Copenhagen Business School

MSc. Business Administration and Data Science

KAN-CDSCO1005U Predictive Analytics

Forecasting Monthly Danish Pharmaceutical Exports

Statistical forecasting of Danish pharmaceutical exports using SARIMA and ETS

Characters: 30.547

Number of pages: 15

Group number: Fri-174161-2

Mikolaj Sapek: S185880

Julia Nowak : S160787

Yasemin Pagano: S185872

Peter Emil Larse Have: S160614

Supervisor: Dimitar Yordanov

Submission date: Spring 2026

Table of Contents

Abstract

Danish pharmaceutical exports roughly doubled between 2022 and 2026, driven by GLP-1 demand and Novo Nordisk's output. We forecast the next 24 months from monthly DST data (2007–2026) using a seasonal ARIMA model, with exponential smoothing (ETS) as a benchmark. We recommend ARIMA(0,1,2)(1,0,1)[12] with drift. It passes the residual diagnostics at lags 12 and 24 and forecasts about as well as the ETS benchmark out of sample. On the single 24-month holdout ETS is in fact marginally more accurate on RMSE, MAE and MASE, but SARIMA wins the rolling cross-validation across 19 origins and has the wider diagnostic safety margin, which is why we prefer it. Secondly, on a strict 24-month holdout (April 2024–March 2026), SARIMA produces RMSE ≈ 1.73 Bn DKK, MAPE ≈ 8.9%, MASE ≈ 1.28. Over the 24-month horizon the central forecast points to monthly exports near 17Bn DKK by early 2026, with an 80% prediction interval of roughly 15.0–19.5 Bn DKK and a 95% interval of 13.5–21.0 Bn DKK. We tested further with SARIMAX variants with COVID and GLP-1 dummies. They fit the past slightly better but produced substantially worse holdout accuracy, and were therefore excluded from the final model.

1 Introduction

Denmark is considered to be the leader in Biotechnology in Europe. It dominates in industries like the insulin market, and the booming weight loss drug market. This has made the small country a powerhouse in Pharma in Europe, and Copenhagen in particular is now considered the "Silicon Valley of BioTech" (Bregenholt, 2015). GLP-1 demand has recently spiked across the globe, and monthly exports have surged to an all-time high. Denmark is the 8th largest exporter of pharmaceutical products in the EU, at €13.7 billion in 2024 (Charoux, 2025). Pharmaceuticals are Denmark's largest export area, accounting for roughly 30% of total goods exports (Pharma Boardroom, 2025). This makes Danish pharmaceutical exports one of the most dynamic trade series in Europe right now.

The time series is statistically interesting; the series exhibits structural acceleration post-2022, non-constant variance, and evidence of a regime shift, properties that make extrapolation both important and non-trivial. It creates the question of how much recent acceleration a model can capture central to any forecast. We ask whether the standard fpp3 toolkit can produce a forecast trustworthy for decisions over the next 24 months, estimating seven specifications and selecting the final model through residual diagnostics, holdout accuracy, and rolling cross-validation. The dataset comprises monthly Danish pharmaceutical exports (SITC 54) from the DST Statbank, covering 2007 to March 2026 (Danmarks Statistik, 2026). The series is log-transformed for variance stabilisation, tested for structural breaks and unit roots, and four ARIMA specifications and three ETS specifications are estimated and gated on residual diagnostics before holdout evaluation, with the best from each class compared in a final head-to-head. The paper will compare these through 2 models (ARIMA and ETS) to determine which model is best. This paper finds that ARIMA(0,1,2)(1,0,1)[12] with drift forecasts about as accurately as ETS(A,A,A) out of sample and produces more reliable prediction intervals over the 24-month horizon, which is why we prefer it. We do not rank the two on AIC, because ARIMA and ETS information criteria are computed on different bases and are not directly comparable across model classes. ARIMA and ETS are well-established frameworks for univariate forecasting, widely applied across macroeconomics and trade series (Hyndman & Athanasopoulos, 2021). One caveat shapes how these results should be read. Danish pharmaceutical exports are concentrated in a single firm, Novo Nordisk, so this aggregate series behaves in large part like one company's GLP-1 output. The lag-12 seasonality we model may therefore reflect production and shipment cycles as much as a genuine macroeconomic export pattern, and the forecast is exposed to single-firm events, such as capacity changes or competitor entry, that a univariate model cannot see. We return to this in the limitations.

Therefore, our research question is: *Can a univariate time series model, estimated using the fpp3 framework, produce a diagnostically valid and accurate 24-month forecast of Danish pharmaceutical exports?*

2 Data description

2.1 Source and definition

The dataset came from the official Danish statistics database, Danmarks Statistik (Danmarks Statistik, 2026). The data gives the monthly value of Danish Pharma exports. The period that was obtained was from January 2007 to March 2026 giving us 231 observations. The variable used is the nominal value of Danish pharmaceutical exports (SITC 54, Medicinal & pharmaceutical products). The SITC 54 captures multiple indicators, such as Insulin, Weight loss, and Dermatology treatment, etc….

Export rises from 2.6-4.0 Bn DKK/month in 2007 to 15-19 Bn DKK/month in 2025-2026 (Figure 1). The data dataset also highlights a mean pre 2022 of 6.69 and post having a mean of 14.10 highlight a strong upward trend (Table 1). The seasonal plot (Figure 2a) confirms a consistent recurring annual pattern across the full sample: exports are slightly elevated in January-February and softer around July-August, with a modest autumn recovery. This shape remains stable over the years, even as the level rises sharply. The subseries plot (Figure 2b) shows the post-2022 acceleration is present in every calendar month, not concentrated in a single season, indicating a broad-based level shift rather than a change in seasonal structure.

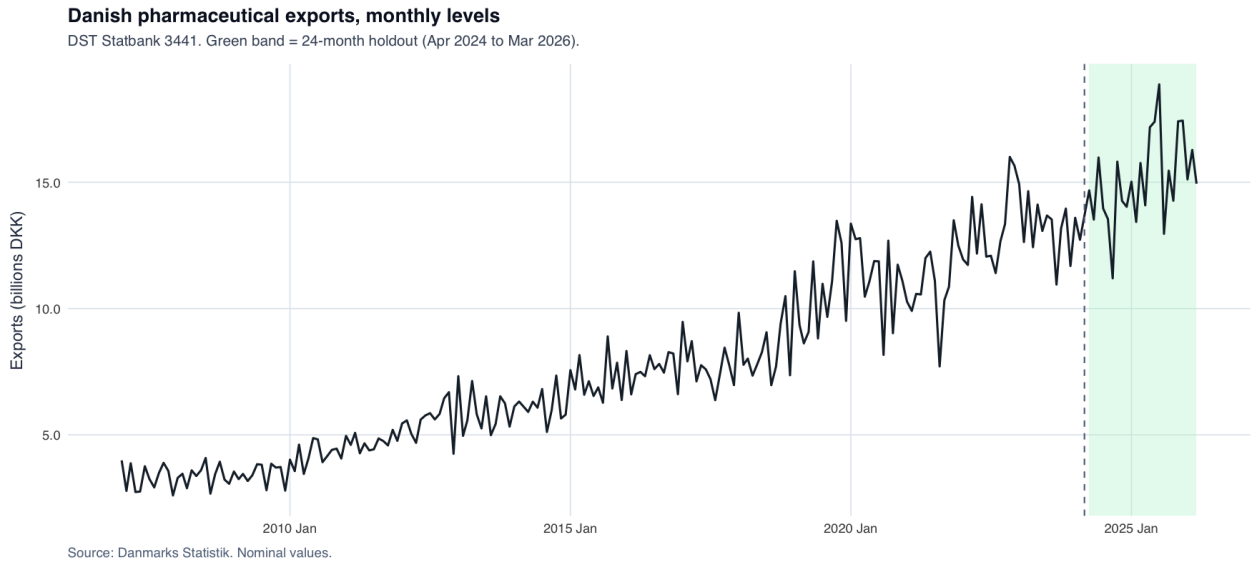


***Figure 1**: Time Series of Monthly export levels (billions DKK). Green shading = 24-month holdout.*

Period	Mean	Min	Max	Std. dev.
Full sample (2007–2026)	**8.33**	**2.60**	**18.87**	**4.03**
Pre-2022 (2007–2021)	**6.69**	**2.60**	**10.89**	**1.82**
Post-2022 (2022–2026)	**14.10**	**8.91**	**18.87**	**2.41**

***Table 1:** Descriptive statistics (levels, DKK billions)*

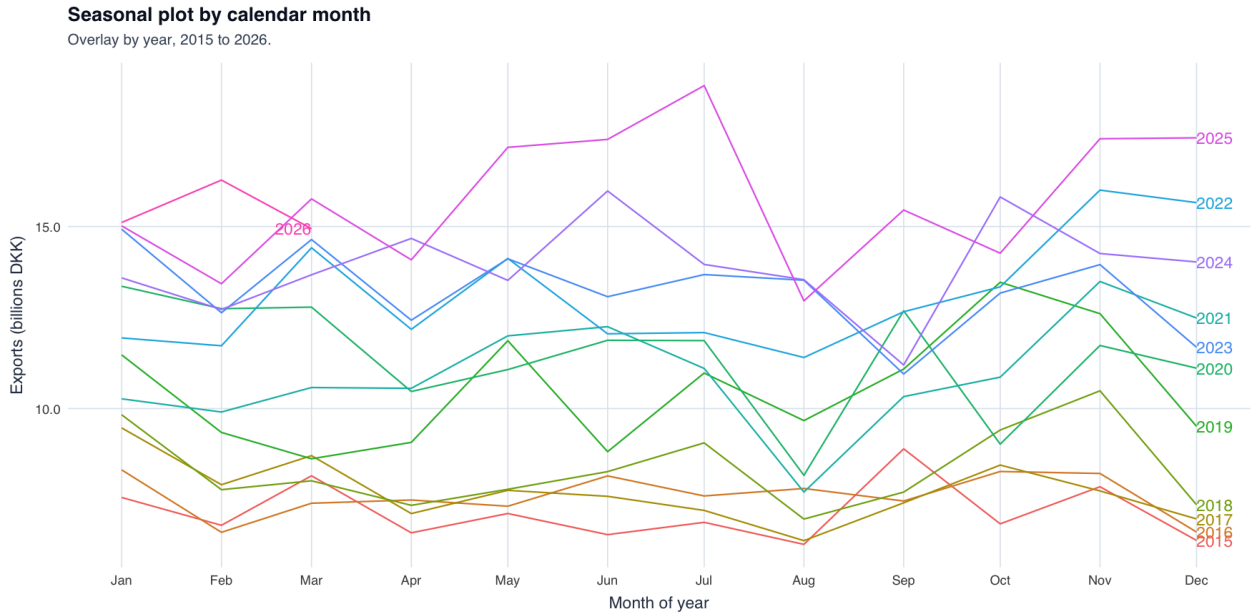


Figure 2a: *Seasonal plot by calendar month*

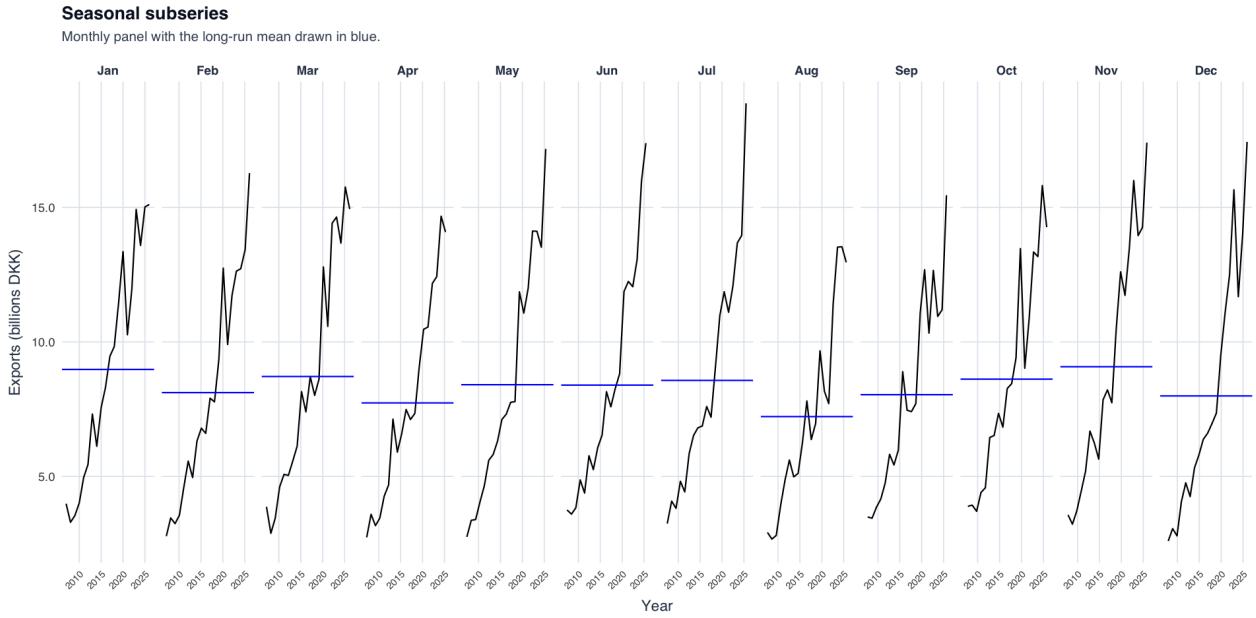


Figure 2b: *Seasonal subseries by calendar month*

2.2 Transformation and stationarity

The Guerrero method (Guerrero, 1993)returns $\lambda = -0.0177$, which is approximately zero (Figure 3). When $\lambda \approx 0$, the Box-Cox transformation reduces to a natural log, so the series is log-transformed throughout the analysis. This compresses the growing seasonal swings visible in the raw series, producing approximately constant variance across the sample.

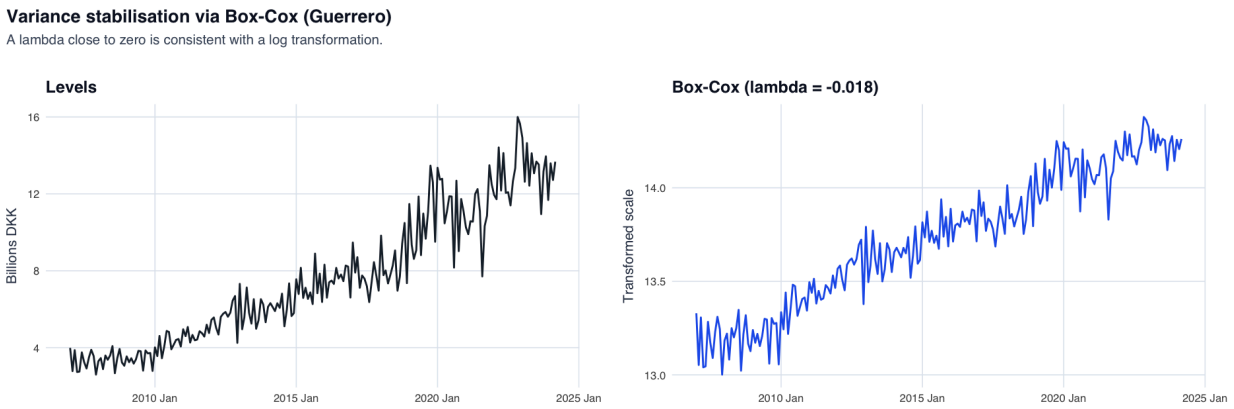


Figure 3: *Box-Cox lambda (guerrero)*

To determine the integration order, we tested for unit roots with the Augmented Dickey-Fuller or ADF (Dickey & Fuller, 1979), Phillips-Perron (Phillips & Perron, 1988), and KPSS (Kwiatkowski et al., 1992) applied on both the level and differenced series. KPSS returns 0.01 on both raw and log rejecting the null of stationarity and indicating a trending mean (Table 2). ADF fails to reject the null of unit roots at the 5% level (p=0.21 and p=0.31), consistent with the KPSS results. After one round of differencing, it yielded p=0.01, confirming that the differenced series is stationary. KPSS also indicates ndiffs = 1 and nsdiffs = 0, confirming that one difference is sufficient to achieve stationarity, with no seasonal differencing required. The seasonal pattern is stable in amplitude, as confirmed by the STL decomposition in Figure 4, with the seasonal component remaining consistent across all years, further justifying an additive seasonal specification.

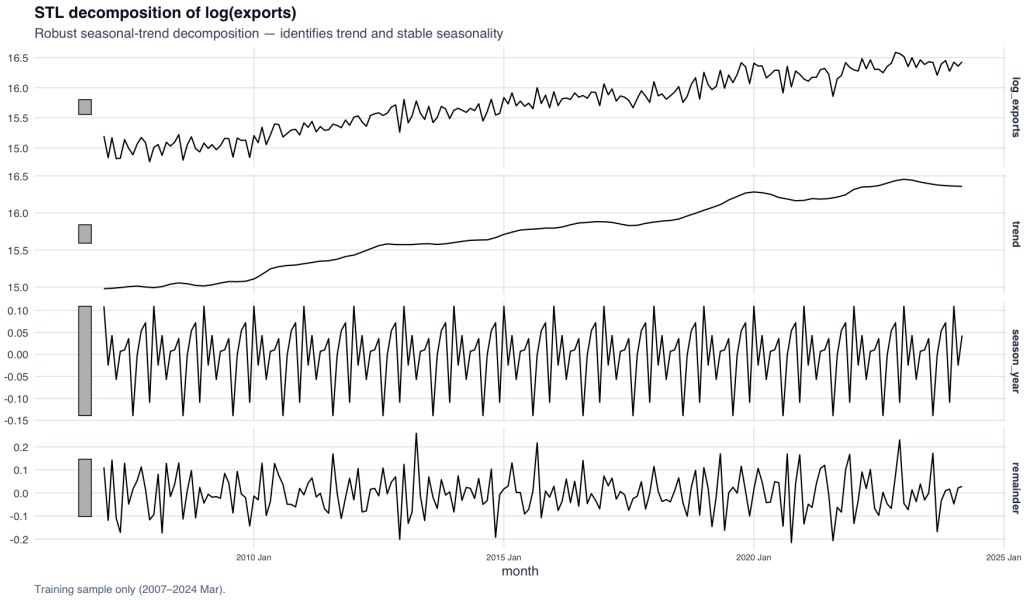


Figure 4: *STL decomposition of log(exports) on training data trend and stable seasonality.*

Test	Series	p-value	Interpretation
KPSS	levels (Box-Cox)	0.01	Non-stationary → $d = 1$
KPSS	log(exports)	0.01	Confirms $d = 1$
ADF	levels	0.21	Unit root not rejected
ADF	log(exports)	0.31	Unit root not rejected

ADF	diff(log)	0.01	Stationary after differencing
PP	levels / log	0.01	Mixed evidence; differencing justified

Table 2: *Unit-root tests on the transformed series.*

Figure 5 shows the ACF and PACF of the first-differenced log series. The slowly decaying pattern present in the undifferenced series has disappeared, confirming that one round of differencing is sufficient to remove the trend. A noticeable spike at lag 12 in the ACF confirms that a recurring 12-month seasonal pattern remains present after trend removal. The PACF shows negative spikes at lags 1 and 2, with gradual decay at subsequent lags, reflecting short-run autocorrelation consistent with a moving-average structure. The seasonal pattern at lag 12 is confirmed by the ACF spike rather than the PACF.

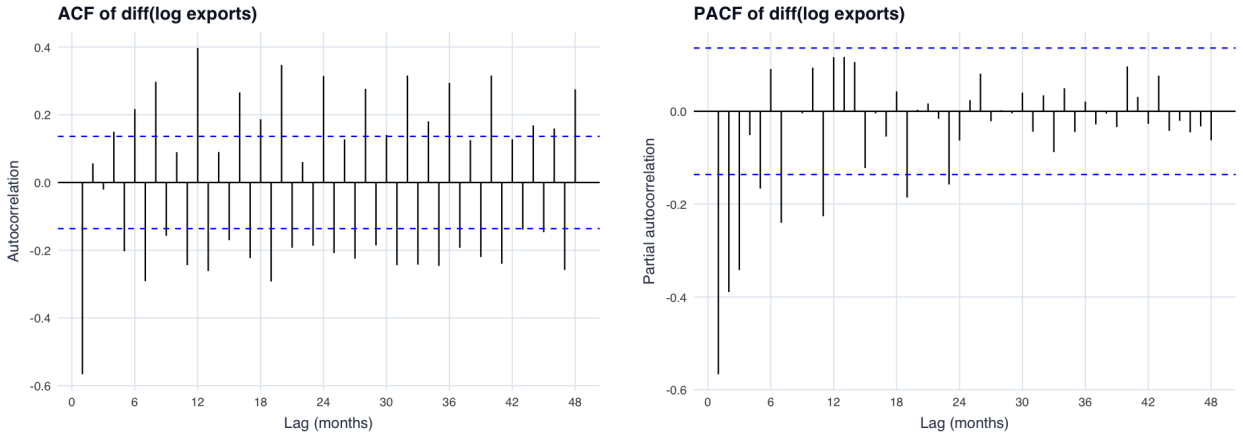


Figure 5: *ACF/PACF of diff(log exports)*

2.3 Structural breaks and outliers

We test for a structural break in diff(log(exports)) using the Quandt Likelihood Ratio (QLR) / sup-F test (Quandt, 1960; Andrews, 1993) from R package strucchange (Zeileis et al., 2002). The test does not reject the no-break null (p = 1.00) within the 15%/85% trimming window. This is what we should expect rather than evidence against a regime change. The test runs on the differenced series, so it looks for a break in the growth rate, while the obvious shift in this series is in the level, and first-differencing removes that level shift. The acceleration after 2022 is therefore better described as a smooth, sustained change than as a single discrete break.

Indicator saturation (Pretis, Reade & Sucarrat, 2018) via gets::isat test every observation as a potential outlier. Interestingly, December 2012 and January 2013 at strict thresholds (1%), also scattered points around the 2008 financial crisis, COVID-19 (2020), and 2021, appear only at the lenient 5% level and are not considered robust outliers. These are documented as data points rather than modelled further in the final specifications. This resilience is consistent with the nature of pharmaceutical products classified as "necessary goods". Demand for insulin and essential medicines is largely inelastic with respect to macroeconomic events, which helps explain why events like the 2008 financial crisis and COVID-19 were not classified as strict thresholds. The QLR test found no significant break, though it has limited power against gradual shifts. As an additional test, Bai-Perron's multiple break point test was applied; the BIC selects zero breakpoints (m=0, BIC = −109.7). Both tests confirm the absence of a discrete structural shift in the series. Model performance on the holdout is therefore the main test.

2.4 Train-test split

The test covers the recent export boom. The April 2024 cutoff was chosen so that the test period captures the GLP-1 boom and a more present representation (Table 3). With 207 training observations and 24 months of holdout. All transformation parameters, model orders, and hyperparameters were fixed on the training set

before the holdout was evaluated. A forecast horizon of h = 24 reflects the planning window relevant for trade and pharmaceutical export decisions.

Set	Period	*N*
Training	2007M01 – 2024M03	207
Holdout (test)	2024M04 – 2026M03	24
Forecast horizon	$h = 24$	

Table 3: *Train Test Split*

3 Methodology

3.1 Model classes

The seasonal naïve serves only as an accuracy baseline and reference point. Exponential smoothing is represented by ETS as the competing benchmark class. Multiple ARIMA specifications were tested and compared before selecting the final model. As additional checks, three SARIMAX extensions were estimated incorporating a COVID-19 dummy, a GLP-1 demand dummy, and both combined, alongside a Fourier-augmented ARIMA variant.

3.2 ARIMA Specification

ARIMA is selected as one of the model classes because pharmaceutical exports exhibit a stochastic upward trend and strong seasonal structure characteristics that ARIMA handles through differencing and autoregressive moving-average terms. The unit root tests confirm one round of differencing is required (d=1) to remove the upward trend, with no seasonal differencing needed (D=0) as the seasonal pattern is stable and captured through seasonal AR and MA terms instead. Candidates are identified from the ACF and PACF inspection of the differenced log series, with auto.ARIMA included as a data driven benchmark. A drift term is included in the three manually specified candidates to account for the long-run upward trend visible in the series while auto.ARIMA determines this automatically. The three candidates include seasonal AR(1) and MA(1) terms at lag 12. The best specification is chosen based on which model produces the most reliable residuals and the lowest forecast error on data it has never seen.

3.3 Justification of the ETS Benchmark

ETS is included as the competing benchmark class because it reaches a forecast without any unit-root pre-testing, which makes it a useful, easy-to-read contrast to the Box-Jenkins route. We estimate three variants: additive errors, an additive trend and additive seasonality. Modelled on the log scale, additive seasonality here is close to multiplicative seasonality on the original scale, which is consistent with the growing seasonal swings we stabilised with the log transform and confirmed by the stable amplitude in the STL decomposition.

3.4 SARIMAX and Fourier Limitations

SARIMAX extensions incorporating COVID-19 and GLP-1 dummies were estimated to explore whether the outliers identified in the structural break analysis could be captured as explicit regressors. A fourier-augmented ARMIA variant was also included to test alternative approaches to seasonality. Both approaches are treated as additional checks rather than primary candidates.

3.5 Two-Stage Model Selection Criteria

We select the final model in two stages. The first stage is diagnostic validity: a model must return a Ljung-Box (Ljung & Box, 1978) p-value above 0.05 at lags 12 and 24. We treat this as a gate because autocorrelated residuals make the 80% and 95% prediction intervals unreliable, and for this series the forecast is used mainly through its intervals rather than its single point value.

The second stage is out-of-sample accuracy. Among models that pass the diagnostic gate the two class winners one from ARIMA and one from ETS are therefore compared directly in a head to head on holdout RMSE, MAPE, MASE. Residual diagnostics are examined for both finalists, including ACF plots and residual histograms, to confirm the winning model carries the stronger diagnostic profile. Rolling-origin

cross-validation over 19 origins confirms the result holds beyond the single test window. Information criteria are compared only within each model class, as ARIMA and ETS are computed on different bases and are not directly comparable across classes.

4 Results

4.1 ARIMA candidates

Four ARIMA specifications were estimated on Box-Cox-transformed training data: three manually chosen orders guided by ACF and PACF inspection, plus auto.ARIMA as a search benchmark. Each candidate includes a drift term to account for the long-term upward trend visible in the series (Figure 1). The table below reports information criteria (AIC, AICs, BIC) and holdout accuracy (RMSE, MAPE) side by side, with the Ljung-Box p-value at lag 24 as the diagnostic gate (Table 4).

Model	AIC	AICc	BIC	LB p (lag24)	RMSE (Bn DKK) (mean)	MAE (Bn DKK) (mean)	MAPE
ARIMA(0,1,1)(1,0,1)[12]	**−415**	**−415**	**−398**	**0.093**	**1.62**	**1.34**	**8.65%**
ARIMA(0,1,2)(1,0,1)[12]	**−417**	**−417**	**−397**	**0.363**	**1.73**	**1.39**	**8.85%**
ARIMA(1,1,1)(1,0,1)[12]	**−416**	**−416**	**−396**	**0.271**	**1.69**	**1.37**	**8.76%**
auto.ARIMA	**−399**	**−399**	**−379**	**0.155**	**2.02**	**1.54**	**9.58%**

All four candidates pass Ljung-Box at the 5% level. ARIMA(0,1,2)(1,0,1)[12] has the largest Ljung-Box margin at 0.363

Table 4: *Comparison of ARIMA candidates.*

The three manual candidates were chosen as follows. The ACF of diff(log exports) (Figure 5) shows a sharp negative spike at lag 1 that cuts off a clear signature of a moving-average (MA) process. This supports ARIMA(0,1,1)(1,0,1)[12] as the simplest starting specification, using one MA term to capture the dominant spike. However, a smaller but visible spike at lag 2 in the ACF does not fall inside the confidence bands, suggesting a second MA term may be influential. For this reason, a second candidate was selected, specifically ARIMA(0,1,2)(1,0,1)[12]. The PACF shows a significant negative spike at lag 1, with decay at lags 2 and 3 outside the confidence bands, which supports an autoregressive component as an alternative to a second MA term. Because of this, ARIMA(1,1,1)(1,0,1)[12] was chosen as the third candidate. All three include a seasonal AR(1) and seasonal MA(1) at lag 12, motivated by the prominent spike at lag 12 in the ACF, confirming a seasonal MA component, while the overall decay pattern in the PACF at seasonal lags supports a seasonal AR term. Auto.ARIMA was added purely as a search-based benchmark with no manual identification involved.

We test all four candidates on the Ljung-Box test at the 5% level and all pass the Ljung-Box test (Table 4) so the decision comes down to a combined reading of fit and diagnostic safety margin (Table 4). ARIMA(0,1,2)(1,0,1)[12] has the lowest AIC and AICc values and by far the largest Ljung-Box margin at 0.363, compared to a borderline 0.093 for the (0,1,1) variant. Auto.ARIMA performs worst on every metric and is eliminated. ARIMA(0,1,2)(1,0,1)[12] with drift is carried forward as the final ARIMA model for comparison against ETS.

The selected model is ARIMA(0,1,2)(1,0,1)[12] with drift, the specification that combines the manually identified non-seasonal MA(2) structure with the seasonal AR(1) and MA(1) terms confirmed by the ACF at lag 12. The estimated coefficients on the Box-Cox scale are (Table 5): MA(1) = −0.926 (s.e. 0.073), MA(2) = +0.166 (0.074), SAR(1) = 0.974 (0.024), SMA(1) = −0.845 (0.073), drift = 1e-04(1e-04). When exports

experience an unexpected shock, the MA terms ensure the effect fades. MA(1) reverses most of it the following month, MA(2) applies a small correction two months later. The seasonal terms confirm that the export pattern repeats strongly each year, with all MA and seasonal coefficients statistically significant at the 5% level (|t| > 2). The drift term has a t-ratio of around 1 and is not individually significant, but is retained because exports show a clear and sustained upward trend throughout the sample period (Figure 1).

```
Coefficients:
          ma1     ma2    sar1     sma1  constant
      -0.9257  0.1664  0.9744  -0.8446     1e-04
s.e.   0.0727  0.0738  0.0236   0.0729     1e-04
```

Table 5: Selected ARIMA model output

4.2 ETS candidates

Three ETS specifications were estimated on Box-Cox-transformed training data: ETS(A,A,A) (additive trend with additive seasonality), ETS(A,Ad,A) (additive damped trend with additive seasonality), and Auto ETS as a cross-check.

Model	AIC	AICc	BIC	LB p (lag 24)	RMSE (Bn DKK) (mean)	MAE (Bn DKK) (mean)	MAPE
ETS(A,A,A)	**103**	**106**	**160**	**0.124**	**1.69**	**1.37**	**8.92%**
ETS(A,Ad,A)	92	96	152	0.037	2.36	1.84	11.71%
Auto ETS	92	96	152	0.037	2.36	1.84	11.71%

Table 6: *Comparison of ETS candidates.*

ETS(A,A,A) is the only candidate that passes Ljung-Box at the 5% level, with a p-value of 0.124 (12.4%). Auto ETS converges on ETS(A,Ad,A), which fits the training data more tightly (better AIC) but degrades both diagnostics and holdout performance when asked to forecast through the post-2022 boom.

The selected ETS(A,A,A): $\alpha = 0.206$ depicting that the baseline updates gradually to new observations, $\beta = 0.000155$ reflecting an essentially fixed long-run drift, $\gamma \approx 0.036$ indicating a stable and persistent seasonal pattern, and $\sigma^2 = 0.0073$ suggesting little residual noise after accounting for all components.; while the AIC ≈ 102.9, the AICc ≈ 106.2 and the BIC ≈160.0 (Table 6).

4.3 Head-to-head and the diagnostic gate

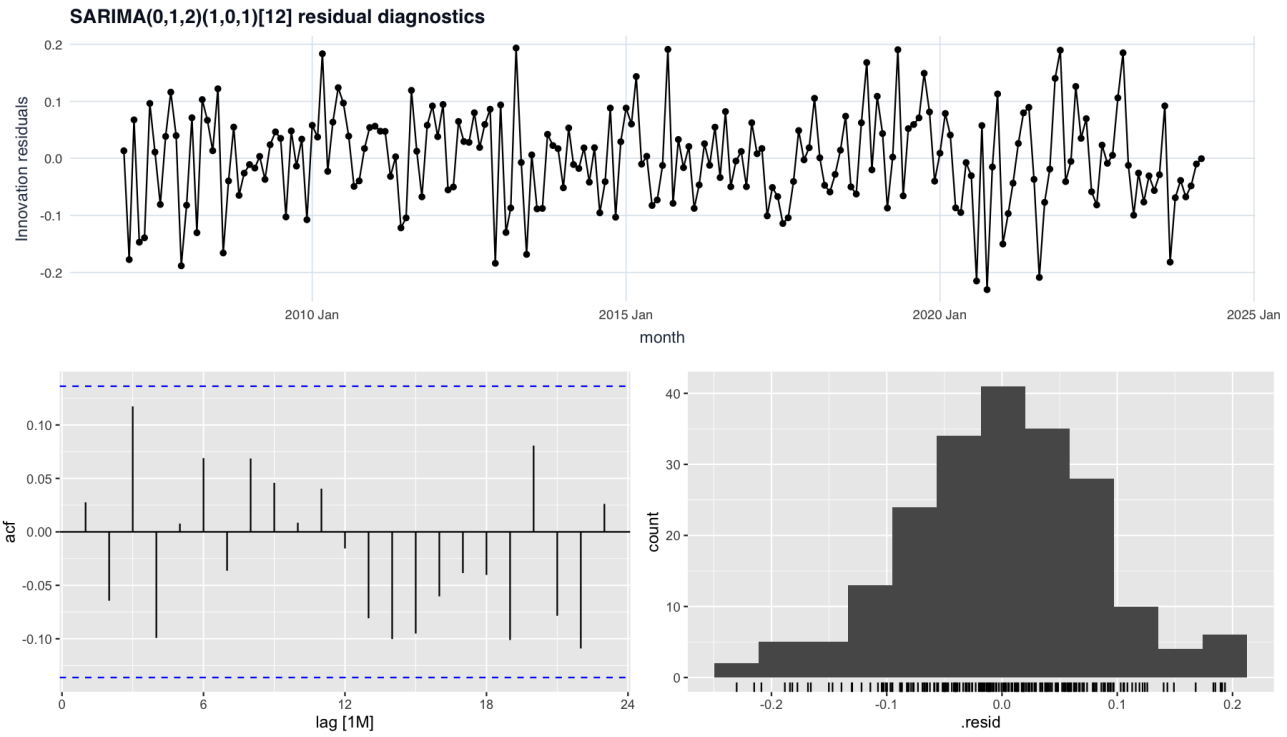
The two within-class winners are compared directly (Table 7). Per the two-stage gate (Hyndman & Athanasopoulos, 2021), before comparing accuracy, each model must first pass a residual diagnostic check. A model whose residuals show remaining autocorrelation produces unreliable prediction intervals and is excluded from the comparison, regardless of its holdout performance. Both models clear the Ljung-Box test at lag 24 returns p = 0.363 for ARIMA(0,1,2)(1,0,1)[12] and p = 0.124 for ETS(A,A,A), both well above the 5% threshold.

Model	AIC	AICc	BIC	LB p (lag 24)	RMSE (Bn DKK) (mean)	MAE (Bn DKK) (mean)	MAPE	MASE
ARIMA(0,1,2)(1,0,1)[12]	−417	−417	−397	0.363	1.73	1.39	8.85%	1.28
ETS(A,A,A)	103	106	160	0.124	1.69	1.37	8.92%	1.26

Table 7: *Head-to-head - ARIMA(0,1,2)(1,0,1)[12] vs ETS(A,A,A).*

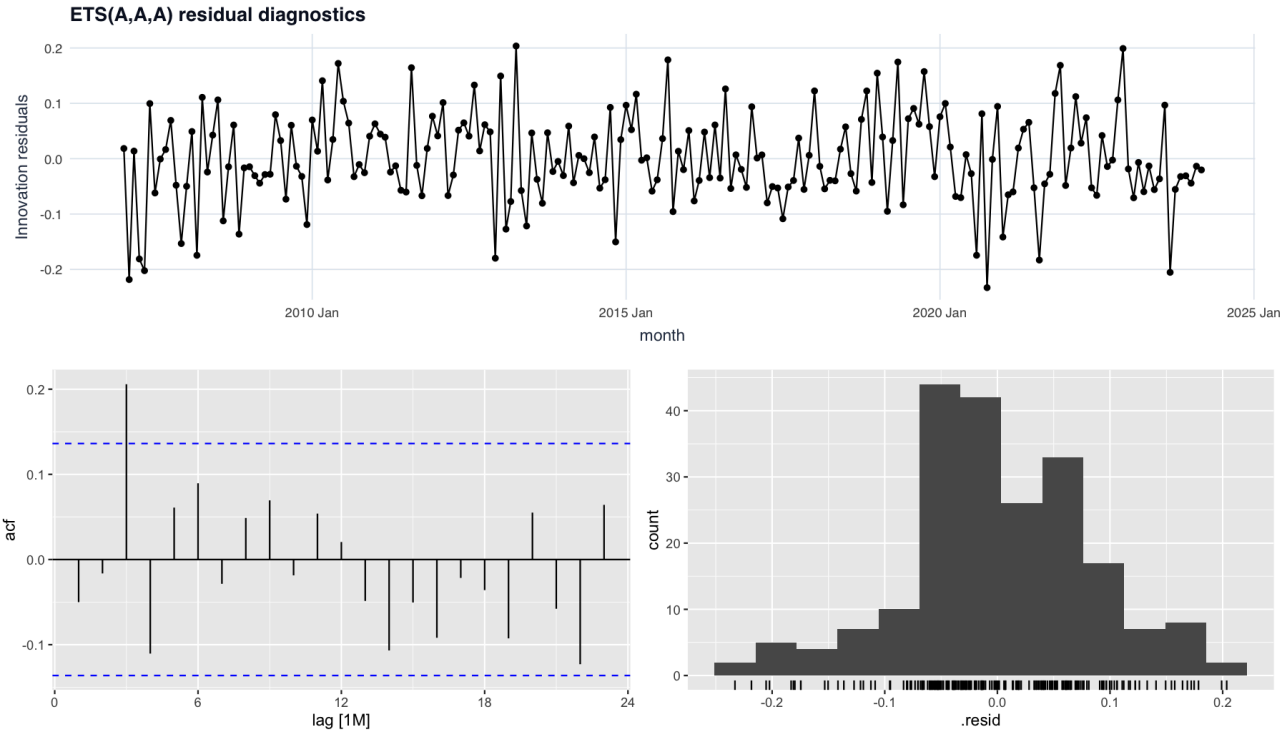
On the single holdout the two models are close. ETS(A,A,A) has marginally lower RMSE (1.69 vs 1.73 Bn DKK) and MAE, the gap is around 2%, MAPE is almost the same (8.92% vs 8.85%), and the MASE scores are 1.26 for ETS against 1.28 for ARIMA. So on point accuracy over this window ETS is, if anything,

slightly ahead. We still select ARIMA(0,1,2)(1,0,1)[12] with drift for two reasons: it wins the rolling cross-validation (Section 4.5), and it has the wider residual-diagnostic margin (Ljung-Box p = 0.363 against 0.124), which matters here because the forecast is used through its prediction intervals. We do not use AIC in this comparison, since the ARIMA and ETS values (−417 and 103) are computed differently and are not comparable across classes (Hyndman and Athanasopoulos, 2021).



ARIMA residual diagnostics

***Figure 6**: ARIMA(0,1,2)(1,0,1)[12] residuals, ACF, histogram (gg_tsresiduals).*



ETS residual diagnostics

***Figure 7**: ETS(A,A,A) passes Ljung-Box*

Residual Diagnostic supports this choice (Figures 6 and 7), both models produce residual centred around zero with no systematic variance drift. The ARIMA ACF shows all bars within the 95% confidence bands, confirming clean white noise, while ETS(A,A,A) displays a marginal spike at lag 1, though the joint Ljung-Box test still passes (p = 0.124). The ARIMA residual histogram is approximately symmetric, whereas the ETS shows a slight left skewness. The ARIMA's Ljung-Box p-value of 0.363 provides a substantially wider safety margin for the prediction interval used in the final forecast.

Four additional specifications were considered and excluded by the diagnostic gate. ARIMA + Fourier(K=2) achieves the lowest holdout RMSE (1.55 Bn DKK) and MAPE (8.27%) but failed Ljung-Box at both lags ($p \approx 0$), so its prediction intervals would be unreliable. Three SARIMAX variants were also tested (a GLP-1 dummy, a COVID dummy, and both combined). Using the Ljung-Box test with degrees of freedom equal to the number of estimated parameters, all three fail at lag 24 (p = 0.029, 0.047 and 0.034), and their holdout RMSE of roughly 2.02–2.85 Bn DKK offers no improvement over the simpler ARIMA despite the added complexity. This is likely because the ARIMA drift term had already adjusted to the rising trend thus the dummy added no new information. Seasonal naïve serves as the MASE scaling benchmark only, with a scaling factor of 1,088,384,000 DKK.

4.4 Forecast comparison

The ETS(A,A,A) holdout forecast (Figure 8) captures the general upward level of the series but fails to follow the sharp acceleration from mid-2024 onward, producing a point forecast that rises more slowly than the actual series. This is reflected in a positive mean error of around 574.358.000 DKK, indicating under-prediction during the holdout period.

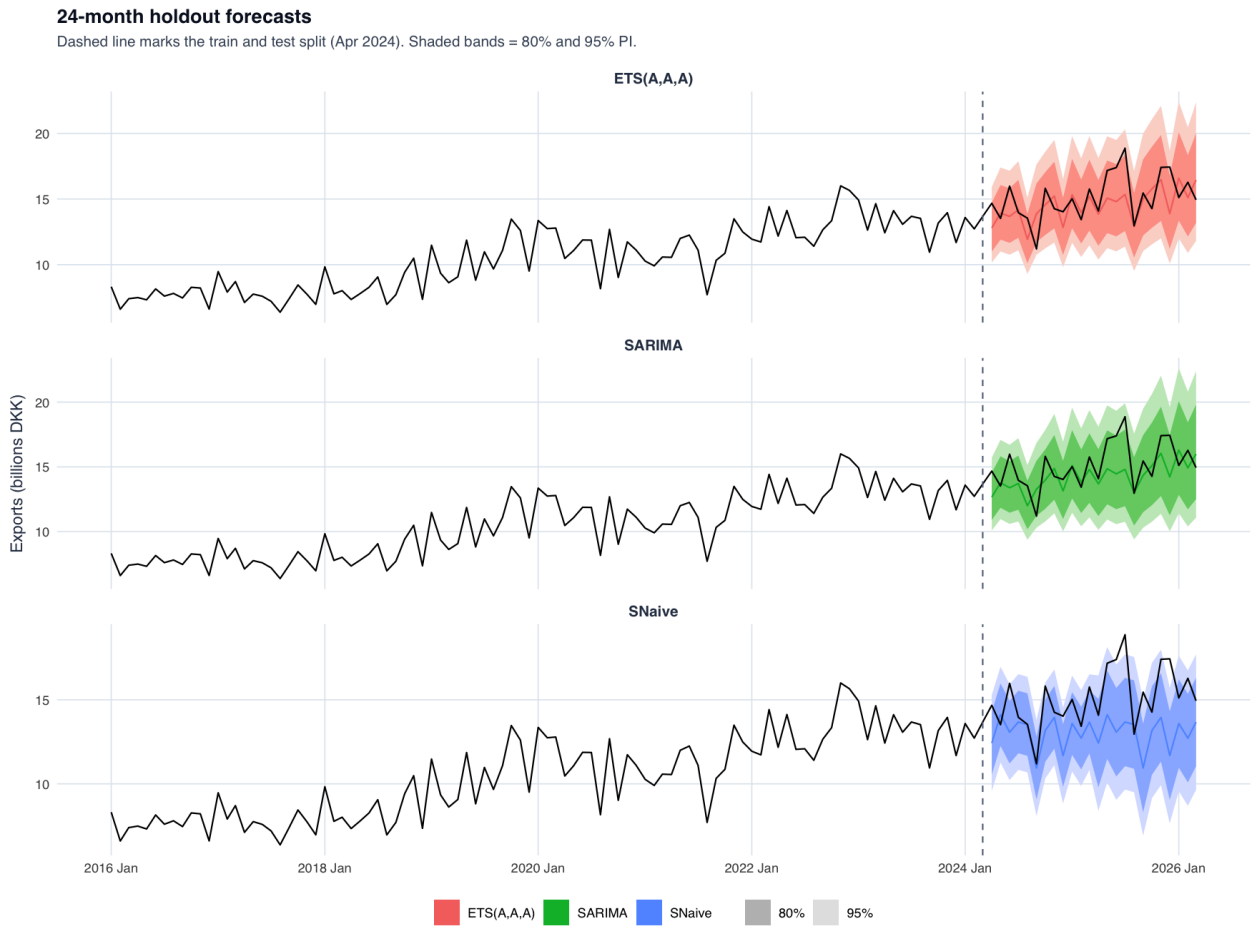


Figure 8. *24-month holdout forecasts - model comparison (SARIMA, ETS, SNaive).*

Both models under-predict the 2024 to 2026 holdout, and SARIMA under-predicts a little more than ETS: its mean error is about 846.196.000 DKK against ETS's 574.358.000 DKK (Figure 8). On the holdout, the two are close overall, with SARIMA at MASE 1.28 against ETS's 1.26 and a marginally lower MAPE (8.85 vs 8.92). SARIMA's advantage does not show up on this single split; it shows up in the cross-validation and in its cleaner residuals.

Lastly, Figure 9 summarises forecast accuracy across all candidate models. This figure is necessary because it explains why SARIMA is selected over models with lower RMSE: ARIMA+Fourier achieves the best raw accuracy (1.55 Bn DKK) but fails the Ljung-Box test, indicating it is statistically invalid. SARIMA (1.73Bn DKK) is the best model among those passing all diagnostics. The comparison also positions SARIMA against the seasonal naive and ETS baselines, showing the gain from the full modelling effort.

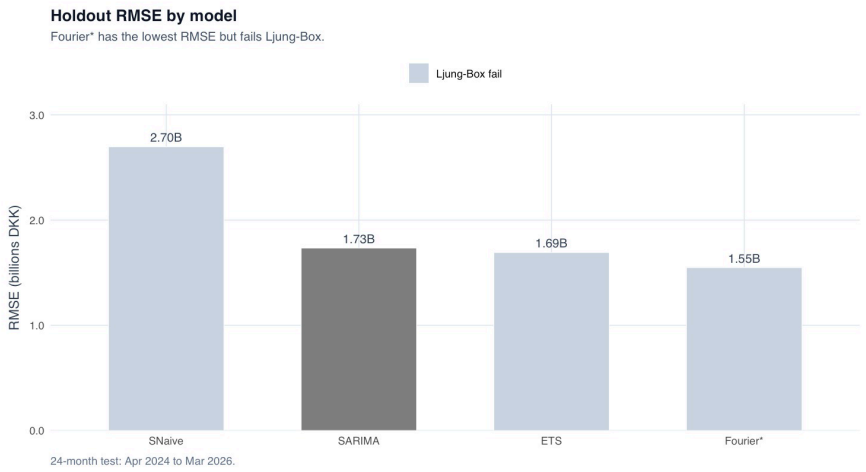


Figure 9. *Holdout RMSE by model: 24-month test*

4.5 Rolling cross-validation

The single holdout split could favour SARIMA by chance. To test whether its advantage generalises, a rolling cross-validation was run over 19 origins using a stretch window with a step size of 6 months (Table 8).

Model	RMSE (Bn DKK) (mean)	MAE (Bn DKK) (mean)	MAPE (mean)
SARIMA	**1.66**	**1.35**	**11.6%**
ETS(A,A,A)	1.70	1.43	12.3%
SNaive	1.92	1.60	13.1%

Table 8: *Rolling cross-validation: mean accuracy over 19 origins*

SARIMA wins on all three averaged metrics. Its mean RMSE of 1.66 Bn DKK is around 2.5% lower than ETS(A,A,A) and 14% lower than seasonal naive, with consistent gains on MAE and MAPE as well. The advantage holding across 19 different evaluation windows confirms it is not an artifact of the particular test split used in section 4.4.

5 Conclusions

5.1 Main findings

This paper asked whether a standard univariate time series model could produce a diagnostically valid and accurate 24-month forecast of Danish pharmaceutical exports. The answer to this research question is yes, with important qualifications.

Among the seven models we tested, ARIMA(0,1,2)(1,0,1)[12] with drift gives the best combination of clean residuals and out-of-sample accuracy. ETS(A,A,A) also passes the diagnostic gate and matches SARIMA on the single holdout, but SARIMA wins the rolling cross-validation and has the wider Ljung-Box margin at lags 12 and 24.

It produces a holdout RMSE of 1.73Bn DKK, a MAPE of 8.85%, and a MASE of 1.28 on the demanding April 2024 to March 2026 test period. Rolling cross-validation over 19 origins confirms this advantage is consistent across different evaluation windows.

The Fourier regression model achieves slightly lower point-forecast errors but fails residual diagnostics, making its prediction intervals unreliable. Adding structural dummies for GLP-1 or COVID did not improve out-of-sample performance, suggesting the level shift was already absorbed by recent observations before a step function could help.

5.2 Practical implications

For a non-technical reader: exports will probably stay high, but the month-to-month forecast band is wide. The prediction intervals matter more than the point forecast. A univariate model captures trend and seasonality but cannot anticipate when GLP-1 demand will plateau or when competitors will enter the market, which is why qualitative scenario analysis is needed.

For a CFO or trade analyst: SARIMA is a reasonable 12 to 24-month central case, with ETS serving as a reference point. When the two models agree, that lowers the risk of a specification mistake, but it does not remove the risk they share: both are univariate, both miss demand-side information, and on the holdout, both under-predicted the elevated export levels in the same direction. Where the disagreement arises, we lean on SARIMA, mainly because it gave more reliable intervals and held up better under cross-validation. GLP-1 step dummies did not improve the holdout, so product-level expectations belong in qualitative planning rather than in the model equation itself.

For a statistician reviewing the work, the key checkpoints are whether Box-Cox was applied consistently, whether Ljung-Box degrees of freedom match estimated parameters, whether the holdout was truly withheld during estimation, and whether CV folds respect temporal ordering. The analysis satisfies all four.

5.3 Discussion

First, the failure of external regressors. Adding GLP-1 and COVID step dummies worsened holdout performance despite improving in-sample fit. By the time the GLP-1 acceleration was large enough to model as a structural break, the drift term in SARIMA had already absorbed much of it through recent observations. A step dummy fires once and stays fixed; the model's level had already been adjusted. For series undergoing smooth, sustained regime shifts, external dummies tend to add noise rather than signal out-of-sample.

Second, the MASE above 1 needs context. A MASE of 1.28 means that, on the holdout, SARIMA is less accurate month by month than a seasonal-naive forecast. The holdout is the most unpredictable period in the whole sample, with GLP-1 demand pushing exports to new highs month after month, so even a good model still trails a naive one that simply repeats last year's level. More importantly, seasonal-naive gives neither a trend-extrapolating central path as it copies last year's values. It cannot show where exports are heading or provide a forecast range for planning. Those are exactly what a decision-maker needs. SARIMA captures the upward trend and produces meaningful forecast bands, which is where its value lies. In rolling cross-validation, where the comparison is fairer, SARIMA also outperforms seasonal-naïve, with a mean RMSE of 1.66B DKK against 1.93B DKK.

Third, the near-identical holdout performance of SARIMA and ETS (around 2.5% on RMSE and MAE) validates the benchmark design. When two structurally different model classes produce nearly the same forecast, it suggests that both models have captured most of what the data can tell us. The data itself is noisy, so neither model can do much better in monthly pharmaceutical trade data. The preference for SARIMA therefore rests on its diagnostic margin and cleaner residual structure, not on a decisive accuracy advantage that ETS cannot reach.

5.4 Limitations

Several limitations should be noted. Novo Nordisk's dominance means the aggregate pharmaceutical export series effectively behaves like a single-firm proxy, and adding extra variables to capture this did not help out-of-sample. The analysis uses nominal DKK values; real exports adjusted for price or mix shifts would require additional data not available at this stage. The Ljung-Box p-value at lag 24 of 0.124 passes the 5% threshold but is not particularly large, so ongoing monitoring is advisable as new observations arrive.

Finally, the holdout covers an unusually fast-growth period, so these accuracy results may not generalise to calmer conditions.

5.5 Future work

The clearest next step follows from the concentration issue: moving from the aggregate series towards firm- or product-level data, or modelling capacity and supply events directly, since these are the things a univariate macro model cannot capture. A variable that keeps changing over time, such as a production index might also work where the GLP-1 and COVID step dummies failed, because it keeps moving after the level shift instead of firing once and staying fixed. Bayesian structural time series would handle the regime change more explicitly, and forecasting growth rates rather than levels is worth testing. Any extension should be judged with the same diagnostics-first, accuracy-second framework used here.

References

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica, 61*(4), 821–856.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological), 26*(2), 211–252.

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. Holden-Day.

Bregenholt, S. (2015, September 6). The valley of life: South Scandinavia's Silicon Valley. *The Copenhagen Post*.
https://cphpost.dk/2015-09-06/life-in-denmark/opinion-old/the-valley-of-life-south-scandinavias-silicon-valley/

Charoux, V. (2025, April 14). Which are the top 10 EU countries exporting pharmaceutical products in 2024? *Portugal Business News*.
https://www.portugalbusinessesnews.com/post/which-are-the-top-10-eu-countries-exporting-pharmaceutical-products-in-2024

Danmarks Statistik. (2026). *Foreign trade in goods (SITC 54, medicinal and pharmaceutical products)* [Data set]. Statbank. Retrieved May 2026, from https://www.statbank.dk

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74*(366), 427–431.

Guerrero, V. M. (1993). Time-series analysis supported by power transformations. *Journal of Forecasting, 12*(1), 37–48.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. https://otexts.com/fpp3/

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics, 54*(1–3), 159–178.

Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika, 65*(2), 297–303.

Pharma Boardroom. (2025, September 29). Denmark: Life sciences leadership in the Nordic powerhouse. https://pharmaboardroom.com/country-reports/denmark-country-report-2025/

Phillips, P. C. B., & Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika, 75*(2), 335–346.

Pretis, F., Reade, J. J., & Sucarrat, G. (2018). Automated general-to-specific (GETS) regression modeling and indicator saturation for outliers and structural breaks. *Journal of Statistical Software, 86*(3), 1–44. https://doi.org/10.18637/jss.v086.i03

Quandt, R. E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association, 55*(290), 324–330.

Zeileis, A., Leisch, F., Hornik, K., & Kleiber, C. (2002). strucchange: An R package for testing for structural change in linear regression models. *Journal of Statistical Software, 7*(2), 1–38. https://doi.org/10.18637/jss.v007.i02