# Nintendo sales analysis

Valerio Ferdinando Calà

25/03/2021

## Analisi esplorativa dei dati

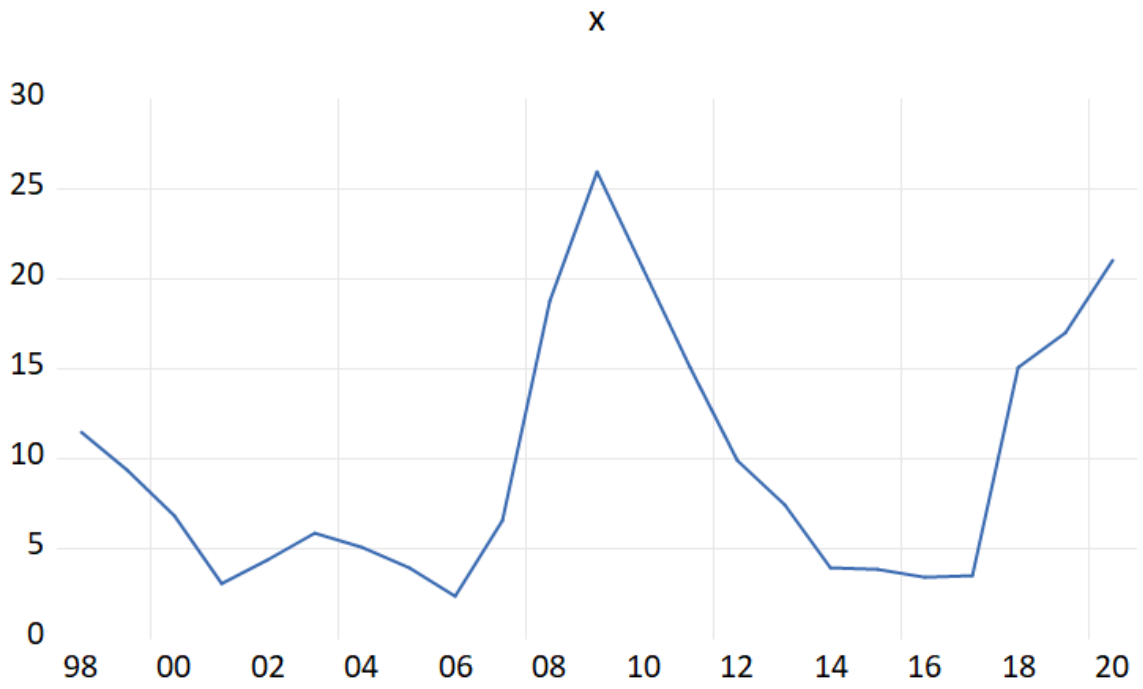



Figure 1: X = serie storica annuale delle unità (in milioni) di console domestiche Nintendo vendute

Dall'analisi grafica dei dati grezzi, possiamo vedere come le unità vendute hanno seguito un andamento altalenante negli anni: nel 1998 c'erano più di 10 milioni di unità vendute, ma questo numero è diminuito per tutto il periodo 1998-2006. È solo a partire dal 2007 che inizia una nuova tendenza positiva ed il numero di unità vendute ogni anno è costantemente superiore ai 10 milioni di unità fino al 2011, con un picco di oltre 25 milioni di unità vendute nel 2009.

Dopo il picco del 2009, a livello grafico vediamo che la serie ha seguito una tendenza negativa fino a toccare un plateau di meno di 4 milioni di unità vendute per 4 anni consecutivi (2014-2017) e 'schizzare' oltre le 15 milioni di unità vendute negli ultimi tre anni osservati (2018-2020).

È dunque ragionevole chiedersi se è successo qualcosa prima dell'inizio delle due tendenze positive, cioè prima del 2007 e prima del 2018, che giustifica un aumento così eccessivo del numero di consoles domestiche vendute.

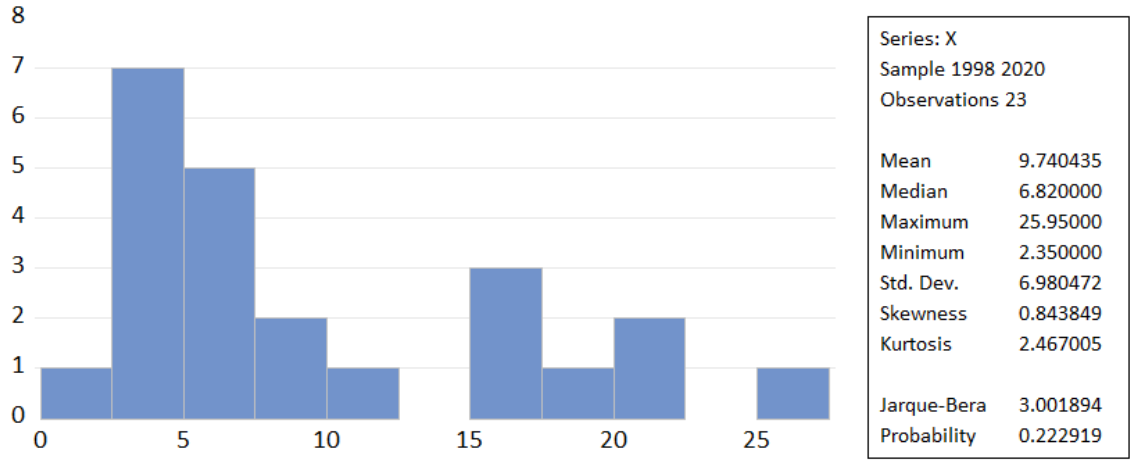



Figure 2: Tabella dei principali indici di posizione, dispersione, asimmetria e curtosi. Test di normalità.

Data la tabella di cui sopra, non possiamo rifiutare l'ipotesi nulla: x ha distribuzione Normale.

Infatti, il test di Jarque-Bera confronta gli indici di asimmetria e di curtosi con quelli che si otterrebbero nel caso di dati normalmente distribuiti. In caso di normalità, abbiamo un'asimmetria nulla ed un indice di curtosi vicino a 3. In questo caso i dati esibiscono una leggera asimmetria positiva, cosa che si può evincere anche dall'istogramma e dal confronto dei valori di media (9.74 milioni) e mediana (6.8 milioni). Sebbene i dati non siano distribuiti in modo simmetrico, non c'è abbastanza evidenza empirica per rifiutare l'ipotesi nulla di normalità.

Dalle altre statistiche descrittive possiamo fare le seguenti considerazioni:

- Su un totale di 23 anni osservati, nella metà degli anni abbiamo visto un numero di consoles domestiche vendute pari o inferiore a 6.82 milioni di unità.
- Su un totale di 23 anni osservati, la media delle unità di consoles domestiche vendute è di 9.74 milioni con una deviazione standard pari a 6.98 milioni. Il coefficiente di variazione è quindi pari a circa 0.72 (valore utile se si vuole confrontare la variabilità del fenomeno, con quella di fenomeni espressi in unità di misura diverse).
- I dati variano in un range di valori molto ampio: il massimo è pari a 25.95 milioni, il minimo è pari a 2.35 milioni e quindi il range è pari a 23.60 milioni.

Il box-plot è un grafico particolarmente utile per analizzare la dispersione dei dati e le principali misure di posizione (quantili e media), nonché per cercare di individuare ad occhio valori anomali (troppo estremi, cioè troppo elevati o troppo bassi). Le conclusioni di questo grafico sono identiche a quelle fatte sulla base della tabella delle statistiche descrittive, ma in questo caso sono desumibili a colpo d'occhio e la proporzione tra la scatola ed i baffi suggerisce subito un'alta dispersione (che prima abbiamo quantificato con il range e

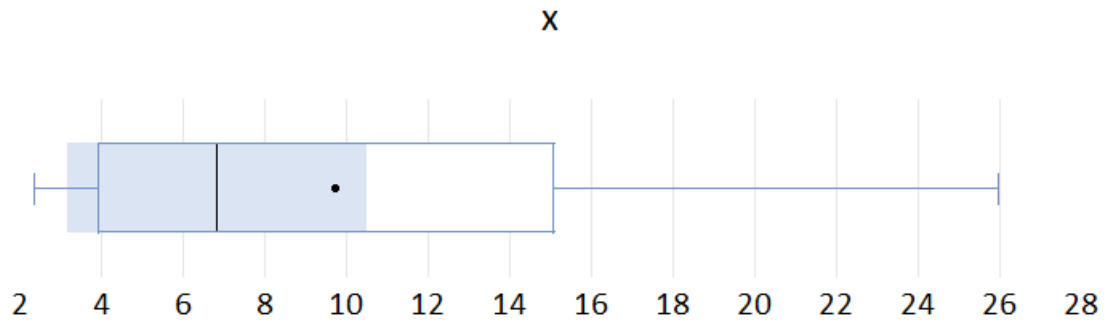
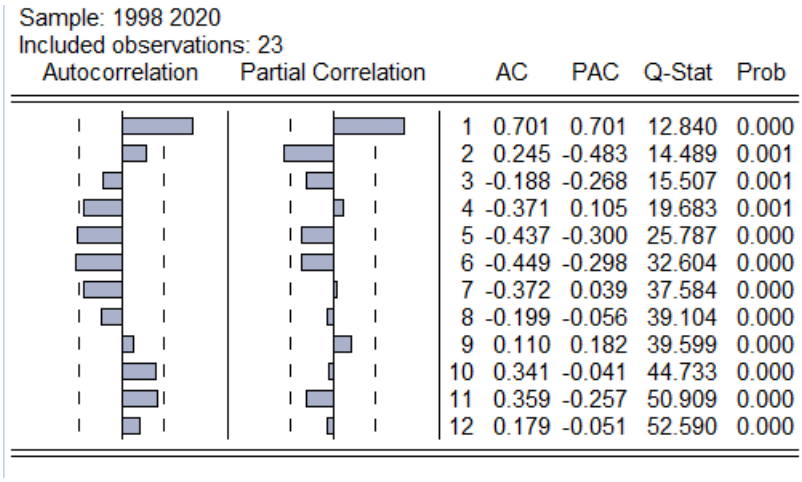



Figure 3: Box-plot

con il coefficiente di variazione).

## Analisi dell'autocorrelazione




| | AC | PAC | Q-Stat | Prob |
|---|---|---|---|---|
| 1 | 0.701 | 0.701 | 12.840 | 0.000 |
| 2 | 0.245 | -0.483 | 14.489 | 0.001 |
| 3 | -0.188 | -0.268 | 15.507 | 0.001 |
| 4 | -0.371 | 0.105 | 19.683 | 0.001 |
| 5 | -0.437 | -0.300 | 25.787 | 0.000 |
| 6 | -0.449 | -0.298 | 32.604 | 0.000 |
| 7 | -0.372 | 0.039 | 37.584 | 0.000 |
| 8 | -0.199 | -0.056 | 39.104 | 0.000 |
| 9 | 0.110 | 0.182 | 39.599 | 0.000 |
| 10 | 0.341 | -0.041 | 44.733 | 0.000 |
| 11 | 0.359 | -0.257 | 50.909 | 0.000 |
| 12 | 0.179 | -0.051 | 52.590 | 0.000 |

Figure 4: Correlogramma

Dall'analisi grafica dell'ACF e della PACF possiamo vedere come, oltre al primo ritardo, l'effetto di autocorrelazione nel tempo svanisce rapidamente. Questo ci farebbe pensare ad un modello $AR(1)$ per la serie storica $\{X_t\} \quad t \in 1998, .., 2020$.

Purtroppo, però, l'analisi non è così semplice perché bisogna prima accertarsi che la serie storica non presenti radici unitarie. Questo viene svolto in letteratura con il test di Dickey-Fueller aumentato (test ADF), il cui output è riportato nella tabella seguente:

Non possiamo rifiutare con una confidenza del 95% l'ipotesi nulla di radici unitarie: in altre parole, l'effetto marginale del lag di primo ordine di $X$ è esattamente pari a 1 e per questo dobbiamo modellare la serie in differenze prime, $DX_t = X_t - X_{t-1}$.

A questo punto, bisogna fare un nuovo test di radice unitaria; se non si può rifiutare l'ipotesi nulla, bisogna

Null Hypothesis: X has a unit root
Exogenous: Constant
Lag Length: 1 (Automatic - based on AIC, maxlag=4)

| | | t-Statistic | Prob.* |
|---|---|---|---|
| Augmented Dickey-Fuller test statistic | | -2.695375 | 0.0914 |
| Test critical values: | 1% level | -3.788030 | |
| | 5% level | -3.012363 | |
| | 10% level | -2.646119 | |

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(X)
Method: Least Squares
Date: 03/25/21 Time: 15:01
Sample (adjusted): 2000 2020
Included observations: 21 after adjustments

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| X(-1) | -0.386961 | 0.143565 | -2.695375 | 0.0148 |
| D(X(-1)) | 0.651345 | 0.197610 | 3.296108 | 0.0040 |
| C | 3.916214 | 1.558130 | 2.513406 | 0.0217 |

| | | | |
|---|---|---|---|
| R-squared | 0.420759 | Mean dependent var | 0.556667 |
| Adjusted R-squared | 0.356399 | S.D. dependent var | 4.997911 |
| S.E. of regression | 4.009561 | Akaike info criterion | 5.746804 |
| Sum squared resid | 289.3784 | Schwarz criterion | 5.896022 |
| Log likelihood | -57.34144 | Hannan-Quinn criter. | 5.779188 |
| F-statistic | 6.537579 | Durbin-Watson stat | 2.112383 |
| Prob(F-statistic) | 0.007341 | | |

Figure 5: Test ADF

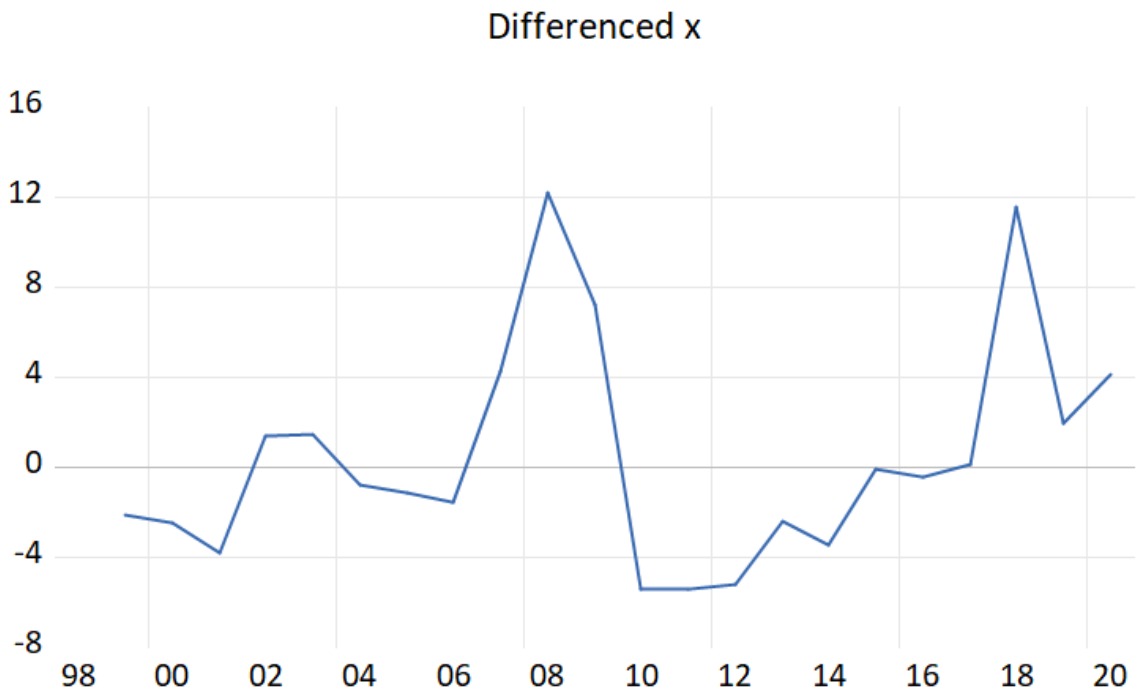



Figure 6: DX = differenza annuale delle unità (in milioni) di console domestiche Nintendo vendute

prendere la differenza di secondo ordine $DX_t^2 = X_t - X_t - 2$. Fortunatamente, dall'output del test possiamo vedere che l'ipotesi nulla di radice unitaria è rifiutata con una probabilità di errore di primo tipo inferiore all'1%.

Null Hypothesis: DX has a unit root
Exogenous: None
Lag Length: 0 (Automatic - based on AIC, maxlag=4)

| | | t-Statistic | Prob.* |
|---|---|---|---|
| Augmented Dickey-Fuller test statistic | | -2.746561 | 0.0085 |
| Test critical values: | 1% level | -2.679735 | |
| | 5% level | -1.958088 | |
| | 10% level | -1.607830 | |

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(DX)
Method: Least Squares
Date: 03/25/21 Time: 15:17
Sample (adjusted): 2000 2020
Included observations: 21 after adjustments

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| DX(-1) | -0.559666 | 0.203770 | -2.746561 | 0.0124 |

| | | | |
|---|---|---|---|
| R-squared | 0.271495 | Mean dependent var | 0.296667 |
| Adjusted R-squared | 0.271495 | S.D. dependent var | 5.306597 |
| S.E. of regression | 4.529312 | Akaike info criterion | 5.905465 |
| Sum squared resid | 410.2934 | Schwarz criterion | 5.955204 |
| Log likelihood | -61.00738 | Hannan-Quinn criter. | 5.916260 |
| Durbin-Watson stat | 1.791487 | | |

Figure 7: Test ADF

## Identificazione e stima del modello: metodo di Box-Jenkins

Definiamo due variabili **dummy ausiliarie** $D_1$ e $D_2$, che rispettivamente si accendono durante gli anni in cui vengono vendute le due consoles più recenti di successo (Wii e Nintendo Switch).

Viene stimato il **Modello ARIMAX(k, 1, 0)** per diverse scelte di $k$:

$$DX_t = X_t - X_{t-1} = \beta_0 + \beta_1 X_{t-1} + \beta_2 X_{t-2} + ... + \beta_k X_{t-k} + \gamma_1 D_{1t} + \gamma_2 D_{2t} + \epsilon$$

In tutti i casi, l'intercetta $\beta_0$, il coefficiente $\beta_1$ e tutti i ritardi di ordine pari o superiore a 3 non sono statisticamente significativi. Il miglior modello in termini di significatività statistica e bontà di adattamento ($R^2$ aggiustato e $AIC$) si ha per $k^\star = 2$.

Si riporta a titolo esemplificativo l'output del modello stimato per $k = 3$, seguito dall'output del modello stimato per $k^\star$.

Dependent Variable: DX
Method: Least Squares
Date: 03/25/21 Time: 16:02
Sample: 1998 2020 IF 2008
Included observations: 20

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| @LAG(X,1) | 0.051839 | 0.232084 | 0.223361 | 0.8263 |
| @LAG(X,2) | -0.535700 | 0.322745 | -1.659824 | 0.1177 |
| @LAG(X,3) | 0.082803 | 0.205429 | 0.403075 | 0.6926 |
| DUMMY1 | 5.427906 | 2.178370 | 2.491729 | 0.0249 |
| DUMMY2 | 8.848484 | 2.688751 | 3.290927 | 0.0050 |

| | | | |
|---|---|---|---|
| R-squared | 0.573493 | Mean dependent var | 0.710500 |
| Adjusted R-squared | 0.459758 | S.D. dependent var | 5.076484 |
| S.E. of regression | 3.731274 | Akaike info criterion | 5.683695 |
| Sum squared resid | 208.8361 | Schwarz criterion | 5.932628 |
| Log likelihood | -51.83695 | Hannan-Quinn criter. | 5.732289 |
| Durbin-Watson stat | 1.961052 | | |

Figure 8: Modello per k=3

Dependent Variable: DX
Method: Least Squares
Date: 03/25/21 Time: 15:57
Sample: 1998 2020 IF 2008
Included observations: 21

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| @LAG(X,1) | 0.014558 | 0.194080 | 0.075008 | 0.9411 |
| @LAG(X,2) | -0.411273 | 0.174224 | -2.360599 | 0.0305 |
| DUMMY1 | 5.301216 | 1.965830 | 2.696681 | 0.0153 |
| DUMMY2 | 8.676100 | 2.540937 | 3.414528 | 0.0033 |

| | | | |
|---|---|---|---|
| R-squared | 0.567736 | Mean dependent var | 0.556667 |
| Adjusted R-squared | 0.491454 | S.D. dependent var | 4.997911 |
| S.E. of regression | 3.564132 | Akaike info criterion | 5.549362 |
| Sum squared resid | 215.9517 | Schwarz criterion | 5.748318 |
| Log likelihood | -54.26830 | Hannan-Quinn criter. | 5.592540 |
| Durbin-Watson stat | 1.821830 | | |

Figure 9: Modello per k=2

Dall'output del modello di regressione temporale, possiamo vedere che non c'è un grosso problema di autocorrelazione degli errori (statistica Durbin Watson vicina a 2), che la bontà di adattamento per la differenza di primo ordine $DX$ è di circa il 50% (dato non allarmante: interessa valutare la bontà di adattamento della serie $X$).

Interessante vedere come entrambe le ultime due consoles abbiano avuto un effetto medio positivo rispetto ai livelli medi di vendita pre-uscita della Wii.

La regressione stimata è quindi:

$$\widehat{X_t - X_{t-1}} = \widehat{DX} = -0.41X_{t-2} + 5.30D_{1t} + 8.67D_{2t}$$

Visualizziamo nel grafico seguente il confronto tra: serie **actual** della differenza $DX = X_t - X_{t-1}$, serie dei valori **fitted** e serie dei **residui**.

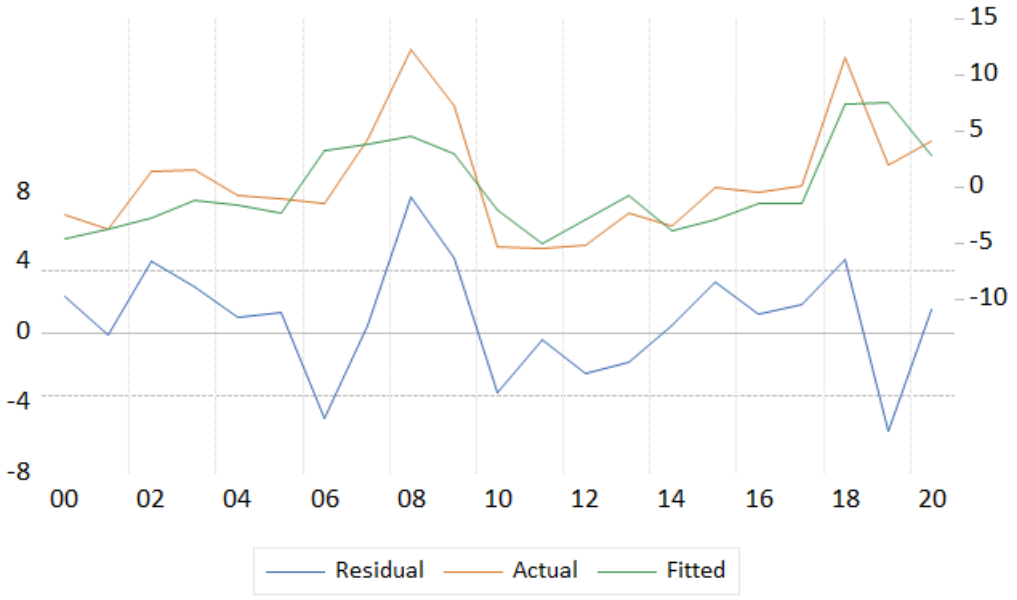



Figure 10: Confronto tra DX actual, DX fitted e residui

Per generare infine la serie dei valori previsti dal modello, sarà sufficiente sommare ai valori previsti per la differenza prima $\widehat{DX}$ il livello della serie al passo precedente $X_{t-1}$, come nell'equazione seguente:

$$\widehat{X_t} = X_{t-1} - 0.41X_{t-2} + 5.30D_{1t} + 8.67D_{2t}$$

**Misure di performance**:

$$R^2 = 1 - \frac{\sum_{t=2000}^{2020}(x_t - \widehat{x_t})^2}{\sum_{t=1998}^{2020} x_t^2} = 1 - \frac{\sum e_t^2}{\sum x_t^2} = 0.9335$$

$$MAE = T^{-1} \sum_{t=1}^{T} |e_t| = 2.5425$$

$$RMSE = \sqrt{T^{-1} \sum_{t=1}^{T} e_t^2} = 3.2104$$

**Modello finale scelto con approccio Box-Jenkins: ARIMAX(2,1,0)**

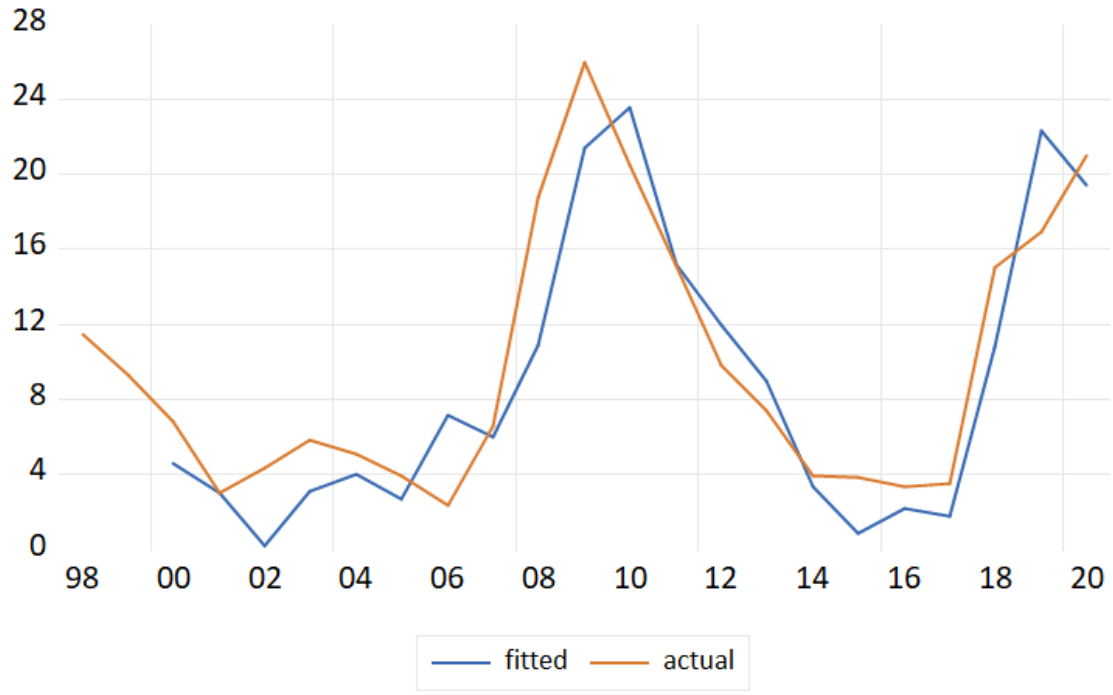



Figure 11: Confronto tra X actual e X fitted