חיזוי שינויים במדד מחירי הפירות והירקות בעזרת נתוני מאגר המחירים הקמעונאים

דור גולדנברג ויונתן רוזן*

תקציר

מדד המחירים לצרכן מתפרסם אחת לחודש על ידי הלשכה המרכזית לסטטיסטיקה ובפיגור של 15 יום מתום החודש הנמדד. בתום כל חודש מופקת בבנק ישראל עוד במועד מוקדם יותר ושקרוב לתום החודש, תחזית למדד המחירים לצרכן שמטרתה לאפשר את הניתוח של המידע. סעיף הפירות והירקות, על אף משקלו הנמוך במדד המחירים לצרכן, הוא תנודתי מאוד ועל כן תרומתו הפוטנציאלית לטעות בתחזית למדד היא משמעותית. בשנים האחרונות החל בנק ישראל באיסוף נתונים של מחירי המוצרים שנמכרים בסניפים של קמעונאי המזון הגדולים בארץ. מידע זה שמדווח על פי חוק[1], מתקבל מקבצים שנשלחים לאתרי אינטרנט מיוחדים של הקמעונאים הגדולים ונקלט במאגר ייעודי שהוקם בבנק בשם "מאגר המחירים הקמעונים". בעבודה זו מוצגת תחזית לסעיף הפירות והירקות במדד המחירים לצרכן, שמבוססת על המחירים שנקלטים ב"מאגר הנתונים הקמעונים". שילוב של תחזית זו עם התחזיות האחרות שקיימות כיום בבנק ישראל, שיפרה בתקופת הזמן שנבחנה את טיב החיזוי בכ-23%. פרויקט זה, שעושה שימוש בנתוני עתק (ביג דאטה) לשיפור תחזיות קיימות, הוא הראשון מסוגו בבנק ומהווה תכנית הרצה לשימוש נרחב יותר בנתוני עתק לצרכי מדיניות.

* חטיבת המחקר והחטיבה למידע ולסטטיסטיקה, בנק ישראל.

[1] החוק לקידום התחרות בענף המזון.

1. הקדמה

היעד העיקרי במדיניות המוניטרית בבנק ישראל, כמו בבנקים מרכזיים רבים בעולם, הוא שמירה על יציבות המחירים. בנק ישראל מחויב לחתור להשגת יעד האינפלציה שנקבע ע"י הממשלה בהתייעצות עם הנגיד ומתבסס על מדד המחירים לצרכן שמפרסמת הלמ"ס מידי חודש. מדד המחירים לצרכן משמש למדידת אחוז השינוי שחל על פני זמן ב"סל קבוע" של מוצרים ושירותים. סל זה כולל סעיפים, סעיפי משנה ותתי סעיפים שונים. סעיפים עיקריים הם מזון, פירות וירקות (פו"י), הלבשה והנעלה, דיור, בריאות ועוד. ככלי עזר בניהול המדיניות המוניטרית, מפיקים בבנק ישראל מידי חודש תחזית למדד המחירים ברמת סעיפיו, שמקדים את פרסום המדד ב-15 יום.

תת סעיף הפירות והירקות הטריים מהווה כ-2%[2] מהמדד. החשיבות של חיזוי מדויק של סעיף הפירות והירקות במדד זניחה, לכאורה, בשל משקלו הנמוך. יחד עם זאת, מחירי הפירות והירקות נוטים להיות תנודתיים ביותר וקשים לחיזוי, כך שהטעות בחיזוי סעיף הפו"י גדולה לעיתים בהרבה מהטעויות בסעיפים האחרים. בעקבות זאת תרומת הטעות בחיזוי פו"י לטעות בחיזוי המדד הכללי עולה בהרבה על חלקו במדד ומגיעה ל-0.05 נקודות אחוז בממוצע ולכ-0.1-0.25 נקודות אחוז במקרים הקיצוניים[3].

השימוש במקורות נתונים חדשים, ובפרט בנתוני עתק, לשם חיזוי מדדי המחירים הרשמיים של הלשכות המרכזיות לסטטיסטיקה הפך נפוץ בשנים האחרונות. דוגמה מרכזית היא "פרויקט מיליארד המחירים" (Billion prices project), מיזם אקדמי משותף של שני חוקרים מהאוניברסיטאות הרווארד ו-MIT (Cavallo and Rigobon, 2016). פרויקט זה שעוסק בחיזוי מדדי המחירים הרשמיים של קבוצה רחבה של מדינות באמצעות נתוני עתק בתדירות גבוהה, מתמקד באיסוף מחירים של מוצרים שנמכרים באופן מקוון (אונליין) וזאת משתי סיבות עיקריות: ראשית, מדובר לרוב בנתונים זמינים מאוד שניתן לאסוף אותם בתדירות גבוהה, בקלות יחסית ובעלות נמוכה. שנית, קיימת הנחה שמחירים אלה מהווים קירוב טוב למדדי המחירים הרשמיים, שמבוססים לרוב על מחירים בחנויות פיזיות. בעבודה זו עשינו שימוש בנתוני מאגר המחירים הקמעונים שמספק נתונים על מחירי פו"י בחנויות פיזיות בתדירות יומית ובכך אנו מתגברים למעשה על הצורך באיסוף מחירים מקוונים[4].

עבודה זו מצטרפת לתכנית הרצה שנערכה לאחרונה בבנק המרכזי של שבדיה ושכלל ניסיון לחיזוי שינוי המחיר של סעיף הפו"י בשבדיה על בסיס מחירים שנאספו באופן מקוון (Hull et al., 2017). התוצאות בתכנית ההרצה מעודדות ומצביעות על תועלת מסוימת שיש בשימוש במחירים מקוונים לחיזוי סעיף הפו"י.

בעבודה זו נתאר את "מאגר הנתונים הקמעונים" ואת השימוש בו לחיזוי שיעורי השינוי בסעיף פירות וירקות במדד המחירים לצרכן. נתחיל בתיאור הנתונים, בהצגת האתגרים שניצבים בפנינו בעבודה עם המאגר, נמשיך באמידה ובתוצאות ונסיים בדיון בכיוונים להמשך.

2. הנתונים

א. משתנה המטרה: מדד פירות וירקות של הלמ"ס

הלמ"ס אוספת מדי חודש את המחירים של פירות וירקות טריים בהתאם לסל הצריכה של משקי הבית (שנדגם במסגרת סקר הוצאות משקי הבית). בהתבסס על מחירים אלה מרכיבה הלמ"ס את "מדד פירות וירקות" שבאמצעותו מחושב שיעור השינוי של סעיף הפירות והירקות במדד הכללי[5]. המחירים נדגמים על ידי סוקרים ששולחת הלמ"ס ל"שכבות" השונות: ירקניות, רשתות קמעונאיות ושווקים. מדד הפו"י מורכב מתוך המחירים שנדגמים בשכבות השונות. המדד נותן משקל לכל פרי וירק באופן שתואם את משקלו בסקר הוצאות משקי הבית. המדד מתחשב כמו כן בהיבטים שונים של עונתיות.

ב. נתוני מאגר המחירים הקמעונים

ועדת קדמי, הוקמה בקיץ 2011 לאחר "מחאת הקוטג'", במטרה לבחון את רמת התחרותיות בשוק המזון. בעקבות ממצאי הוועדה, עבר בכנסת בשנת 2014 "חוק קידום התחרות בענף המזון" (להלן: "חוק המזון") אשר מחייב את הרשתות הקמעונאיות הגדולות

[2] סעיף הפירות והירקות כולל רכיב נוסף במשקל של 1% מסך המדד – ירקות ופירות קפואים, כבושים ומשומרים.
[3] "תרומת הטעות" מוגדרת כטעות בחיזוי הסעיף כפול המשקל שלו במדד כולו.
[4] למעשה, המאגר כולל גם מידע על "חנויות" מקוונות.
[5] ראו לדוגמה, עלון סטטיסטי-קל 153, מדד המחירים לצרכן, אוקטובר 2016, הלשכה המרכזית לסטטיסטיקה.

לפרסם את מחירי כל המוצרים שהן מוכרות בכל סניפיהן בכל יום. לאור זאת, החליט בנק ישראל להקים מאגר הנקרא מאגר המחירים הקמעונאיים (להלן: "המאגר") אשר יקבץ את הנתונים שמפרסמות הרשתות במקום אחד שיאפשר עיבוד וניתוח של הנתונים.

3. אתגרים

המאגר כולל את המחירים של כל המוצרים בכל הסניפים של רשתות המזון הקמעונאיות הגדולות בארץ ומאפשר לראות נתוני אמת של שכבה זו. יחד עם זאת, חיזוי מדד הפו"י באמצעות המאגר מציג מספר אתגרים שנתאר להלן.

א. נתונים חסרים

המאגר כולל רק את מחירי המוצרים שנמכרים ברשתות הקמעונאיות. בתור שכזה, המאגר לא כולל את כל המחירים שמהם מורכב המדד, שכן הוא מורכב גם ממחירי פירות וירקות בחנויות שלא מופיעות במאגר, כמו שווקים, מכולות וירקניות. לפיכך, במידה שהשכבות האחרות מתנהגות באופן דומה לשכבת הרשתות הקמעונאיות, השינוי במחירי המאגר ישקף את כל השינוי שמשתקף במדד. לחילופין, אם המחיר בשווקים יורד כשהרשתות הקמעונאיות מעלות את מחיריהן – השינוי במחירי המאגר ישקף את העלייה במחירי הרשתות הקמעונאיות ולא את הירידה במחירי השווקים. מכיוון שיש תחרות מסוימת בין השכבות השונות, סביר להניח שיש מתאם חיובי בין פעולותיהן. יחד עם זאת, התנהגות שונה של המחירים בשכבות האחרות, מהווה מקור לטעות בחיזוי.

ב. עומק היסטורי

תדירות פרסום המדד היא חודשית ומכיוון שהמאגר כולל נתונים של כ-4 שנים בלבד, מספר התצפיות שבהן יש לבנק נתונים, הן על מדד הפו"י והן על המחירים של המוצרים שמרכיבים את המדד, עומד כעת על כ-54 בלבד[6]. הכמות המועטה של הנתונים לא מאפשרת לתקף הנחות במודל פרמטרי על ידי שימוש בנתונים מחוץ למדגם (Test Data) ומעלה חשש ממשי להתאמת יתר (overfitting). לאור זאת, לא ניתן בשלב זה לאמוד באופן מעשי מודל סטטיסטי שדורש כיול של פרמטרים. כדי להתגבר על בעיה זו, השתמשנו בשיטה שמבוססת רק על מידע מלכתחילה.

ג. נתוני עתק

פרויקט זה, שמבוסס על נתוני עתק, הוא הראשון מסוגו בבנק. ככזה הוא מציב אתגרים שלא קיימים בעבודה עם נתונים שגרתיים.

המאגר כולל כמות עצומה של נתונים (כ-10 מיליארד תצפיות), שלא מאפשרים עבודה בכלים הרגילים. כלים של ענן מאפשרים אחסון כמות כזו של נתונים וכן יכולות שליפה ועיבוד מבוזרים שנדרשים לעבודה יעילה על כמות כזו של נתונים.

הנתונים מגיעים בצורות שונות ועיבוד הנתונים הקיימים לנתונים שאפשר לעבוד אתם, זו משימה מאתגרת. נתונים חסרים, החלפה בין העמודות, טקסטים בעייתיים – כל אלה מקשים את העבודה עם הנתונים הקיימים. אמנם, כל הנתונים מתקבלים באותה תבנית, אך יש שדות מרכזיים שבהם אין אחידות בדיווח בין רשתות ואף בתוך רשת. לדוגמה, לכל רשת יש מק"ט שונה לכל פרי וירק. יתירה מזו, בתוך רשת יש מספרי קטלוג (מק"ט) שונים לאותו מוצר ולפעמים יש מוצרים שונים שחולקים את אותו המק"ט[7]. בפרט, זיהוי מוצר בתוך רשת ועל פני רשתות הוא משימה קשה [8].

[6] בפועל אנו משתמשים ב-40 תצפיות בלבד, עקב איכות נמוכה של המאגר לפני שנת 2017.

[7] סוג המוצר נקבע לפי המלל שמתאר את המוצר, אולם אין לשלול את האפשרות שהמלל הוכנס בטעות והמק"ט באמת לא מקודד את המוצר המתואר. כמו כן, פעמים רבות שם של פרי או ירק מופיע במאות מוצרים שונים.

[8] מה גם שאין לשלול מצב שבו אותו מק"ט משמש בחודש אחד למוצר מסוים ובחודש אחר למוצר אחר.

4. אמידה

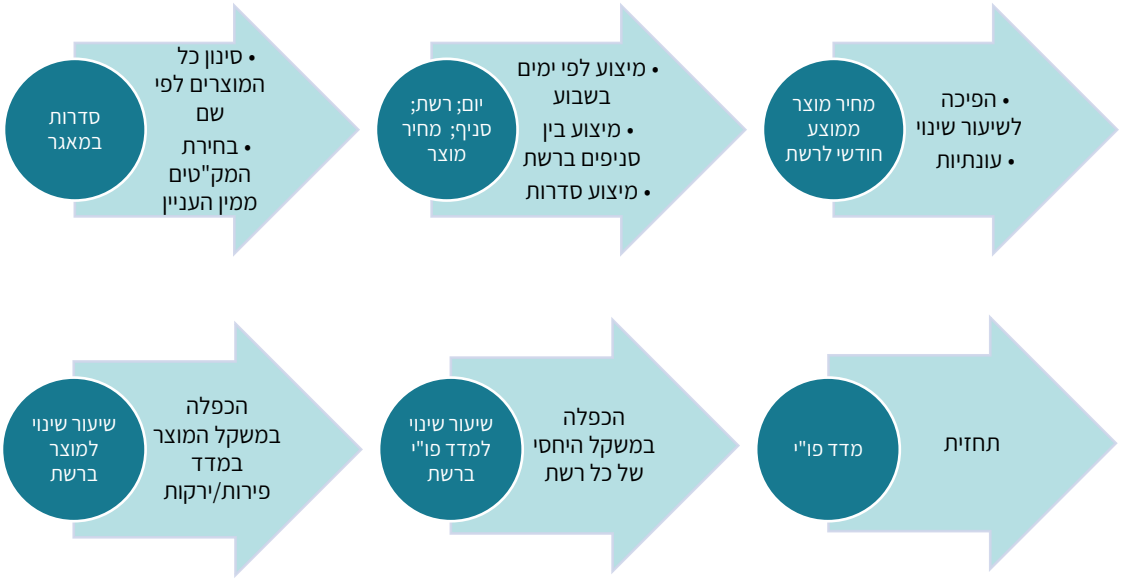
בחלק זה נסביר את שיטת האמידה. שיטה זו משקללת את המחירים הפרטניים בסניפים לכדי מדד פו"י אחד באופן הבא. התחזית נבנתה על בסיס מחירי פירות וירקות של הרשתות הקמעונאיות הגדולות[9].

השלב הראשון בתהליך הוא זיהוי הפירות והירקות הספציפיים בכל רשת, דהיינו, מציאת המק"טים ממין העניין שבאמצעותם אפשר להרכיב סדרת מחירים אחת לכל רשת. שלב זה נעשה על ידי סינון כל המוצרים לפי שם ובחירת המק"טים שתיאור המוצר שלהם מתאים לפרי או לירק הרצויים[10]. בשלב השני, המחירים היומיים אוחדו לכדי סדרה חודשית לכל מק"ט על ידי מיצוע משוקלל בין ימות השבוע, כשסוף השבוע קיבל משקל גדול יותר מתחילתו. הסדרות החודשיות של המק"טים אוחדו לכדי סדרה אחת לכל פרי וירק, על ידי מיצוע על פני כל הסניפים שבתוך הרשת ומיצוע על פני כל המק"טים שמתאימים לתיאור של המוצר בתוך הרשת. התוצאה של שלב זה היא סדרת מחיר חודשי ממוצע לכל פרי וירק בכל רשת.

סדרת מחיר חודשי מאפשרת לחשב את שיעור השינוי של המחיר של כל פרי וירק בכל רשת. שיעורי השינוי של כל המוצרים אוחדו לכדי שיעור שינוי של מדד פו"י על ידי ממוצע משוקלל בהתאם למשקל של כל פרי וירק במדד. כדי למנוע הטיה של המדד על ידי מוצרים שאינם בעונה[11], שיעור השינוי של המדד נקבע רק על פי המוצרים שנמצאים בעונה[12].

בשלב האחרון, מדדי פו"י שחושבו לכל רשת בנפרד אוחדו לכדי תחזית אחת, כאשר כל רשת קיבלה משקל שונה[13]. התהליך כולו מתואר בתרשים 1 להלן[14]. לבסוף, תחזית זו מוצעה עם התחזיות הקיימות כיום בבנק לכדי תחזית סופית למדד פו"י. הסיבה לכך היא שחיזוי סטטיסטי מטבעו איננו מדויק ובמקרה הטוב מספק קירוב טוב למשתנה המטרה. באופן טיפוסי, גם אין חזאי בודד אחד שבאופן עקבי מדייק ביחס לאחר. יש מחקרים שמראים אמפירית שמיצוע תחזיות פותרת באופן חלקי בעיות אלה במובן שהיא עמידה בפני טעויות של מודל חיזוי ספציפי (ראו למשל:Altavilla and De grauwe 2006).

תרשים 1. תיאור מהלך אמידת התחזית



[9] כאמור, למוצרים אין שם או מק"ט אחידים בין רשתות ובתוך רשת ועל כן זיהוי פרי או ירק אינו פשוט כלל ועיקר. עקב קושי זה, תהליך בחירת המוצרים לא נעשה בצורה אוטומטית ועל כן התחזית נעשתה על ידי רשתות השיווק הגדולות תחת ההנחה ששינוי המחירים ברשתות הגדולות יקרב בצורה טובה את שינוי במחירים בכל הרשתות הקמעונאיות.

[10] כדי להקטין את מרחב החיפוש נעשה סינון נוסף באמצעות יחידת המידה של המוצר. בניגוד למוצרים אחרים, רוב הפירות והירקות הטריים נמכרים ביחידות של ק"ג.

[11] העונה של פרי או ירק נקבעה לפי כמות החנויות שבהן היא נמכרת על פני הרשת.

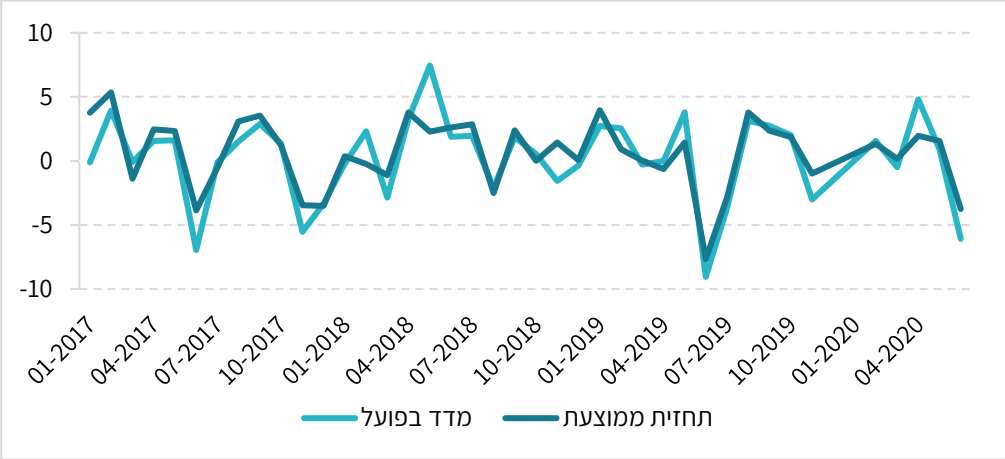
[12] בניסוח אחר, שיעור השינוי שניתן למוצרים שאינם בעונה הוא שיעור השינוי של שאר המוצרים שכן בעונה וכך מוצרים שאינם בעונה לא יטו את המדד המתקבל.

[13] לפי גודל יחסי.

[14] כל האיורים והתרשימים מצורפים בסוף המסמך.

5.תוצאות

איור 1. מדד פו"י מול התחזית הממוצעת (קמעוניים ומשרד החקלאות)



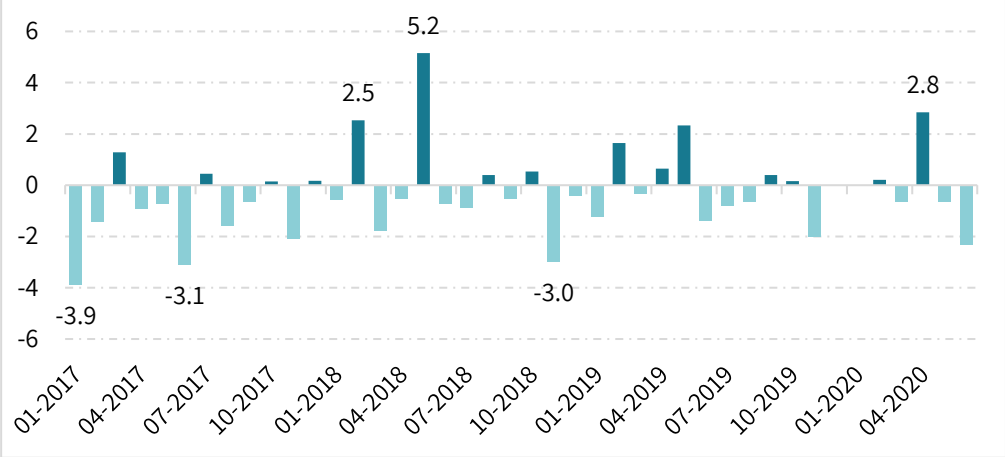
המקור: מאגר המחירים הקמעונאיים ועיבודי בנק ישראל

התחזית שמופקת מהמאגר ומהתחזיות שקיימות כיום בבנק הושוו למדד פו"י שמפרסמת הלמ"ס על פני התקופה 1/2017-6/2020. באיור 1 ניתן לראות את התחזית שממצעת בין התחזית שקיימת בבנק לתחזית שמופקת מהמאגר, לעומת מדד פו"י שפרסמה הלמ"ס. טיב החיזוי נבחן באמצעות שורש הטעות הריבועית הממוצעת (RMSE), שמחושב באופן הבא:

$$\sqrt{\sum_{t=1:T}(y_t - \hat{y_t})^2}$$

הסבר של הנוסחה: y_t הוא ערך מדד פירות וירקות האמיתי בזמן t ו- (y_t)^הוא ערך חזוי (מיצוע התחזיות) בזמן t. חישוב זה מראה שמיצוע התחזיות שקיימות בבנק, עם התחזית שמופקת מהמאגר, משפר בתקופה שנבחנה בכ-23%. להלן מספר היבטים של התוצאות. ראשית, באיור 1 ניתן לראות את התחזית שממצעת בין התחזית שקיימת בבנק לתחזית שמופקת מהמאגר, לעומת מדד פו"י מהלמ"ס. אם בוחנים את ה-RMSE של התחזית שקיימת בבנק ושעומד על כ-2.2 לעומת RMSE של 2.3 בתחזית מהמאגר - הרי שאין שיפור[15].

איור 2. שגיאות התחזית המוצעת ביחס למדד פו"י של הלמ"ס

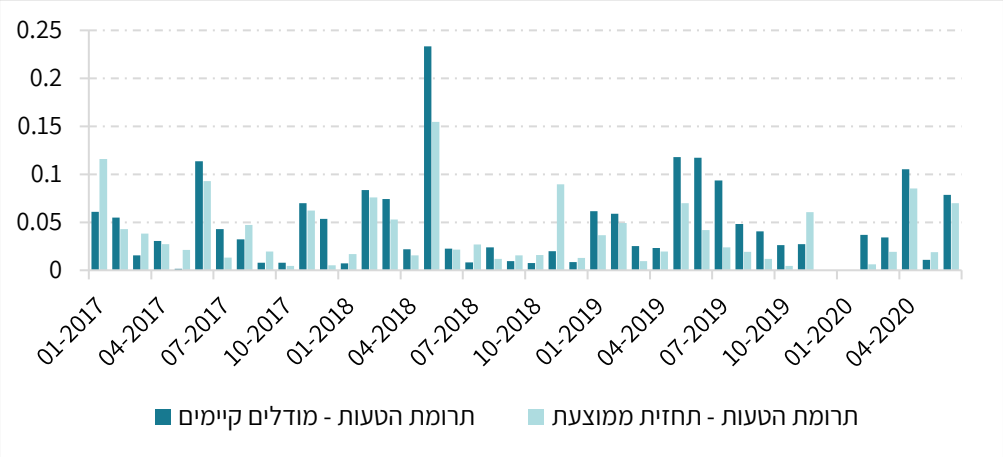


המקור: מאגר המחירים הקמעונאיים ועיבודי בנק ישראל

[15] מדדים אלה חושבו על נתונים מינואר 2017 ועד יוני 2020 (ללא חודשי דצמבר 2019 וינואר 2020). מדד ה-RMSE "מעניש" את התחזית על טעויות גדולות ולכן נבחר.

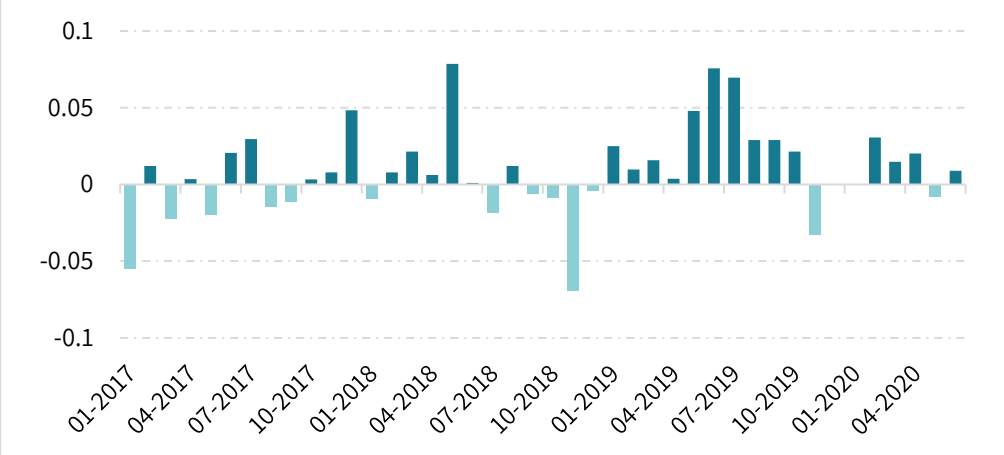
אך כשממצעים את שתי התחזיות, מקבלים RMSE של 1.7 – שיפור של כ-23% ביחס לתחזית שקיימת בתקופה שנבחנה. כפי שניתן לראות באיור 2, כמעט כל הטעויות של התחזית המוצעת הן בטווח של פחות מסטיית תקן אחת של מדד פו"י (3.3) ולרוב הרבה פחות מכך. סיכום של הביצועים של התחזיות השונות, כולל טעות ממוצעת וטעות חציונית, מובא להלן בטבלה 2.

איור 3. התרומה לטעות בחיזוי מדד המחירים לצרכן



המקור: מאגר המחירים הקמעונאיים ועיבודי בנק ישראל

איור 4. השוואה בין הטעות של התחזית הקיימת לתחזית המוצעת – הפרש הערך המוחלט בין התחזיות



המקור: מאגר המחירים הקמעונאיים ועיבודי בנק ישראל

השיפור מתבטא גם בתרומה של סעיף הפו"י לטעות הכוללת בחיזוי מדד המחירים לצרכן. באיור 3 ניתן לראות שהתרומה לטעות של התחזית המוצעת קטנה יותר (בערך מוחלט) מהתרומה לטעות של התחזית הקיימת ולעיתים בפער גדול. לדוגמה, ביולי 2019, התרומה לטעות של התחזית הקיימת היא 0.09 נקודות אחוז ואילו התרומה לטעות של התחזית המוצעת באותה התקופה היא 0.02 נקודות אחוז, שיפור של 77%. התרומה לטעות של התחזית המוצעת לא עולה על 0.15 נקודות אחוז ובממוצע היא 0.038 נקודות, שיפור בשני הפרמטרים ביחס לתחזית הקיימת (טעות מרבית 0.23 נקודות אחוז, וממוצעת 0.048 נקודות אחוז). באיור 4 ניתן לראות את ההפרש שבין הערך המוחלט של התרומה לטעות של התחזית הקיימת לבין הערך המוחלט של התרומה לטעות של התחזית המוצעת[16]. ניתן לראות שברוב המקרים התחזית המוצעת טובה יותר וגודל הטעות משמעותי יותר בתחזית הקיימת.

[16] נסביר ביתר פירוט את משמעות הגרף: ערך חיובי בגרף פירושו שהתחזית המוצעת טובה יותר ולהפך. גובה העמודה משקף את גודל הטעות. לדוגמה, ערך חיובי גבוה פירושו טעות גדולה של תחזית הבנק ביחס לתחזית המוצעת וערך שלילי גבוה פירושו טעות גדולה של התחזית המוצעת ביחס לתחזית הבנק.

טבלה 2. ביצועי התחזיות השונות

מודל	RMSE	טעות ממוצעת	טעות חציונית
מודל מחירי משרד החקלאות (שקיים במחקר כיום)	2.2	1.6	1.0
מאגר קמעוניים	2.3	1.9	1.6
ממוצע של השניים (התחזית המוצעת)	1.7	1.3	0.8

המקור: מאגר המחירים הקמעונאיים ועיבודי בנק ישראל

באופן דומה, ניתן להשוות בין התחזיות ביחס לטעות במדד הכללי. שימוש בתחזית המוצעת משפר את התחזית של המדד הכללי ב-5% במונחי RMSE. בפרט, ביולי 2019 הטעות של התחזית הקיימת היא 0.23 נקודות אחוז, לעומת 0.16 בלבד של התחזית המוצעת, שיפור של 30%.

6. מחשבות להמשך וסיכום

הראינו שהתחזית שמתקבלת ממיצוע הנתונים מהמאגר עם התחזית שקיימת כיום בבנק, משפרת את יכולת החיזוי של סעיף פו"י במדד המחירים לצרכן בתקופה שנבחנה. יתר על כן, ניתן לשפר את התחזית שמופקת מהמאגר בכמה דרכים וביניהן שימוש בנתונים מטויבים, שימוש במחירים לאחר מבצע, שילוב תחזיות ממאגרים נוספים ועוד.
לסיכום, ראינו שניתן לשפר את התחזית למדד פו"י על ידי מידע פרטני ורציף. זו תחילתה של מגמה של שימוש הולך ומתרחב בנתוני עתק, כזו שתאפשר מתן מענה לשאלות חדשות וליצירת דיוק רב יותר של תחזיות קיימות.