	2.13	2.20
infl_proc_food	3.58	3.78	4.75	7.29	9.36	11.58	20.49	75.31	***
infl_ind_pr	2.37	3.05	4.73	5.71	6.57	7.54	8.93	10.92	12.78
infl_services	2.28	2.43	3.67	5.40	7.32	6.57	7.37	18.04	41.90
infl_core	6.40	7.29	11.69	10.51	10.13	10.18	15.97	26.21	***

The results presented in Table 3 lead to the following conclusions:

1. The fastest mean reversion is observed at the 0.1 and 0.2 quantiles, approximately one month for unprocessed food prices, while it takes the longest – around six months – for core inflation;

2. Processed food prices and core inflation exhibit the slowest mean reversion. For the higher quantiles, it exceeds one year by a considerable margin, while at the highest (0.9) quantile, the reversion is undefined;

3. In the case of headline inflation, at the higher quantiles, inflation takes between 8 and 12 months to revert to its average level.

Since the presence of structural breaks may alter the dynamics of inflation, the analysis was extended by applying the Bai and Perron (2003) test, which allows for testing multiple structural breaks in a time series. The test identified two structural breaks in the headline inflation series in Serbia, occurring in November 2012 and March 2021. The first structural break marks the transition from a period of high and volatile inflation to a period of low inflation, supported by both subdued cost pressures from the international environment and the effects of relatively weak domestic demand, including the period of fiscal consolidation, underpinned by relative exchange rate stability and anchored inflation expectations. The second structural break relates to the period of rising inflation globally, driven by a sharp increase in demand following the reopening of economies after the initial phase of the COVID-19 pandemic, as well as by disruptions in global supply chains and increases in global energy and food prices, further exacerbated by the outbreak of the conflict in Ukraine. In the case of core inflation, two structural breaks were also detected – February 2013 and September 2021 – and they are, in principle, related to the same factors. However, headline inflation typically responds first, followed by core inflation.

The quantile unit root test is subsequently applied, allowing for the presence of these structural breaks, which are incorporated into the model through appropriate dummy variables. The headline inflation series is therefore adjusted as follows: $y_t = \pi_t - \widehat{\mu_t} - D_{2012M11} - D_{2021M}$, and core inflation series: $y_t = \pi_t - \widehat{\mu_t} - D_{2013M2} - D_{2021M}$.

Table 4 **Bai and Perron (2003) test for structural breaks in headline and core inflation**

infl_cpi			infl_core		
Structural break test	F-statistic	Critical value**	Structural break test	F-statistic	Critical value**
0 vs. 1*	11.44	8.580	0 vs. 1*	17.13	8.580
1vs. 2 *	15.89	10.130	1vs. 2 *	33.26	10.130
2 vs. 3	10.95	11.140	2 vs. 3	6.57	11.140
Structural break date:			Structural break date:		
1	2012M 11		1	2013M 02	
2	2021M 03		2	2021M 09	

* Significance at the 5% level, Newey-West estimator of variance was applied.

The results of the quantile unit root test indicate that, for all analysed quantiles and for both core and headline inflation, the null hypothesis of a unit root is rejected. Nevertheless, the results also confirm that larger positive shocks have a greater impact on the persistence of both headline and core inflation than negative shocks. However, the estimated autoregressive coefficients are lower than those obtained from the model that does not account for the effects of structural breaks. This clearly indicates that inflation persistence increased after the COVID-19 pandemic, while it was also higher during the initial years of the analysed period.

Table 5 **Results of the quantile unit root test for headline and core inflation adjusted for the effects of structural breaks**

		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
infl_cpi	$\beta_0(\tau)$	-6.416***	-5.096***	-3.013***	-1.551***	-0.566	0.840***	2.427***	4.509***	7.325***
	$\beta_1(\tau)$	0.071***	0.073***	0.226***	0.319***	0.401***	0.482***	0.553***	0.493***	0.65***
	QKS					11.44***				
infl_core	$\beta_0(\tau)$	-7.053***	-4.897***	-3.274***	-1.229**	-0.015	1.052**	2.871***	4.978***	6.865***
	$\beta_1(\tau)$	0.007***	0.134***	0.253***	0.376***	0.452***	0.508***	0.583***	0.552***	0.624***
	QKS	10.63***								

Note: *** denotes statistical test significance at the 1% level; ** statistical significance at the 5% level, and * statistical significance at the 10% level. The significance of $\beta0(\tau)$ is tested using the t test, while the significance of $\beta1(\tau)$ is tested using $t_n(\tau)$.

In the final step of the analysis, in order to examine the dynamic dependence of headline inflation persistence on its components, the correlation between headline inflation and its components – processed food prices, energy prices and core inflation – was estimated using the RWWQC method. The analysis employed a 48-month rolling window and four scales of the wavelet transformation, which separate short-term from long-term frequencies. Specifically, D1 corresponds to a period of one to two months, D2 to two to eight months, D3 to eight to 16 months, and D4 to a period of more than 16 months up to 32 months. The correlation was examined for $\tau = 0{,}1$, $\tau = 0{,}5$ and $\tau = 0{,}8$. The results are presented graphically using a so-called heat map.

The results show that the correlation between headline inflation and processed food prices is strongest at the higher quantiles of the distribution, i.e. when positive inflationary shocks are stronger. The correlation is also strongest at the D4 resolution level, i.e. when longer-term frequencies are considered. The lowest degree of correlation was recorded in the 2017-2020 period; notably, during the initial phase of the pandemic, i.e. in 2020, the correlation even became negative. By contrast, the strongest correlation was observed in the post-pandemic period, i.e. during 2022-2024.

Chart 3 **RWWQC results for headline inflation and changes in processed food prices**

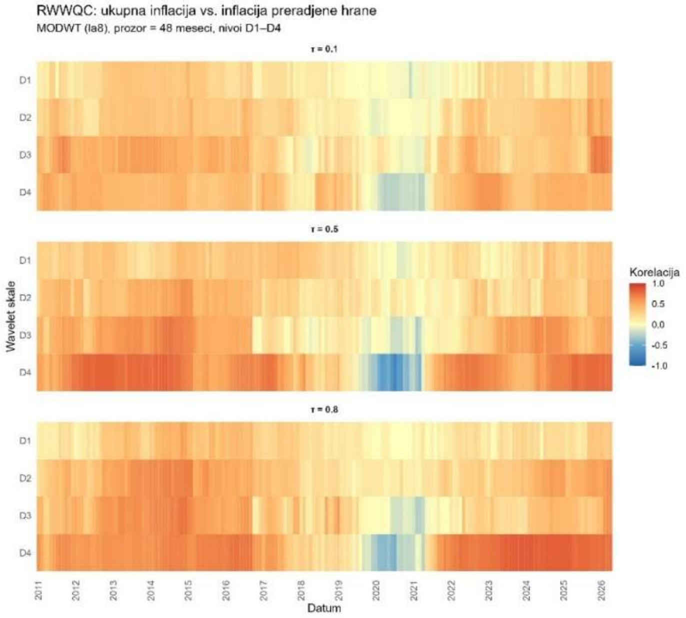


Chart 4 **RWWQC results for headline inflation and changes in energy prices**

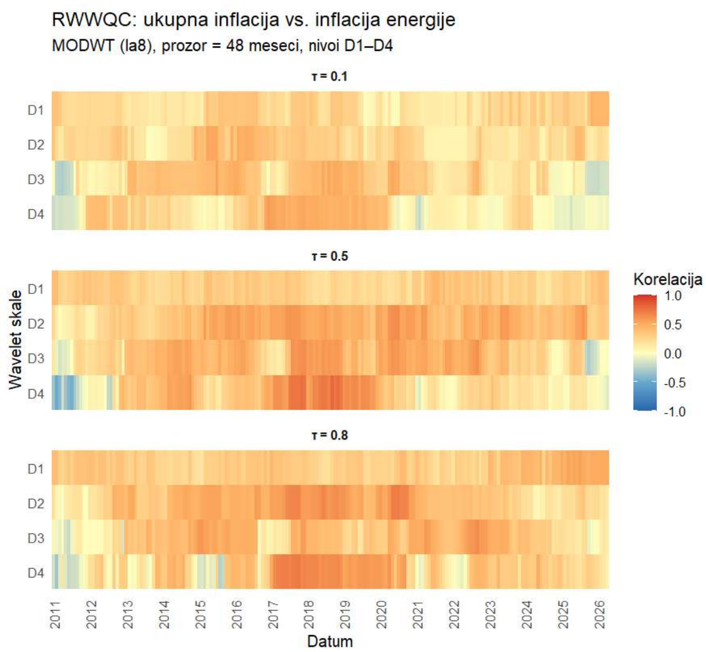
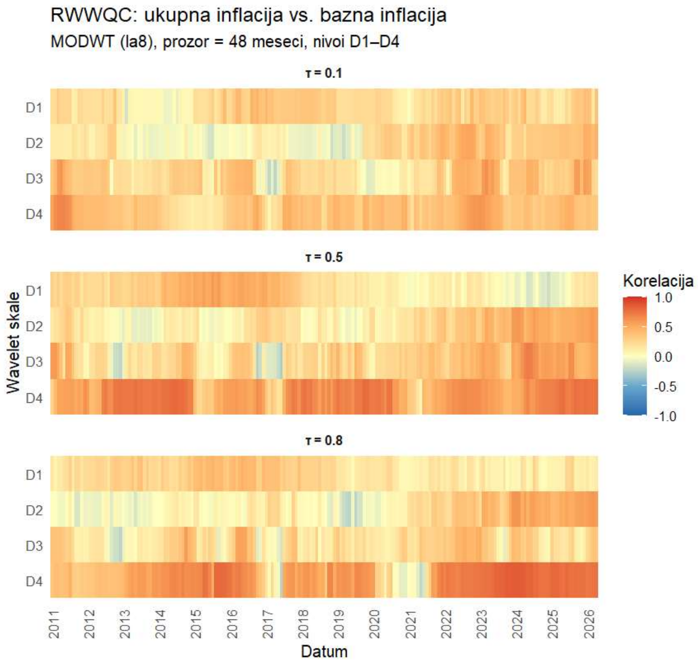


Chart 5 **RWWQC results for headline inflation and core inflation**



The correlation between headline inflation and energy prices (Chart 4) is lower than that observed with processed food prices. In this case, however, there is no pronounced asymmetry in the sense that the strongest correlation is associated with higher quantiles of the distribution, although the correlation is also stronger at longer-term than at shorter-term frequencies. For core inflation, a very similar pattern is observed as in the case of processed food prices: a higher correlation between headline and core inflation is recorded at higher quantiles of the distribution, as well as at longer-term frequencies. In addition, the highest degree of correlation is observed in the period after 2022.

4 Conclusion

This paper investigates the dynamics of inflation in Serbia over the period from January 2007 to April 2026, employing a quantile autoregression model and quantile unit root tests. This methodology identifies certain asymmetries in the mechanism of inflation adjustment towards its equilibrium level, depending on the intensity and direction of the underlying shock.

The analysis shows that both the sign and magnitude of shocks significantly affect the speed of inflation adjustment towards its long-run equilibrium level. Negative inflation shocks revert to the mean more quickly, whereas large positive shocks result in greater inflation persistence.

The analysis confirms that, in the case of Serbia over the observed period, inflation tended to revert to its equilibrium level, although the intensity of this adjustment differed across inflation categories. For example, unprocessed food prices exhibit the lowest degree of persistence, i.e. the period required for mean reversion is relatively short. By contrast, core inflation is characterised by a higher degree of persistence, particularly at the higher quantiles of the distribution. With regard to headline inflation, the null hypothesis of a unit root cannot be rejected for the 0.7 and 0.8 quantiles, whereas at the highest inflation rates inflation nevertheless reverts to its equilibrium level more quickly. This represents an important finding for monetary policy-makers in terms of determining the appropriate strength of the monetary policy response during periods of rising inflation.

The Bai and Perron test identified two structural breaks in the dynamics of headline inflation in Serbia – November 2012 and March 2021. The first structural break marked a transition from a period of high and volatile inflation to a period of low inflation, while the second break corresponds to a period of rising global inflation, driven by a sharp increase in demand following the reopening of economies after the initial phase of the COVID-19 pandemic. The results of the analysis indicate that inflation persistence increased after the COVID-19 pandemic.

The overall conclusion of the analysis of inflation persistence in Serbia using quantile autoregressive analysis is that inflation tends to revert to its equilibrium level, with the duration of this process depending on the magnitude of the inflation shock. At the same time, inflation persistence has increased in recent years, not only as a result of the crisis associated with the COVID-19 pandemic, but also due to a prolonged environment of global uncertainty, reflected in higher energy and primary commodity prices.

The importance of the findings of this analysis for monetary policy can be considered from several perspectives. The results confirm that the monetary policy response does not necessarily have to be the same in the case of positive and negative shocks. They indicate that a more forceful response is generally required when inflation is above the target, given the greater persistence of positive shocks. Therefore, central banks should also act pre-emptively in response to positive shocks, before risks materialise through higher inflation. Such a response does not necessarily have to take the form of interest rate increases, but may instead involve enhanced communication, highlighting risks and containing inflation expectations.

The particular importance of the model lies in its ability to demonstrate the persistence of an asymmetric pattern in inflation persistence, suggesting that such dynamics may represent an inherent feature of inflation. This finding is consistent with evidence from other studies conducted for various countries and regions. Consequently, the results add to the arguments in favour of considering an asymmetric monetary policy response depending on the nature and direction of the inflation shock.

The results of the correlation analysis based on one of the most advanced methods – the rolling-window wavelet quantile correlation (RWWQC) method – applied to monthly annualised headline inflation and monthly annualised changes in processed food prices and core inflation indicate that the correlation is stronger at higher quantiles, i.e. in the presence of larger positive inflation shocks, and that it is also stronger when longer-term rather than short-term changes are considered. The lowest degree of correlation was recorded in the 2017–2020 period; notably, during the initial phase of the pandemic, i.e. in 2020, the correlation even became negative, whereas the strongest correlation was observed in the post-pandemic period, i.e. during 2022–2024. The relationship between headline inflation and energy prices is less pronounced and does not display such a strong association with positive shocks.

Literature

Altissimo, F., M. Ehrmann and F. Smets (2006)., "Inflation persistence and price-setting behaviour in the euro area: a summary of the IPN evidence", Occasional paper series, no 46, June

Bai, J. and P. Perron (2003)., "Computation and analysis of multiple structural change models." Journal of applied econometrics, 18, 1–22

Benati, L. (2008)., "Investigating inflation persistence across monetary regimes", Quarterly Journal of Economics 123, 1005-1060.

Cecchetti, S. G. and G. Debelle (2006)., "Has the inflation process changed?", Economic Policy April 2006, 311-35

Cicek, S. and C. Akar (2013)., "The asymmetry of inflation adjustment in Turkey", Economic Modelling, 31, 104–118

Darvas, Z., and B. Varga, (2014)., "Inflation persistence in central and eastern European countries", Applied Economics, 46(13), 1437–1448.

Elliott, G., T. J. Rothemberg, and J. H. Stock (1996), "Efficient Tests for an Autoregressive Unit Root." Econometrica, 64, 813–836.

Franta, M., B. Saxa, and K. Šmidkova, (2009)., "The Role of Inflation Persistence in the New EU Member States. " In What Drives Inflation in the New EU Member States, 11–24. Brussels: European Commission

Fuhrer, J. C. (2011): "Inflation persistence", in B. Friedman and M. Woodford (eds.), Handbook of Monetary Economics, Volume 3A, Elsevier, North-Holland

Gaglianone, W. P., O. T. de Carvalho Guillén, and F. M. R. Figueiredo, (2018)., "Estimating inflation persistence by quantile autoregression with quantile-specific unit roots", Economic Modelling, Volume 73, June 2018, pages 407-430

Galvao, A. F. (2009)., "Unit root quantile autoregression testing using covariates", Journal of Econometrics 152, 165-178

Ghosh, T. and Y. Ajit (2025)., "A Historical Perspective on India's Inflation Persistence: A Quantile Analysis", https//doi/10.1515/snde-2024-0078/html

Koenker, R. and G. W. Bassett (1978)., "Regression quantiles", Econometrica 46, 33-50. [35]

Koenker, R. and Z. Xiao (2004)., "Unit root quantile autoregression inference", Journal of the American Statistical Association 99, 775-787.

Koenker, R. and Z. Xiao (2006)., "Quantile autoregression", Journal of the American Statistical Association 101, 980-1006

Levin, A. T. and J. Piger, "Is Inflation Persistence Intrinsic In Industrial Economics?" Federal Reserve Bank of St. Louis working paper no. 2002-03, 2002.

Lopez C, Papell DH (2012), "Convergence of euro area inflation rates". J Int Money Finance 31(6):1440–1458

Meller, B. and D. Nautz (2012), "Inflation persistence in the Euro area before and after the European Monetary Union", Economic Modelling, Volume 29, Issue 4, July 2012, Pages 1170-1176

Mladenović, Z. A. Nojković (2012) "Inflation Persistence in Central and Southeastern Europe: Evidence from Univariate and Structural Time Series Approaches", Panoeconomicus, 2012, 2, Special Issue, pp. 235-266 Received: 28 July 2011; Accepted: 9 March 2012.

Quineche, R., J. Aguilar and R. Garibay, (2026)., “Asymmetric Inflation Persistence in Latin America: Evidence from Chile, Colombia, and Peru Using Quantile Autoregression Analysis”, Latin American Economic Review, Vol 37

Phiiri, A., (2018)., “Inflation persistence in BRICS countries: A quantile autoregressive (QAR) approach”, Business and Economic Horizons, Prague Development Center (PRADEC), vol. 14(01), January

Pivetta, F. and R. Reis (2007).,“The persistence of inflation in the United States”, Journal of Economic Dynamics and Control 31, 1326-1358.

Taylor, J. B. (2000)., “Low inflation, pass-through, and the pricing power of firms”, European Economic Review 44, 1389-1408.

Tillmann P. and M. H. Wolters (2014)., “The changing dynamics of US inflation persistence: a quantile regression approach”, Kiel Working Paper No. 1951 | August 2014

Tsong, C-C and C. Lee (2011)., “Asymmetric inflation dynamics: evidence from quantile regression analysis.” Journal of Macroeconomics, 33, 668–680

Stock, J.H. and M. Watson, (2007), “Has inflation become harder to forecast?”, Journal of Money, Credit, and Banking, 39, 3-34.