

התממה של קבצים עם מידע פרטני

אריאל מנצורה*

תקציר

החטיבה למידע ולסטטיסטיקה בבנק ישראל אוספת ומנהלת קבצים ממקורות מגוונים, אשר חלקם מכילים מידע פרטני. כדי לאפשר את חופש המידע ויחד עם זאת לשמור על סודיות המידע בונה החטיבה תהליכי התממה (אנונימיזציה) של קובצי המידע – תהליך מורכב, שמטרתו למנוע זיהוי או חשיפה של מידע רגיש/סודי על פרטים שנתוניהם מופיעים בקבצים. עבודה זו מתארת את תהליך ההתממה של נתונים פרטניים, מגדירה את המושגים הבסיסיים בנושא, מציגה גישות מקובלות להערכת סיכון החשיפה בקבצים ומדגימה את יישום התהליך.

* החטיבה למידע ולסטטיסטיקה, בנק ישראל.

1. מבוא

לאחר המשבר הפיננסי העולמי של 2008 החלו בנקים מרכזיים, וביניהם בנק ישראל, לנהל מדיניות מקרו-יציבותית, שמטרתה לזהות סיכונים מערכתיים בשלבי התהוותם ולקדם פעולות שיטפלו בהם ויגבילו את השפעתם על היציבות הפיננסית של המשק. האתגרים החדשים מניעים את הבנקים המרכזיים לנהל בסיסי נתונים עקביים ואינטגרטיביים שיתמכו במדיניות זו. לצד התפתחות הטכנולוגיה, המאפשרת לאחסן ולעבד כמויות גדולות מאוד של מידע, גבר הצורך במאגרי נתונים פרטניים, שיאפשרו השלמת מידע על זרימת ההון במשק, ועל בסיסם יהיה ניתן לקבל תמונה מפורטת וזמינה של רמת יציבותו ואיתנותו הפיננסית. על רקע מגמות אלו, ובמקביל להתפתחותם של חוקי חופש המידע, שהדגישו את החשיבות של הגברת השקיפות ושיתוף מידע, גופים שונים המנהלים מידע סטטיסטי נוטים לאפשר גישה גם למידע פרטני למטרות ניהול מדיניות, ניתוח כלכלי ומחקר. כדי לאפשר גישה למידע כזה בתוך הארגון או מחוץ לו נדרש, על פי חוק הגנת הפרטיות, לשמור על סודיות המידע כשזה מתייחס לאנשים פרטיים. נוסף על כך החוק מחייב לשמור על סודיות מסחרית של ישויות עסקיות – משימה מורכבת, בפרט כשמדובר במידע פיננסי, המאופיין לעיתים בריכוזיות גבוהה.

החטיבה למידע ולסטטיסטיקה בבנק ישראל, האוספת ומייצרת סטטיסטיקה פיננסית, מנהלת מאגרי מידע, המכילים, בין היתר, מידע פרטני בנושאים שונים: שוק ההון, שוק מטבע החוץ, בנקאות, שוק האשראי ועוד. בהקשר זה בונה בנק ישראל בימים אלה מאגר אשראי, המכיל מידע פרטני על היסטוריית האשראי של הלווים במשק וישמש את לשכות האשראי¹ לצורך בניית מודלים לדירוג אשראי של לוויים. על בסיס מאגר זה תנהל החטיבה למידע ולסטטיסטיקה מאגר סטטיסטי שהמידע הפרטני הכלול בו אינו מזוהה; זאת לשימושים פנימיים בבנק ישראל לשם מילוי תפקידיו על פי החוק.

כדי לאפשר גישה למידע ויחד עם זאת לשמור על סודיותו מעצב בנק ישראל תהליך הנקרא "אנונימיזציה של קובצי המידע" (להלן התממה). מטרתו של תהליך ההתממה להגן על המידע כך שלא יתאפשר לזהות או לחשוף את הפרטים שנתונים מופיעים בקבצים, ובפרט מידע רגיש/סודי עליהם². תהליך זה יתייחס הן לנתונים המיועדים לשימוש בתוך הבנק – אף שרק כלכלנים מועטים מתוכו יורשו לגשת אליהם – והן למידע שיותר להנגיש לחוקרים מחוץ לבנק, בכפיפות למגבלות של הגנה על הפרטיות ושמירה על סודיות מסחרית.

מאגר המכיל מידע פרטני כולל, באופן טבעי, מידע המזוהה את הפרט ישירות, כלומר שדה החושף כשלעצמו את זהות הפרט גם ללא צורך במידע נוסף, המצוי בשדות אחרים. דוגמאות לכך הן תעודת הזהות ושמו המלא של פרט. לכן תנאי הכרחי להתממה של מאגר הוא השמטת כל המזהים הישירים. ואולם תנאי זה אינו מספיק כדי להגן על המאגר, משום שגם בהעדר מידע זה, ניתן לעיתים לגלות מידע על פרטים באמצעות חיבור מספר שדות, או הצלבה עם מידע ממאגרים אחרים, שהגישה אליהם מותרת. כמו כן ניתן לזהות פרט על ידי חיפוש צירופים שאינם שכיחים באוכלוסייה הרלבנטית המאפיינים רק אותו או רק פרטים מועטים.

תהליך ההתממה מתחיל בהגדרה מדויקת של תרחישי החשיפה. (ראו מושגים בפרק 2). תרחישים אלו כוללים את האפשרויות העומדות לרשות המשתמשים כדי לחשוף מידע על פרטים, ומפניהן נרצה להתגונן. בהינתן תרחישים אלו ניתן להפעיל, בין היתר, שיטות שמטרתן לטשטש את הזהות ולהגן על המידע. בסוף התהליך נדרש להעריך את הסיכון שעדיין נותר וכן לכמת את המידע שאבד כתוצאה מהתהליך. ברור ששורת תחלופה בין מידת ההתממה, כלומר מידת ההגנה על הקובץ, לבין מידת השימושיות של המידע, שכן מדובר באובדן מידע.

עבודה זו תתאר תהליך ההתממה של נתונים פרטניים, תגדיר את המושגים הבסיסיים בנושא, תציג גישות מקובלות להערכת סיכון החשיפה ותדגים את יישום התהליך על טבלאות-לדוגמה של נתונים המכילים מידע פרטני³.

¹ חוק נתוני אשראי סעיף 16 חוק זה עתיד להכנס לתוקף.

² תהליך ההתממה המתואר בעבודה זו אינו מתייחס לסדרות שהחטיבה למידע ולסטטיסטיקה מפרסמת באתר בנק ישראל. סדרות אלו מציגות מידע מצרפי ולא מידע פרטני.

³ לשם שמירה על הפשטות ועל הדיוק ככל שניתן נעסוק כאן בנתונים הכוללים את כל הרשומות באוכלוסייה הרלבנטית ("קובץ מפקדי") ולא בקובץ מדגמי, כגון סקר הכולל רק חלק מהרשומות באוכלוסייה הרלבנטית, ששיטת הטיפול בו מורכבת יותר. כן נניח שאין בקובץ מבנה הירארכי, מבנה המאפיין, לדוגמה, קובץ נתונים המכיל משתנה המסמן את משק הבית שאליו שייך הפרט.

2. מושגים הרלבנטיים לתהליך התממה

בקרה סטטיסטית לחשיפה (Statistical Disclosure Control) – שם כללי המתאר את קבוצת השיטות להפחתת הסיכון של חשיפה (להלן חשיפה) של פרטים בקובץ. באופן כללי השיטות נחלקות לשתיים: א. שיטות המוסיפות לנתונים רעש (perturbative methods) ב. שיטות שאינן מוסיפות רעש, כגון קיבוץ קטגוריות של שדות או השמטת ערכים משדות בעלי סיכון גבוה מן המותר והכנסת ערכים חסרים במקומם.

התממה (Anonymization) – תהליך שבו הופכים קובץ לא מוגן לקובץ מוגן לפי רמת הגנה שנקבעת מראש באמצעות שיטות של בקרה סטטיסטית לחשיפה.

חשיפה⁴ (Disclosure) – גילוי מידע שלא היה ידוע ומפורסם קודם לכן על פרט באמצעות קובץ מידע שהופץ. ישנם שלושה סוגים של חשיפה:

- **חשיפת זהות** – קישור בין זהות ידועה של פרט, כגון שמו הפרטי ושם משפחתו, לבין רשומה בקובץ. משנעשה קישור כזה נחשף המידע המצוי בשאר השדות שבקובץ על פרט זה. לדוגמה: הצלבת רשומה בעלת שני שדות – תעודת הזהות של פרט והכנסתו החודשית – עם רשומה מקובץ חיצוני שבה מופיע גם שמו המלא של הפרט (או עם מידע אישי על הפרט בעל מספר הזהות האמור) תגרום לחשיפת זהותו והכנסתו החודשית של פרט זה.

רשומה מקובץ הכנסות	
מספר הזהות	ההכנסה החודשית בשקלים
123456	5,000

רשומה מקובץ חיצוני	
השם המלא	מספר הזהות
ישראל ישראלי	123456

- **חשיפת מאפיין** – חשיפת מאפיין מסוים של פרט גם בלי לקשר בין זהות הפרט לבין רשומה מסוימת. לדוגמה: אם בקובץ מסוים הכנסתם של כל הפרטים בגילים 70–74 ללא יוצא מן הכלל היא בטווח של 5,000–10,000 שקלים ניתן לדעת את טווח הכנסתו של פרט שגילו בטווח הגילים האמור הכלול באותו קובץ גם בלי לדעת את זהותו.

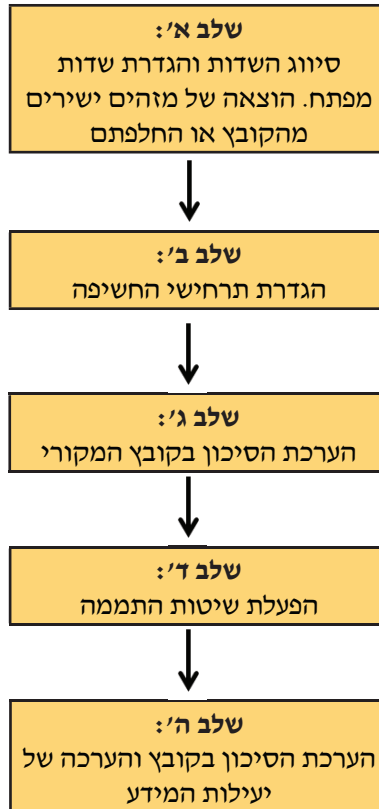
המין	טווח הגילים	טווח ההכנסה החודשית בשקלים
זכר	70–74	5,000–10,000
נקבה	70–74	5,000–10,000

- **חשיפה סטטיסטית** – זיהוי מאפיין של אדם באמצעות ניתוח סטטיסטי של הקובץ. לדוגמה: ניחוש מדויק למדי – באמצעות מודל חיזוי טוב – של הכנסת אדם מסוים על בסיס מאפיינים ידועים של אותו אדם המופיעים בקובץ.

⁴ ראו למשל [4].

3. תיאור השלבים של תהליך ההתממה

תרשים זרימה של תהליך ההתממה



שלב א': סיווג השדות והגדרת שדות מפתח

סוגי שדות – מקובל לחלק את השדות בקובץ לשלושה סוגים. חלוקה זו אינה בהכרח אקסקלוסיבית: שדה יכול להשתייך ליותר מסוג אחד:

- **מזהים ישירים** – שדות המזהים פרטים בקובץ ללא שימוש בשדות אחרים. דוגמאות לשדות כאלה הן תעודת הזהות, השם המלא והכתובת המדויקת. שדות מסוג זה מושמטים מהקובץ כשלב ראשון של תהליך ההתממה, או מוחלפים באופן חד-חד ערכי בשדות אחרים, שאינם מזהים.
- **שדות מפתח**⁵ – שדות שניתן להצליבם עם מידע חיצוני – למשל כאלה המצויים בקובץ מפקדי שמפורסם לציבור או לחלק ממנו – וכך לחשוף את זהותם של הפרטים שמאחורי רשומות מסוימות בקובץ.

⁵ ראו למשל [7].

- **שדות רגישים** – שדות שמפאת רגישותם אסור שערכם, לגבי כל אחד מהפרטים שזהותם ידועה בקובץ, יתגלה. דוגמאות לשדות כאלה הן מצבו הרפואי של אדם והכנסתו.

נוסף על חלוקה זו אפשר לחלק את השדות לשני סוגים:

- **שדות קטגוריאליים** – שדות הכוללים מספר סופי (בדרך כלל קטן) של קטגוריות/ערכים. קבוצה זו ניתן לחלק לשדות סדורים ולא סדורים.
- **שדות רציפים** – שדות נומריים שניתן לבצע עליהם פעולות אריתמטיות. שדות אלה יכולים לקבל מספר גדול של ערכים.

שלב ב': הגדרת תרחישי החשיפה

תרחישי חשיפה⁶ הם קבוצת הנחות המתארות את האופן שבו משתמש, או אדם אחר שנחשף לקובץ, יכול לחשוף מתוכו מידע על פרטים. לדוגמה: משתמש יכול להצליב את המידע מהקובץ עם מידע אחר הנמצא ברשותו באמצעות מספר מאפיינים משותפים, או באמצעות מידע על פרט שהוא מכיר וידוע לו שפרט זה נמצא בקובץ. כך הוא יוכל לגלות באמצעות מאפיינים הידועים לו מידע נוסף, רגיש, על אותו פרט.

את תרחיש החשיפה ניתן לסכם לרוב באמצעות קביעת קבוצות של שדות מפתח שדרכן ניתן להצליב מידע בקובץ עם מידע חיצוני אחר (קובץ או ידע אישי) ולגלות מידע על פרטים באמצעות צירופים המאפיינים פרטים מעטים שבקובץ.

קביעת תרחישי החשיפה הכרחית בתהליך ההתממה, שכן מפניהם אנו מבקשים להגן על המידע. גם הערכת רמת הסיכון לגילוי המידע תלויה כמובן בקביעת תרחישים אלו, משום שזו אינה כללית אלא מתייחסת לתרחישי חשיפה מסוימים. תרחישי החשיפה נקבעים בעזרת מומחים בעולמות התוכן הרלבנטיים. אלה יודעים כיצד ובאלו אמצעים משתמש וכל מי שניגש למידע יכול לחשוף מידע על פרטים בקובץ. אף על פי כן, גם מומחים בעולמות התוכן אינם מכירים את כל האפשרויות לחשיפת מידע, ולכן במקרים מסוימים הנטייה היא להניח את האפשרות המחמירה ביותר (worst case scenario).

תרחישי החשיפה יכולים להיות מקלים או מחמירים מהאפשרויות האובייקטיביות הקיימות לחשיפת מידע – וזאת בהתאם למדיניות החשיפה, התלויה באופן השימוש בנתונים, במטרת השימוש, בזהות המשתמשים, בחומרת הנזק הכרוך בחשיפה וכו'. בהקשר זה מקובל להבחין בין קבצים לצורכי מחקר (SUF – Scientific Use Files), שמטרתם לשמש חוקרים על פי חוזה, בכפיפות להרשאות והגבלות כגון עבודה בחדר מחקר פיזי או בחדר מחקר וירטואלי באמצעות גישה מרחוק (remote access), ובין קבצים המפורסמים לציבור (PUF – Public Used Files) ללא כל הגבלה או בקרה. המדיניות לגבי קובצי מידע המוצאים לציבור לרוב מחמירה מאוד וכרוכה באיבוד מידע משמעותי.

שלב ג': הערכת סיכון החשיפה בקובץ

סיכון חשיפה מתייחס כאמור ישירות לתרחישי החשיפה, כלומר לקבוצות שדות המפתח (קטגוריאליים או רציפים) שמוגדרים עבור קובץ מסוים. לאחר שהגדרנו את קבוצות שדות המפתח ניתן להתייחס למספר ממדים של סיכונים.

- **סיכון של רשומה בקובץ** – סיכון של רשומה הוא ההסתברות שיהיה ניתן לקשר בין רשומה מסוימת בקובץ לבין פרט מסוים שזהותו ידועה. בהקשר זה יש להבחין בין שדות מפתח קטגוריאליים לשדות מפתח רציפים. בהתייחס לתרחיש שבו תיעשה הצלבה של שדות מפתח קטגוריאליים ישנן שתי דרישות מקובלות.

- **דרישת k אנונימיות (k-anonymity)**⁷ – דרישה שבכל צירוף של שדות המפתח הקטגוריאליים בקבוצות שהוגדרו בתרחיש החשיפה תהיינה לפחות k רשומות בעלות אותו צירוף. כדי לבדוק זאת ניתן לבנות טבלה (או טבלאות לכל תרחיש חשיפה) רב-ממדית, שבה מספר תאים השווה למספר הצירופים האפשריים. על בסיס טבלה זו ניתן

⁶ ראו למשל [7].

⁷ ראו למשל [7].

לחשב את ההסתברות-לסיכון של כל רשומה. דרישה זו מטרתה להגן בפני חשיפת זהות, כי במקרה שצירוף מסוים מתוך הטבלה מתייחס לפרט אחד בלבד ניתן להצליב צירוף זה עם אותו צירוף מתוך טבלה אחרת עם אותם שדות מפתח וכך לחשוף את זהותו של הפרט.

- **דרישת I שונות (I-diversity)⁸** – דרישה נוספת שבאה להגן מחשיפת מאפיינים. ייתכן שבכל תא בטבלת השכיחויות יש מספיק רשומות, אבל לגבי שדה רגיש מסוים אין שונות בקרב אותן רשומות השייכות לאותו צירוף. הדרישה של I-diversity היא שבכל הצירופים האפשריים יהיו לפחות I ערכים שונים. במצב שבו אין שונות מספיק לדעת על פרט איזה צירוף מתייחס אליו כדי לזהות בוודאות מאפיין זה שלו גם בלי לדעת שהרשומה מתייחסת אליו.

בטבלה להלן מוצגות שתי הדרישות האלה באמצעות דוגמה פשוטה של תרחיש שבו רק שני שדות מפתח – המין והגיל.

מספר הערכים השונים	שדה רגיש – אחוז הריבית (מעוגל) על הלוואה	השכיחות של הצירוף בטבלה	שדה מפתח 2 – טווח הגילים	שדה מפתח 1 – המין	רשומה
2	2%	3	60–50	זכר	1
2	4%	3	60–50	זכר	2
2	4%	3	60–50	זכר	3
1	2%	3	50–40	נקבה	4
1	2%	3	50–40	נקבה	5
1	2%	3	50–40	נקבה	6

בטבלה רואים שהפרט הראשון (רשומה 1) שייך לתא עם הצירוף מין = זכר, טווח גילים = 60–50. בצירוף זה ישנם שלושה פרטים בטבלה (רשומות 1–3). ואולם עבור המשתנה הרגיש ישנן שני אפשרויות (ריבית 2%, 4%). כל הרשומות בטבלה מקיימות את דרישת ה-3 אנונימיות, ואילו רשומות 1–3 בלבד מקיימות את דרישת 2 שונות.

- **סיכון בשדות מפתח רציפים** – לגבי שדות המפתח הרציפים לא ניתן לבנות טבלת שכיחויות, שכן רוב הערכים מופיעים פעם אחת בלבד. באופן כללי נהוג להעריך את הסיכון במשתנים אלו בהתבסס על המידה שבה מתאפשר קישור רשומות (record linkage) בין הקובץ שבו שיינו את נתוני המשתנים הרציפים, למשל על ידי הוספת רעש, לבין הקובץ המקורי.

- **סיכון גלובלי של כל הקובץ** – מדד שנותן ציון לרמת הסיכון של הקובץ כולו, המחושב על בסיס אגרגציה מסוימת של ההסתברויות לזיהוי הרשומות בקובץ. דוגמה למדד כזה היא סכום ההסתברויות לזיהוי בקובץ, השווה לתוחלת מספר הזיהויים בו.

שלב ד': הפעלת שיטות התממה

להלן נתאר מספר שיטות מקובלות להתממה.

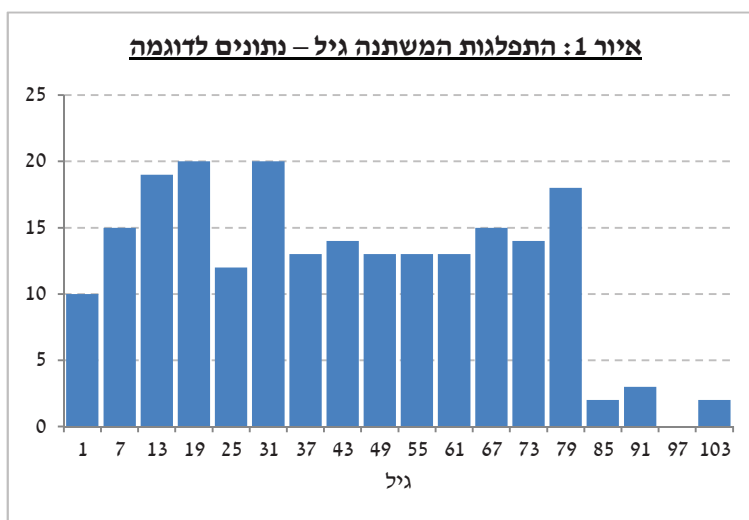
- **קידוד גלובלי (Global recoding)** – שיטה שמצמצמת את רמת האינפורמציה בשדה ומתאימה לשדות קטגוריאליים ולשדות רציפים. לגבי שדה קטגוריאלי קידוד גלובלי משמעותו צירוף של מספר קטגוריות לקטגוריה משותפת. לדוגמה: בשדה המקצוע של פרט ניתן לאחד את המקצועות סטטיסטיקאי ומתמטיקאי לקטגוריה אחת משותפת, אם בצירופים

⁸ ראו [5].

מסוימים של שדות המפתח הכוללים את אחת מהקטגוריות הללו יש מספר קטן מדי של רשומות. דוגמה נוספת היא שינוי שדה הגיל לטווחים של 5 או 10 שנים. קידוד גלובלי בשדה רציף הוא למעשה החלפה של שדה רציף בשדה קטגוריאל. לדוגמה: שדה שהוא סכום הלוואה ניתן להחליף במספר קטגוריות שהן בטווחים של 100,000 שקלים. בטבלה להלן מוצגת דוגמה לקידוד גלובלי של שדה ההכנסה (רציף).

מספר הרשומה	ההכנסה החודשית בשקלים	ההכנסה החודשית לאחר קידוד
1	8,365	עד 10,000
2	16,569	10,000–20,000
3	100,200	100,000–200,000
4	5,750	עד 10,000

- קידוד עליון ותחתון** – שיטה זו היא מקרה פרטי של קידוד גלובלי, והיא מטפלת בקצוות ההתפלגות. לגבי שדה רציף היא מקבצת את הקטגוריות הקיצוניות מעבר לסף עליון לקטגוריה אחת, ואותו דבר ניתן לעשות לגבי קטגוריות נמוכות. בשדה רציף השיטה מקבצת את כל הערכים מעבר לסף עליון או/ו תחתון לשתי קטגוריות – עליונה ותחתונה ובשאר הטווח הנתונים מקובצים כבסעיף הקודם. שיטה זו מתאימה לשדות שבהם מעבר לסף מסוים יש מספר קטן של מקרים. באיור להלן ניתן לראות דוגמה להתפלגות של שדה הגיל, שבו יש מעט פרטים מעל גיל 80. אם ישנם מעט מדי רשומות בצירופים הכוללים את משתנה הגיל בגילים הגבוהים ניתן, בשיטה זו, לקבץ את כל הגילים שמעל גיל 80 לקטגוריה אחת – 80+.



- השמטה מקומית (Local suppression)** – שיטה זו מכניסה ערכים חסרים בשדות מסוימים של רשומות מסוימות, והיא מתאימה לשדות קטגוריאלים ולא לשדות רציפים. כאשר ישנם צירופים של שדות המפתח שבהם יש מספר מועט של רשומות ניתן להכניס באחד מהשדות ערך חסר. יתרונה של שיטה זו שהיא מטפלת רק ברשומות בסיכון. מצד שני היא יוצרת חוסר אחידות בשדה מסוים, כי ברשומות מסוימות מופיע ערך חסר בשדה זה. בטבלה להלן מוצגת דוגמה של השמטה מקומית והכנסת ערך חסר (NA) ברשומה 4 לגבי צירוף של המין זכר וטווח הגיל 20–30, צירוף שבו מצוי פרט אחד בלבד.

לאחר השמטה מקומית		לפני השמטה מקומית		הרשומה
שדה המפתח 2 – טווח הגיל	שדה המפתח 1 – המין	שדה המפתח 2 – טווח הגיל	שדה המפתח 1 – המין	
50-60	זכר	50-60	זכר	1
50-60	זכר	50-60	זכר	2
50-60	זכר	50-60	זכר	3
NA	זכר	20-30	זכר	4

- **הוספת רעש (חיבורי)**⁹ – שיטה זו משנה את הערכים המספריים בשדה ומתאימה לשדות רציפים ולא קטגוריאליים. לכך מספר דרכים מקובלות, ושתיים מהן מוצגות להלן.
- הוספת רעש לבן (בלתי מתואמים) – בשיטה זו מוסיפים לשדה מסוים, שנסמן ב- X , רעש בלתי מתואם באופן הבא:

$$Z = X + \varepsilon$$

כאשר ε הוא וקטור של רעשים מפולגים נורמלית ובלתי מתואמים (רעש לבן).

ניתן להראות, ששיטה זו משמרת (בקירוב) את התוחלת ואת השונות המשותפת בין כל שני משתנים, אך אינה משמרת את השונות ואת מקדמי המתאם. בפרט היא מגדילה את שונות המשתנים מצד אחד ומהצד האחר מקטינה את המתאם, בערך מוחלט, בין כל שני משתנים, וזאת בגלל רכיב הרעש, שהתווסף.

- **הוספת רעשים מתואמים** – בשיטה זו מגרילים (randomize) רעשים מתואמים לגבי מספר משתנים. ניתן להראות שבשיטה זו המתאמים בין כל זוג משתנים משתמרים.

בעיה נפוצה בהוספת רעש חיבורי היא שלערכים גבוהים ולערכים נמוכים של המשתנה מוסיפים רעשים בעלי אותו קנה מידה. המשמעות היא שערך גבוה יחסית משתנה מעט ואילו ערך נמוך משתנה הרבה, יחסית. דרך אחת לפתור זאת היא הוספת רעש כפלי, שהוא פרופורציוני לגודל הערך בשדה, במקום רעש חיבורי. דרך זו משמרת מאפיינים שונים של השדות, כגון התוחלת והשונות.

- **מיקרו-אגרגציה (Micro-Aggregation)**¹⁰ – שיטה שמצמצמת את רמת האינפורמציה בשדה ונועדה בעיקר לשדות רציפים. שיטה זו ניתן להפעיל על שדה אחד או על מספר שדות בו-זמנית. שיטה זו לוקחת שדה אחד או מספר שדות ומחלקת את הרשומות למספר קבוצות עם לפחות k רשומות בכל קבוצה. בכל קבוצה מחליפים את ערכיהם של השדות במוצע הקבוצה. העיקרון בחלוקה הוא ליצור קבוצות עם מקסימום הומוגניות בתוך הקבוצה. שיטה זו באה להבטיח שבקובץ המופץ יהיו רשומות הממלאות את דרישת ה- k -anonymity. לצורך המחשה על שדה אחד מוצגת להלן טבלה מספרית המחלקת את הרשומות לקבוצות, אשר כל אחת מהן מכילה לפחות שתי רשומות שערכיהן המקוריים מוחלפים במוצע הקבוצה.

⁹ ראו למשל [8].

¹⁰ ראו למשל [3].

מספר הרשומה	הערך הישן	הערך החדש
1	25	21.5
2	12	9
3	18	21.5
4	10	9
5	105	109
6	99	109
7	5	9
8	122	109

- החלפת קטגוריות מקרית (PRAM¹¹) – שיטה המתאימה לשדות קטגוריאליים. קטגוריות בתוך שדה מסוים מוחלפות באמצעות מטריצה עם ההסתברויות להחלפת ערכים. נסמן ב- X את המשתנה הקטגוריאלי בקובץ המקורי וב- Y את המשתנה החדש שנייצר. נניח שלשני המשתנים K קטגוריות דומות $1, \dots, k$. המעבר בין המשתנה X למשתנה Y ייעשה באמצעות מטריצת מעברים עם איבר כללי המוגדר לכל $i, j = \{1, \dots, K\}$.

$$P_{ij} = P(Y=j | X=i)$$

ביטוי זה מציג את ההסתברות שהקטגוריה i תוחלף לקטגוריה j .
להלן דוגמה למטריצה כזאת, המתאימה לשדה עם שלוש קטגוריות.

במטריצה זו ניתן לראות שההסתברויות באלכסון הראשי, כלומר ההסתברויות שלא יהיה שינוי בקטגוריה, הן הגבוהות ביותר. אם מטריצה זו ידועה ניתן ללמוד על מאפייניו של המשתנה המקורי, כגון התוחלת והשונות שלו.

מטריצה עם הסתברויות להחלפת קטגוריות

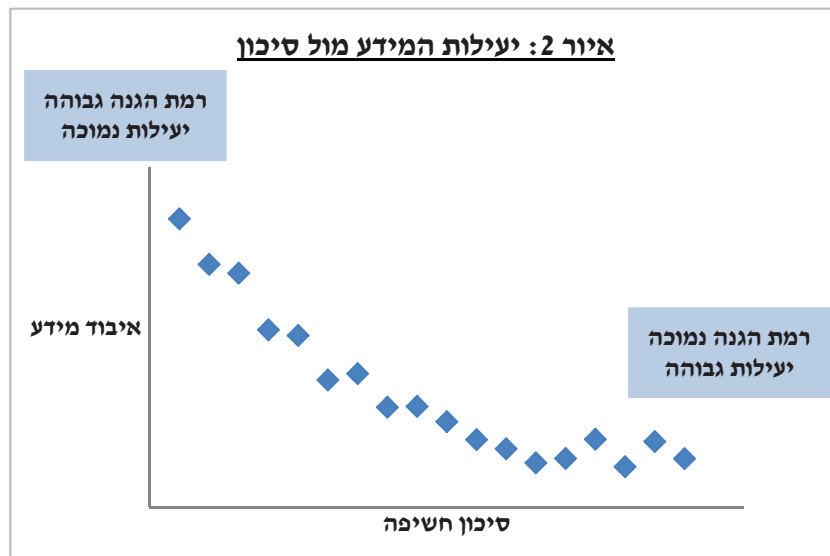
הקטגוריה החדשה			הקטגוריה המקורית
3	2	1	
0.1	0.1	0.8	1
0.05	0.9	0.05	2
0.6	0.3	0.1	3

¹¹ PRAM – Post Randomization Method. ראו למשל [6 ו-2].

- **יצירת קובץ עם נתונים סינתטיים** – קובץ סינתטי הוא קובץ שמכיל נתונים שונים מהקובץ המקורי, אך נבנה באופן המשמר בקירוב מאפיינים סטטיסטיים של הקובץ, כגון ההתפלגויות השוליות של השדות בו וכן מתאמים בין השדות. אף ששיטה זו אינה חביבה על החוקרים, שכן אלה מעדיפים בדרך כלל גישה לנתונים האמיתיים, היא יכולה לשמש לכיול מודלים של חוקרים, לביצוע ניסוי וטעייה בהעדר גישה לנתונים אמיתיים – לדוגמה עבור סטטיסטיקאים המבצעים התממה ונזקקים לקובץ בעל מאפיינים דומים. אף שכל התצפיות שונות, שיטה זו לא תמיד מספקת הגנה מלאה לקובץ. בדומה למצב שבו מוסיפים רעש לנתונים נהוג להעריך את הסיכון בקובץ הסינתטי – באיזו מידה ניתן לבצע קישור רשומות (record linkage) בינו לבין הקובץ המקורי.

שלב ה': הערכת סיכון החשיפה בקובץ ושמירה על יעילות המידע

שמירה על יעילות המידע ומזעור הסיכון – המטרה בתהליך ההתממה היא להנגיש קובץ מוגן של נתונים כך שהוא יגלם סיכון נמוך לזיהוי הפרטים, ובד בבד, בכפיפות למגבלה זו, ישמר מקסימום אינפורמציה בקובץ (יעילות/שימושיות המידע). שוררת תחלופה בין רמת ההגנה על המידע לבין שימושיותו: ככל שרמת ההגנה גבוהה יותר אובד יותר מידע (איור 2)¹². המטרה היא למצוא את השיטות שיביאו את התחלופה הזאת לאופטימום בהינתן חשיבות השימוש במידע והנזק שיכול להיגרם מזיהויו. ישנן מספר שיטות למדידת השמירה על יעילות המידע בקובץ – ביניהן השוואה ישירה בין נתוני הקובץ המקורי לנתוניו לאחר ההתממה והשוואה של סטטיסטיים מחושבים (הממוצע, סטיית התקן וכו') ביניהם.



¹² ראו [4].

4. סיכום

החטיבה למידע ולסטטיסטיקה משתמשת בשיטות שונות ומורכבות, שתוארו לעיל, לשם התממה של נתונים פרטניים במגוון עולמות תוכן עבור משתמשי המידע. תהליך התממה אפקטיבי מגן על הנתונים הפרטניים ויחד עם זאת משמר את שימושיות המידע גם לאחר איבוד חלק ממנו. מידת ההתממה נקבעת בהתאם לתרחישים של חשיפת המידע שמפניהם רוצים להתגונן. בניית תרחישים אלו היא תהליך מורכב, המצריך מומחיות בתוכן ומביא בחשבון גם את הימצאותם של מאגרים משלימים הזמינים למשתמשים ומאפשרים הצלבת מידע וזיהוי של הפרטים.

בעידן שבו ניתוח המידע מתבסס יותר ויותר על מאגרים עצומים של נתונים פרטניים יידרש בנק ישראל להמשיך ולבצע תהליכי התממה מורכבים, כדי לאפשר את חופש המידע לצורכי מדיניות ומחקר כלכלי ובד בבד לשמור על סודיות המידע הפרטני כמתחייב על פי החוק.

ביבליוגרפיה

- [1] Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer and Peter-Paul de Wolf (2012), Statistical Disclosure Control, First Edition.
- [2] Dalenius, T. and Reiss S.P. (1978), Data-swapping: a technique for disclosure control. Proceedings of the ASA Section on Survey Research Methods, pp. 191–194. American Statistical Association, Washington DC.
- [3] Defays D. and Nanopoulos P. (1993), Panels of enterprises and confidentiality: the small aggregates method. Proceedings of 92 Symposium on Design and Analysis of Longitudinal Surveys, pp. 195–204. Statistics Canada, Ottawa.
- [4] Duncan G., Keller-McNulty S. and Stokes S. (2001), Disclosure risk vs. data utility: the r-u confidentiality map. Technical Report LA-UR-01-6428, Los Alamos National Laboratory, Statistical Sciences Group, Los Alamos, New Mexico.
- [5] Gehrke J., Kifer D., Machanavajjhala A., and Venkatasubramaniam M., (2006), “L-diversity: privacy beyond k-anonymity,” 22nd International Conference on Data Engineering (ICDE'06), Atlanta, GA, USA, pp. 24-24.
- [6] Gouweleeuw J.M., Kooiman P., Willenborg L.C.R.J. and de Wolf P.P. (1997), Post randomization for statistical disclosure control: Theory and implementation. Technical report, Statistics Netherlands. Research paper no. 9731.
- [7] Samarati P. (2001), Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13(6), 1010–1027.
- [8] Sullivan G.R. (1989), The Use of Added Error to Avoid Disclosure in Microdata Releases. PhD thesis Iowa State University.