

שימוש בשיטות סטטיסטיות לתחקור מאגר עסקות שוק מט"ח

אמיר חטיב ואראל מזוז¹*

תקציר

עבודה עם מאגר נתוני עתק (Big Data) מצריכה, בשל היקף הנתונים הרחב שמקשה על השימוש בשיטות אופייניות לניתוח המידע, שימוש בכלים סטטיסטיים מתקדמים לצורך בקרת איכות הנתונים והפקת תובנות מהמידע שמתקבל. החטיבה למידע ולסטטיסטיקה בבנק ישראל מנהלת מאגר נתוני עתק על מכשירים פיננסיים נגזרים בשוק המט"ח והריבית ("מפ"ן"), שמכיל מידע אודות עסקות בנגזרים אלה שמבוצעות בשוק ה-OTC. מאגר הנתונים משמש בעיקר לצורך הכרת השוק והפקת תוצרים תומכי החלטה לקובעי המדיניות בבנק ישראל. בעבודה זו נסקרים חלק מהכלים ומהשיטות שמשמשים את החטיבה לצורך טיוב, עיבוד והפקת תובנות מהמידע שקיים במאגר מפ"ן וכן מתוארים בה כמה מודלים מעולם מדע הנתונים, שמשמשים לאיתור דיווחים חריגים, להשלמת המידע במאגר הנתונים, לסיווג שחקנים ולהפקת תובנות תומכות מדיניות.

* החטיבה למידע ולסטטיסטיקה, בנק ישראל.

¹ שותפים נוספים בתהליך החשיבה והיישום של המודלים והשיטות שמתוארות בעבודה זו: היחידה לשיטות סטטיסטיות ומדע הנתונים (מידע וסטטיסטיקה)

1. הקדמה

החטיבה למידע ולסטטיסטיקה בבנק ישראל (להלן – החטיבה) מנהלת מערכת נתונים פרטנית על מכשירים פיננסיים נגזרים בשוק המט"ח והריבית (להלן "מערכת מפ"ן²). החטיבה מקבלת נתונים ממתווכים פיננסיים בארץ ובעולם על עסקות בנגזרי מט"ח וריבית שמבוצעות בשוק המסחר מעבר לדלפק (שוק ה-OTC - Over the counter). הנתונים הם תוך-יומיים ומתקבלים ברמה יומית. המתווכים הפיננסיים שמחויבים בדיווח (להלן – "הגופים המדווחים") הם תאגידים בנקאים ישראלים וכן מתווכים פיננסיים שהיקף פעילותם בשוק המט"ח השקלי עולה על 15 מיליון דולר ביום (להרחבה, ראו צו בנק ישראל³).

המידע שמתקבל במערכת הנתונים מגיע לכדי מיליוני רשומות בשנה ויוצר מאגר נתוני עתק (Big Data) בשל עומקו ההיסטורי והמידע הרב שיש אודות כל עסקה. לכך נדרשים כלים סטטיסטיים מתקדמים לשם בקרת איכות הנתונים והפקת משמעויות מהם. מטרת עבודה זו היא לסקור מספר כלים שמשמשים את החטיבה לצרכים אלה.

לצורך הרכבת תמונה כוללת שמשקפת את האירועים שמתרחשים בשוק ה-OTC ושמאפשרת הפקת תובנות מהנתונים שמתקבלים, הוטמעו כלים סטטיסטיים שמאפשרים סיווג שחקנים בהתאם למאפיינים ייחודיים שנקבעו באמצעות המודלים שיתוארו להלן.

הפרק הראשון בעבודה יכלול סקירה קצרה אודות הנתונים שמתקבלים לצורך היכרות כללית עם המערכת. בפרק השני נביא מספר דוגמות לתובנות שהופקו באמצעות שימוש בכלים סטטיסטיים על מסד הנתונים. לשם המחשה – נראה שניתן להבחין בדפוס פעילות מסוים שמאפיין חברות שעיקר עיסוקן הוא בייבוא של סחורות ושירותים ודפוס זה שונה באופן ניכר מדפוס הפעילות שמאפיין חברות שעוסקות בייצוא של סחורות ושירותים.

הפרק השלישי יעסוק בהפקת תובנות מהנתונים. המערכת נבנתה כדי לספק לבנק ישראל מידע מפורט ככל האפשר על שוק המט"ח, כולל הפקת תובנות שיעזרו לו לקבל החלטות מיטביות בנוגע לשוק זה. זוהי המטרה הסופית שלשמה מיועדת מערכת המפ"ן – הקניית כלים לבנק ישראל לצורך ביצוע עבודתו.

בפרק הרביעי נציג בקצרה מספר בקורות סטטיסטיות שהוטעמו במערכת המפ"ן לצורך איתור נתונים חריגים. במהלך הפרק נדון בבעיות שאיתם נדרשה החטיבה להתמודד, נתאר את המודלים הסטטיסטיים שנבחנו ואת המודל שנבחר בסופו של דבר. למרות שאין זה נושא העבודה, אנו רואים חשיבות בהצגה קצרה של בקורות אלה, שכן תחילתה של כל עבודה סטטיסטית היא בבדיקת איכות הנתונים

2. רקע על מערכת הנתונים

מערכת מפ"ן (מכשירים פיננסיים נגזרים) היא מערכת נתונים למעקב אחר המסחר השקלי בשוק ה-OTC. מערכת הנתונים משלבת מספר ממדים עיקריים – המכשירים הפיננסיים, נכסי הבסיס, הזמן והמגזרים, שמסייעים לבנק ישראל לאפיין את אופי הפעילות בשוק השקל/מט"ח ולקבוע את מדיניותו בהתאם.

נתוני העסקות במערכת מגיעים משלושה סוגים של מקורות מידע:

- תאגידים בנקאים מקומיים שמדווחים על עסקותיהם החל משנת 2008.
- מוסדות פיננסיים זרים שמדווחים על עסקותיהם החל משנת 2017.
- מתווכים פיננסיים מקומיים שמדווחים על עסקותיהם החל משנת 2017.

הנתונים מתקבלים בתדירות יומית ומכילים פירוט של כל העסקות שנעשו ביום העסקים הקודם. כיוון שהדיווח מתקבל מאזורי זמן שונים, נקבע, לשם האחידות, שמסגרת הדיווח על העסקות תהיה לפי זמן בין-לאומי (UTC) ולכן הדיווח היומי כולל עסקות שנקשרו במהלך 23:59-00:00 זמן UTC ביום העסקים (T) והמידע שנדרש מתקבל ביום העסקים העוקב (T+1).

² להרחבה ראו [מבט סטטיסטי 2018](#)

³ [צו בנק ישראל](#)

המערכת כוללת כ-40 מדווחים ועשרות אלפי שחקנים שפעילים בשוק ה-OTC. כפי שצוין לעיל, היקף העסקות שמתקבלות למערכת נאמד במיליוני רשומות בשנה.

מתכונת הדיווח החדשה הותאמה לתקן בין-לאומי שגיבש ה-ISDA⁴, הארגון שמוביל את קידום הנהלים לתקינה (סטנדרטיזציה) של המסחר בעסקות OTC. המתכונת מכילה כ-40 שדות דיווח לפי שלושה פרמטרים:

1. פרטי המדווח.
2. פרטי הלקוח/שחקן.
3. מאפייני העסקה – זמן הביצוע (ברמת השנייה), המכשיר, תאריך פקיעת החוזה, שער העסקה, הסכום הנקוב ושווי העסקה.

בנק ישראל מפיק תוצרים רבים באמצעות נתונים במערכת המפ"ן, שחלקם מתפרסמים במסגרת ההודעות העיתיות של הבנק על פעילותו בשוק המט"ח. דוגמאות לתוצרים שמופקים באמצעות המערכת הם אומדן לרכישות המט"ח המצטברות של מגזרים עיקריים, נפח המסחר היומי בשוק המט"ח בפילוח לפי סוגי מכשירים, סטיית התקן הגלומה באופציות שקל/דולר, ועוד.

3. מודלים סטטיסטיים ככלי לסיווג

למסד הנתונים מתקבלים נתוני העסקות לאחר שעברו את כל הבקורות הלוגיות שנדרשות. בשלב זה ניתן להפעיל עליהם ניתוחים שונים. היות שמדובר במסד נתונים של מיליוני רשומות, יש צורך להשתמש בשיטות ואלגוריתמים (תהליכים חישוביים) מתאימים מתחום מדע הנתונים, שמתחלקות בעיקר לשני סוגים:

למידה מונחית (Supervised) - מספקים למודל נתונים עם תיוג או סיווג קיים והמטרה היא לסווג תצפיות חדשות על סמך המידע שהיה ברשותנו אודות סיווג קיים של תצפיות קודמות.

למידה לא-מונחית (Unsupervised) - במקרה זה אין סיווג ראשוני והמטרה היא לחלק נתונים אלה לאשכולות על סמך המבנה הפנימי והדפוסים המשותפים שהאלגוריתם יכול לזהות.

היתרון של למידה לא-מונחית טמון בכך שאין קיבעון לגבי התוצאה והחוקר שמשתמש בה, בא עם "ראש פתוח" לתוצאות השונות. יחד עם זאת, ייתכן שלא תימצא אף חלוקה שמתיישבת עם ההיגיון שעולה מהנתונים (כפי שיתואר להלן). לעומתה, למידה מונחית אמנם "מקובעת" יותר, שכן התוצאה הרצויה מוגדרת מראש, אך היא מאפשרת מתן "סיוע" למודל כדי לאפיין את המשקלות של המשתנים המסבירים בצורה טובה יותר.

לשם המחשת השיטות, הדוגמה שתלווה פרק זה היא ניסיון לסווג את השחקנים השונים במגזר העסקי שעשו עסקות בשוק המט"ח לקבוצות שונות.

בשלב ראשון מבוצעת באמצעות למידה לא-מונחית חלוקה של המגזר העסקי למספר אשכולות שונים בשיטה שנקראת PCA (ניתוח גורמים ראשיים) בשילוב עם k-MEANS clustering (ניתוח אשכולות).

א. ניתוח אשכולות (Cluster Analysis) באמצעות למידה לא-מונחית

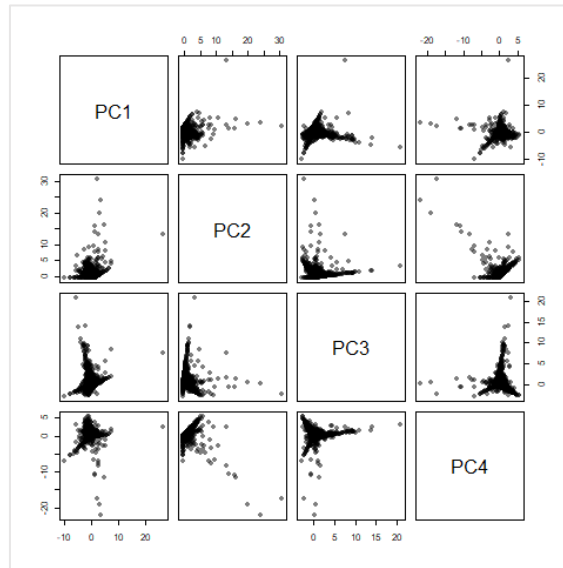
במסגרת הניסיון לסווג את השחקנים לקבוצות לפי מאפייני עסקות, בוצע שימוש בשיטת ה-PCA (Principal components analysis), שמאפשרת תצוגה גרפית של התצפיות במסד נתונים באמצעות צמצום מספר המשתנים שמוצגים בגרף. השיטה לוקחת את כל המשתנים הראשיים (משתנים שהוגדרו על ידי המשתמש, כגון, נפח מסחר, רכישות מט"ח, ועוד) ויוצרת מהם משתנים חדשים, כך שכל משתנה חדש הוא צירוף לינארי⁵ של המשתנים הראשיים. כך יוצרת השיטה מספר מצומצם של משתנים שטומנים בחובם את עיקר השונות שקיימת במסד הנתונים שנבחר.

⁴ ארגון ISDA יצר חוזה אחיד לביצוע עסקות בנגזרים. לפרטים נוספים: <https://www.isda.org/#>

⁵ צירוף לינארי הוא משוואה פשוטה של משקלות לכל משתנה. לדוגמה, A הוא צירוף לינארי של המשתנים c-1 b במשוואה $A=0.5b+1.5c$

לצורך הבחנה בין קבוצות שונות במגזר העסקי בשיטת ה-PCA, הוכנסו למודל 8 מאפייני העסקות העיקריים, ביניהם מחזור המסחר היומי הממוצע, רכישות המט"ח נטו בממוצע ליום, הטווח הממוצע/החציוני של העסקות ועוד. כך נוצרו 4 משתנים חדשים (שהם צירוף לינארי של 8 המשתנים הקודמים), שמכסים כ-87% מהשונות הקיימת. איור 1 להלן מראה את השילובים השונים של 4 המשתנים ברמה דו-מימדית על מערכת צירים. כל גרף מציג את החלוקה של התצפיות במדגם לפי שני משתנים מסבירים, לדוגמה, הגרף הימני העליון מראה את החלוקה לפי PC1 ו-PC4. באמצעות גרפים אלה ניתן להבחין בצורה פשוטה יותר באשכולות השונים במגזר העסקי (אם יש כאלה) על בסיס הנתונים הגולמיים. חשוב לציין שהחלוקה לאשכולות לפי שיטת ה-PCA מבוססת על שונות סטטיסטית ולכן היא לא תהיה בהכרח מתואמת עם החלוקה הכלכלית.

איור 1: מערכות צירים של הגורמים הראשיים (PCA)



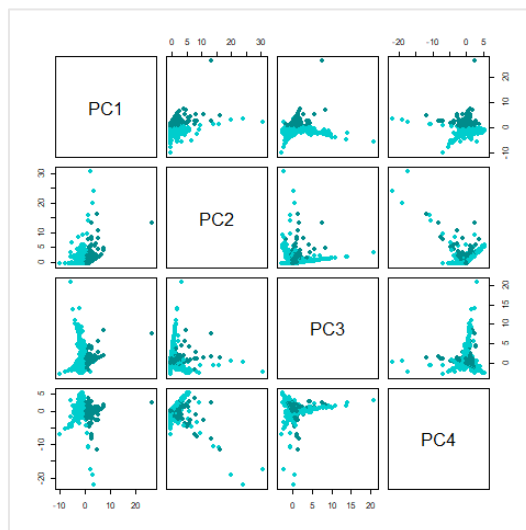
המקור: דיווחי הבנקים ועיבודי בנק ישראל

בבחינת הפיזור של השחקנים השונים במרחבים הדו-ממדיים, ניתן אמנם לבצע חלוקה למספר קבוצות, אך אין הפרדה ברורה בין הקבוצות באף חלוקה. לדוגמה, התבוננות בגרף שמציג את הפיזור בין PC3 ו-PC4, מראה שיש מרכז של ההתפלגות ושממנו יוצאות 3 "זרועות" לכיוונים שונים וניתן להגיד שהם מחלקות את הנתונים, אך חלוקה זו אינה מובהקת ולכן נפנה לשיטה נוספת של למידה לא-מונחית - k-MEANS.

שיטת k-MEANS היא שיטה נפוצה עבור ניתוח אשכולות (clustering) במדע הנתונים. מטרת השיטה היא לבצע חלוקה של הנתונים לפי מרכזי כובד, כשמספר מרכזי הכובד נקבע על ידי המשתמש וכל מרכז כובד מייצג אשכול של נתונים. על ידי בחירה נכונה של מספר מרכזי הכובד, ניתן לאתר קבוצות שונות בתוך הנתונים.

לצורך חלוקת השחקנים במגזר העסקי לקבוצות, נבחרו מספר האשכולות הרצויים בכל פעם ובוצעה החלוקה לקבוצות (פעם לפי 2 אשכולות, פעם לפי 3 ופעם לפי 4 אשכולות). בכל בחירה של מספר מרכזי כובד בוצעה חלוקה לקבוצות באמצעות שיטת k-MEANS (כפי שניתן לראות באיור 2 להלן, שמתאר חלוקה לשני מרכזי כובד), אך לא נמצא היגיון כלכלי בחלוקות השונות, למשל, חלוקה בין יבואנים לייצואנים. לכן נפנה כעת לתיאור תהליך הסיווג באמצעות למידה מונחית, בשיטה בשם LDA (Linear discriminant analysis).

איור 2: מערכות צירים של הגורמים הראשיים בחלוקה לקבוצות לפי שיטת KMEANS



המקור: דיווחי הבנקים ועיבודי בנק ישראל

ב. סיווג לקבוצות באמצעות למידה מונחית

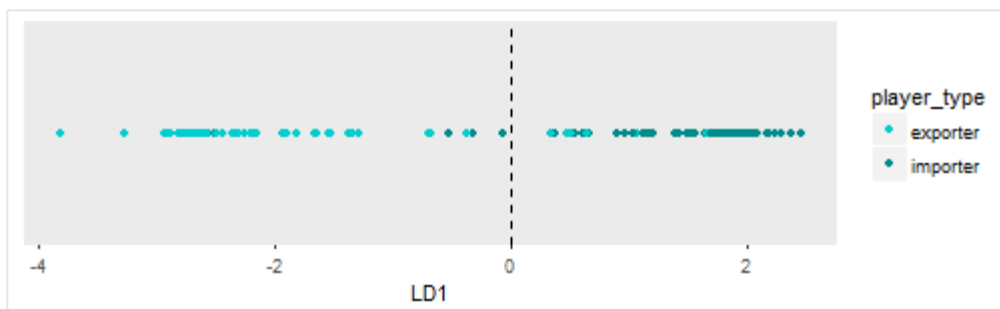
כשקיימת חלוקה לקבוצות שידועה מראש, אפילו אם מדובר על חלוקה של מדגם מהנתונים, ניתן להשתמש בשיטת LDA לצורך איתור המאפיינים שמבחינים בין הקבוצות השונות. שיטת ה-LDA מחשבת צירוף לינארי של המשתנים המסבירים, שמאפשר את ההפרדה הטובה ביותר בין הקבוצות שהוגדרו (משתנה המטרה). באמצעות הצירוף הלינארי שמתקבל, ניתן לבצע סיווג קבוצות לנתונים שלא היו במדגם הראשוני, לרבות נתונים חדשים.

בדוגמה של חלוקת השחקנים במגזר העסקי שמלווה את הפרק, ניתן היה להגיע לחלוקה ראשונית של חברות שעוסקות בייבוא לעומת חברות שעוסקות בייצוא. באמצעות מאגרי מידע אחרים שקיימים בבנק ישראל ובשילוב מאפיינים כלכליים מסוימים, נבנתה רשימה של חברות שנחלקות לחברות שעוסקות בייבוא ולעומתן חברות שעוסקות בייצוא. למשל, מצופה שחברה עסקית שעוסקת בפעילות יבוא בעיקר, תרכוש יותר מט"ח מאשר תמכור אותו ולעומתה חברה שעוסקת בפעילות יצוא בעיקר, תמכור יותר מט"ח מאשר תרכוש אותו. בהתאם לכך, חברה יצואנית הוגדרה כחברה שמייצאת סחורות יותר מאשר שהיא מייבאת וחברה יבואנית היא המקרה ההפוך, חברה שמייבאת יותר סחורות מאשר שהיא מייצאת.

לאחר מכן בוצע שימוש באותם משתנים מסבירים שמאפיינים כל שחקן, כגון, מחזור מסחר שנתי, ימי פעילות מתוך סך ימי העסקים, רכישות נטו, חלק השימוש בדולר לעומת מטבעות אחרים, ממוצע של טווח העסקות, ועוד. על נתונים אלה הופעל אלגוריתם לצורך מציאת ההפרדה הטובה ביותר בין שתי הקבוצות שמוגדרות ונמצא הצירוף הלינארי של המשתנים המסבירים, שמפריד בין שתי הקבוצות בצורה הטובה ביותר.

לפי אותו צירוף שמורכב מממד אחד, כפי שניתן לראות באיור 3 להלן, יש הפרדה די ברורה באופי הפעילות של יצואנים וייבואנים, דבר שמאפשר כאמור לסווג פעילים חדשים ולא מזהים בשוק באופן אוטומטי ובצורה מהימנה, על בסיס מאפייני פעילותם בשוק.

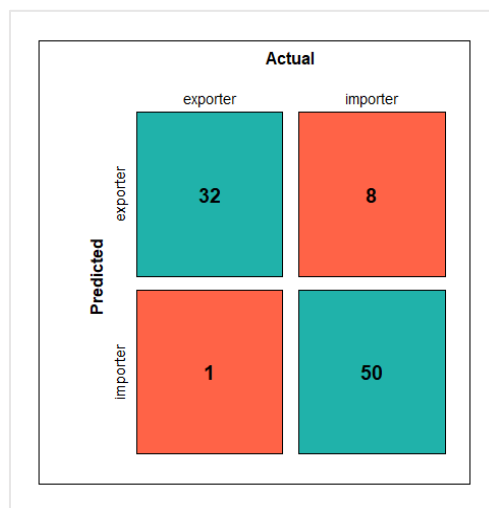
איור 3: חלוקה ליבואן\יצואן לפי הצירוף הלינארי שחושב בשיטת LDA



המקור: דיווחי הבנקים ועיבודי בנק ישראל

לאור התוצאות משביעות הרצון של המודל, הופעל הצירוף שנמצא על קבוצת מבחן שהוגדרה מראש, חברות מהמגזר העסקי שיש להם סיווג יבואן/יצואן (לפי סיווג שבוצע באמצעות מאגרי מידע אחרים ובדומה לקבוצת המדגם) ולא היו חלק מהמדגם בעת חישוב הצירוף והמודל סיווג אותם בהתאם למקדמים שחושבו ב-LDA עבור כל משתנה. תוצאות הסיווג באמצעות המודל הושוו לחלוקה הידועה מראש ונמצא שהמודל חזה באופן טוב (ראו איור 4) את החלוקה של החברות לענפי הייבוא והייצוא, בהתאם למאפייני הפעילות שלהם ולכן הוחלט להפעילו באופן שוטף, לצורך סיווג השחקנים החדשים, שלא ניתנים לזיהוי מראש במגזר זה.

איור 4: Confusion Matrix



המקור: דיווחי הבנקים ועיבודי בנק ישראל

4. הפקת תובנות מהנתונים

עד כה עסקה העבודה בסיווג הנתונים והשבחתם, אך העיקר חסר מן הספר – אין משמעות למסד נתונים, איכותי ככל שיהיה, אם לא ניתן להפיק ממנו תובנות. לכן, כדי להפיק תוצרים ותובנות ממערכת עם נתונים רבים, כמו מערכת מפ"ן, הוטמעו מספר תהליכים סטטיסטיים נוספים במערכת.

א. איתור הבדלים בין מגזרים

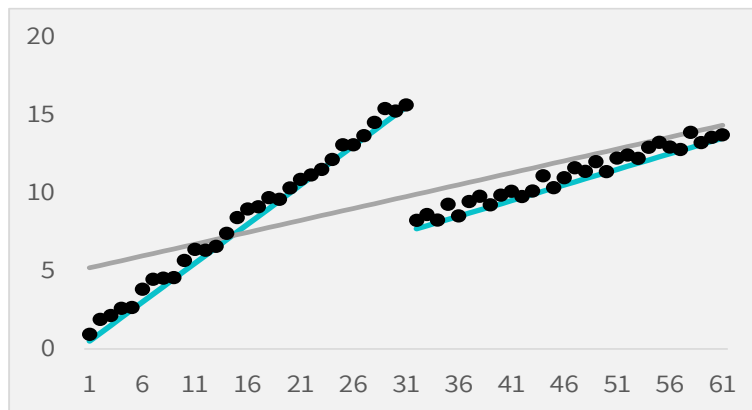
לאחר סיווג הקבוצות באמצעות שיטת ה-LDA שתוארה בפרק הקודם, נבחנו המקדמים של הצירוף הלינארי שלפיו בחר המודל לסווג את השחקנים. הבחינה הראתה שיש מספר הבדלים בין הקבוצות ושהעיקרי שבהם הוא, שייצואנים פעילים בשוק ימים רבים יותר מאשר יבואנים, אך בימים עם שינויים חריגים בשע"ח, הייבואנים פעילים יותר מאשר הייצואנים. כמו כן הם קושרים עסקות לטווח ארוך יותר. זאת בנוסף לרכישות המט"ח נטו - בהתאם לאופי הפעילות העסקית של כל קבוצה - יצואנים מוכרים מט"ח בעוד שייבואנים רוכשים אותו.

ב. איתור שברים מבניים בסדרות

אחד הכלים המשמעותיים שעשויים לסייע בהתוויית מדיניות הוא האפשרות להבחין בין תקופות זמן שונות בהתאם למאפיינים מוגדרים (למשל, להבחין בין תקופת מיתון לתקופת צמיחה). באופן דומה, הבחנה בין תקופות שונות עבור מגזר מסוים, עשויה ללמד רבות על אופי הפעילות של אותו מגזר. לשם המחשה – במידה וניתן היה לקבוע שבעתות של ייסוף בשער הדולר־שקל, המגזר העסקי הוכש יותר מט"ח (בגלל הצורך של יבואנים להגן על עצמם מפני ייסוף נוסף), ניתן היה להעריך באופן טוב יותר כיצד ישפיע ייסוף־פיחות על פעילות המגזר העסקי.

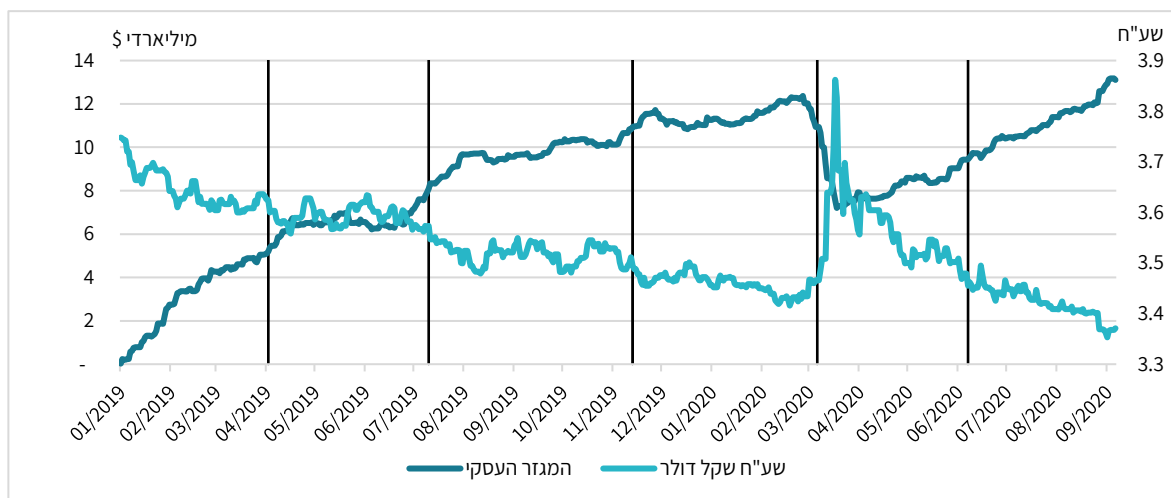
לצורך איתור המגמות השונות, הוטמעה שיטה סטטיסטית בשם Structural Break. (השיטה מתבססת על אלגוריתם Bai and Perron, 2003). – האלגוריתם מחשב את החלוקה המיטבית של הסדרה העתית למספר רגרסיות שונות, חלוקה שנקבעת באופן אנדוגני ושמובילה למודל שממזער למינימום את סכום השאריות בריבוע ובכך מסביר יותר טוב את הסדרה העתית. לדוגמה, ניתן לראות באיור 5 ששני קווי רגרסיה במקרה הזה מסבירים טוב יותר את הסדרה מאשר קו הרגרסיה שנפרש על כל המקטע.

איור 5: המחשה גרפית של שיטת Bai & Perron



השיטה הופעלה על סדרות עתיות של רכישות מט"ח מצטברות של המגזר העסקי, להלן התוצאות:

איור 6: שע"ח היציג שקל/דולר ורכישות מט"ח נטו מצטברות של המגזר העסקי⁶ (1/2019 - 9/2020)



המקור: בנק ישראל

⁶ חשוב לציין שבניתוח זה כללנו במגזר העסקי רק חברות שניתן לזהות את פעילותן באופן פרטני.

התקבלה חלוקה מובהקת עבור כל קבוצה ונמצאו שברים מבניים שמחלקים את התקופה למספר תתי-תקופות. בנוסף, השברים המבניים זהים ברכישות המצטברות של המגזר הריאלי ובשער החליפין שקל\דולר, קרי, אותן שש תקופות מאפיינות מגמות בשתי הסדרות, מה שמעיד כנראה על קשרי גומלין בין הפעילות של המגזר העסקי לשינויים בשער החליפין.

ג. זיהוי פעילות חריגה בשוק

לצורך איתור פעילות חריגה של מגזר או שחקן מסוים, נבנתה מערכת התראות שלוקחת בחשבון את מאפייני המסחר המרכזיים בפעילות של כל קבוצה או שחקן (מחזורי המסחר, רכישות מט"ח נטו וכו'). בכל יום מבוצעת בדיקה שמבוססת על שיטת ה-MAD כדי לאתר פעילות. במידה שאותרה פעילות חריגה, מתבצעת בדיקה שלא מדובר בטעות דיווח, אלא בחריגה אמיתית בפעילות של אותו שחקן או של אותה קבוצה ואז מבוצע ניתוח של הפעילות החריגה, בהתאם למשתנים כלליים שקשורים לשוק המט"ח ושעשויים להשפיע על הפעילות של השחקנים/הקבוצות באותו יום. הניתוח מאפשר למצוא את "הסיפור" שמאחורי החריגות כדי להבין עוד על השוק ועל הדינמיקה שמתרחשת בו.

למשל, שחקן א' שעוסק בפעילות יצוא ואופי הפעילות שלו הוא מכירה מתונה של דולרים לאורך השנה (למשל, 10 מיליון דולר בכל פעם), כנראה בהתאם לתקבולים שהוא מקבל מחו"ל. במידה וביום מסוים א' מוכר דולרים בכמות גדולה (לדוגמה, 100 מיליון דולר), המערכת תתריע שיש פעילות חריגה של א', בצירוף משתנים מרכזיים אודות הפעילות שלו (ממוצע, חציון, סטיית תקן ועוד), לצורך בחינה ידנית של הנתונים ופנייה לא' במידת הצורך. באמצעות שיטה זו ניתן לאתר ימים חריגים או שינויים בדפוסי ההתנהגות של השחקנים בשוק, דבר שתורם רבות למעקב אחר השוק ולהיכרותו.

4. טיוב הנתונים

בשולי עבודה זו אנו רואים חשיבות בהוספת סקירה קצרה על חלק מהבקורות הלוגיות שקיימות במערכת. חשיבותה של סקירה זו נובע מכך שמודל סטטיסטי, טוב ככל שיהיה, אינו מועיל במאומה אם מסד הנתונים שעליו נבדק המודל, מכיל מידע שגוי וחסר.

בעת קבלת נתוני הדיווח היומיים מהגופים המדווחים, מבוצעות מספר בקורות לוגיות שמוודאות שכל שדות החובה דווחו, שהנתונים תואמים למבנה השדה המיועד ושמתקיימים תנאים לוגיים בסיסיים בעסקה (לדוגמה, לא ניתן לדווח על עסקה שתאריך הקשירה שלה התרחש לאחר תאריך הפירעון). בנוסף, לאחר קליטת הנתונים, מבוצעת הצלבה של הדיווחים של גופים מדווחים – עסקה שנקשרה בין שני גופים מדווחים, אמורה להיות מדווחת על ידי שניהם עם פרמטרים זהים (תאריך, סכומים, שע"ח וכו'). בדיקה זו מאפשרת גילוי של דיווחים חסרים או שגויים ובכך מסייעת אף היא לשיפור איכות הנתונים במערכת.

אולם, כפי שיתואר להלן, אין זה מחייב שנתונים אשר "עברו" את כל הבקורות הנ"ל הם נתונים תקינים. לכן נתאר כעת בקצרה מספר כלים לאיתור חריגים בנתונים.

א. איתור נתונים חריגים

• **סכום העסקה** - כפי שצוין לעיל, ייתכן ועסקות שעומדות בכל הכללים שהוגדרו בבקורת, נכנסו למערכת, למרות היותן עסקות שגויות. לדוגמה, עסקת המרה של 1 מיליארד דולר אמנם אפשרית, אך כשמדובר בגוף שהיקף השנתי שלו מסתכם בכמה עשרות מיליוני דולרים, מדובר בעסקה שחשודה כשגויה.

לאחר בחינה מעמיקה של הנתונים ומספר שיטות סטטיסטיות, נמצא שהתפלגות סכומי העסקות שנקשרות לאחר התמרה של Log, הינה נורמלית בקירוב. בנוסף, נראה שהשימוש בממוצעים וסטיות תקן, הוביל לסטיית תקן גדולה בגלל תצפיות חריגות, דבר שהקשה על קביעת מספר סטיות התקן לצורך הגדרתו של נתון כחריג. לכן בוצע שימוש ב-MAD לצורך איתור עסקות בסכומים חריגים, שכן שיטה זו לא מושפעת מתצפיות חריגות.

בכל יום מבוצעת בדיקה של עסקות חריגות ואלה נבחנות בהתאם למאפייני השוק, השחקן וגורמים מקרו-כלכליים נוספים, לצורך הקביעה האם העסקה היא עסקה הגיונית או שכנראה מדובר בטעות דיווח. במקרה שמתקבל הרושם שמדובר בטעות דיווח, מבוצעת פנייה לגורם המדווח, לצורך אישור העסקה או תיקונה.

• **בדיקת כמות רשומות שדוחו** - כאמור לעיל, כל גוף מדווח שולח דיווח יומי על העסקות שקשר בשוק ה-OTC ביום העסקים שקדם לו. לעיתים קיימת תקלה אצל אחד הגופים המדווחים ורק חלק מהעסקות מדווחות בקובץ הדיווח היומי. לצורך איתור

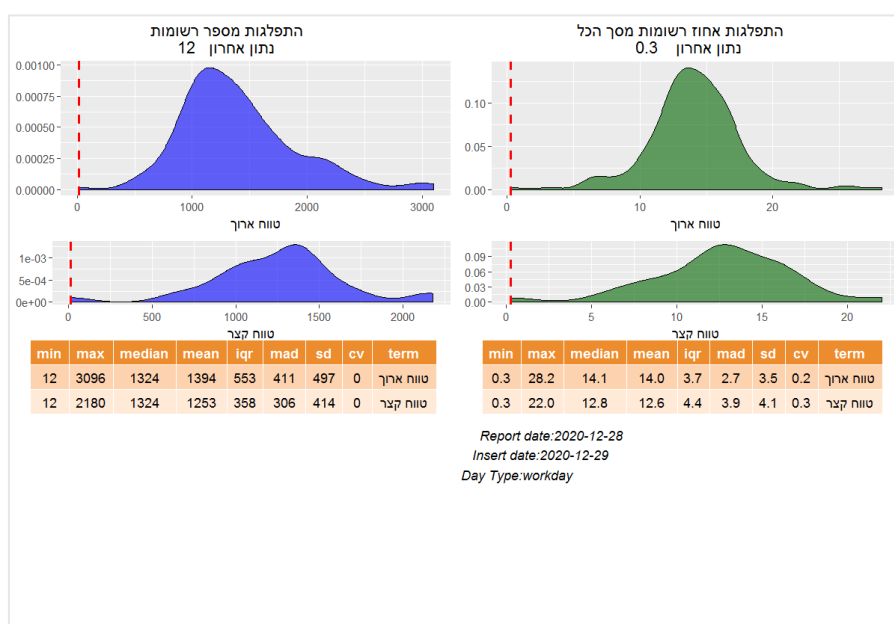
דיווחים מסוג זה (שקשים יותר לאיתור בהשוואה לאי-דיווח), נדרשה החטיבה להתמודד עם מספר סוגיות, בהן, כיצד להבחין בין ימי מסחר ערים לימי מסחר דלילים, התחשבות בשינויים אקסוגניים שמשפיעים על המסחר השקלי, מהו מדד הבסיס להשוואה, ועוד.

לצורך בחירת המודל הנכון ביותר, נבחנו במשך תקופה שתי גישות שונות - בחינת כמות העסקות שדווחו על ידי גוף מדווח בהשוואה למספר העסקות שדווחו על ידו בטווח קצר והארוך, לעומת שיעור העסקות של הגוף המדווח כאחוז מסך העסקות שדווחו, בחלוקה בין מדווחים מקומיים וזרים (גישה זו מניחה שיש קשר בין השינוי בכמות העסקות המדווחות על ידי גוף אחד לבין השינוי בכמות העסקות אצל כלל המדווחים).

כפי שניתן לראות באיור 7 להלן, התפלגות הדיווח של הגופים המדווחים הינה התפלגות עם זנב ימני ארוך (תופעה שמכונה צידוד - Skewness). ממצא זה אינו מפתיע, כיוון שמספר העסקות המדווחות הוא בהכרח מספר אי-שלילי ובד"כ הוא יהיה מרוכז סביב ממוצע\חציון, כשקיימות תצפיות גבוהות שמרכיבות את הזנב הימני. בשל המבנה של ההתפלגות, לא ניתן היה להשתמש בממוצע ובסטיית התקן לצורך איתור חריגים.

המודל הסטטיסטי שהוחל על הנתונים הוא MAD, כשדיווח חריג הוגדר כסטייה של $2 * MAD$ מהחציון. כפי ניתן לראות באיור 7 להלן, הקו האדום הוא הדיווח האחרון של הגוף המדווח ומאחוריו ניתן לראות את ההתפלגות על פני זמן (טווח קצר ומעליו טווח ארוך), כשבחלקו הימני של האיור מופיעה ההתפלגות של שיעור העסקות של המדווח כאחוז מסך העסקות ובחלקו השמאלי מופיעה התפלגות כמות העסקות שדווחו על ידו. בנוסף, לכל ניתוח קיימת טבלה עם הנתונים מרכזיים (ממוצע, חציון, מינימום, מקסימום, ועוד).

איור 7: חריגים בדיווח - התפלגות מספר הרשומות וחלקן היחסי



המקור: דיווחי הבנקים ועיבודי בנק ישראל

לאחר תקופה שבה נבחנו שתי הגישות, נמצא שיש יציבות רבה יותר בנתון של שיעור העסקות כאחוז מסך העסקות בשוק ולכן שיטה זו נבחרה. יצוין שהטמעת תהליך זה הוכחה כיעילה. מאז שהוטמע התהליך, אותרו דיווחים חריגים רבים, שללא התהליך הזה הם היו מתגלים באיחור ניכר, דבר שהיה פוגע באיכות הנתונים במערכת. לדוגמה, במהלך חודש יוני 2020 התגלה דיווח חריג וביורור מול הגוף המדווח העלה שבשל תקלה, דווחו רק כ-3% מהעסקות שנקשרו.

5. סיכום

בעבודה זו הוצגו חלק מהכלים ומהשיטות שמשמשים את החטיבה למידע וסטטיסטיקה בבנק ישראל לצורך טיוב, עיבוד והפקת תובנות מהמידע שקיים במערכת הנתונים הייחודית על מכשירים פיננסיים נגזרים. כמות המידע שנאסף מצריכה שימוש בכלים סטטיסטיים מתקדמים, לשם בקרה והפקת משמעויות מהנתונים שמתקבלים.

בחלק הראשון תוארו מספר מודלים שמשמשים להשלמת המידע ולהפקת תובנות ממנו, לצורך קבלת תמונה שלמה ומקיפה על הפעילות של השחקנים השונים. השיטות שתוארו מגיעות מעולם מדע הנתונים, שיטות מסוג PCA בשילוב עם k-MEANS, למציאת אשכולות של שחקנים ושיטת LDA, לסיווג שחקנים שונים במגזר העסקי. השתמשנו בשיטת LDA כדי למצוא את המאפיינים השונים של תתי-קבוצות בתוך המגזר ומצאנו הבדלים מובהקים בין יבואנים לייצואנים.

בחלק השני סקרנו חלק מהבקורות הסטטיסטיות שהוטמעו במערכת לצורך איתור חריגים, תוך סקירה של הסוגיות שעמדו בבסיס בניית הבקורות ובחירת הכלי הסטטיסטי המתאים ביותר בכל בקרה.

מטרת העבודה היא לתת סקירה כללית על חלק מהשיטות הסטטיסטיות שמשמשות את החטיבה לצורך בקרה ועיבוד של הנתונים.