Progetto SMS2

Silviu Filote 1059252,
Nicolò Carissimi 1069015,
Jonathan Bommarito 1068755

June 11, 2021

Contents

1 Prima analisi

1.1 Descrizione del dataset

Il set di dati contiene 9358 istanze di risposte medie orarie da una serie di 5 sensori chimici di ossido di metallo incorporati in un dispositivo multi-sensore chimico per la qualità dell'aria. Il dispositivo è stato localizzato sul campo in un'area notevolmente inquinata, a livello stradale, all'interno di una città italiana. I dati sono stati registrati da marzo 2004 a febbraio 2005 che rappresentano le registrazioni delle risposte di dispositivi di sensori chimici della qualità dell'aria utilizzati sul campo. Le concentrazioni orarie medie di Ground Truth per CO, idrocarburi non metanici, benzene, ossidi di azoto totali (NOx) e biossido di azoto (NO2) sono state fornite da un analizzatore di riferimento certificato. I valori mancanti sono contrassegnati con il valore -200.

Date	Date (DD/MM/YYYY)
Time	Time (HH.MM.SS)
CO(GT)	True hourly averaged concentration CO in mg/m^3 (reference analyzer)
PT08.S1(CO)	PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
NMHC(GT)	Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)
C6H6(GT)	True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)
PT08.S2(NMHC)	PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
NOx (GT)	True hourly averaged NOx concentration in ppb (reference analyzer)
PT08.S3(NOx)	PT08.S3 (tungsten oxide) hourly averaged sensor response
NO2(GT)	True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)
PT08.S4(NO2)	PT08.S4 (tungsten oxide) hourly averaged sensor response
PT08.S5(O3)	PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)
T	Temperature in Â°C
RH	Relative Humidity (%)
AH	AH Absolute Humidity

Figure 1: variabili presenti nel dataset

1.2 Quesiti che ci siamo posti

- Modellizzare la risposta del sensore *PT08.S1(C0)* mediante i valori registrati degli altri sensori;
- Modellizzare i valori di ground truth *C0* in funzione degli altri inquinanti;
- Confronto dei due modelli;

1.3 Metodi statistici utilizzati

- Metodo OLS;
  - Pulizia dataset;
  - Dataset di validazione ("Holdout");
  - Stima del modello utilizzando tutti i regressori presenti nel dataset;
  - Analisi covariate con rimozione dei regressori non significati;
  - Analisi residui;
  - Test F sui coefficienti dei regressori;
  - Verifica di non distorsione dei beta;
- Metodo GLS;
  - Tecnica di crossvalidazione, "K-fold";
  - Analisi residui;

1.4 Risultati

Dopo aver pulito ed eliminato i dati inconsistenti, si è deciso di utilizzare, inizialmente, il modello lineare. Prima di procedere si è suddiviso il dataset in due parti:

- Training set = 70%
- Test set = 30%

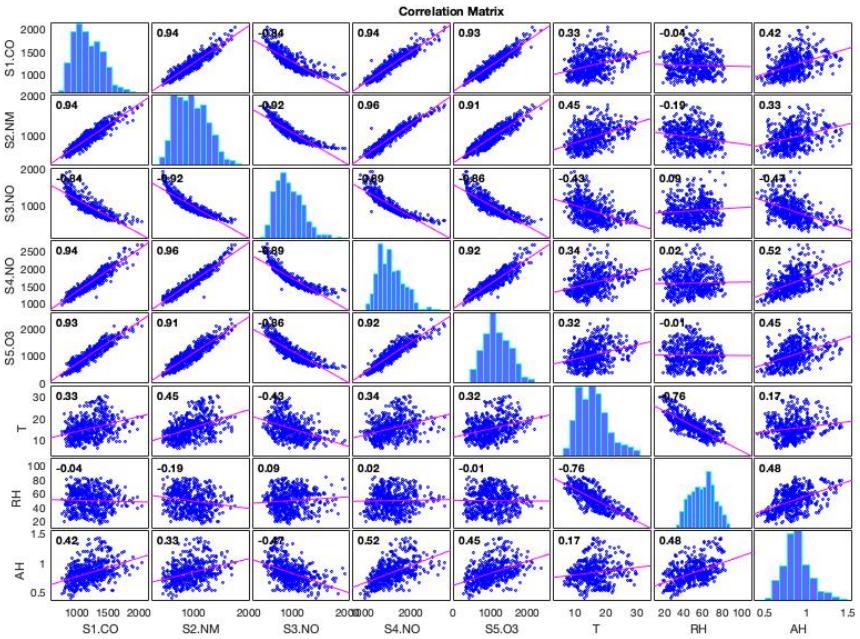


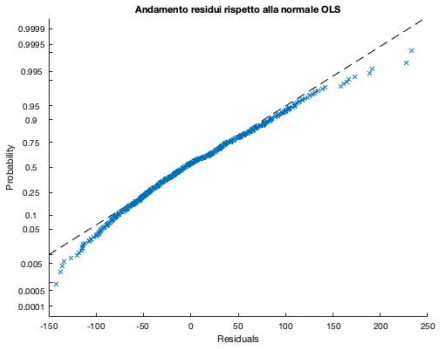
Figure 2: correlazione tra sensori

Dopo aver stimato il modello utilizzando il training set si è deciso di eliminare le covariate che non risultassero significative, ottenendo così il seguente modello:

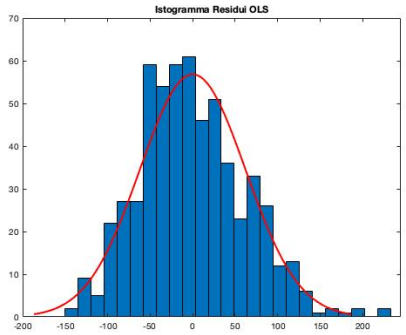
```
modelSensori =


Linear regression model:
    S1.CO ~ 1 + S2.NMHC + S3.NOx + S4.NO2 + S5.O3

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    _______    __________

    (Intercept)     -52.609      50.187    -1.0483       0.29496
    S2.NMHC         0.39652    0.039251     10.102    3.3965e-22
    S3.NOx           0.2035    0.024438     8.3272    6.0947e-16
    S4.NO2          0.24974    0.031975     7.8105    2.7238e-14
    S5.O3           0.26874    0.017026     15.784    7.0342e-47


Number of observations: 579, Error degrees of freedom: 574
Root Mean Squared Error: 63
R-squared: 0.931,  Adjusted R-Squared: 0.931
F-statistic vs. constant model: 1.94e+03, p-value = 0
```

Figure 3: training set sensori

Fatto ciò è stato possibile procedere con l'analisi dei residui verificando che siano *iid* normalmente distribuiti



(a) andamento normale dei residui



(b) istogramma dei residui

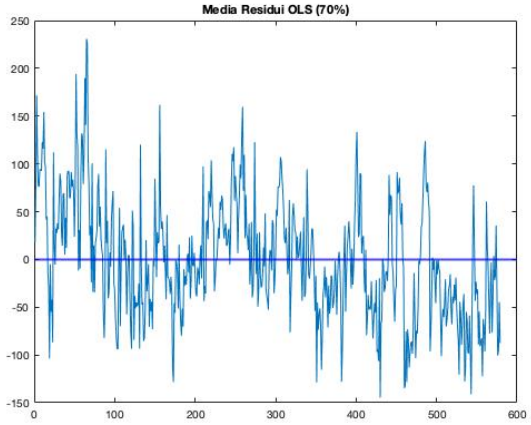


Figure 4: media dei residui

Conclusa l'analisi del modello stimato si è deciso di inserire i dati presenti nel "test set", verificando che il modello si adatti bene anche alla restante parte dei dati del dataset totale. Ciò consente di affermare che il modello non insegue gli errori e che quindi si comporta bene anche in fase di validazione

```
modelValidazione =


Linear regression model:
    S1.CO ~ 1 + S2.NMHC + S3.NOx + S4.NO2 + S5.O3

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    _______    __________

    (Intercept)    -142.66       69.041    -2.0663      0.039864
    S2.NMHC        0.30822     0.060298     5.1116    6.4711e-07
    S3.NOx         0.20317     0.034375     5.9105    1.1475e-08
    S4.NO2         0.39588     0.051834     7.6374    5.0794e-13
    S5.O3          0.21453     0.026588     8.0684    3.2743e-14


Number of observations: 248, Error degrees of freedom: 243
Root Mean Squared Error: 59.6
R-squared: 0.943,  Adjusted R-Squared: 0.942
F-statistic vs. constant model: 1.01e+03, p-value = 3.41e-150
```

Figure 5: test set sensori

Avendo riscontrato eteroschedasticità all'interno del modello si è deciso di procedere utilizzando il metodo GLS, per fare in modo che nella stima dei beta si tenesse conto della diversa varianza dei residui.

```
modelTotale =


Linear regression model:
    S1.CO ~ 1 + S2.NMHC + S3.NOx + S4.NO2 + S5.O3

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    _______    __________

    (Intercept)    -78.833       40.621    -1.9407      0.052638
    S2.NMHC        0.37654     0.032843     11.465    2.4694e-28
    S3.NOx         0.20492     0.019938     10.278    2.1539e-23
    S4.NO2         0.28676     0.027159     10.559    1.5877e-24
    S5.O3          0.25486     0.014328     17.788    3.9353e-60


Number of observations: 827, Error degrees of freedom: 822
Root Mean Squared Error: 62
R-squared: 0.935,  Adjusted R-Squared: 0.934
F-statistic vs. constant model: 2.93e+03, p-value = 0
```

Figure 6: dataset completo sensori OLS

```
modelGLS =


Linear regression model (robust fit):
    S1.CO ~ 1 + S2.NMHC + S3.NOx + S4.NO2 + S5.O3

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    _______    __________

    (Intercept)    -67.269       41.525    -1.6199      0.10563
    S2.NMHC        0.36754     0.033573     10.947    3.9364e-26
    S3.NOx         0.19298     0.020382     9.4681    2.9325e-20
    S4.NO2          0.2975     0.027764     10.715    3.6191e-25
    S5.O3           0.2443     0.014646      16.68    5.0876e-54


Number of observations: 827, Error degrees of freedom: 822
Root Mean Squared Error: 63.4
R-squared: 0.932,  Adjusted R-Squared: 0.931
F-statistic vs. constant model: 2.8e+03, p-value = 0
```

Figure 7: dataset completo sensori GLS

Nella stima del modello Ground Truth si è deciso di utilizzare il regressore *S5.O3*, anche se non vi è il corrispettivo GT presente nel dataset, a causa dell'enorme significatività riscontrata nei modelli precedentemente stimati. Fatto ciò, però, si può osservare che in questo caso la covariata non ha la stessa importanza di prima, il che consente di eliminarla definitivamente con giustificato motivo.

```
modelGT =


Linear regression model:
    CO ~ 1 + NMHC + NOx + NO2 + S5.O3

Estimated Coefficients:
                    Estimate        SE         tStat       pValue
                    ________    _________    _______    __________

    (Intercept)      -0.22306     0.046552    -4.7915     1.9641e-06
    NMHC            0.0021622     9.19e-05     23.528     5.4913e-94
    NOx              0.009245   0.00036372     25.418    1.2682e-105
    NO2             0.0051943   0.00072189     7.1955      1.403e-12
    S5.O3          0.00021961   6.5355e-05     3.3602     0.00081469


Number of observations: 827, Error degrees of freedom: 822
Root Mean Squared Error: 0.318
R-squared: 0.949,  Adjusted R-Squared: 0.949
F-statistic vs. constant model: 3.85e+03, p-value = 0
```

Figure 8: dataset completo sensori GLS

1.5 Conclusioni

Come si nota dai due modelli qui sopra illustrati l'utilizzo del metodo *GLS* per stimare i beta non comporta un miglioramento nella statistica del $RMSE_{\text{test}}$, il che evidenzia una sostanziale omoschedasticità dei residui.

Come ci si aspettava il modello di ground truth presenta un grado di precisione molto elevato dovuta all'affidabilità dei valori di Ground truth.

Entrambi i modelli presentano dei PV significativi per le medesime covariate, ma con ordini di significatività differenti.

In seguito sono riportate le covariate dei due modelli in ordine descrescente di significatività:

Significativà covariate dei due modelli	
Sensori	Ground truth
S5.O3	*NOx*
S2.NMHC	*NMHC*
S3.NOx	*NO2*

2 Analisi delle serie storiche

2.1 Obiettivi

Il focus di questa seconda parte della relazione è quello di descrivere la serie storica del $S1.CO$ attraverso l'implementazione di modelli per studiare la sua evoluzione storica e di individiare quello che meglio spieghi la variabile in questione.

2.2 Operazioni preliminari

Osservare l'andamento di $S1.CO$ nel tempo:

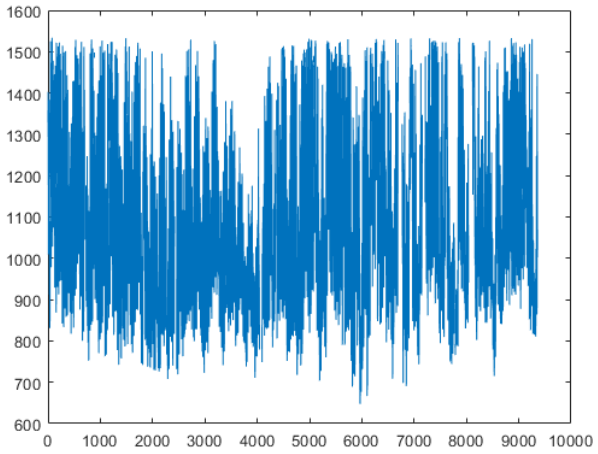


Figure 9: *S1.CO* non stazionario

- **Gestione dei valori mancanti**

  Per come era strutturato il dateset le informazioni mancanti erano identificate dal valore -200, perciò si è deciso di sostituire tali valori con la notazione NaN.

- **Outliers**

  Il metodo che si è utilizzato per identificare gli outlier consiste nel verificare che le singole osservazioni non distino dalla media per valori superiori a $2 \cdot \sigma$ e in caso contrario rimuoverle.

$$outliers = (y - \mu) > 2 \cdot \sigma$$

- **Sostituzione valori NaN**

  Le osservazioni che presentano valori NaN sono state sostiuite da una stima ricavata da un algoritmo di interpolazione localmente lineare.

2.3 Stazionarietà della traiettoria

Dopo una prima analisi la traiettoria della variabile di studio è risultata essere **non stazionaria** e si sono applicate le seguenti trasformazioni:

- **Trasformazione logaritmica**

  L'applicazione della funzione logaritmica permette di ridurre la variabilità dei dati nel tempo.

- **Rimozione trend lineare**

  Si è rimosso il trend lineare applicando la differenziazione di primo ordine per poter osservare nel dominio delle frequenze le armoniche più rilevanti.

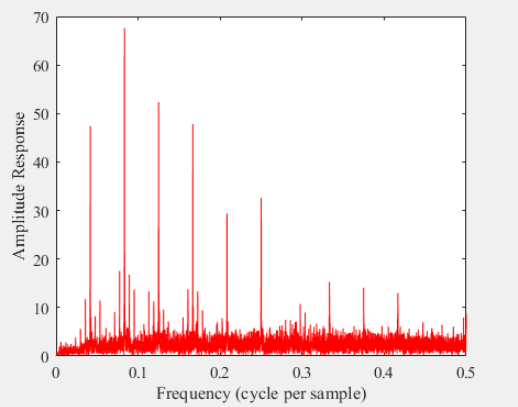


Figure 10: visualizzazione armoniche

- **Rimozione seasonality**

  Tramite l'utilizzo del toolbox TSAF si sono marcati gli spyke più evidenti per stimare il periodo della seasonality.

$$\rightarrow y = \cancel{Trend} + \cancel{Seasonality} + \ random\ fluctuations$$

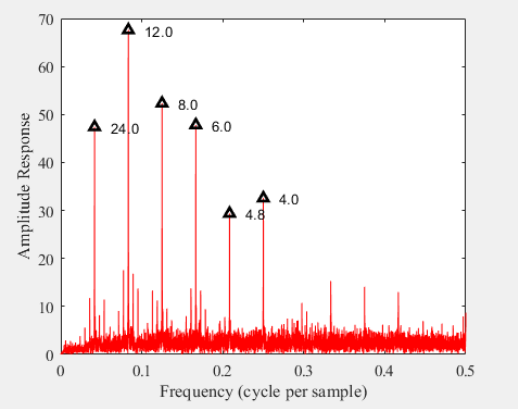
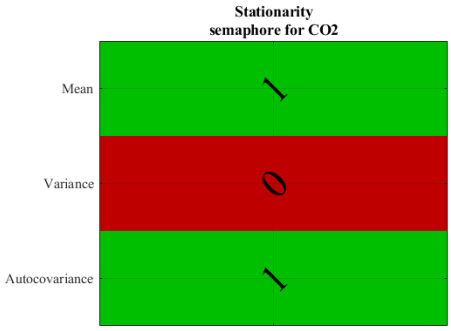


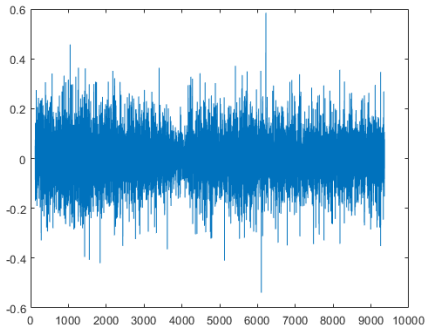
Figure 11: armoniche più rilevanti

Dopo aver eseguito tutte le operazione sopra citate si sono rieseguiti i test per verificare la stazionarietà delle osservazioni e si sono ottenuti i seguenti risultati:

- media costante nel tempo;
- eteroschedasticità;
- covariaza costante nel tempo;



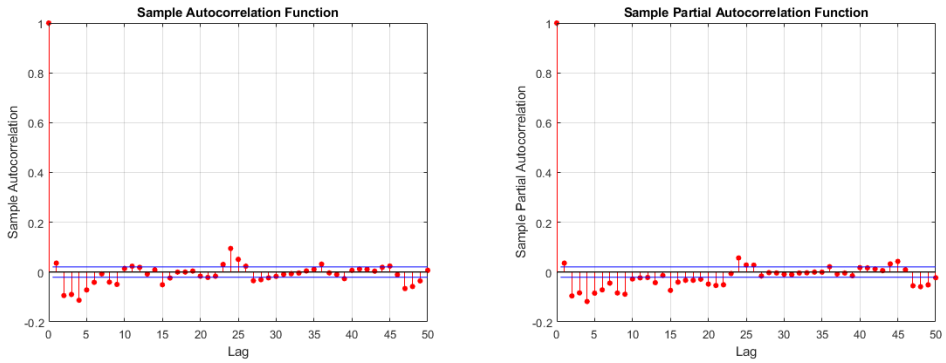
(a) test stazionarietà



(b) $S1.CO$ dopo le operazioni

2.4 Stima modello

La correlazione parziale e l'autocorrelazione del $S1.CO$ risultano essere:



Osservando i grafici si è optato per la stima del modello in maniera iterativa includendo le covariate più significative (*S2.NMHC, S3.NOx, S5.O3*). Nello specifico l'algortimo prevedeva di calcolare iterativamente il modello passando alla funzione i seguenti parametri:

- p = 1:4
- q = 1:10

```
temp_AIC=0;
temp_model=arima('Constant',NaN,'ARLags',1:1,'D',0,'MALags',1:1,'Distribution','Gaussian');
for p = 1:4
    for q = 1:10

        model = arima('Constant',NaN,'ARLags',1:p,'D',0,'MALags',1:q,'Distribution','Gaussian')
        est_model = estimate(model, y_final,'X',X,'Display','params');
        if(temp_AIC> summarize(est_model).AIC)
        temp_model=summarize(est_model);
        end
        model_matrix(p,q) = est_model;
    end
end
```

Figure 12: algoritmo iterativo

Il modello che si è reputato essere più adatto alla nostra analisi è risultato essere:

ARIMAX(3,0,6) Model (Gaussian Distribution)

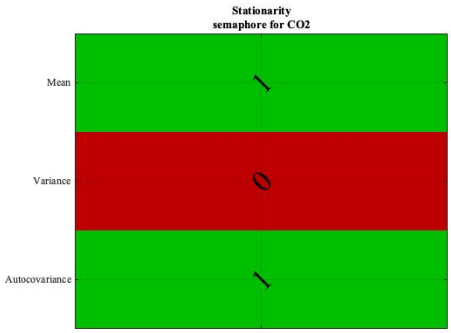
Effective Sample Size: 9236
Number of Estimated Parameters: 14
LogLikelihood: 9867.06
AIC: -19706.1
BIC: -19606.3

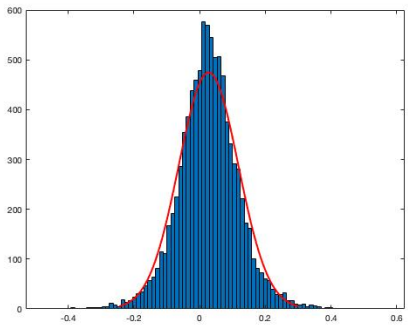
	Value	StandardError	TStatistic	PValue
Constant	-0.097238	0.015339	-6.3393	2.308e-10
AR{1}	-1.2208	0.019619	-62.225	0
AR{2}	-0.99012	0.025354	-39.052	0
AR{3}	-0.46739	0.020085	-23.27	8.8828e-120
MA{1}	1.3376	0.021586	61.967	0
MA{2}	1.0951	0.030943	35.392	2.2577e-274
MA{3}	0.47055	0.028804	16.336	5.4483e-60
MA{4}	-0.1091	0.019539	-5.5837	2.3541e-08
MA{5}	-0.14001	0.016601	-8.4338	3.3453e-17
MA{6}	-0.074096	0.010254	-7.2257	4.9835e-13
Beta(1)	0.0001026	1.2418e-05	8.2625	1.4262e-16
Beta(2)	-9.7907e-05	1.0251e-05	-9.5511	1.2828e-21
Beta(3)	7.9489e-05	9.6094e-06	8.272	1.3171e-16
Variance	0.0069118	7.9025e-05	87.463	0

Figure 13: modello arma definitivo

2.5 Analisi residui

I residui ottenuti con il modello arma(3,6) sopra descritto denotano una evidente distribuzione normale, come tra l'altro si evince dal test di Jarque-Bera.





Tuttavia è ancora presente eteroschedasticita negli stessi, che non possono quindi definirsi iid nel tempo.

2.6 Conclusioni

Come si può osservare dal grafico seguente il modello in questione si comporta in maniera discreta anche in ottica previsiva:

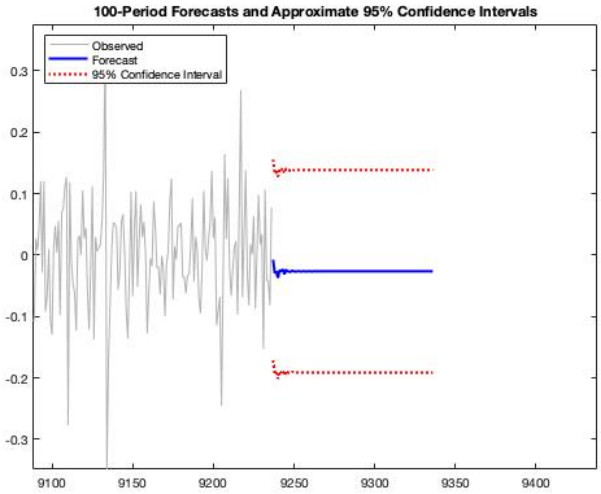


Figure 14: forecast

In conclusione va precisato che il modello scelto non sia il migliore in termini di indici di validazione interna (AIC e BIC), bensì si è deciso di preferire quello che presentasse una significatività di covariate maggiore, a discapito di una leggera variazione dei parametri suddetti.

ARIMAX(4,0,10) Model (Gaussian Distribution):

	Value	StandardError	TStatistic	PValue
Constant	−0.080736	0.013658	−5.9113	3.3949e−09
AR{1}	−1.386	0.019515	−71.024	0
AR{2}	−1.3543	0.030965	−43.735	0
AR{3}	−0.87963	0.03112	−28.266	8.9965e−176
AR{4}	−0.35964	0.019927	−18.048	8.2346e−73
MA{1}	1.4877	0.02124	70.04	0
MA{2}	1.4283	0.035941	39.739	0
MA{3}	0.8175	0.038564	21.199	9.8284e−100
MA{4}	0.10514	0.029888	3.5178	0.00043513
MA{5}	−0.37398	0.024931	−15	7.3157e−51
MA{6}	−0.37262	0.026121	−14.265	3.6007e−46
MA{7}	−0.24423	0.026241	−9.3073	1.3118e−20
MA{8}	−0.12435	0.024203	−5.138	2.7772e−07
MA{9}	−0.045223	0.018259	−2.4768	0.013258
MA{10}	−0.0083118	0.010099	−0.82306	0.41048
Beta(1)	7.2626e−05	1.1691e−05	6.2123	5.2216e−10
Beta(2)	−0.00010748	9.0689e−06	−11.852	2.1073e−32
Beta(3)	9.85e−05	9.1754e−06	10.735	6.9511e−27
Variance	0.0066051	7.4592e−05	88.549	0

ARIMAX(3,0,6) Model (Gaussian Distribution)

Effective Sample Size: 9236
Number of Estimated Parameters: 14
LogLikelihood: 9867.06
AIC: −19706.1
BIC: −19606.3

	Value	StandardError	TStatistic	PValue
Constant	−0.097238	0.015339	−6.3393	2.308e−10
AR{1}	−1.2208	0.019619	−62.225	0
AR{2}	−0.99012	0.025354	−39.052	0
AR{3}	−0.46739	0.020085	−23.27	8.8828e−120
MA{1}	1.3376	0.021586	61.967	0
MA{2}	1.0951	0.030943	35.392	2.2577e−274
MA{3}	0.47055	0.028804	16.336	5.4483e−60
MA{4}	−0.1091	0.019539	−5.5837	2.3541e−08
MA{5}	−0.14001	0.016601	−8.4338	3.3453e−17
MA{6}	−0.074096	0.010254	−7.2257	4.9835e−13
Beta(1)	0.0001026	1.2418e−05	8.2625	1.4262e−16
Beta(2)	−9.7907e−05	1.0251e−05	−9.5511	1.2828e−21
Beta(3)	7.9489e−05	9.6094e−06	8.272	1.3171e−16
Variance	0.0069118	7.9025e−05	87.463	0