

תהליך אמידה במפקד האוכלוסין 2022

תהליך האמידה הוא השלב שבו אומדים ומעריכים את גודל האוכלוסייה ואת תכונותיה ברמה גאוגרפית מפורטת באמצעות שיטות סטטיסטיות לפי הנתונים שהתקבלו מהמדגם. במפקדים מסורתיים שבהם ספרו את כל האוכלוסייה בסריקת שטח מדינת ישראל ועברו מבית לבית, הייתה פעולה זו יחסית פשוטה ואף לא סטטיסטית. בסיום המפקד היה צורך יחיד: להעריך את מספר האנשים שלא הצליחו לספור (ספירת חסר) ואת מספר האנשים שנספרו אף שאינם שייכים לאוכלוסייה (ספירת יתר). כאשר מדובר במדגמים ובמיוחד בכאלה שהוצאו לפי שיטות דגימה מורכבות, אמידה (קביעת משקל) הופכת להיות משימה מאתגרת ומורכבת. אי-השבה (נדגמים שלא משיבים) מקשה עוד יותר על קביעת משקל לכל פרט ומשק בית שהשיב, משקל שמשקף את מספר הנפשות או משקי הבית שהוא מייצג.

לפני תהליך האמידה, נעשו מספר פעולות כדי לוודא שהמשיבים שייכים לאוכלוסיית המטרה (אוכלוסיית המפקד). לשם כך הרשומות האלה, שאינן שייכות לאוכלוסיית המטרה (מקרי אפס), נמחקו מתוך רשימת המשיבים:

- פרטים במסגרת ברשימת הנדגמים שנפטרו או ירדו מהארץ לפני המועד הקובע (לפי נתונים מינהליים עדכניים למועד הקובע ולפי מידע מהמשיבים)
- שיוך פרטים לאוכלוסיות מתאימות – רגילה, מוסדית ודרכוניסטים.¹ האמידה נעשית באוכלוסיות אלו בנפרד ובתהליכים שונים.
- שיוך ילדים עד גיל 18 אשר נמצאים במוסדות חינוך למשקי בית של הוריהם באוכלוסייה רגילה

בסופו של התהליך התקבל קובץ שמכיל כ-662,000 נדגמים ששייכים לאוכלוסיית המפקד ומתוכם כ-528,300 משיבים שענו על השאלה של יישוב כתובת מגורים (השבה של כ-80%) וכ-523,800 שהשיבו על השאלה של כתובת מגורים שאפשר לעגן לאזור סטטיסטי בתוך מועצה מקומית או עירייה.

1. אמידת הסתברות להשבה

כאשר יישוב מגורים של המשיב לא ידוע, הנדגם מוגדר כלא-משיב ושאלון שלו אינו נכנס לאמידה כלל. אם יישוב מגורים ידוע אך הכתובת אינה מעוגנת לאזור סטטיסטי, זה נחשב כאי-השבה של אזור. כדי לצמצם את ההטיה הנובעת מאי-השבה ולהגיע לסך כל היישובים וכן לכלל אוכלוסיית מדינת ישראל האמידה נעשתה במספר שלבים.

בשלב הראשון נבנו מודלים להסתברות השבה ליישוב ולאזור בעזרת משתנים דמוגרפיים וחברתיים-כלכליים (לא גאוגרפיים) הידועים הן לגבי המשיבים והן לגבי הלא-משיבים. המשתנים המסבירים הם: דת מקובצת (6), התעודה הגבוהה ביותר (8), מקום לידה: בארץ או בחו"ל (2), מגדר (2), מצב משפחתי דה יורה (2), קבוצת גיל (18), מספרי טלפון (שיש בידי הלמ"ס להתקשרות) ומדד חברתי-כלכלי של יישוב או אזור רציף כאשר בסוגריים מופיע מספר הקטגוריות של המשתנה.

אחרי בחינת שיטות אמידה שונות להסתברות ההשבה כמו Logistic Regression, Regression tree,

XGBoost, Random forest (דרך SAS Viya) בחרנו את השיטה Logistic Regression, אשר נתנה תוצאות דומות ל-XGBoost, אבל פשוטה יותר והתוצאות ניתנות לאינטרפרטציה ברורה יותר ביחס לשיטות אחרות.

¹ יהודים ללא אזרחות ישראלית שנכנסו לישראל ברישיון סטודנט לרוב כדי ללמוד בישיבות, ובני משפחה הנלווים אליהם.

בשלב שני בוצעה שכבתיות לאחר מעשה (Post-Stratification) באמצעות חישוב גורם תיקון השבה בקבוצות גאוגרפיות – **ביישוב או באזור סטטיסטי**.

ביישוב דגימה חושב אומד למספר הנדגמים בו באמצעות סכום של $1/P_y$ לכל המשיבים המתגוררים ביישוב, כאשר P_y היא הסתברות ההשבה ביישוב הדגימה לפי המודל. לכן גורם תיקון לאי-השבה ביישוב הדגימה PY הוא סכום של $1/(P_y \times n)$ עבור המשיבים המתגוררים ביישוב, ו- n הוא גודל המדגם ביישוב. אם יישוב הדגימה הוא מעורב, נעשתה אמידה בנפרד עבור אזורים שבהם אוכלוסייה של ערבים ויהודים מכיוון ששיעורי ההשבה בהם שונים מאוד.

באופן דומה חושב גורם תיקון לאזור סטטיסטי. סימנו ב- Pas את הסתברות ההשבה על אזור סטטיסטי שפרט גר בו לפי המודל. אומד למספר הנדגמים באזור הוא סכום של $1/Pas$ לכל המשיבים באזור, וגורם התיקון לאזור דגימה PAS הוא סכום $(1/Pas \times nas)$ לכל המשיבים המתגוררים באזור כאשר nas הוא גודל מדגם באזור סטטיסטי.

גורמי תיקון PY , PAS עשויים להיות קטנים או גדולים מ-1 ומשקפים את מצב ההשבה בכל יישוב/אזור דגימה. לכן הסתברות השבה מתוקנת עבור אוכלוסיית היישוב ואזור סטטיסטי מחושבת באמצעות המכפלה $(PY \times PY)$ ו- $(PAS \times Pas)$, בהתאמה.

עבור נדגמים במוסדות אך בפועל שייכים לאוכלוסייה במשקי בית לא נעשה תיקון השבה וגורמי התיקון הם PY ו- Pas , בהתאמה.

ביישובים בדוויים בדרום פועלים אחרת מאשר באזורים אחרים. שתי בעיות עיקריות בתהליך אמידת בדוויים בדרום הצריכו תיקון של הסתברות ההשבה. הבעיה הראשונה היא עיגון לא טוב במסגרת הדגימה, דהיינו, בקמ"מ. לעיתים אנשים שרשומים ביישובים גרים בפזורה ולהפך, לעיתים אנשים שרשומים בפזורה גרים ביישובים. הבעיה השנייה היא השבה כמעט אפסית בפזורה ושיעורי אי-השבה גבוהים מאוד בחלק מהיישובים הבדוויים. בעקבות שתי הבעיות שעלו יחד הסתברות ההשבה שחושבה עבור יישובי דגימה גרמה לתיקון ענק ולא מוצדק של המשיבים, אשר בפועל גרים ביישובים. כדי למנוע את התופעה נעשו התיקונים האלה:

- כאשר משיב היה רשום בפזורה (סמל 1997) או ביישובים אבו קורינאת (1342) או אבו תלול (1375), אך התגורר ביישוב אחר, חושב לו גורם תיקון השבה לפי יישוב מגורים (ולא יישוב דגימה כפי שנעשה לעיל).
- כאשר נדגם היה רשום בפזורה (סמל 1997) והתגורר ביישובים אבו קורינאת (1342) או אבו תלול (1375), חושב לו גורם תיקון השבה לפי יישוב מגורים (ולא יישוב דגימה).
- כאשר נדגם היה רשום לא בפזורה, אך נמצא בה, הוגדר לו גורם תיקון השבה כאילו נדגם בפזורה.

2. חישוב משקל מפקדי למתגוררים ביישוב

לכל פרט משיב i מחושבת הסתברות להידגם וגם להשיב כמכפלה הזאת:

$$p_i = \pi_i \times P_y \times PY$$

כאשר π_i היא הסתברות של פרט i להידגם, P_y הסתברות השבה של הפרט לפי מודל ו- PY גורם תיקון אי-השבה ביישוב שבו הפרט נדגם.

עבור כל משק בית k מחשבים אומדן ראשוני למשקל של משק בית

$$WY_k = \sum_i P_i$$

כאשר i עובר על כל בני משק בית k .

אומד ראשוני זה הוא חסר הטיה ונבחן בסימולציות ובחזרה כללית של מפקד יחד עם אומדנים אחרים. הוא נבחר כאומד יציב ביותר לתופעות העלולות לקרות במפקד הנוכחי, כאשר דוגמים פרטים במשק בית גרעיני/מינהלי מהמסגרת ואוספים מידע על משקי בית בפועל. דבר זה גרם לדיווח של אותם פרטים במשקי בית שונים. נציין כי אומד זה נבחן גם במצב אי-השבה דיפרנציאלית ואינפורמטיבית.

להלן רשימת אוכלוסיות שאינן נאמדות מהסקר המפקדי:

- דיירי מוסדות (כ-154,000 פרטים)
- אוכלוסיית הזרים (כ-212,000 פרטים)
- יישובים מוסדיים (כ-4,500 תושבים) לפי סמלים
(2039,1951,1948,1934,1896,1325,1094,926,921,890,886,566,427,)
(412,392,338).
- יישובים בדווים שלא ענו לסקר מפקדי: תרבין א-צאנע – 1346, סעווה – 1360
(כ-5,800 תושבים).
- פזורה בדווית שגם לא שיתפה פעולה במפקד, סמל 1997 (כ-70,000 תושבים)
- בדווים שגרים בפזורה פנימית של יישובים יהודיים בדרום (כ-11,100 תושבים):
סמלי יישוב (2560,2200,9000,1268,1271,831,666), אזור סטטיסטי – 9999.

אוכלוסיות אלו אינן משתתפות בתהליך הכיול ונאמדות מנתונים מינהליים או ממפקד מוסדות. הקובץ הסופי לאמידה כלל כ-2,034,300 פרטים וכ-527,400 משקי בית.

3. אמידה ברמת יישוב

כדי לקבל אומדני מפקד עקביים שבהם התכונות הידועות של האוכלוסייה ממקורות מינהליים הוגדרו משתנים דמוגרפיים שלהם כדאי לכייל את המשקלים. להלן מופיעים האילוצים שהוכנסו לחישוב המשקלים הסופיים:

- אילוצים דמוגרפיים לפי משתנים בודדים (סוג 1):
 - מין: 2 קטגוריות.
 - גיל בודד עד 84 ו-85+: 86 קטגוריות.
 - דת מקובצת: 6 קטגוריות
 - אוכלוסייה של מוסלמים במחוז הדרום (בדווים)
- אילוצים דמוגרפיים לפי מכפלה של המשתנים דת, מין וגיל (סוג 2):
 - גיל בזוגות שנים (החלוקה השונה נוצרה כדי לא ליצור קבוצות כיול קטנות):
0-1, 2-3, ..., 84-85, +86 עבור דתות 1, 2, 3, 6
0-1, 2-3, ..., 68-69, +70 עבור דתות 4, 5
○ מין: 2 קטגוריות
 - דת מקובצת: 6 קטגוריות

- אילוח לגבי סך כל הבדווים: מוסלמים שגרים במחוז הדרום (סוג 3)

בסך הכל עד כה נוצרו 592 קבוצות כיול לאמידת משקל של יישוב. אילוצים לכל 592 קבוצות הכיול הם תלויים, אך תהליך הכיול פתיר רק עבור מערכת אילוצים לא-תלויים. לכן מכל האילוצים בחרנו חלק שהם לא-תלויים ובאותו זמן מספקים כיול לכל הקבוצות הנדרשות. בסופו של דבר נוצרה מערכת של 547 אילוצים לא-תלויים ושלמים והוכן קובץ שבו נתונים חיצוניים שמכילים סה"כ פרטים לכל אחד מקבוצות הכיול. לכל משק בית חושב מספר הפרטים השייכים ל-547 קבוצות הכיול. כלומר, נוספו 547 משתנים כל אחד עבור אילוח מסוים, והוא משייך כל פרט במשק בית לקבוצת הכיול המתאימה.

שלב ראשון של כיול נערך בשיטה Generalized Regression (GREG) בקובץ של משקי בית שבו כל האילוצים המובאים לעיל, כאשר המשקל ההתחלתי הוא WY_k כפי שהוגדר בפרק 2. תהליך הכיול התכנס וסיפק פתרון שמקיים את כל האילוצים בו-זמנית באופן מדויק.

לאחר מכן התקיימה הערכה משרדית בידי מומחים (אגף דמוגרפיה ומפקד יחד עם אגף מתודולוגיה סטטיסטית) כאשר הושאו מספר האומדנים ברמת יישוב:

- אומדן מפקדי ליישוב לפי סכום המשקלים של משיבים ביישוב;
- אומדנים אלטרנטיביים לגודל היישוב מהסוג Po/Pu מהמפקד, מנתונים מינהליים כגון של חברת החשמל, ומסקרים;
- אומדנים דמוגרפיים שוטפים יחד עם נתונים מינהליים נושאים כגון של בנייה חדשה, מוסדות.

נבחנה ההתאמה והשוני בין האומדנים ליישובים תוך בחינת ידע נושאי עבור כל יישוב בנפרד. במקרה של התאמה טובה, נבחר אומדן שהתקבל בתהליך אמידה של המפקד. במקרים של אי-התאמה משמעותית נבחר אומדן לפי החלטת מומחים (או אומדן אלטרנטיבי, או ממוצע של האומדנים, או נתון מינהלי). נציין כי במקרים לא ודאיים ניתנה עדיפות לאומדן מפקדי. בסיום הבדיקות והדיונים התקבלה החלטה על שיטה לקביעת גודל יישוב ונוצרו עוד 184 אילוצים של גודל יישובים שקבעו מומחים בשיטה לא רגילה. רוב היישובים האלה (170) הם חד-אזוריים ו-14 מחולקים לאזורים סטטיסטיים. התפלגות היישובים לפי דת: 154 יהודיים, 29 ערביים ואחד מעורב.

שלב שני של כיול נערך גם בשיטת GREG בקובץ משקי בית עם כל האילוצים, דהיינו 731 (547+184). התהליך התכנס וסיפק בסופו פתרון מדויק. לכל משק בית מתקבל משקל יישוב סופי, WY , אשר מקיים כל 731 האילוצים וההתפלגות שלו קרובה להתפלגות של המשקל הראשוני. המשקל תקף לאמידת אוכלוסייה וגם משקי בית.

4. אמידה ברמת אזור סטטיסטי

כדי לאמוד משקל אזורי למשיבים ברמת אזור סטטיסטי חוזרים כמעט על כל הסעיפים של אמידת משקל יישוב תוך שינויים מסוימים. לפני אמידת המשקל האזורי נעשתה **זקיפת אזורים** לכ-4,800 משקי בית (כ-16,000 פרטים), שכתובותיהם לא עוגנו לאזור סטטיסטי או שהיו מעוגנות לאזורים אשר הוגדרו לא לאמידה (קטנים ושלא למגורים). הזקיפה נעשתה כדי לא לייצר שתי מערכות משקלים, דהיינו, משקל יישוב ומשקל אזור, לכן לכל משק בית ניתן משקל אחד: משקל אזור, מטעמי נוחות לניגשים לנתוני מפקד.

הזקיפה נעשתה ביישובים המחולקים לאזורים פרופורציונלית למספר משקי בית משיבים המעוגנים באזורים סטטיסטיים. ביישובים מעורבים משקי בית יהודיים נזקפו לאזורים יהודיים ומשקי בית ערביים לאזורים ערביים. משקלי האזורים חושבו בדיוק באותו אופן כפי שנעשה עבור משקלי יישובים: חושבה הסתברות הדגימה וההשבה וחושב משקל אזורי למשקי בית.

קבוצות כיוול לאומדנים אזוריים:

- כל 592 קבוצות הכיוול הדמוגרפיות שהיו באמידת משקל יישוב.
- אילוף נוסף: אומד ליישוב לפי משקל אזור צריך להיות זהה לאומד לפי משקל יישוב, לכן התווספו 1,214 אילוצים עבור היישובים.
- שני אילוצים נוספים: אחוז רצוי של משקי בית בגודל 1 באוכלוסייה הערבית (15%) ובאוכלוסייה של יהודים ואחרים (25%). האילוצים הוגדרו על סמך משקל לפני כיוול כדי להשיג עקביות עם אומדנים ידועים ממקורות אחרים.

בסך הכל הוגדרו 1,808 (2+1,214+592) קבוצות כיוול תלויות. מתוך 1,808 אילוצים נבחרו 1,762 אילוצים לא תלויים ושלמים לתהליך הכיוול. עבור קבוצות כיוול דמוגרפיות נוצל אותו קובץ המכיל 547 נתונים חיצוניים. נוסף לכך, נוצרו השדות המתאימים לאילוצים עבור גודלי יישוב ושני אילוצים של משקי בית בגודל 1. הקובץ הנוצר הכיל עבור כל משק בית משתני שייכות ל-1,762 קבוצות כיוול עבור אילוצים לא תלויים.

כמו קודם, הכיוול נעשה בשיטה Generalized Regression (GREG) בקובץ של משקי בית עם שבו אילוצים של נתונים חיצוניים, כאשר המשקל ההתחלתי הוא WAS_k אשר הוגדר לאזורים סטטיסטיים בדומה למה שהוסבר בפרק 2. בסוף התהליך, עבור כל משק בית k התקבל משקל W_k , אשר מקיים את כל האילוצים, מיועד לאמידה ברמה של אזורים סטטיסטיים וגם ברמה של יישוב. חשוב להדגיש כי המשקל מחושב למשקי בית והוא זהה למשקל של כל בני משק הבית, דבר שמשרת את האמידה הן של תכונות האוכלוסייה והן של תכונות משקי הבית.

5. אמידת שונות האומדנים

יש חשיבות רבה להערכת מהימנות האומדנים של מפקד 2022 המבוססת על מדגם שהוצא בצורה מורכבת ותהליכי אמידה מורכבים עוד יותר כפי שהובאו בפרקים קודמים. קשה מאוד להעריך את הטעויות הלא-מדגמיות הנובעות מתהליכי עיבוד שונים כמו קביעת שייכות לאוכלוסייה, קישור רשומות (במסמך המתאים הובאה הערכה כללית), זקיפת נתונים חסרים, בניית מודל להסתברות השבה ברמות שונות, הערכת מומחים, טעויות תשובה. כאן נניח שאין הטיה משמעותית באומדנים של המפקד ונסה לחשב אומדן לשונות לגבי כל אומדן של אוכלוסייה מסוימת. אומדן לשונות מאפשר למשתמש להתייחס לאי-דיוק באומדן המחושב וליצור רווח סמך מסביב לאומדן.

במפקד 2022 נעשתה הדגימה בשכבות קטנות של משפחות מינהליות/גרעיניות בתוך אזורים סטטיסטיים, שבהם רשומים בני המשפחות. לצורך חישוב שונות, נשתמש באזורי דגימה (אזורים סטטיסטיים לאחר טיפול) כשכבות דגימה (Strata), ולא בשכבות קטנות בתוך האזורים. הסיבה לכך היא שהשימוש בשכבות קטנות הביא לבעיות טכניות הקשורות לאי-השבה ולחוסר יציבות באומדני השונות. עם זאת, בשימוש באזורי דגימה מתקבל **אומדן-על** לשונות ולכן הוא שמרני יותר מבחינת המשתמש. **חישוב השונות של אומדני המפקד נעשה בשיטת פיתוח טיילור לדגימה בשכבות, שמקובלת גם בסקרים אחרים בלמ"ס.**

כל משק בית שייך לכתובת אחת (אזור סטטיסטי מגורים, יישוב וכו'), לכן הוא אשכול (Cluster). כדי לאמוד שונות של אומדני אוכלוסייה ביחידה גאוגרפית (Domain) מתייחסים לנתונים של משקי בית

ולתכונות של הפרטים בתוך משקי הבית. לצורך זה יש לחשב עבור כל משק בית מאפיינים לפי המשתנים של קבוצות אוכלוסייה אשר מעוניינים לאמוד. למשל, גודל משק בית, מספר פרטים השייכים לקבוצות גיל, מספר אקדמאים, מספר גברים/נשים. בעזרת המשתנים האלה מחשבים את האומדנים לשונות של אומדני סה"כ, ממוצעים או אומדני יחס. המשקל שישמש לחישוב האומדן לשונות הוא משקל אזורי W_k ואומדן לשונות התקבל באמצעות פרוצדורה של מערכת SAS (SurveyMeans).

בפרסום האומדנים מתייחסים לטעות הדגימה היחסית CV, כלומר, היחס בין סטיית התקן של האומדן לאומדן עצמו. לפי הנוהל הקיים בלמ"ס אומדנים עם CV עד 15% מפורסמים בצורה רגילה ללא שום סימן. כאשר ה-CV המחושב הוא בין 15% ל-30% מציגים אותו עם אזהרה (בדרך כלל בתוך סוגריים), ואין מפורסמים אומדנים כלל כאשר הוא גבוה מ-30%. נציין כי לגבי אומדני יחס ופרופורציות ישנם כללים שמתייחסים גם למספר התצפיות במונה ובמכנה.