

U.S. International Air Traffic - Final project

מקצע – סדרות עתיות 00960425

אלעד נחליאלי – 319000725

מטר רבץ – 207036211



הקדמה :

במסגרת עבודה זו בחרנו לנתח נתונים בתחום התחבורה האוירית הבינלאומית, המתמקדים בטיסות בין שדות תעופה בארצות הברית לשדות תעופה מחוץ לארה"ב. התחום מהו מרכיב מרכזי בעולמות הכלכליות הגלובליות, וניתוח הנתונים עשוי לשקף מגמות בתעבורה אוירית, השפעות עונתיות, וכן תנועת נוסעים ומטען בין מדינות.

חלק 1: הנתונים שנבחרו

הנתונים שנבחרו לעבודה מגיעים מטור דוח הסטטיסטי של נסעים ומטען בינלאומיים באוויר של ארצות הברית, חלק מתוכנית ממשתנית לאיסוף נתונים תעבורה מפורטים מחברות תעופה אמריקאיות ובינלאומיות.

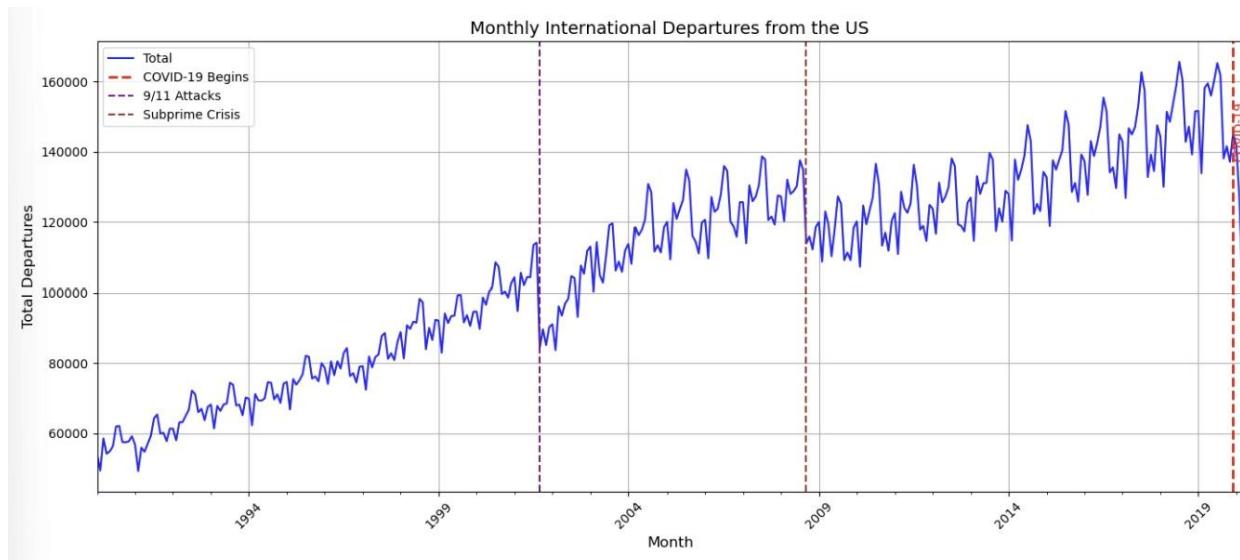
בחרנו להתמקד בקטgoriyת המראות, הכוללת נתונים על כל הטיסות בין שערי כניסה אמריקאים לשערים מחוץ לארה"ב, ללא קשר למוצא או ליעד הסופי של הטיסה. כל רשומה בקובץ מייצגת חברות תעופה מסויימת וטיסה בין זוג שדות תעופה – אחד בתחום ארה"ב ואחד מחוצה לה.

*אפי' הנתונים הוא יומי, כאשר כל רשומה מייצגת טיסה בין זוג שדות תעופה מסוים, בתאריך מסוים.

ישן 930808 רשומות החל מינואר 1990 עד מרץ 2020 כאשר ביצעו ארגזיה חודשית (סכמו את כל הטיסות שבוצעו באותו החודש). לאחר הארגזיה קיבלנו 363 רשומות והשารנו רק את העמודות של החדשים בשנה וכמות הטיסות.

הסיבה לביצוע הארגזיה היא לצורך ייעול העבודה עם הנתונים. ארגזיה מסייעת לצמצם את נפח הנתונים ולפשט את מבנה הטבלה, כך שניתן לבצע ניתוחים והסקת מסקנות בצורה נוחה, ברורה ויעילה יותר. בנוסף, היא תורמת לשיפור ביצועים ולהפחית העומס החישובי, במיוחד כאשר עובדים עם כמויות גדולות של מידע.

ויזואליזציה של הנתונים (עם סימון של תחילת הקורונה, משבר הסaab פריים ואסון התאומים):



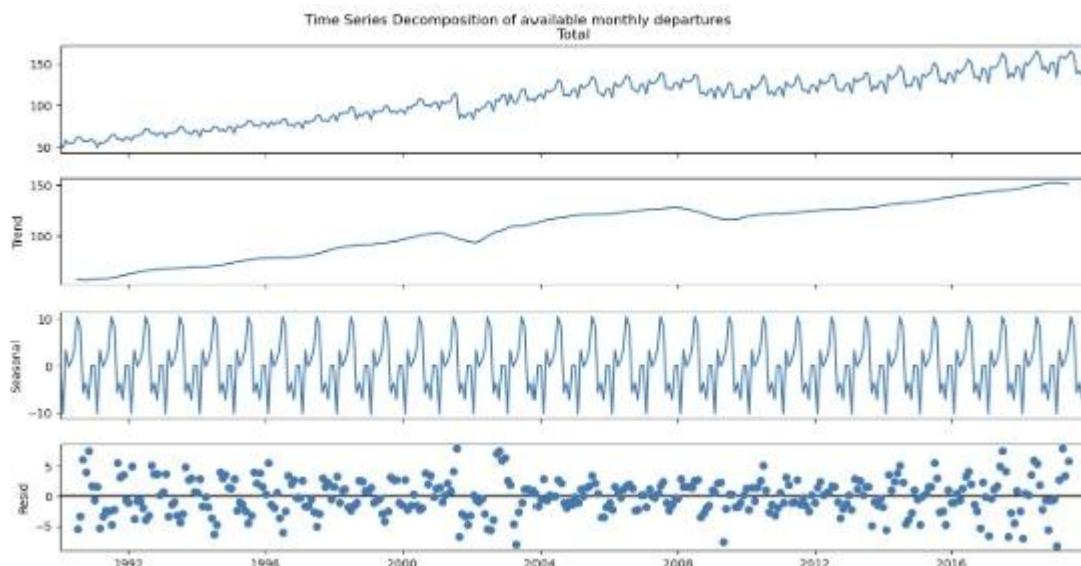
כדי לשפר את היציבות והפרשנות של המודל, חילקו את מספר הטיסות היוצאות החודשיות ב 1,000 (בתחילת pre-processing מצאנו כי מספר הטיסות המינימלי הוא 49264 ומספר הטיסות המקסימלי הוא 165616 ואלו כפונן מספרים מאד גדולים ולכן חשבנו שמתאים לחלק ב1000). עבדה עם ערכים גבוהים במיוחד עלולה להכניס חוסר יציבות נורנית או להשפיע על הרגשות של המודל (לא מוצג כאן אך נעשה על הנתונים בהמשך). בנוסף, הסרנו את כל נתונים הנזקים לאחר דצמבר 2019, מאחר שמגפת הקורונה שיבשה באופן ממשוני את דפוסי הנטיות האויריות. הכללת התקופה הזאת יוצרת שינוי חד בוגמה ללא נתונים על התאוששות לאחר מכן, מה שהיא מקשה על המודלים ללמידה דפוסים משמעותיים ולבצע חיזויים בצורה מדויקת.

ויזואלית ניתן לראות שיש מגמת עלייה (טרנד עולה) ונitinן לראות כי בספטמבר 2001 ובספטמבר 2008 יש ירידת משמעותית במספר הטיסות שנכראה בעקבות מיגעון מגדי התאומים וממשבר ה"סאב פרימס" בארץ הארץ בהתאם. (תוצאות אלה נצפו גם בהמשך חלק 4 של העבודה ומעבר לזה, אף מצאו באותו החלק אונומליה נוספת ונטו נתיחה לכך בהמשך).

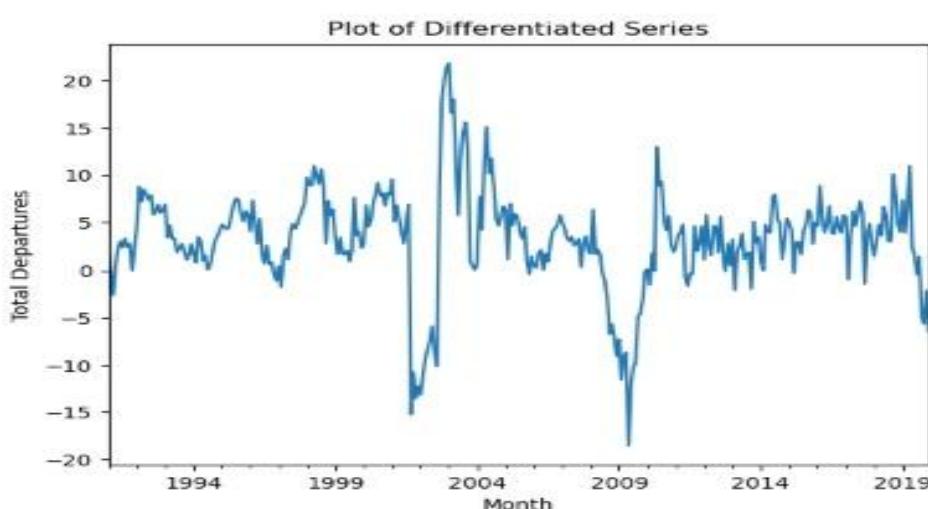
מעבר לכך, בדקנו קיום של ערכים חסרים, והאם קיימים חודשיים ללא טיסות ונמצא כי אין ערכים חסרים ואין חדש ללא נתונים על טיסות.

מגמה, עונתיות ורעש:

אלו גרפים של התפלגות הנתונים, מגמה, עונתיות ורעש.



תרשים של הסדרה לאחר הסרת העונתיות (לאחר גזירה- כאשר $D=12$ = קלומר עונתיות שנתית):



שאלות נייחיות שניתנו לשאול:

- מתי התרחשו נקודות חריגות בתחילת, והאם ניתן לקשר אותן לאירועים בעולם?
- האם ישנו אירוע גלובלי (change point) שלא ניתן להזהות ויזואלית שהופיע על הסדרה (על כמה הטיסות?).
- איזה מודל מתאר בצורה טובת את הנתונים שלנו.

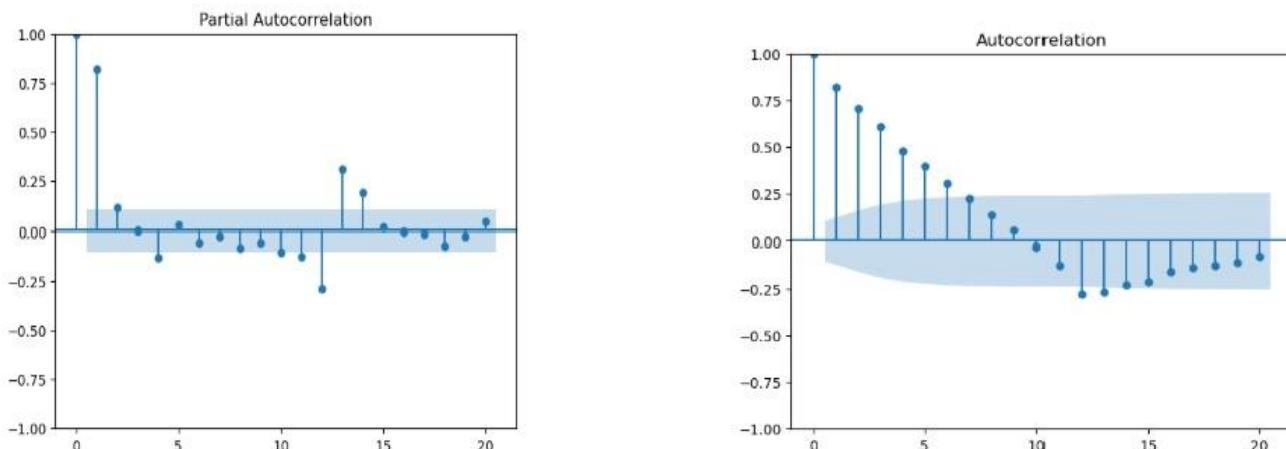
*בפועל קובץ הנתונים שהוגש נערך כדי לשלב את הסדרה המקורי והאקסוגנית ולהן עבר עריכה מבחינת הסרט העמודות הלא רלוונטיות והאגרגציה החודשית שתוארה.

חלק 2 - מתודולוגיה והתאמת מודלים:

בחרנו לבחון את המודלים : סרימה, פרופט, החלקה אקספוננציאלית (holt winters) ורגסיה לינארית עם טור פוריה.

מודל סרימה:

בחירה הפרמטרים: סרטטנו את הגראפים של ACF, PACF של הסדרה לאחר גזירה:



בהתבסס על גראפי ה- ACF (Autocorrelation) ו- PACF (Partial Autocorrelation), ניתן להסיק את המסקנות הבאות לגבי מבנה המודל:

गראף ה- ACF מראה דעיכה הדרגתית בצורה גיאומטרית, תופעה שמאפיינת תחילת של AR (Auto-Regressive) מסדר ראשון, קלומר AR(1).

גראף ה- PACF מציג פיק מובהק בລג הראשון ושני, ולאחר מכן נחתר בחודות – זה תומך במודל AR(2) או SARIMA(2,0,0).

בנוסף, נראה פיק סביר לג 12, המעידים על עונתיות שנתית (Seasonality) – ולכן יש טעם לבחון מודלים עם מרכיב עונתי.

במונחים של מודלים עונתיים, נראה שמתאים המודלים:

SARIMA(2,0,0)(0,1,0)_12

SARIMA(1,0,0)(1,1,0)_12

SARIMA(2,0,0)(1,1,0)_12

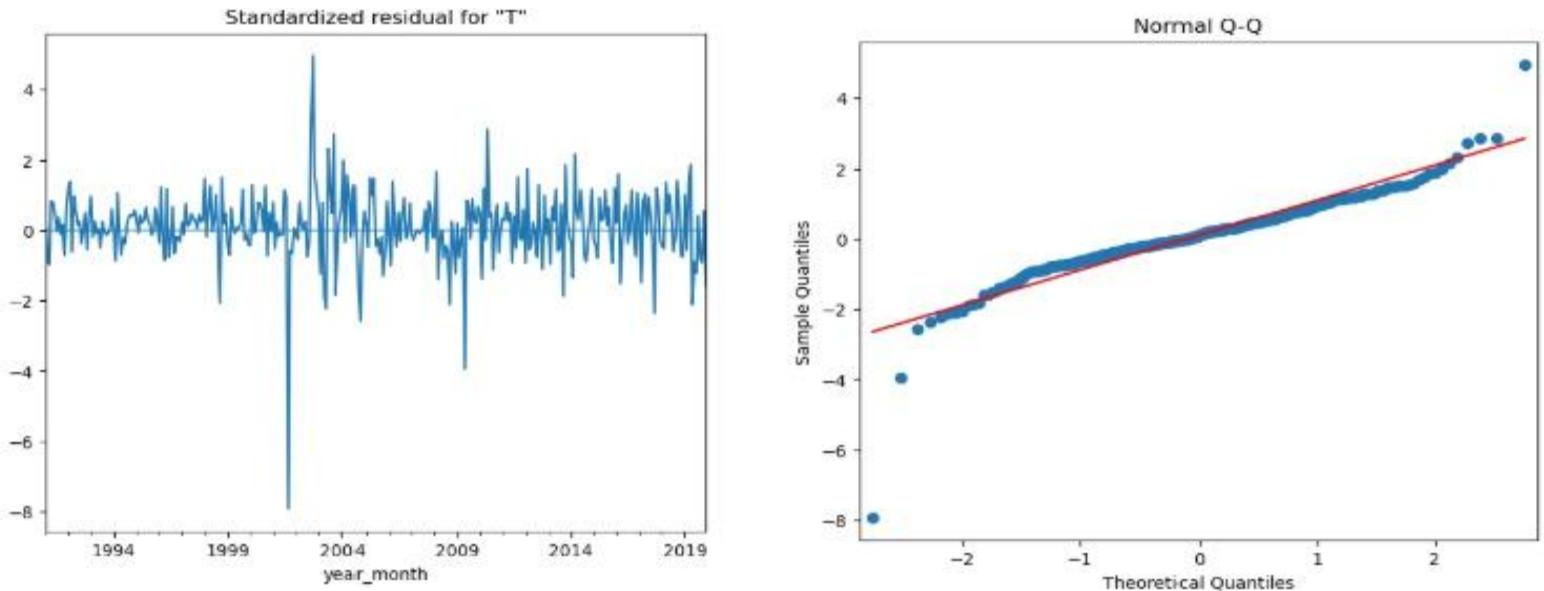
כלומר, כדי לבדוק את השילובים של $1 = k \text{ או } 2$, יחד עם רכיב עוני בעזרת $D=12=s$, כאשר $Q=0$ או 1 .

את בחירת הפרמטרים ביצעו לפי קרייטריון BIC כאשר מצאו שלמודל SARIMA(2,0,0)(1,1,0)_12 יש את הערך היכי נמוך (1695) ולכן בחרנו אותו כמתאים ביותר לתאר את הנתונים שלנו מבין מודלי הסרימה שבדקנו.

מעבר לכך, הצגנו את גרפי השאריות של כל המודלים והם היו מאוד דומים, וכך החלטנו להסתפק בקריטריון ה CIC – 2.1.4 – 2.1.1 – 2.1.4.

מעבר לכך, הרצינו דיאגנוזטיקה על המודל,

להלן גרפ' השאריות והplot-QQ כתוצאה מהרצת הדיאגנוטיקה על המודל סרימה הנבחר.



:Q-Q Plot .1

ניתן לראות שהשאריות נצמדות יחסית טוב לגו האדם, שהוא קין הנורמליות התאורטית, בעיקר בחלק המרכזי של ההתפלגות.

עם זאת, יש חריגות בקצוות - שאריות קייזוניות, במיוחד מצד השמאלי התיכון.

המשמעות: באופן כללי, התפלגות השאריות קרובת לנורמלית, אבל קיימים מספר ערכים קיצוניים (שכנראה תואמים לאיורים חריגים כמו 11/9 או המשבר הכלכלי של 2008).

:Standardized Residuals Plot .2

ניתן לראות שהshares מתחנגוות באופן ייחודי יציב לאורר רוב התקופה.

ישנו קפיצות קיצונית במינוח בסביבות השנים 2001–2003 ו-2008, שנראתה תואמת לאיורים כלכליים/גיאופוליטיים משמעותיים.

בוך הכל נראה שמודל סרימה טוב את המבנה הכללי של הנתונים.

מודל פרופט:

ניסינו להתאים את מודל פרופט לנ נתונים שלנו, אך מצאנו שהוא אינו התאמה טובה - בהשוואה למודל סרימה שעליו דנו קודם, ניתן לראות בגרפים המוצגים במחברת בחלק 2 (גרפים 2.3.1, 2.3.3) שהוא טוב יותר מאשר טוב את המבנה של הנתונים שלנו בהתאם לשגשוג של גרפ השאריות והQQ-Q. מעבר לכך, כשביצענו השוואה בין המודלים, פרופט קיבל תוצאות מאוד טובות על סט האימון ותוצאות גורועות ביחס לשאר המודלים שנבדקו על סט המבחן וכנראה ביצע overfitting.

התוצאות של פרופט על סט האימון:

Prophet Performance on Training Set:
MSE: 15.74
MAPE: 3.25%

ועל סט המבחן, קיבלנו MSE של 271.579 ו-MAPE של 9.99%.

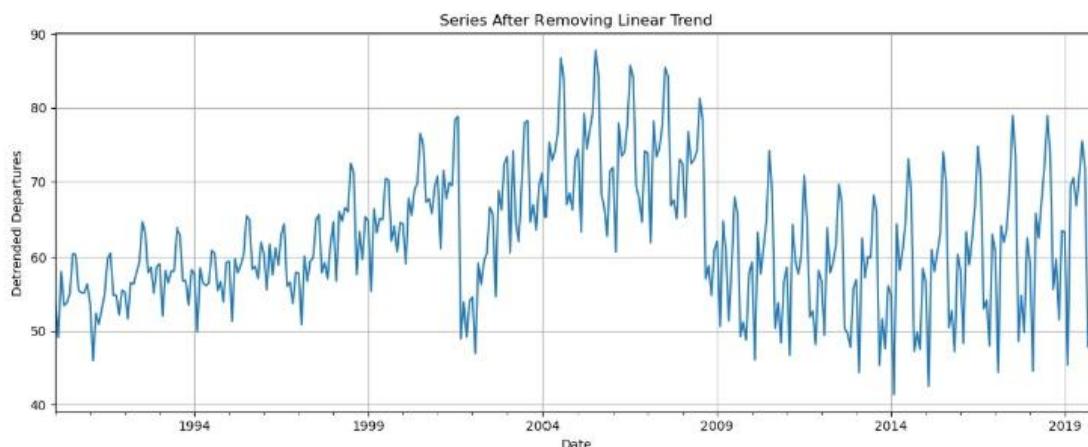
יתכן ונitin לננות את ערכי היפר הפרמטר של המודל שאחראי על הרגולרייזציה של הטרנד כדי לקבל התאמה טובה יותר, אך החלהנו להתמקד במודלים האחרים המתוארים, שהניבו תוצאות טובות יותר.

החלק אקספוננציאלית:

בחרכנו להשתמש בהחלקה אקספוננציאלית מאחר שמדובר במודל פשוט, שמתאים היטב לנ נתונים הכלולים מגמה ועונתיות יחסית יציבה – כפי שנצפה בנ נתונים הטיסות (יותר טיסות בקיץ ופחות בחורף, ועליה כללית לאורך השנים). המודל מאפשר תחזית טובה לטווח קצר-בינוני כי שם דגש על הנתונים האחרונים ביחס לעבר.

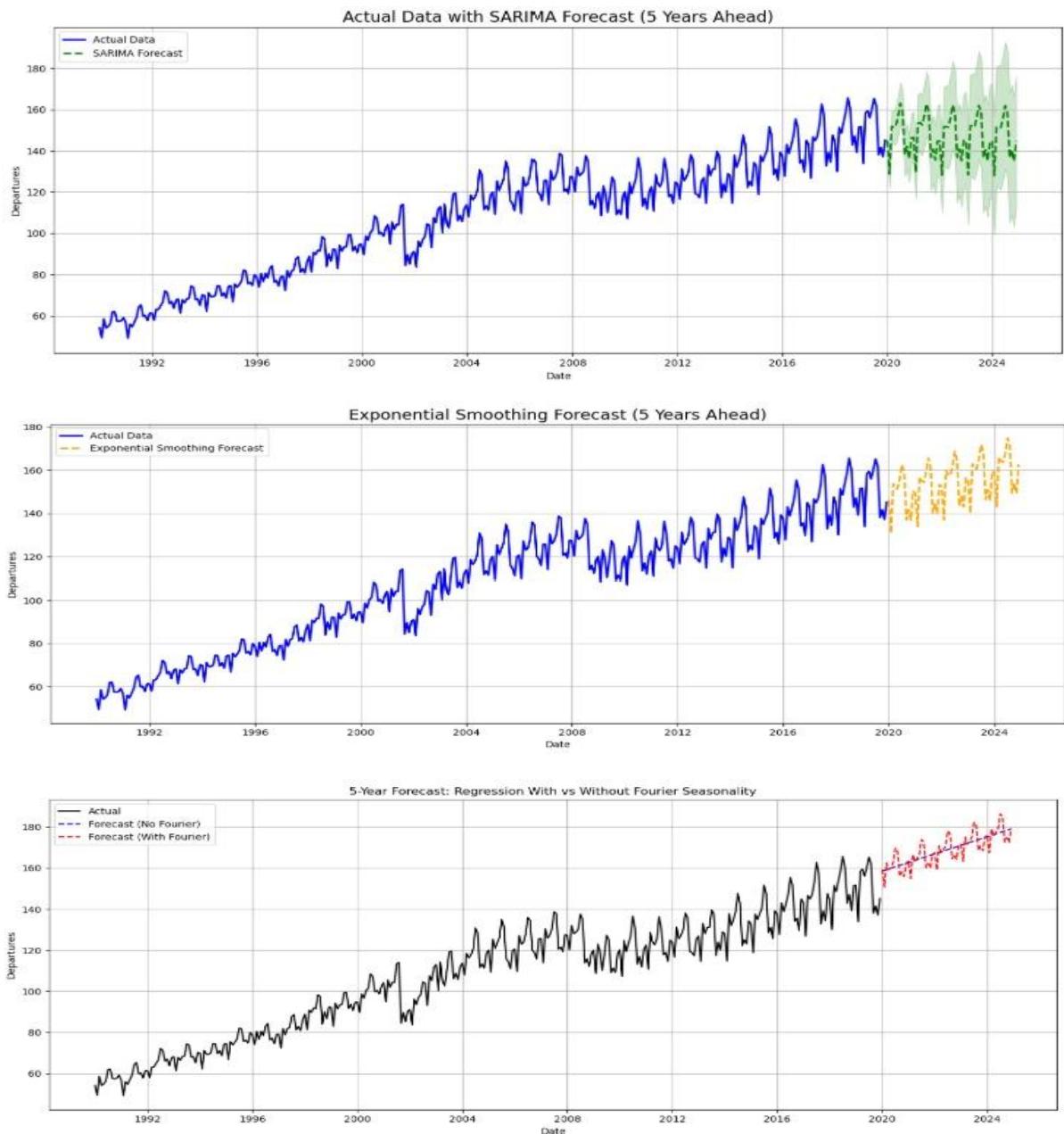
**רגסיה עם טורי פוריה:

בהתחלת חשבונו להשתמש בקירוב עם טורי פוריה בלבד ללא רגסיה, אבל לאחר הסורה של הטרנד מהסדרה ראיינו כי עדין קיימת מגמה והטרנד לא מסור לגמרי. זה נראה נובע מהה�� change敏锐 המגמה בנתונים. להלן גרפ של הסדרה לאחר הסורה הטרנד:



כדי להתמודד עם שינוי המגמה בחרנו לשלב את הרגרסיה עם טורי פוריה. השתמשנו בעונתיות שנתית ($D = 12$) כאשר לקחנו בחשבון את שני שינוי המגמה שעיליהם דיברנו בחלק המבוא ואוותם ראיינו בצורה ויזואלית (אsoon התואמים וה"סאב פרימ").

כעת נציג את החיזויים של המודלים – 5 שנים קדימה:



נראה שמלבד מודל סרימה, כל המודלים ממשיכים את מגמת העליה. לעומת זאת, סרימה חוזה מגמה "ممותנת" ביחס למגמה הקודמת.

הערכת המודלים: ביצענו חלוקה לסת אימון – 80% וסת מבחן – 20%.

החלוקת מוצגת בתרשימים 2.8 במחברת.

(סת האימון כלל 288 תצפיות וסת המבחן כלל 72 תצפיות).

כאשר סט המבחן היווה את התצפיות האחרונות בנתונים.

השתמשנו בשני מדדים: MSE ו-MAPE.

ה-MSE מוגיב לשגיאות גדולות בצורה חזקה יותר בשל הריבוע של השאריות, מה שהופך אותו לשימושי במילויים בziehungי מודלים שנוטים לבצע טעויות גדולות.

מצד שני, MAPE מבטא את השגיאות בהתאם להערכת האמיטיות. מכיוון שבמערך הנתונים שלנו יש ערכים גדולים ולא ערכים אפסיים, ניתן להשתמש ב-MAPE באופן בטוח והוא נותן ממד אמין למדד היחס של התוצאות מהותזאה בפועל (ולא משתנה בעקבות חלוקה ב-1000 את מספר הטיסות)

אלו התוצאות בפועל של המודלים לסת המבחן:

Combined Model Performance:

Model	MSE	MAPE (%)
Exponential Smoothing	19.694172	2.396369
Linear (Fourier + AR(1))	22.277165	2.676541
Linear (with Fourier)	38.052373	3.522995
Linear (no Fourier)	101.741288	5.819154
Sarima	189.175099	8.052918
Prophet	271.579794	9.998873

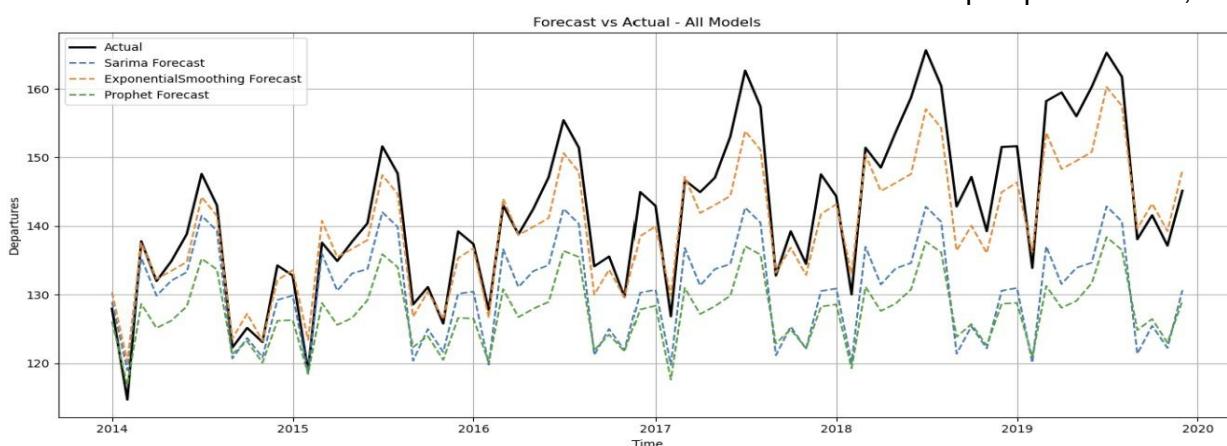
כמו שניתן לראות, החלקה אקספוננציאלית השיגה את התוצאות הטובות ביותר ואחריה וגרסיה עם טורי פוריה עם פיצר נוסף של ar1.

כפי שנאמר קודם, המודל שהציג את התוצאות הנמוכות ביותר הוא מודל פרופט.

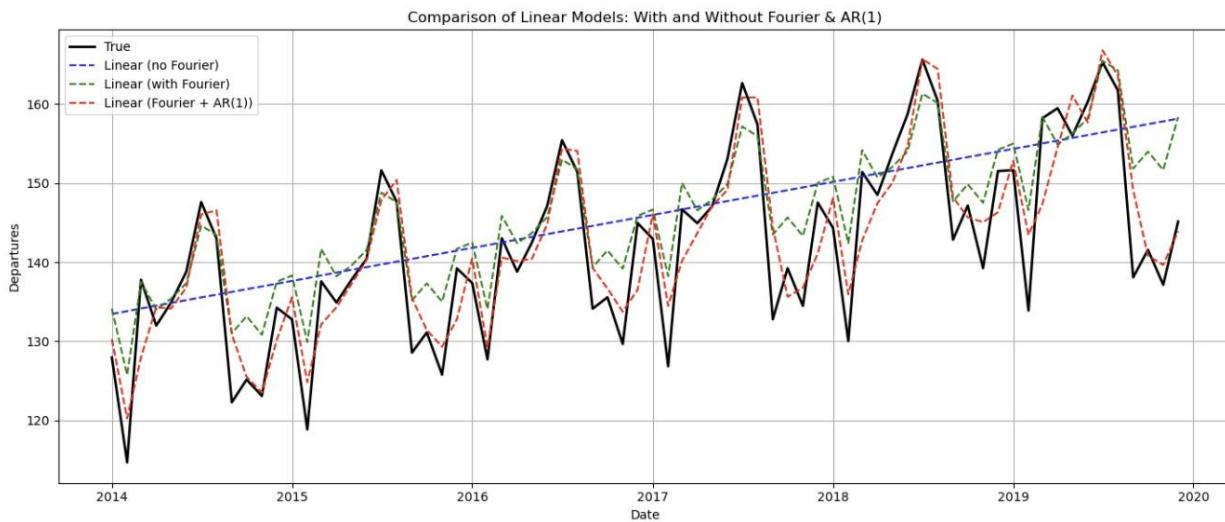
להלן גרפים של תוצאות על סט המבחן:

הערה: כדי לא להעמיס הכל על גרף אחד, חילקנו את זה למודלים ללא הרגשות בנפרד.

סריימה, פרופט והחלקה אקספוננציאלית:



רגRESSED, REGRESSION WITH FOURIER AND REGRESSION WITH FOURIER AND AR(1):



ניתן לראות שמודל הרגRESSION הינו יחסית עבודה טובה בחזויים.

*בפועל התאמנו גם מודל רגRESSION לינארית פשוט לשם השוואה בחלק של מודל הרגRESSION.

חלק 3: משתנים אקסוגניים:

בחרנו את מדדי ה-*I*-CPI עבור אנרגיה, תחבורה ודלק (בנ"ז) מכיוון שהם קשורים ישירות לגורמים המשפיעים על

פעילות הטיסות.

עלויות האנרגיה והדלק משפיעות על הוצאות התפעול של חברות התעופה, מחירי הדלק ועלויות הכספיים. מדד ה-*I*-CPI לתחבורה משקף את הביקוש לתחבורה ציבורית ופרטית.

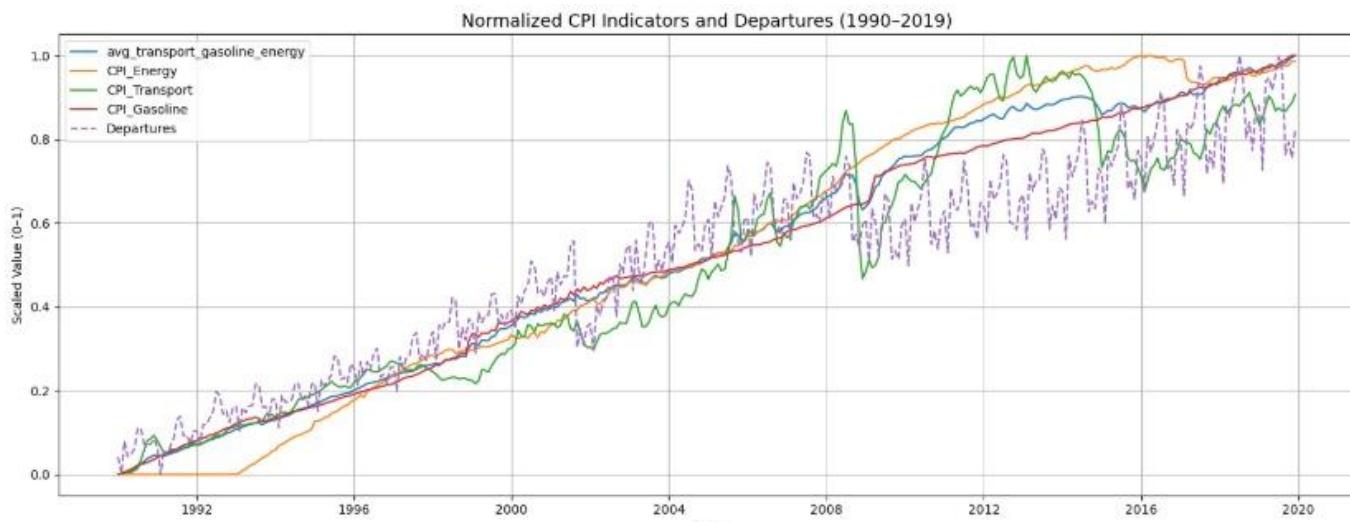
מצאנו שבסדרת ה-*I*-CPI לאנרגיה אין כל נתונים בין השנים 1990 ל-1992, שקלנו שלוש אפשרויות להתחמಡות עם הערכים החסרים:

- **השלמה עם אפסים** – נשלה מכיוון ש-*I*-CPI הוא מדד יחסי, והכנסת אפסים הייתה מעוותת את הסקאללה.
- **ויתור על הסדרה של האנרגיה** – הייתה מביאה לאובדן של משתנה מסויר שעשוי להיות ממשמעותי.
- **השלמת הערכים החסרים על בסיס התצפית הראשונה הדמינה בינואר 1993**

לכן, בחרנו למלא את הערכים החסרים בין השנים 1990 ל-1992 בעזרת הערך של ינואר 1993, מתוך הנחה שהערכים המוקדמים היו יציבים יחסית ודומאים לנקודת הנתון הראשונה הזרמינה ובכך ניסינו לשמור על הרציפות במדד הנתונים תוך מזעור העיוותים האפשריים.

הערה: היינו צריכים לחוץ את המדדים הרלוונטיים מתוך הקובץ של כלל המדדים לצריך. הקובץ המקורי כולל 276,773 רשומות, ולאחר סינון המדדים, ארגזציה וסינון השנים הרלוונטיות (הנתונים היו משנת 1913-2025) כל סדרה בלבד הסדרה של האנרגיה כולו 363 רשומות, ולאחר ההשלמה גם היא כוללת 363 רשומות.

להלן גרפ' של כל הסדרות לאחר סקילינג:



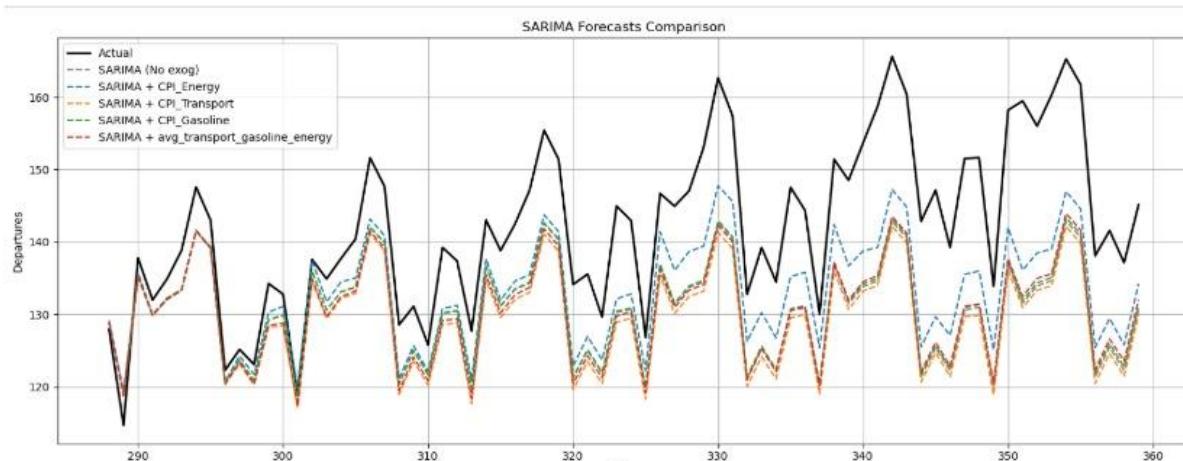
מבחן ויזואלי, נראה שקיים קשר בין הסדרות של מדדי המוצרים, סדרת כמות הטיסות היוצאות ומגמה דומה. שני המודלים שיכולים להשתמש בסדרה אקסוגנית מבין המודלים שהציגו בחלק 3 הם סרימקס ורגסיה עם טורי פוריה (כפי'ר נוסף) – שוב, כאן לא המשכנו עם פרופט.

להלן תוצאות המודלים של סרימקס:

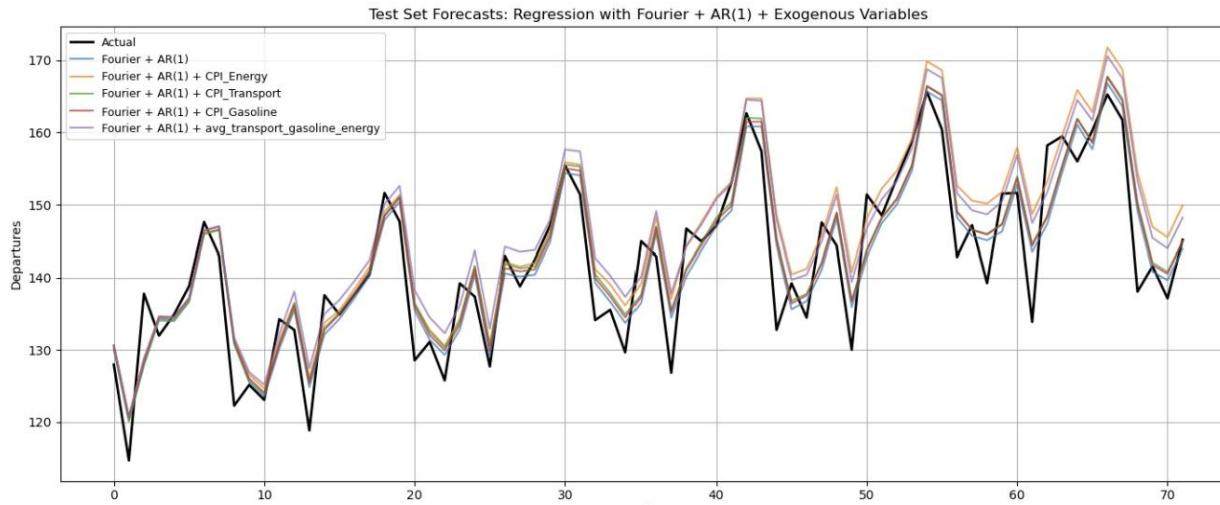
Final comparison including SARIMA without exog:			
Exogenous Variable	MSE	MAPE (%)	
CPI_Energy	114.593500	0.062431	
CPI_Gasoline	182.614331	0.079173	
avg_transport_gasoline_energy	185.475302	0.081478	
SARIMA (no exog)	189.175099	0.080529	
CPI_Transport	213.736219	0.087497	

תחילה התחילנו עם מודל סרימקס, כאשר בדקנו שימוש בכל אחת מהסדרות בנפרד ובממוצע שלהן. ניתן לראות שמודל סרימקס עם סדרת האנרגיה הניב את התוצאות הנמנוכות ביותר (אך צריך לזכור בחשבון שהזיהו המודל שהשתמשנו בהשלמת ערכי חסרים). הבא אחורי זהו סרימקס עם סדרת מדד מחירי הדלק.

להלן גרפ' תוצאות החיזויים של מודל סרימקס על סט המבחן:



להלן גרפ' תוצאות של רגסיה עם פורייה ו-AR(1) עם ובל' המשתנים האוקסיגנים:



תוצאות על סט המבחן של רגסיה עם פורייה ו-AR(1) עם ובל' המשתנים האוקסיגנים:

Model	MSE	MAPE (%)
Fourier + AR(1)	22.277165	2.676541
Fourier + AR(1) + CPI_Gasoline	23.162465	2.733909
Fourier + AR(1) + CPI_Transport	23.734862	2.766375
Fourier + AR(1) + avg_transport_gasoline_energy	34.331349	3.404959
Fourier + AR(1) + CPI_Energy	35.484918	3.342692

נראה שהמשתנים האוקסיגנים אינם מושפרים את החיזויים, אבל עדין מתקבלות תוצאות טובות.

***בפועל קובץ הנתונים שהוגש נערך כדי לשלב את הסדרה המקורית והאקסוגנית ולהן עבר עריכה מבחינת הסרת העמודות הלא רלוונטיות והאגרציה החודשית שתוארה. השלמת הערכים החסרים בוצעה במחברת עצמה.

חלק 4: change point detection

תחילה, השנוו את ערך ה-BIC של כל מודלי סרימה שבדקנו לאורכ' הפרויקט כולל אלו של המשתנים האקסוגניים. המודל שקיבל את ערך ה-BIC הנמוך ביותר היה מודל סרימה ללא המשתנים האקסוגניים, עם אותן הפרמטרים חלק 2 בעובדה. להלן התוצאות:

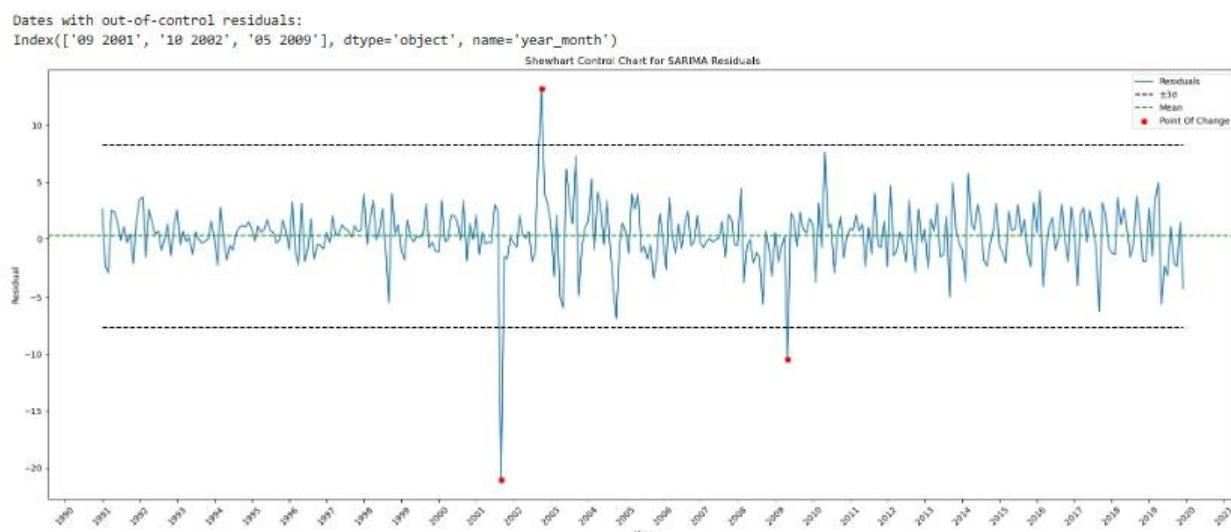
```
SARIMA Model BIC Comparison:
      Model           BIC
4      No Exogenous Variable  1695.621102
1          CPI_Transport    1700.844353
3  avg_transport_gasoline_energy  1700.868276
0          CPI_Energy     1701.378532
2          CPI_Gasoline    1701.460299

Best model by BIC: No Exogenous Variable (BIC = 1695.62)
```

באמצעות תרשימים בקרה מסוג Shewhart שישם על שאריות מודל SARIMA (שכן סרימה איננו מטשטש את האנומליות ושינוי הטרנדים לעומת דוגמאות אחרות עם פוריה שהשם אינו מכני המגמות כפיצרים נוספים), זיהינו שלוש חריגות משמעותיות: ספטמבר 2000, אוקטובר 2001 ומאי 2008. שתי האחרונות תואמות למקרים ידועים – אסון התאומים והמשבר הכלכלי העולמי. לעומת זאת, לקפיצה בספטמבר 2000 לא צפינו הסבר באופן זהה.

את הנקודה זו לא ראיינו בויזואלייזציה הקודמות להציגו, יתכן בשל המגמה והעונתיות החזקות בסדרת הזמן, אשר יתכן וטשטשו את האנומליה.

תרשיים Shewhart עם התאריכים שנמצאו:



סיכום ומסקנות

בעבודה זו ניתחנו נתונים בינלאומיים יוצאים מארצות הברית בין השנים 1990–2019, ובדקנו את יכולת החיזוי של מודלים לסדרות זמן. התחמךנו בבעיות מגמות, עונתיות וחריגות, תוך שימוש במודלים כמו ARIMA, החלקה אקספוננציאלית, Prophet, ורגסיה עם טורי פוריה.

מודל ההחלקה האקספוננציאלית הציג את ביצועי החיזוי הטובים ביותר על סט המבחן, נראהה בזכות התאמת שלו למבנה הנתוניים הכלול בעונתיות יציבה. מודל Prophet נתה לאובייפטינגן, ואילו ARIMA היה יציב אך פחות מדויק. מודל הרגסיה עם פוריה ו-AR(1) סיפק תוצאות טובות והראה ביצועים דומים לאקספוננציאלית.

בדקנו גם השפעה של משתנים אקסוגניים (מדדי CPI) על תחזיות המודלים. השימוש במודלי SARIMAX עם מדדי האנרגיה הראה שיפור משמעותית יותר ביחס לשאר המודלים אך גם במידדים נוספים נראה שיפור קטן. לעומת זאת, במודלי הרגסיה עם משתנים אקסוגניים לא נראה שיפור כלשהו.

הסביר אפשרי לפער בביצועים הוא שמודלי SARIMAX מצילים לנצל את המשתנים האקסוגניים כדי לשפר את החיזוי דווקא בתקופות חריגות כמו אסון התאומים או משבר הסאב-פרריים, שכן שהם מסיעים למודל לחזור לשינויים פתאומיים שאינם מוסברים על ידי עונתיות או מגמה בלבד. לעומת זאת, במודלי הרגסיה עם טורי פוריה, העונתיות מייצגת בצורה גמישה יותר אך כזו שנייה וגישה לאירועים חד-פעמיים, ולכן השפעת המשתנים החיזוניים מטשטשת. מעבר לכך, יתכן שהרגסיה עם פוריה ו-AR(1) כבר סייפה תחזית מדויקת יחסית – כך שלמודל נותרה פחות "הזרמתות" להשתפר בעקבות הוספה משתנים אקסוגניים.

בנוסף, תרשימים Shewhart על שאריות מודל SARIMA סיעו בזיכרון נקודת חריגה שלא נראהיה קודם לכן ביזואלייזיות. שכן, "תכן שהיא "גבלעה" בשל הטרנד ה"חזק" ועונתיות הסדרה. Shewhart ביזואלייזיות כן ראיינו וציינו את משבר ה"סאב פרימ" ואסון התאומים. כמו כן, הם גם ניצפו בתרשימים

בסק הכל, שילוב של חיזוי סדרות זמן עם בדיקות שאריות וניתוח משתנים חיצוניים מאפשר לזרות תבניות משמעותיות בהתקנות סדרת הטיסות ולהשווות בין מודלים שונים באופן שיטתי.

ביבליוגרפיה:

הסדרה המקורית - [U.S. International Air Traffic data\(1990-2020\)](#)

הסדרה האקסוגנית - <https://www.kaggle.com/datasets/paveljurke/u-s-consumer-price-index-cpi>