

U.S. International Air Traffic data - Final project

מקצע – סדרות עתיות 00960425

אלעד נחליAli – 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 אווירית, השפעות עונתיות, וכן תנועת נוסעים ומטען בין מדינות.

חלק 1: הנתונים שנבחרו

הנתונים שנבחרו לעבודה מгиיעים מטור דוח הסטטיסטי של נסעים ומטען בינלאומיים באוויר של ארצות הברית, חלק מתוכנית ממשתית 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 וטישה בין זוג שדות תעופה – אחד בתחום ארה"ב ואחד מחוצה לה.

העמודות המרכזיות בקובץ הנתונים הן:

- **Scheduled** – טיסות סדירות, הפעולות לפי לוח זמנים קבוע.
- **Charter** – טיסות שכר, שאין סדירות ומוגזנות מראש לצרכים ייעודיים.
- **Total** – סך כל הטיסות (סדירות + שכר).

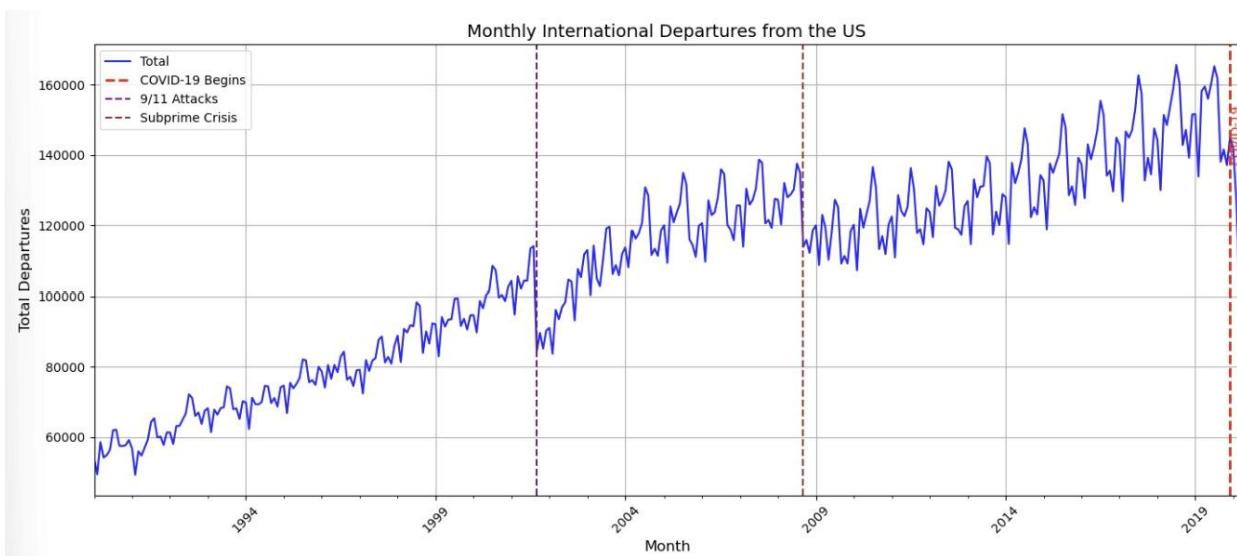
אוף הנתונים הוא יומי, כאשר כל רשומה מייצגת טישה בין זוג שדות תעופה מסוים, בתאריך מסוים.

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

הסיבה לביצוע הארגזיה היא לצורך "על העבודה עם הנתונים".

ארגון מסייעת לצמצם את נפח הנתונים ולפשט את מבנה הטבלה, כך שניתן לבצע ניתוחים והסקת מסקנות בצדקה נוכח, ברורה ויעילה יותר. בנוסף, היא תורמת לשיפור ביצועים ולהפחיתת העומס החישובי, במיוחד כאשר עובדים עם כמות גדלות של מידע.

ויזואлизציה של הנתונים (עם סימון של תחילת הקורונה, משבר הסאב פריים ואסון התאומים):



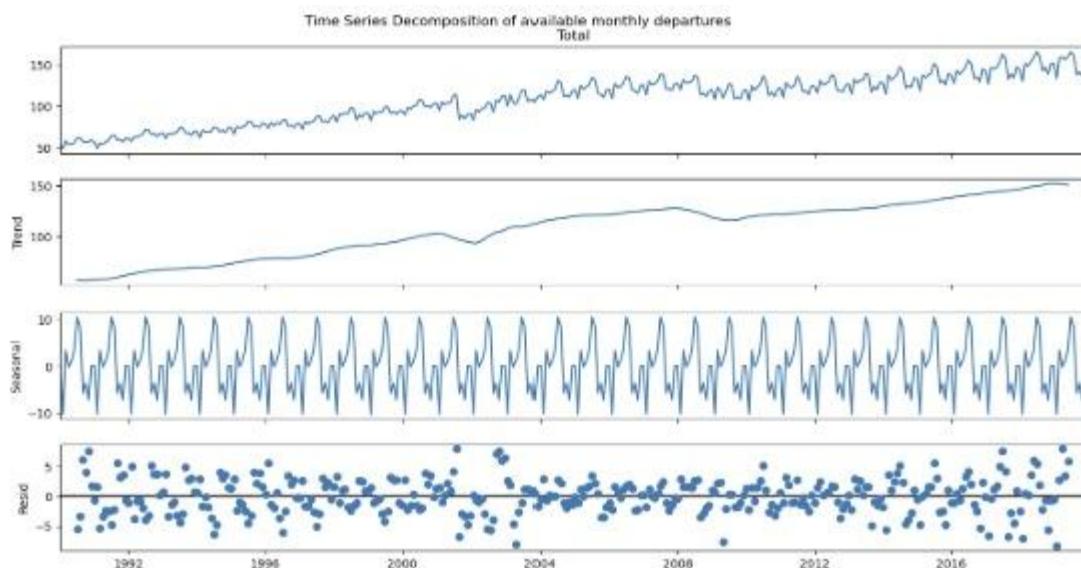
כדי לשפר את היציבות והפרשנות של המודל, חילקנו את מספר הטיסות היוצאות החודשיות ב 1,000 (בתחילה preprocessing מצאנו כי מספר הטיסות המקורי הוא 49264 ומספר הטיסות המקורי הוא 165616 ואלו כמובן מספרים מאד גדולים וכן חשבנו שמתאים לחלק ב1000). עבדה עם ערכים גבוהים במיוחד עלולה להכניס חוסר יציבות נומרית או להשפיע על הרגשות של המודל (לא מוצג כאן אך נעשה על הנתונים בהמשך). בנוסף, היסרנו את כל נקודות הנתונים לאחר דצמבר 2019, מאחר ש兆פת הקורונה שיבשה באופן משמעותי את דפוסי הנסיעות האויריות. הכללת התקופה זו הייתה יוצרת שנייה חד בмагמה ללא נתונים על התאוששות לאחר מכן, מה שהיא מבקשת על המודלים ללמידה דפוסים משמעותיים וביצוע חיזויים בצורה מדויקת.

ויזואלית ניתן לראות שיש מגמת עולה (טרנד עולה) ונitinן לראות כי בספטמבר 2001 ובספטמבר 2008 יש ירידת משמעותית במספר הטיסות שכנהarah נבעה מפגיעה מוגדל התאומות וממשבר ה"סאב פרוי"ם" בארץ הברית בהתאם. (توزאות אלה נצפו גם בהמשך בחלק 4 של העבודה ומעבר לזה, אף מצאו באותו החלק אונומליה נוספת ואני נתיחס לכך בהמשך).

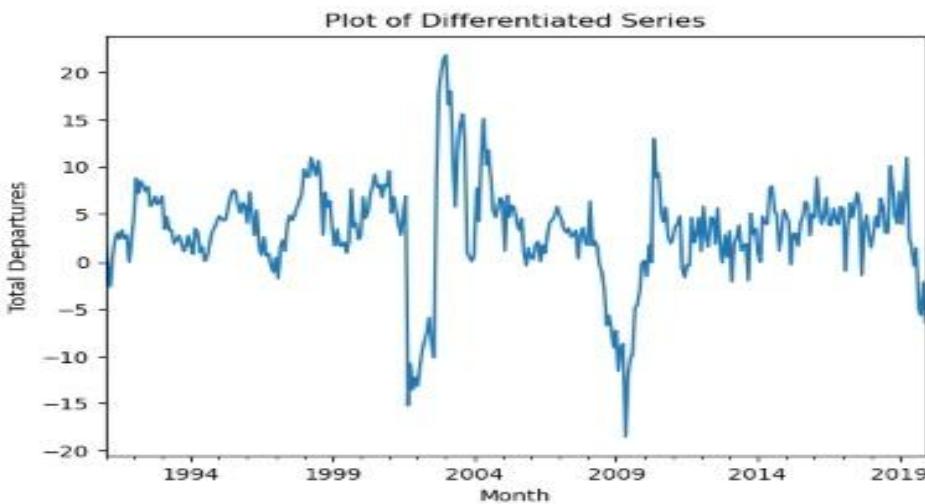
מעבר לכך, בדקנו קיום של ערכים חסרים, והאם קיימים חודשים ללא טיסות ונמצא כי אין ערכים חסרים ואין חודש ללא נתונים על טיסות.

מגמה, עונתיות ורעש:

אלו גרפים של התפלגות הנתונים, מגמה, עונתיות ורעש.



תרשימים של הסדרה לאחר הסרת העונתיות (לאחר גזירה- כאשר $12 = d$ כלומר עונתיות שנתית):



שאלות נייחיות שכיתן לשאול:

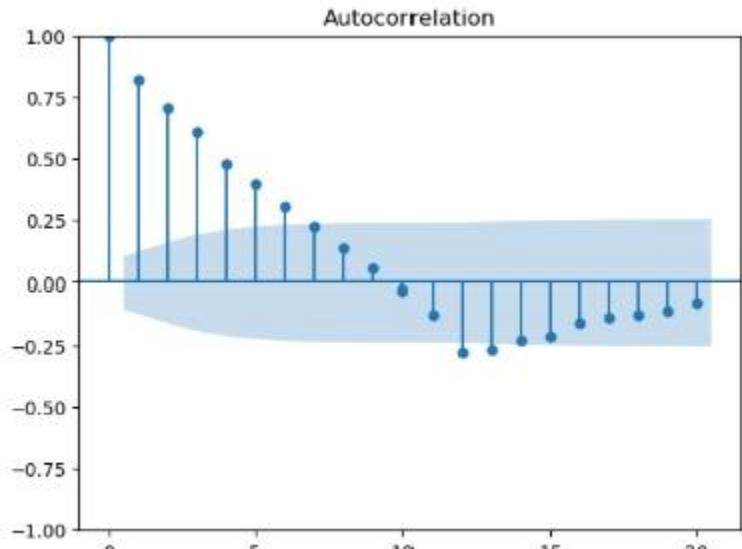
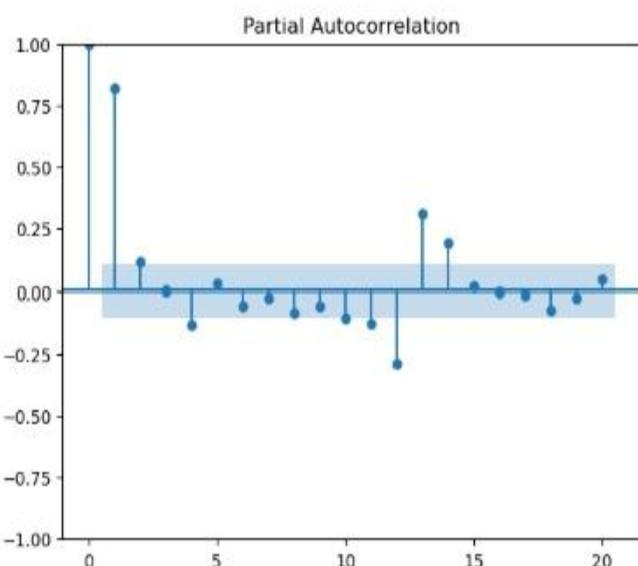
- מתי התרחשו נקודות חריגות בהתהילך, והאם ניתן לקשר אותן לאירועים בעולם?
- האם ישנו אירועים אגלוובלים (change point) שלא ניתן להזותם ויזואלית שהשפיעו על הסדרה (על כמות הטיסות).
- איזה מודל מתאר בצורה טובת את הנתונים שלנו

חלק 2 - מתודולוגיה והתאמת מודלים:

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

מודל סרימה:

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



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

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

גרף ה- PACF מציג פיק מובהק בשלג הראשון ושני, ולאחר מכן נחתך בחודות – זה תומך במודל AR(2) או SARIMA(p,0,0).

בנוסף, נראהים פיקים סביר ל-12, המעידים על עונתיות שנתית (Seasonality) – ולכן יש טעם לבחון מודלים עם מרכיב עונתי.

במנוחים של מודלים עונתיים, נראה שמתאים המודלים:

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

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

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

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

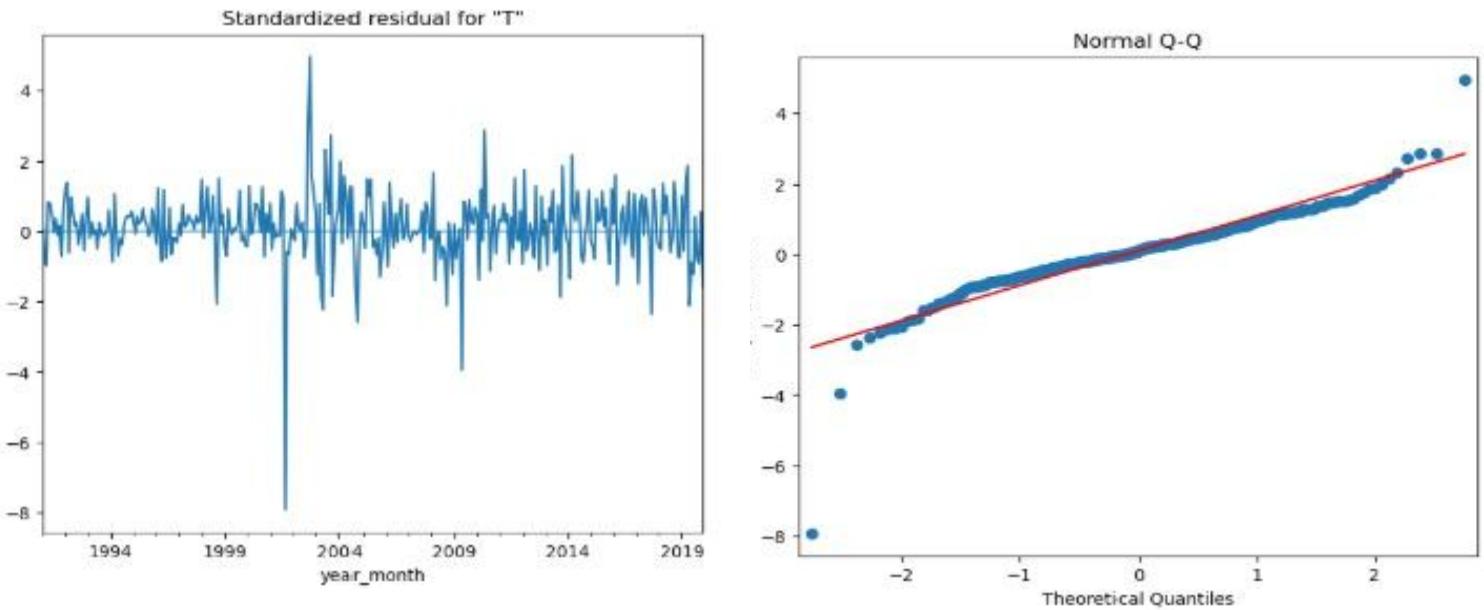
כלומר, כדי לבדוק את השילובים של $1 = k$ או 2 , יחד עם רכיב עונתי בעדרת $1 = D = s$, כאשר $0 = Q$ או 1 .

את בחירת הפרמטרים ביצמנו לפי קритריון BIC כאשר מצאנו למודל 12 SARIMA(2,0,0)(1,1,0)_12 יש את הערך הכי נמוך (1695) וכן בחרנו אותו כמתאים ביותר לתאר את הנתונים שלנו מבין מודלי הסרימה שבדקנו.

מעבר לכך, הצגנו את גראפי השאריות של כל המודלים והם היו מאד דומים, ולכן החלטנו להסתפק בקריטריון ה- BIC – תרשימים 2.1.4 – 2.1.1.

מעבר לכך, הרצינו דיאגנוזטיקה על המודל,

להלן גראף השאריות והQQ-Q plot המציג הדיאגנוזטיקה על המודל סרימה הנבחר.



1. Q-Q Plot:

ניתן לראות שהshareיות נצמדות יחסית טוב לקו האדום, שהוא קו הנורמליות התאורטית, בעיקר בחלק המרכזי של ההתפלגות.

עם זאת, יש חריגות בקצוות - SHAREITS KIIZONIOT, במיוחד לצד השמאלי התיכון.

המשמעות: באופן כללי, ההתפלגות של השאריות קרובה לנורמלית, אבל קיימים מספר ערכים>KIIZONIIM (שכנראה תואמים ליריעים חריגים כמו 11/9 או המשבר הכלכלי של 2008).

2. Standardized Residuals Plot:

ניתן לראות שהshareיות מתנהגות באופן יחסית יציב לאורך רוב התקופה.

ישנו קפיצות>KIIZONIOT במיוחד בסביבות השנים 2001-2003 ו-2008, שכנראה תואמת ליריעים כלכליים/גיאופוליטיים משמעותיים.

בסק הכל נראת שמודל סרימה טוב את המבנה הכללי של הנתונים.

מודל פרופט:

ניסינו להתאים את מודל פרופט לנתונים 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 טובות פחות טוב את המבנה של הנתונים שלנו בהתאם לתוכאה של גרפ השאריות והQQ-Q plot.

מעבר לכך, כש比יצנו השוואה בין המודלים, פרופט קיבל תוצאות מאוד טובות על סט האימון ותוצאות גראעות 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%

על סט המבחן, קיבלנו MSE של 271.579 וMAPE של 9.99%.

יתכן ונitin לננות את ערכי איפר הפרמטר של המודל שאחראי על הרגולריזציה של הטרנד כדי לקבל התאמה טוביה יותר, אך החלתו להתמקד במודלים האחרים המתוארים, שהניבו תוצאות טובות יותר.

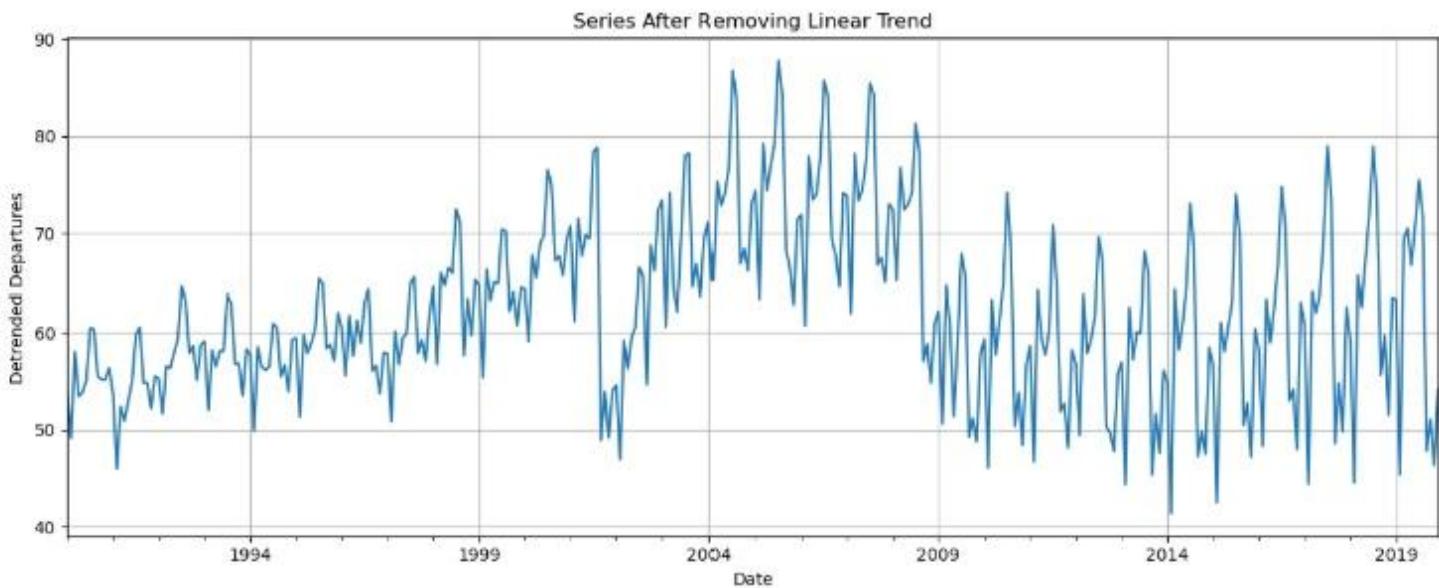
החלוקת אקספוננציאלית:

בחרנו להשתמש בהחלוקת אקספוננציאלית מאחר שמדובר במודל פשוט, שמתאים היטב לנתונים הכלולים מגמה ועוניתות יחסית יציבה – כפי שנצפה בנתוני הטיסות (יותר טיסות בקי' ופחות בחורף, ועליה כללית לאורך השנים). המודל מאפשר תחזית טובה לטוויה קצר-טווחי כי שם דגש על הנתונים האחרונים ביחס לעבר.

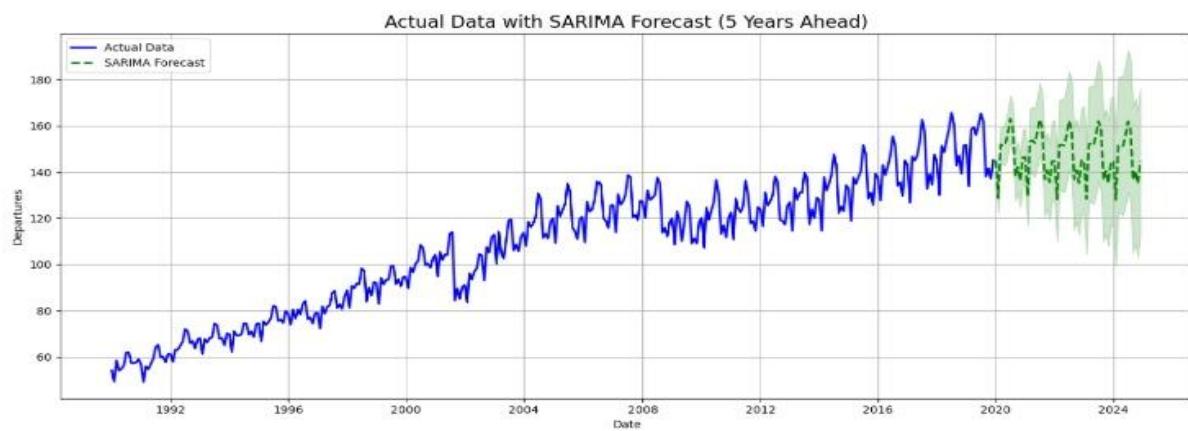
רגRESSED עם טורי פוריה:

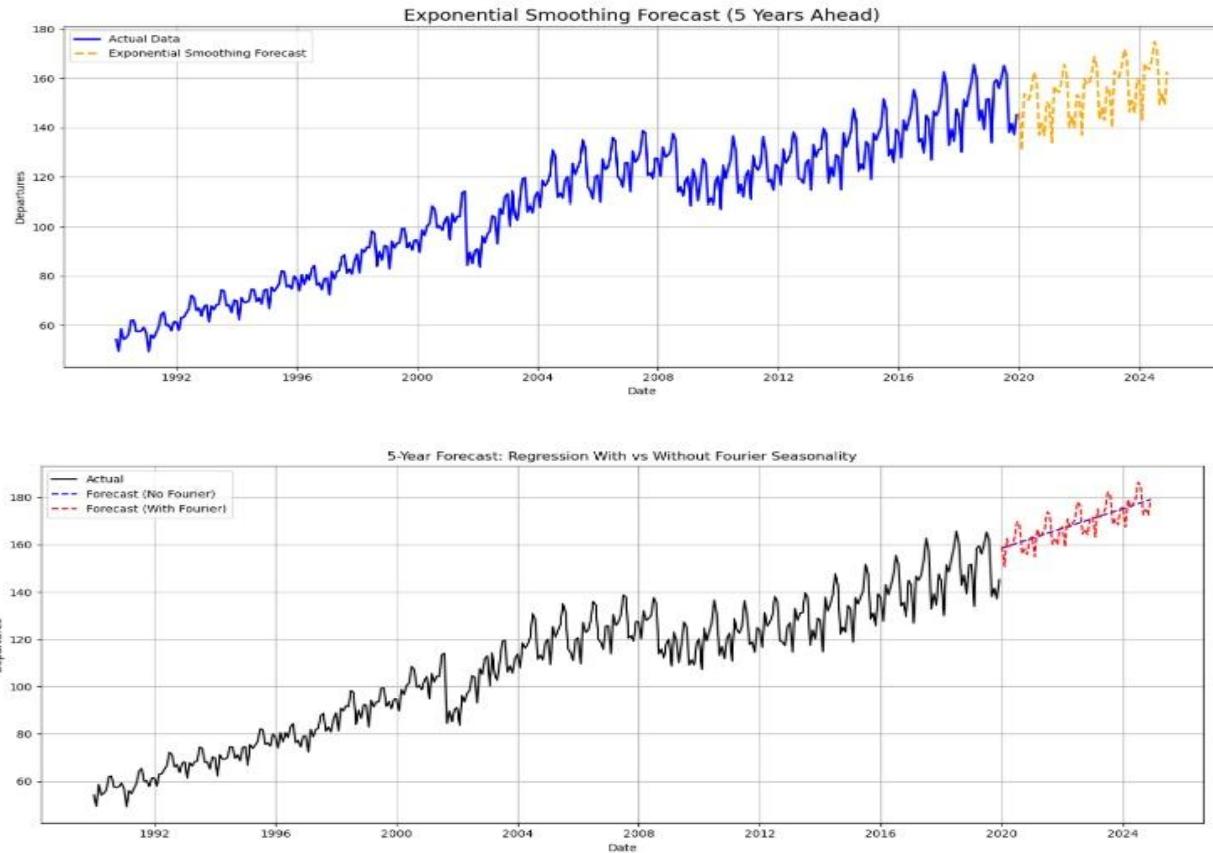
בהתחלת חשבנו להשתמש בק'ירוב עם טורי פוריה בלבד ללא רגרסיה, אבל לאחר הסורה של הטרנד מהסדרה ראיינו כי עדין קיימת מגמה והטרנד לא מסור לגמרי. זה נראה נובע מהpoint change משינוי המגמה בתנומינו. להלן גרפ של הסורה לאחר הסורה הטרנד:

כדי להתמודד עם שינוי המגמה בחרנו לשלב את הרגסיה עם טורי פוריה. השתמשנו בעונתיות שנתית ($D = 12$) כאשר לקחנו בחשבון את שני שינוי המגמה שעלייהם דיברנו בחלק המבוא ואולם ראיינו בצורה ויזואלית (אסוון התאומים וה"סאב פריים").



כעת נציג את החיזויים של המודלים – 5 שנים קדימה:





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

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

חלוקת מוצגת בתרשימים 2.8 במחברת.

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

כאשר סט המבחן היהו את התצפיות האחרונות בנותנים.

השתמשנו בשני מדדים: MSE ו-MAPE.

ה-MSE מgeb לשגיאות גדולות בצורה חזקה יותר בשל הריבוע של השאריות, מה שהופך אותו לשימושי במיוחד בזיהוי מודלים שנוטים לבצע טעויות גדולות.

מצד שני, MAPE מבטא את השגיאות כ אחוז מהערכים האמיתיים. מכיוון שבמערך הנתונים שלנו יש ערכים גדולים

וללא ערכים אפסיים, ניתן להשתמש ב-MAPE באופן בטוח והוא נותן מدد אמין למרחק היחסי של

התחזיות מהתוצאה בפועל (ולא משתנה בעקבות החלוקה ב – 1000 את מספר הטיסות)

אלו התוצאות בפועל של המודלים לסט המבחן:

Combined Model Performance:

Model	MSE	MAPE (%)
ExponentialSmoothing	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

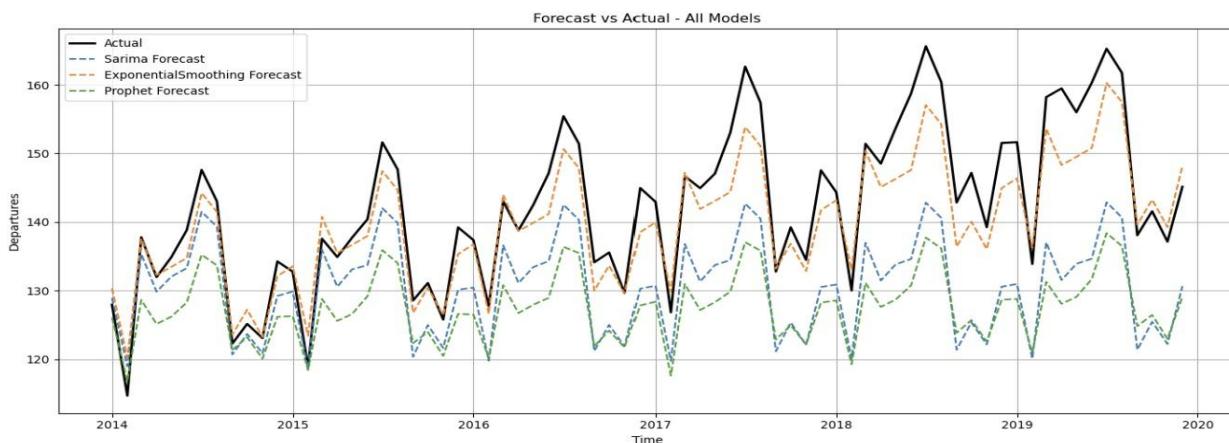
כמו שניתן לראות, החלקה אקספוננציאלית השיגה את התוצאות הטובות ביותר ביותר ואחריה רגסיה עם טורי פוריה עם פיצר נוסף של ar1.

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

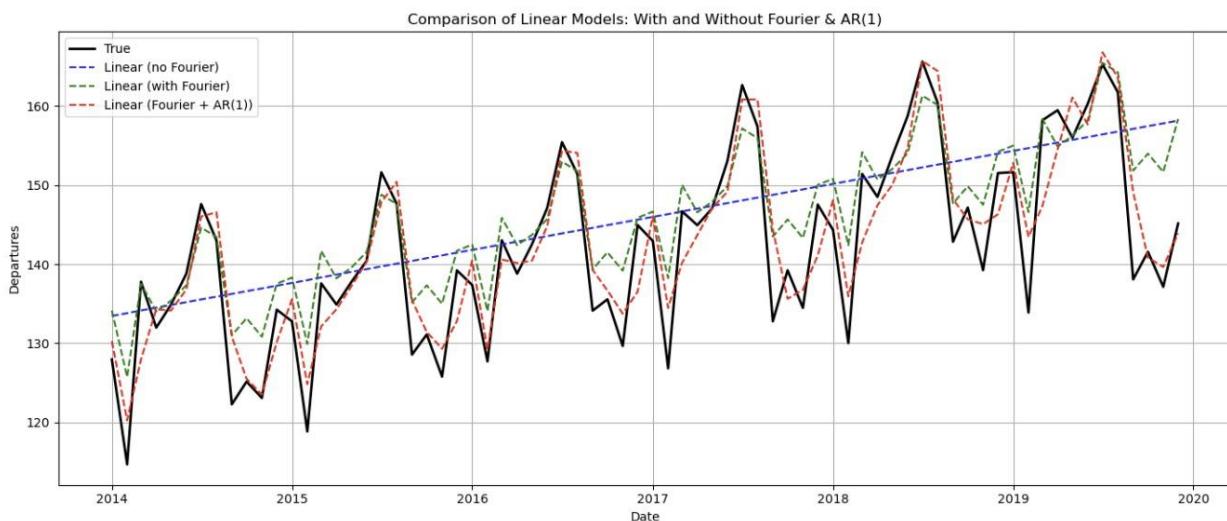
להלן גרפים של תוצאות של סט המבחן,

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

סדרימה, פרופט והחלקה אקספוננציאלית:



רגסיה, רגסיה עם טורי פוריה ורגסיה עם טורי פוריה וAR(1):



ניתן לראות שמודל הרגסיטה עושם יחסית טוביה טוביה בחיזויים.

חלק 3: משתנים אקסוגניים:

בחרנו את מדדי ה-*i*-CPI עבור אנרגיה, תחבורה ודלק (בנzie) מכיוון שהם קשורים ישירות לגורם המשפיעים על פועלות 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.

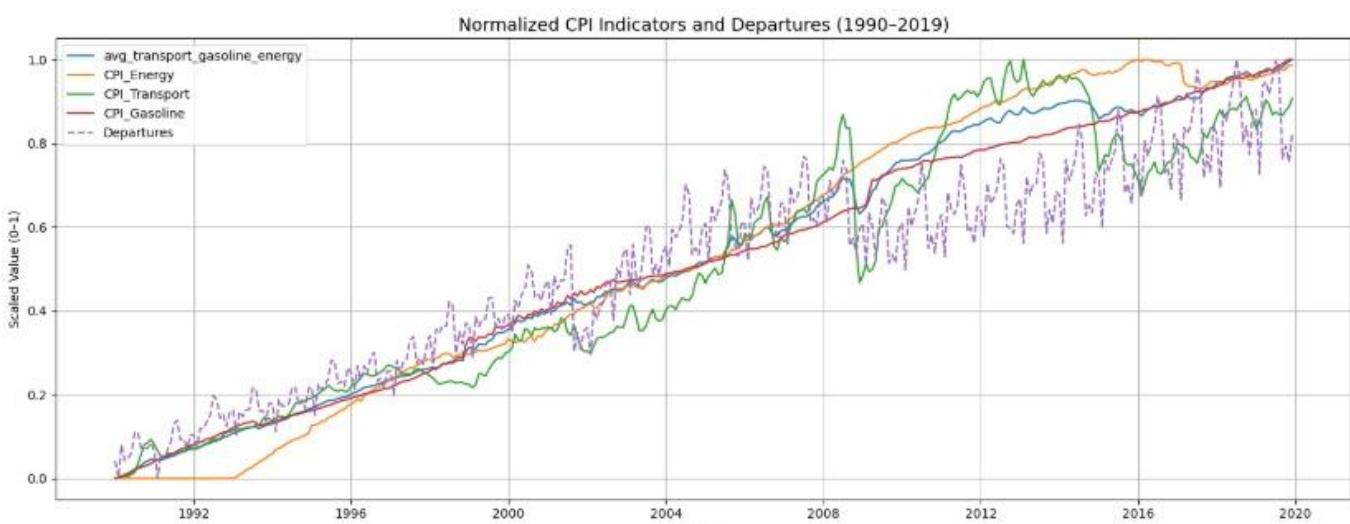
מצאנו שבסדרת ה-*i*-CPI לאנרגיה אין כל נתונים בין השנים 1990 ל-1992, שקלנו שלוש אפשרויות להתחממות עם העריכים החסרים:

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

לכן, בחרנו למלא את הערכים החסרים בין השנים 1990 ל-1992 באמצעות הערך של ינואר 1993, מתוך הנחה שהערכתים המוקדמים היו יציבים יחסית ודומים לנקודת הנتون הראשונה הזמין ובכך ניסינו לשמר על הריציפות במדד הנتونים תוך מזעור העיוותים האפשריים.

הערה: היו צריכים לחוץ את המדדים הרלוונטיים מתוך הקובץ של כלל המדדים לצריך. הקובץ המקורי כלל 276,773 רשומות, ולאחר סינון המדדים, אgregציה וסינון השנים הרלוונטיות (הנתונים היו משנת 1913-2025) נותרו רק 363 רשומות. כל סדרה מלבד הסדרה של האנרגיה כללו 363 רשומות, ולאחר ההשלמה גם היא כללה 363 רשומות.

להלן גרפ של כל הסדרות לאחר סקילינג:



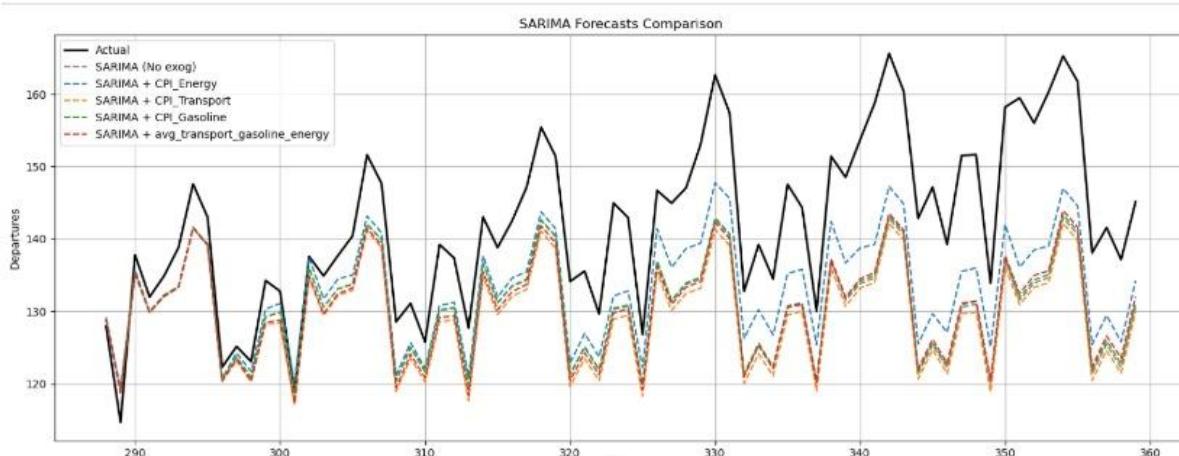
מבחן ויזואלית, נראה שקיים קשר בין הסדרות של מדדי המחיר, סדרת כמות הטיסות היוצאות ומגמה דומה. שני המודלים שיכולים להשתמש בסדרה אקסוגנית מבין המודלים שהציגו בחלק 3 הם סרימקס ורגסיה עם טורי פוריה (כפי'ר נוסף) - שוב, כאן לא המשכנו עם פרופט.

להלן תוצאות המודלים של סרימה:

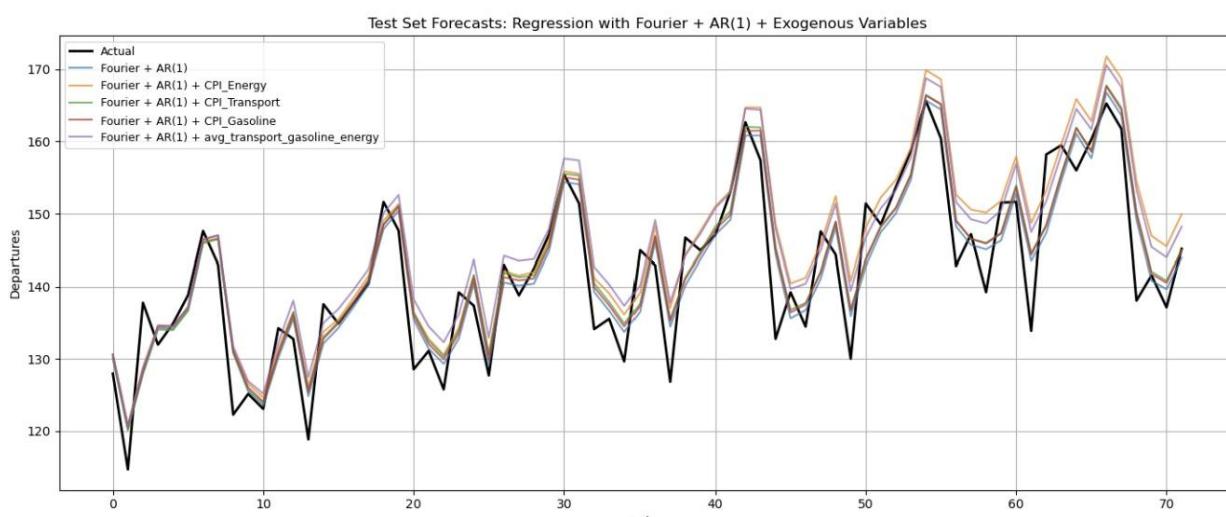
Final comparison including SARIMA without exog:			
Exogenous Variable	MSE	MAPE (%)	
CPI_Energy	114.593500	0.062431	
CPI_Gasoline	182.614331	0.079173	
avg_transport_gasoline_energy	185.475302	0.081470	
SARIMA (no exog)	189.175099	0.080529	
CPI_Transport	213.736219	0.087497	

תחילה התחלנו עם מודל סרימקס, כאשר בדקנו שימוש בכל אחת מהסדרות בנפרד ובממוצע שלהן. ניתן לראות שמודל סרימקס עם סדרת האנרגיה הוביל את התוצאות הנמוכות ביותר (אך צריך לזכור בחשבון שהזו המודל שהשתמשנו בהשלמת ערכים חסרים). הבא אחרי זהו סרימקס עם סדרת מדד מחירי הדלק.

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



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



תוצאות על סט המבחן של גראסיה עם פוריה וARI עם ובלי' 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.

חלק 4: change point detection

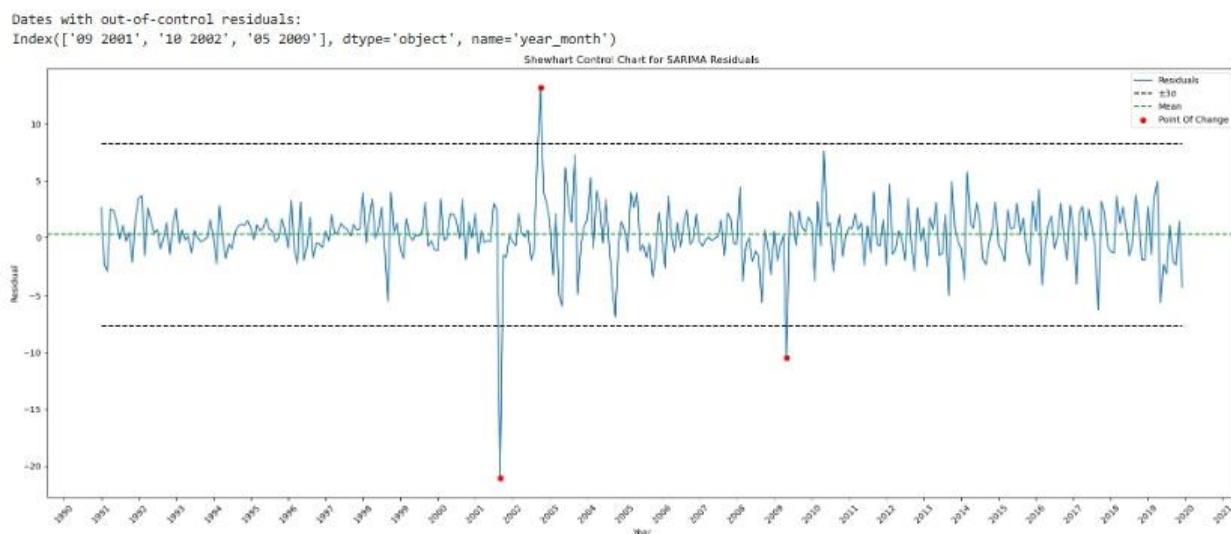
תחילה, השווינו את ערך ה- BIC של כל מודלי סרימה שבדקנו לאור הפרויקט כולל אלו של המשתנים האוקסיגנים. המודל שקיבל את ערך ה- BIC הנמוך ביותר היה מודל סרימה ללא המשתנים האוקסיגנים, עם אותם הפרמטרים מחלק 2 בבדיקה.

באמצעות תרשימים בקרה מסוג Shewhart שיישם על שאריות מודל SARIMA (שכן סרימה איננו מטשטש את האנומליות ושינוי הטרנדים לעומת משל גראסיה עם פוריה שם אנו מכנים את שינוי המוגמות כפיצרים נוספים), זיהינו שלוש חריגות משמעותיות: ספטמבר 2000, אוקטובר 2001 ומאי 2008.

שתי האחרונות תואמות למקדים ידועים – אסון התאומים והמשבר הכלכלי העולמי. לעומת זאת, לקפיצה בספטמבר 2000 לא צפינו הסבר באופן זהה.

את הנקודה זו לא ראיינו בויזואלייזיות הקודמות שהציגו, יתכן בשל המוגמה והעונתיות החזקות בסדרת הזמן, אשר יתכן וטשטשו את האנומליה.

תרשימים Shewhart עם התאריכים שנמצאו:



סיכום ומסקנות

בעובדה זו ניתחנו נתונים נטוי טיסות בינלאומיים יוצאות מארצות הברית בין השנים 1990–2019, ובדקנו את יכולת החיזוי של מספר מודלים לסדרות זמן. הتمקדמו בzychוי מגמות, עונתיות וחריגות, תוך שימוש במודלים כמו SARIMA, החלקה אקספוננציאלית, Prophet, AR(1) ורגסיה עם טורי פוריה.

מודל החלקה האקספוננציאלית הציג את ביצועי החיזוי הטוביים ביותר על סט המבחן, כנראה בזכות התאמת שלו לבניה הנתונים הכלול עונתיות ייצה. מודל Prophet נתה לאוברפיטינגן, ואילו AR(1) 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 (מדד CPI) על תוצאות המודלים. השימוש במודלי SARIMAX עם מדד האנרגיה הראה שיפור משמעותית יותר מאשר המודדים אך גם במדדים נוספים נראה שיפור קטן. לעומת זאת, במודלי הרגסיה עם משתנים אקסוגניים לא נראה שיפור כלל. הסבר אפשרי לפער ביצועים הוא שמודלי SARIMAX מצליחים לנצל את המשתנים האקסוגניים כדי לשפר את החיזוי דווקא בתקופות חריגות כמו אסון התאומים או משבר הסאב-פריים, בכך שהם מסייעים למודל להזות שינויים פתאומיים שאינם מוסברים על ידי עונתיות או מגמה בלבד. לעומת זאת, במודלי הרגסיה עם טורי פוריה, העונתיות מייצגת בצורה גמישה יותר אך צזו שאינה רגישה לאירועים חד-פעמיים, ולכן השפעת המשתנים החיזוניים מטשטשת. מעבר לכך, ניתן שהרגסיה עם פוריה ו-AR(1) כבר סיפקה תחזית מדויקת יחסית – כך שלמודל נותרה פחות "הזמןונת" להשתפר בעקבות הוספה משתנים אקסוגניים.

בנוסף, תרשימים על שאריות מודל SARIMA סייעו בzychוי נקודת חריגה שלא נראתה קודם לכן ביזואלייזיות. שכן, "תכן שהוא 'גבלה' בשל הטרנד 'חזק'" ועונתיות הסדרה. ביזואלייזיות כן ראיינו וצינו את משבר ה"סאב פריים" ואסון התאומים. כמו כן, הם גם ניצפו בתרשימים Shewhart.

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

ביבליוגרפיה:

הסדרה המקורית - [U.S. International Air Traffic data\(1990-2020\)](https://www.kaggle.com/datasets/paveljurke/u-s-consumer-price-index-cpi)

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