

סדרת ניירות עבודה

WORKING PAPER SERIES

מס' 107 No.

דגימה מרחבית

Spatial Sampling

תיאודור יצקוב ואלכסנדרה קצנלנבוגן*

Theodor Itskov and Alexandra Katzenelenbogen*

אדר תשע"ח, מרץ March 2018

***הלשכה המרכזית לסטטיסטיקה –אגף מתודולוגיה סטטיסטית**

***Central Bureau of Statistics –Department of Statistical Methodology**



הלשכה המרכזית לסטטיסטיקה (הלמ"ס) מעודדת מחקר המבוסס על נתוני הלמ"ס , כדוגמת עבודה זו. עבודות מחקר אלו אינן פרסומים רשמיים של הלמ"ס, ומכאן שהדעות והמסקנות הבאות בהן לידי ביטוי, הן של המחברים עצמם ואינן משקפות בהכרח את הדעות והמסקנות של הלמ"ס.

הוצאת הלשכה המרכזית לסטטיסטיקה, רח' כנפי נשרים 66, פינת רח' בקי,

ת"ד 34525, ירושלים 91342

טל': 02-6592666; פקס: 02-6521340

אתר הלמ"ס באינטרנט: www.cbs.gov.il

דואר אלקטרוני: info@cbs.gov.il

תקציר

מבנים הנמצאים בסמיכות גיאוגרפית זה לזה מאוכלסים בדרך כלל על ידי אנשים השייכים לאותה שכבה חברתית-כלכלית. הגורם הגיאוגרפי הוא לכן משתנה דגימה חשוב לצורך ייצוג כל השכבות באוכלוסיית האזור. בעבודה זו נציג שיטת דגימה ייחודית המבוססת על הגורם המרחבי/גיאוגרפי של המבנים במסגרת של האזור הסטטיסטי. השיטה מבוססת על ספירלה המתפתחת במרחב בו נמצאים המבנים תוך שמירה על הסתברות דגימה אחידה ודגימה מייצגת שלהם במרחב. בעבודה נחשב את המרחק המינימלי בין כל מבנה נדגם לשכנו הקרוב ביותר ונציג מדד מסוג מיני-מקס (minimax) המאפיין את המדגם בו ממוצע המרחקים המינימליים הוא הגדול ביותר. המדד יאפיין, במובן מסוים שיוגדר בעבודה, את הפריסה המיטבית של היחידות הנדגמות על פני המרחב. בעבודה זו נתייחס גם לצורת המרחב בו נמצאים המבנים במסגרת הדגימה. נגדיר מדד שאינו תלוי ביחידות המידה (אורך או שטח) ונציג תוצאות המראות כי דגימה מרחבית אופטימלית אינה תלויה כל כך במדד הצורה בטווח העיקרי של אזורים סטטיסטיים. את יעילות השיטה החדשה התלויה בגורמים גיאוגרפיים שונים נציג בעזרת סימולציות תוך השוואה לתוצאות שהיו מתקבלות בדגימת ממ"פ מאותה המסגרת באזור.

מילות מפתח: דגימה מרחבית, ספירלה של ארכימדס, מיני-מקס.

תוכן עניינים

3		תקציר
5		1. הגדרת המשימה
6		2. הגדרת שיטה לדגימה מרחבית
9		3. חקירת תכונות גיאוגרפיות של אזורים סטטיסטיים
10		4. חקירת הגורמים של שיטת הדגימה
12		4.1 צפיפות נקודות
13		4.2 צפיפות נקודות ושטח האזור
15		4.3 פרמטר צורת האזור
16		4.4 שבר דגימה
17		4.5 מדגם מקרי פשוט בהשוואה עם דגימה מרחבית
18		5. סיכום ומסקנות
19		רשימת מקורות

1. הגדרת המשימה

תהליך דגימה בסקרים צריך לספק ייצוג טוב של אוכלוסייה הנמצאת באזורים גיאוגרפיים. לחקירה יעילה של האוכלוסייה עדיף קודם לחלק אותה לשכבות, כך שכל אחת מהן תכיל תת-אוכלוסייה עם תכונות דומות. לאחר מכן מכל שכבה דוגמים פרטים\משפחות בכמות פרופורציונית לגודלה. המצב הרגיל הוא שלפני המעשה אין מידע ודאי על קבוצות אוכלוסייה אחדות פרט למידע גיאוגרפי. לכן, בתהליך הדגימה לוקחים בחשבון משתנים גיאוגרפיים. **המטרה היא לתת לכל יחידה אותה הסתברות להידגם ויחד עם זאת לקבל מדגם פרוס על פני כל השטח.** שילוב של שתי ההתניות מאפשר לקבל מדגם שמייצג הכי טוב את האוכלוסייה באזור, בהינתן שמיקום גיאוגרפי של היחידות מאפיין תכונות שנחקרות בסקר.

המשימה המעשית שעומדת מאחורי הניסוח הכללי מתוארת באופן הבא: מסגרת הדגימה היא מרשם מבנים, אשר מבוסס על שכבת מבנים במערכת ממ"ג. לכל מבנה בין היתר ידוע המזהה החד-חד ערכי שלו, הישוב והאזור הסטטיסטי שהוא שייך לו וכן קואורדינטות Y-X שנמדדות במטרים. עבור כל אזור סטטיסטי או ישוב חד-אזורי צריך לדגום אחוז מסוים של המבנים כך שיתקבל מדגם מקרי המייצג את אוכלוסיית האזור באופן הטוב ביותר. מרשם מבנים נמצא ברשותנו וישרת אותנו לצרכי פיתוח העבודה.

אחת הגישות לדגימה גיאוגרפית הינה חלוקת האזור לתאים ע"י קווים אופקיים ואנכיים. לאחר מכן, דוגמים מכל תא לפי צפיפות של יחידות המסגרת בתוך התא. הגישה הזאת יכולה להיות יעילה כאשר כל תא הוא גדול מספיק כך בהינתן שבר הדגימה של הסקר אפשר לדגום מספר יחידות גם בתאים עם צפיפות נמוכה. מצב כזה יכול להיווצר בחקירת יישובים בגודל בינוני המכילים 1,000 מבנים ויותר, ופרט לזה שבר הדגימה צריך להיות מספיק גבוה (תלוי במספר מבנים). כאשר נחקרים אזורים סטטיסטיים או יישובים קטנים גישה זו לא מתאימה, כיוון שאי אפשר לחלק אותם לתאים גדולים מספיק. נמנה את הבעיות העיקריות של שיטת התאים:

א. שליטה רעה על שבר דגימה בתאים. אי אפשר לדגום 10% מבנים בתאים קטנים (המכילים 3-6 מבנים) ובתאים יותר גדולים (7-14 מבנים) דגימה של מבנה 1 מביא לשבר דגימה שונה מאוד.

ב. משימת הדגימה בתוך תאים הופכת למשימת ההקצאה של מספר נתון המבנים. סיכוי להידגם למבנים יהיה שונה מאוד תלוי בגודל התא.

ג. דגימה מקרית בתוך תא בכלל לא לוקחת בחשבון מרחקים בין בתים, גם בתוך תא גם מעבר להם.

ד. שיטת התאים היא בעייתית באזורים לא מלבניים – ייווצרו תאים ריקים וקטנים בפריפריה. בכוונה למנוע את הבעיות האלה נקטנו בגישה אחרת לדגימה מרחבית.

2. הגדרת שיטה לדגימה מרחבית

אפשר לחלק את השיטות לדגימה גיאוגרפית לשתי קבוצות:

שיטות שמייצרות נקודות על פני שטח נתון.

שיטות שבוחרות חלק ממערך נקודות נתון.

במסגרת העבודה התרכזנו במקרה השני, אשר יותר אופייני לסקרי למ"ס. אכן, בדרך כלל הדגימה מתבצעת על מסגרת של יחידות (בניינים/כתובות) שמיקומן ידוע מראש (ראה Clark et. al. (1954), Bostoen et. al. (2006), ו-Serfling et. al. (1975)), והסוקר נשלח לבניינים מסוימים ולא לנקודות בשטח עם קואורדינטות מסוימות.

למידת היעילות של פריסה מרחבית של מדגם מבנים על פני השטח יש להגדיר **פונקציית מטרה**. דוגמה לפונקציה כזאת – סטיית תקן של מרחק הנקודות ממרכז האזור (או מרכז יחידות של המסגרת). דוגמה שנייה – סכום (או מקסימום) סטיות התקן המחושבות בנפרד לקואורדינטות X ו- Y של היחידות. החיסרון של שתי הפונקציות הוא שהן לא לוקחות בחשבון מיקום של נקודה ביחס לנקודות אחרות של המדגם. דוגמה שלישית לפונקציית המטרה – מרחק ממוצע בין כל n נקודות שנדגמו (סך - $(n-1)/2$ מרחקים). לפונקציה הזאת אין כבר את החיסרון הקודם, אך עדיין יכולים לקבל ערכים גבוהים שלה, כאשר בקרב נקודות שנדגמו קיימות נקודות קרובות אחת לשנייה. על מנת שפונקציה תושפע יותר ע"י נקודות קרובות, עבור כל נקודה i נחשב מרחק מינימאלי r_i לנקודות אחרות. נגדיר את פונקציית המטרה כממוצע של המרחקים המינימאליים האלו (Clark et. al. (1954)):

$$R = \sum_{i=1}^n r_i / n \quad (1)$$

ככל שפונקציית R גדלה, המרחקים המינימאליים בין הנקודות גם גדלים, לכן מטרתנו להביא אותה למקסימום. נקרא ל- R – מדד פריסה מרחבית של המדגם.

הערה: אפשר היה להגדיר את פונקציית המטרה כמינימום של מרחקים מינימאליים r_i , אבל זה לא נראה מוצדק, כיוון שדגש על המרחק המינימאלי ביותר מזניח מרחקים בין נקודות אחרות.

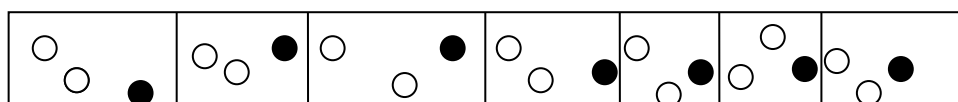
יש לציין נקודה חשובה: מדגם שמביא פונקציית מטרה כלשהיא למקסימום, לא מקיים את ההתניה הראשונה של תהליך הדגימה. אכן, יש רק קבוצה אחת של נקודות שממקסמת את פונקציית (1), לכן

הן ייבחרו בהסתברות 1 ולנקודות אחרות אין סיכוי להיבחר כלל. עובדה את סותרת את ההתניה שלכל נקודה צריך להיות אותו סיכוי להיבחר. כדי לספק אותה צריך לפתח אלגוריתם לדגימת הנקודות בהסתברות שווה שבו זמנית פורס אותם לרחבי השטח ומביא לערכים סבירים של מדד R . במקרה קלאסי, כאשר בונים מדגם מקרי פשוט, ממיינים מסגרת דגימה לפי משתנה מקרי, מתחילים מיחידה שמספרה גם נבחר באופן מקרי מ-1 ועד N/n (סך היחידות במסגרת) ובוחרים יחידות אחרות עם צעד n .

כאשר רוצים לשפר ייצוגיות של הנדגמים לפי משתנה קטגוריאלי (או לפי אוסף משתנים קטגוריאליים), ממיינים את מסגרת הדגימה לפי משתנה זה, ובתוך הקטגוריות שלו לפי משתנה מקרי. לאחר מכן, דוגמים כמו במקרה הקודם. השיטה הזאת תקפה גם כאשר צריך לשפר ייצוגיות לפי משתנה רציף אחד, כי תמיד אפשר למיין את היחידות לפי משתנה כזה.

הבעיה בדגימה מרחבית היא שמרחב מוגדר בשני משתנים רציפים X ו- Y . מיון לפי אחד מהם או לפי כיוון מסוים מוצדק רק כאשר שטח מתוח לאורך הכיוון (תרשים 1). אזי התקדמות לפי מיון כזה מביא לנקודות מרוחקות אחת מהשנייה.

תרשים 1. דוגמה לדגימה של כל נקודה שלישית אחרי מיון נקודות לפי ציר X באזור אופקי



באזורים בעלי צורה שרירותית, מיון הנקודות צריך לקדם אותן בשני הכיוונים כך שההתקדמות הזאת תרחיק אותן אחת מהשנייה. למטרה הזאת אנו מציעים אלגוריתם המבוסס על ספיראלה של ארכימדס. הרעיון הוא – למספר נקודות לפי הזווית שלהם בקואורדינאטות קוטביות, אבל לקדם את המספור בתוך סיבובים של הספיראלה. בתרשים 2 אפשר לראות את אופן המיון של הנקודות. למשל, נקודה 3 קודמת לנקודה 4 למרות שהיא רחוקה יותר מהמרכז, בגלל שספיראלה לא מפרידה אותן ולפי זווית מגיעים לנקודה 3 קודם. כלומר מספור נקודות מתבצע מהמרכז לפי זווית בין סיבובים של ספיראלה.

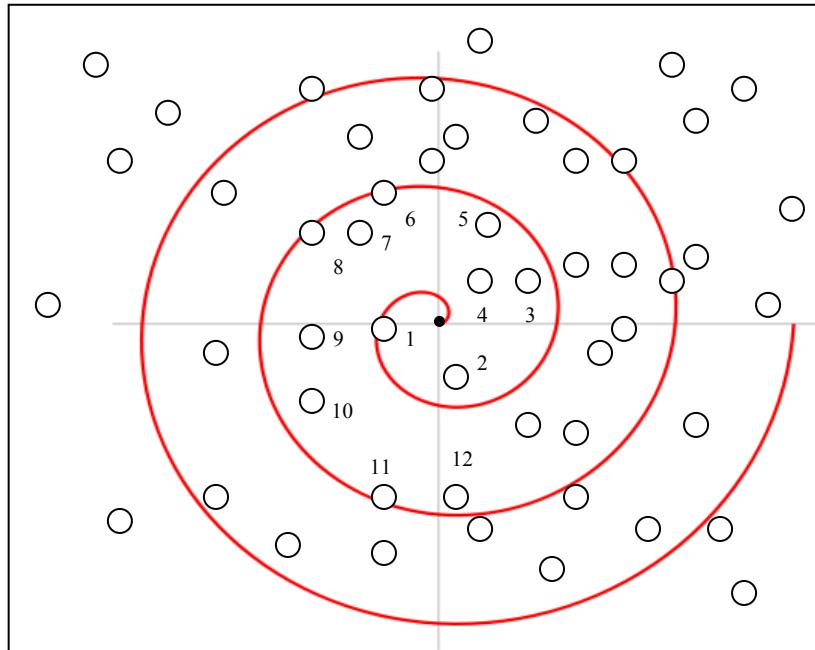
נכון שאחרי סיבוב אחד שלם מגיעים לנקודות אשר יכולות להיות קרובות לאלו ששייכות לסיבוב הקודם, אבל,

א. התקדמות מתבצעת גם לכיוון של זווית וגם לכיוון של רדיוס בו זמני, זאת אומרת בשתי קואורדינאטות;

ב. אורך הסיבובים גדל באופן רציף.

תכונה (ב) מונעת מצב של צעד דגימה "לא מוצלח" כאשר כל הנקודות הנדגמות ימצאו על קו אחד.

תרשים 2. מיון נקודות לפי ספיראלה של ארכימדס



כפי שרואים מהתרשים, נקודות מתקדמות ממרכז השטח לפריפריה בכיוון הספיראלה. חשוב לנו שמרחק בין סיבובי הספיראלה יישמר לאורך כל הדרך. ככול שהמרחק הזה קטן, ההתקדמות (המיון) מתבצעת יותר לפי רדיוס ופחות לפי זווית. ככול שהמרחק בין הסיבובים גדל, המיון מתבצע יותר לפי זווית ולא לפי רדיוס. אפשר לומר ששטח נפרס לרצועה בדומה לדוגמה בתרשים 1. סביר להניח שמיון זה עונה לדרישות תהליך הדגימה. המשך העבודה מיועד לניתוח תוצאות הדגימה אחרי שמיון נקודות נעשה בשיטה זו.

כדאי לחדד שבסט-אפ המוגדר בסעיף 1 "לתת לכל יחידה אותה הסתברות להידגם ויחד עם זאת לקבל מדגם פרוס על פני כל השטח" אי אפשר לדבר על פתרון אופטימלי. קיים פתרון (ובד"כ יחיד) שמביא למקסימום מדד פריסה מרחבית מסוים, אך המדגם לא יהיה מקרי. לשיטה המוצעת יש יתרונות מסוימים, אשר אפשר לנסח עוד לפני הסימולציות שנתאר בהמשך:

א. לשיטה המוצעת אין בעיה קשורה לחלוקה לתאים וצורך להתנהג בתאים קטנים או גדולים.

ב. תהליך הדגימה בין סיבובים של ספירלה מקדמת ממבנה למבנה בשני הכיוונים בו זמנית: ממרכז לפריפריה ולפי זווית. זאת אומרת, כל מבנה הבא שנדגם יורחק ממבנים שנדגמו קודם.

ג. דגימה סיסטמית לפי ספירלה נותנת אותה הסתברות דגימה לכל בית בזכות דגימה מקרית של בית ראשון מתוך מספר מסוים של מבנים ראשונים (המספר הזה שווה להיפוך שבר דגימה).

ד. שיטת הספירלה ישימה באותה מידה וללא סיבוכים נוספים לכל צורות האזור ולכל התפלגויות צפיפות המבנים באזור.

3. חקירת תכונות גיאוגרפיות של אזורים סטטיסטיים

על מנת לנתח את שיטת הדגימה נלמד קודם תכונות גיאוגרפיות של אזורים סטטיסטיים בארץ. מידע זה יעזור להגדיר משתנים החשודים כגורמים לתהליך הדגימה וטווח הערכים שלהם.

נבחרו המשתנים הבאים שמאפיינים את האזור:

מספר מבנים באזור.

שטח האזור - חושב כשטח של גוף קמור שנבנה על מבנים למגורים באזור.

צפיפות מבנים – מספר מבנים באזור חלקי שטח של האזור.

היקף האזור (משתנה עזר) - חושב כהיקף של גוף קמור שנבנה על מבנים למגורים באזור.

פרמטר צורת האזור – היקף האזור בריבוע חלקי שטח.

צפיפות נקודות ופרמטר צורת האזור, הם משתנים יחסיים ולכן פחות מושפעים ביחידות המידה ובגודל מוחלט של האזור. צפוי שהמשתנים האלו ישפיעו על תוצאת הדגימה, אבל ייבדקו גם האחרים.

כל המשתנים הללו חושבו עבור 3,012 אזורים סטטיסטיים השייכים ל-1,236 ישובים ומכילים 10 מבנים ויותר למגורים. בניית גוף קמור לאזורים סטטיסטיים על מערך של מרכזי המבנים נעשה בעזרת פונקציה cvexhull שקיימת בתכנת SAS. שטח והיקף של גוף קמור חושב בעזרת חלוקתו למשולשים מנקודה אחד של הגוף, שתמיד אפשרי לגופים קמורים. בטבלה 1 אפשר לראות פרמטרים סטטיסטיים עבור המשתנים האלו.

טבלה 1. סיכום תכונות של משתנים גיאוגרפיים עבור 2,934 אזורים סטטיסטיים שמכילים 10 מבנים למגורם לפחות

משתנה	עשירון תחתון (10%)	חציון (50%)	ממוצע	עשירון עליון (90%)	מינימום	מקסימום	סטיית התקן
מספר מבנים (N)	50	152	234	519	10	2,126	244
שטח (מ"מ)	101,639	320,658	336,625	1,455,474	6,571	$3.7 \cdot 10^9$	$68.3 \cdot 10^6$
צפיפות מבנים (P, קמ"מ/N)	185	497	564	1,037	< 1	3,372	359
היקף (מ')	1,333	2,371	3,144	5,015	337	284,512	6,088
פרמטר צורה (c)	14.7	16.7	18.4	22.6	13.5	589.7	12.3

4. חקירת הגורמים של שיטת הדגימה

אופן מיון נקודות בעזרת ספיראלה של ארכימדס תלוי בפרמטר ליניארי a של הספיראלה:

$$\rho = a \cdot \varphi \quad (2)$$

כאשר (ρ, φ) - קואורדינטות קוטביות:

ρ - אורך של ווקטור מראשית הצירים לנקודה על הספיראלה (משתנה מ-0 ועד אינסוף).

φ - זווית בין הווקטור לכיוון הראשי (גם מ-0 ועד אינסוף).

בתור ראשית הצירים קובעים את נקודת הממוצע של קואורדינטות Y-X מבנים באזור, בתור הכיוון הראשי לחישוב זוויות קובעים כיוון של ציר X.

ככול שפרמטר a גדול יותר, המרחק בין סיבובי הספיראלה מתרחב וכאשר a שואף לאינסוף נקבל מיון רק לפי זווית φ . ככול שהפרמטר a קטן המרחק בין הסיבובים גם קטן וכאשר a שואף ל-0 נקבל מיון רק לפי מרחק מראשית הצירים ρ . כלומר, במקרה האחרון הספיראלה עושה אינסוף סיבובים ואפשר לתאר אותה כאינסוף מעגלים צפופים.

מטרת סעיף זה – לברר אילו משתנים משפיעים על מדד הפריסה R ועבור משתנים אלה למצוא ערך אופטימאלי של פרמטר a הממקסם את המדד R.

ראשית נתאר את תהליך אמידת מדד R במצב של משתנים קבועים. האמידה נעשתה ע"י 50 סימולציות, שכל אחת מהן מכילה:

יצירת מסגרת דגימה – N נקודות מפולגות באופן אחיד על שטח בגודל מסוים ובצורה מסוימת. אם לא נאמר אחרת, השטח הוא ריבועי בגודל של 500×500 מ"מ (מטבלה 1 רואים שחציון של שטח האזורים הוא 0.32 קמ"מ או 565×565 מ"מ).

חישוב נקודה ממוצעת ומספור נקודות המסגרת לפי שיטת הספיראלה, כפי שהוסבר בסעיף 2 ובתרשים 2. המיון נעשה בהינתן פרמטר a מסוים של ספיראלה.

בתור נקודה ראשונה במדגם לוקחים אחת אחרי השנייה את כל הנקודות מ-1 ועד W מתוך נקודות המסגרת הממוינות כפי שתואר בשלב 2, כאשר W הוא היפוך של שבר הדגימה.

עבור כל נקודה ראשונה, שנבחרה בשלב 3, בוחרים נקודות אחרות למדגם מהמסגרת הממוינת עם צעד W. כך מקבלים W מדגמים שונים.

עבור כל מדגם מחשבים מדד פריסה מרחבית R בעזרת נוסחה (1).

מחשבים ממוצע $\bar{R}$ וסטיית תקן $S(R)$ למדדי הפריסה שהתקבלו עבור W מדגמים שונים.

כיוון שבתהליך דגימה רגיל בוחרים את הנקודה הראשונה למדגם באופן מקרי בהסתברות שווה, רמת פריסת הנקודות מאופיין דווקא ע"י מדד $\bar{R}$ ולא ע"י מדד בודד R, המבוסס על דגימה של נקודה ראשונה מסוימת. כמו כן $S(R)$ היא סטיית התקן שמאפיינת סטייה של מדד הפריסה במצב מסוים של המשתנים. בחירת נקודה ראשונה הוא גורם שמשפיע חזק על פריסת המדגם, לכן סטייה של מדדים בודדים $S(R)$ די גדולה יחסית למדד R (אנו נראה את זה בהמשך). על מנת לקבל אומדן אמין למדד המטרה שלנו $\bar{R}$ ואת הסטייה שלו, כל התהליך מ-1 עד 6 הופעל 50 פעמים. כתוצאה מזה חושבו אומד ממוצע $\bar{\bar{R}}$ של מדד $\bar{R}$ וסטיית התקן שלו $S(\bar{R})$. אלה הפרמטרים שמאפיינים את רמת הפריסה הממוצעת במצב מסוים של גורמים.

ניתוח תלות של פרמטר a אופטימאלי במשתנים שונים נעשה בנפרד לכל משתנה ובמצב קבוע של משתנים אחרים. לא נבנתה תוכנית ניסויים רב-גורמית כיוון שהיה צפוי שרוב הגורמים לא ישפיעו על התוצאה או ישפיעו במידה קלה. המצב הקבוע של הגורמים נבחר לפי החציון שלהם שחושב על נתונים אמיתיים (טבלה 1). טווח הערכים למשתנה הנחקר גם נבחר מטבלה 1 לפי העשירון התחתון (10%) והעשירון העליון (90%).

פרט למשתנים שמאפיינים אזורים סטטיסטיים נחקר גם היפוך שבר הדגימה W . המצב הקבוע שלו נקבע ל-10 (דגימה 1 מ-10 או 10% נדגמים) – היפוך שבר דגימה המתוכנן בינתיים למפקד מתגלגל. הקצוות שנבחרו הם 5 (1 מ-5 או 20% נדגמים) ו-50 (1 מ-50 או 2% נדגמים). בהינתן מספר מבנים ממוצע באזור סטטיסטי - דגימה פחות מ-2% תביא לסטיית תקן יחסית גבוהה שמקשה להבחין בין החלופות.

על מנת למצוא ערך אופטימאלי של פרמטר a , נעשו מספר סימולציות עם ערכים שונים של הפרמטר. טווח של פרמטר a שמכיל ערך אופטימאלי שלו וצעד ההתקדמות הוגדרו דינאמית על סמך הניסיונות. כאשר נצפית תלות מובהקת בין פרמטר a לבין מדד $\bar{R}$, התלות הזאת תמיד מכילה מקסימום אחד. בדרך כלל קיים מקסימום לוקאלי ובתנאים מסוימים המקסימום נמצא בקצה של הערכים.

4.1 צפיפות נקודות

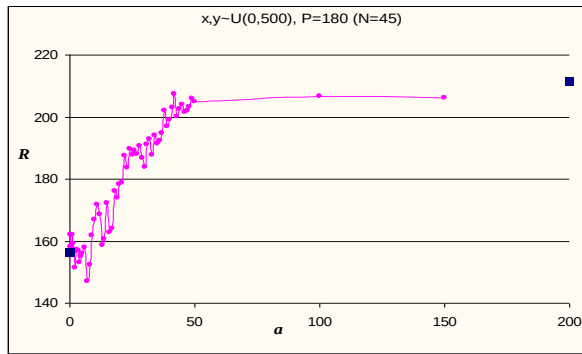
צפיפות נקודות בשטח (P) נראה משתנה חשוב ביותר לקביעת תהליך הדגימה. בסעיף זה הוא נחקר בתנאים קבועים של משתנים אחרים: שטח ריבועי 500×500 מ"מ או 0.25 קמ"מ, התפלגות אחידה של נקודות המסגרת (נכון עבור כל הניתוחים בסעיף זה), היפוך שבר דגימה $W = 10$.

התלות בין מדד פריסה מרחבית R לפרמטר ספיראלה a נחקרה בכמה נקודות של צפיפות מהעשירון התחתון ועד העשירון העליון: $45/0.25=180$, $100/0.25=400$, $200/0.25=800$, $250/0.25=1000$ (קמ"מ / N).

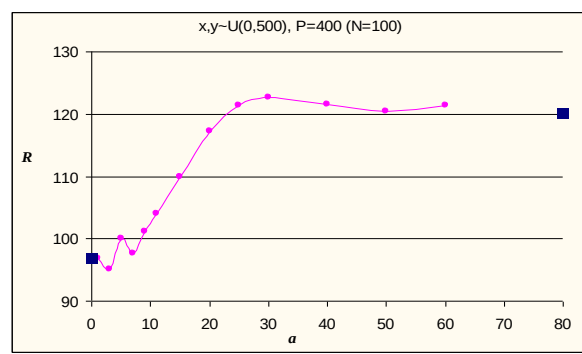
פרט לחקירת התלות של R מ- a , נעשו שני חישובים נוספים – אחד עבור מיון לפי מרחק מראשי הצירים עד הנקודה (מקרה a שואף ל-0) והשני עבור מיון לפי זווית (מקרה a שואף לאינסוף). על הגרפים מוצגות תוצאות חישובים אלה על ידי נקודות מרובעות בצבע כחול בהתחלה ובסוף של ציר ה- X .

כל התוצאות של הסימולציות כולל סטיות התקן נמצאות בקובץ excel שאפשר לקבל לפי צורך. במסמך זה אנו מציגים את הגרפים.

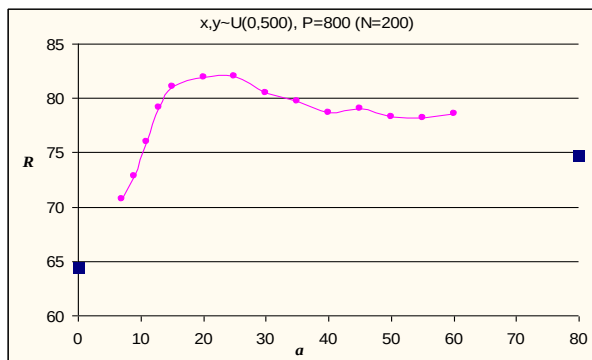
גרף 1



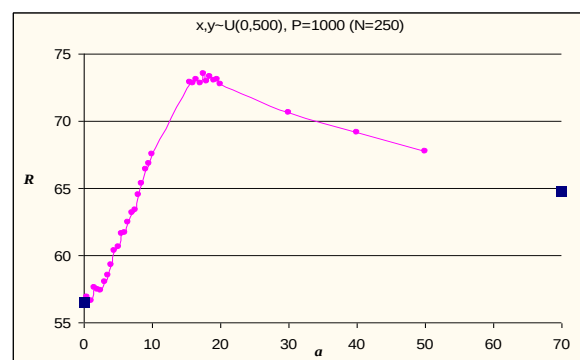
גרף 2



גרף 3



גרף 4

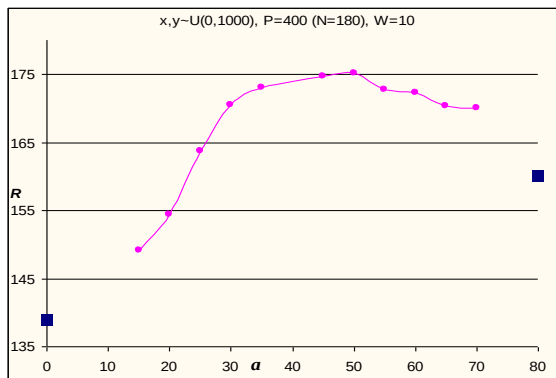


מגרף 1 אפשר להסיק שעבור צפיפות נמוכה - 180 (נקודות) מקסימום מדד R מושג בדגימה לפי זווית (a שואף לאינסוף). כאשר צפיפות עולה עד ל-400 (100 נקודות, גרף 2), מופיע מקסימום לוקאלי בנקודת $a=30$, אשר לא שונה באופן מובהק מאותו מקרה קיצוני. כאשר צפיפות גדלה עד 800 (200 נקודות, גרף 3), קיים מקסימום לוקאלי מובהק בנקודת $a=25$. עבור צפיפות 1000 (1000 נקודות, גרף 4), רואים מקסימום מובהק אחד בנקודת $a=17.5$. נקודות המקסימום נאספו לטבלה מסכמת (תוצג בהמשך העבודה).

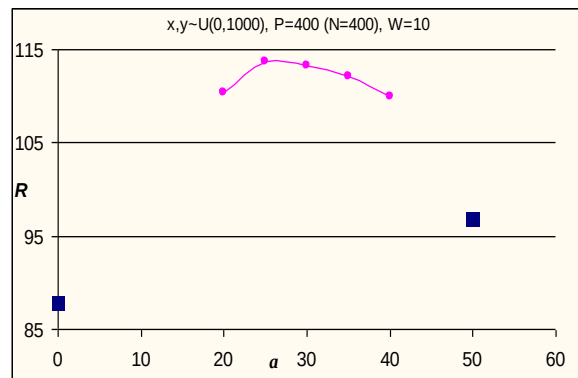
4.2 צפיפות נקודות ושטח האזור

מלכתחילה לא כל כך ברור, האם צריך לשנות את אופן הדגימה באזורים בעלי אותה צפיפות, השונים בגודלם. על מנת לבדוק את השאלה עשינו ניסויים בדומה לסעיף 4.1 אבל עם שטח אחר – 1000×1000 ו- 250×250 מ"מ. להלן מובאים גרפים 5-8 מאותו סוג שבסעיף הקודם עבור שטח 1000×1000 .

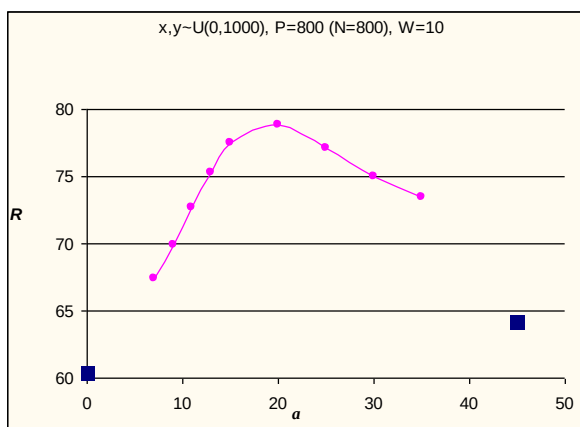
גרף 5



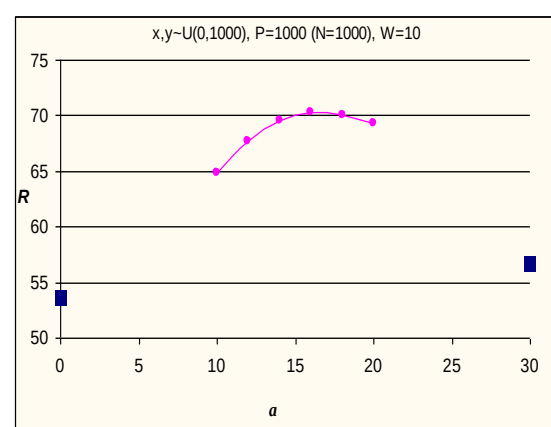
גרף 6



גרף 7



גרף 8



מהגרפים רואים שלכל הצפיפויות שנחקרו קיים מקסימום לוקאלי ובדרך כלל הוא נמצא לא רחוק מהמקסימום עבור אזורים בשטח 500×500 . לצפיפות 400 (גרף 6) המקסימום מושג בנקודה $a=25$ (קודם היה 30), לצפיפות 800 (גרף 7) – $a=20$ (קודם היה 25), ולצפיפות 250 (גרף 8) – $a=16$ (קודם היה 17.5). בצפיפות נמוכה של 180 (גרף 5) הודגם מקסימום בנקודה רחוקה $a=50$ לעומת המצב הקודם, כאשר מקסימום לוקאלי לא הודגם כלל.

אפשר לומר שנצפית תלות שלילית קלה בין גודל שטח לבין ערך אופטימאלי של פרמטר ספיראלה.

ניסוי לבירור התלות הנ"ל עבור שטח 250×250 לא הביא מקסימום לוקאלי באף שכיחות מהסדרה. כל פעם המקסימום הושג במיון לפי זווית (a שואף לאינסוף), לכן לא מוצגים גרפים אלה במסמך (אפשר לצפות בקובץ excel המצורף). אחת הסיבות לזה היא שכדי להשיג ערכים של השכיחויות הנחקרות, בשטח קטן צריך לדגום מספר קטן של נקודות (מ-2 עד 6). סטיית התקן שמתקבלת בדגימה כזאת לא נותנת אפשרות לזהות תלות, אפילו אם היא קיימת.

4.3 פרמטר צורת האזור

נדון על משמעות של פרמטר צורת האזור, אשר הגדרנו כהיקף בריבוע חלקי שטח. פרמטר זה לא תלוי ביחידות המידה והוא שווה עבור אזורים דומים במובן גיאומטרי (שונים בגודל ולא בצורה שלהם). בהינתן אותו שטח, ככול שההיקף גדול יותר הצורה של אזור קמור נהייתה מתוחה יותר בכיוון כלשהוא. הצורה הסימטרית ביותר זה עיגול. ערך הפרמטר שלו הוא $4\pi \approx 12.57$, וזה הערך המינימאלי של פרמטר צורה. עבור ריבוע, שהוא גם בעל סימטריות גבוהה, ערך הפרמטר הוא 16. מטבלה 1 רואים שלאזור הסימטרי ביותר יש ערך פרמטר 13.5, לאזור המתוח ביותר ערך פרמטר – 590 (ערך חריג), החציון – 16.7 מאוד קרוב לצורה ריבועית (!), והעשירון העליון שלו כ- 22.

על מנת לבדוק את השפעתו של פרמטר צורה c , נניח שצורת האזור היא מלבן עם צלעות x ו- y ונרשום את ההגדרה של פרמטר צורה עבור מלבן:

$$4 \cdot (x + y)^2 / (x \cdot y) = c \quad (3)$$

בלי הגבלת הכלליות אפשר להניח ששטח האזור $x \cdot y$ הוא 1 (לכן $y = 1/x$) ולקבל משוואה ריבועית עבור x . פיתרון המשוואה הוא:

$$x_{1,2}^2 = (c - 8 \pm \sqrt{c(c - 16)}) / 8 \quad (4)$$

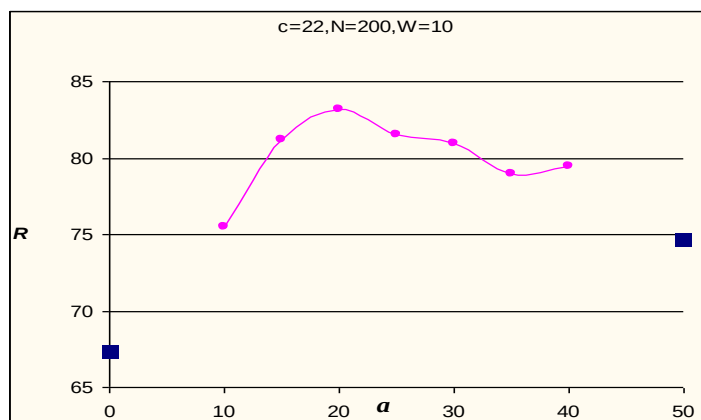
כפי שציינו קודם, פרמטר צורה עבור מלבנים גדולשווה ל-16, לכן הערך תחת שורש תמיד חיובי. אפשר להוכיח שבחירת סימן '+' או '-' במשוואה (4) נותנת פתרונות סימטריים עבור x ו- y (ששווה ל- $1/x$). פתרונות אלה מתאימים למלבן אופקי ולמלבן אנכי, ולכן מספיק לקחת אחד מהם.

כאשר c שווה ל-22 (עשירון עליון), מקבלים $x = 1.78$, $y = 0.56$. כדי לעבור למידות ריאליות צריך להכפיל אותם על שורש של השטח הרצוי. במקרה שלנו, כאשר שטח רגיל הוא $500 \cdot 500$ מ"מ, מכפילים ב-500 ומקבלים מלבן עם אותו שטח וצלעות של 890 ו-280 מטרים בערך.

על מנת לבדוק השפעת צורת האזור על תהליך הדגימה, נזרקו על המלבן 200 נקודות של מסגרת דגימה על ידי התפלגות אחידה. כיוון ששטח המלבן לא השתנה, נשמרה גם צפיפות נקודות של 800 (קמ"מ²). בגרף 9 רואים שמדד הפריסה המקסימאלי מושג בנקודה $a=20$. בגרף 3 רואים שבשטח ריבועי עם אותה שכיחות, שהמקסימום נמצא בנקודה $a=25$, אבל הוא לא חד ובנקודה $a=20$ ערך של מדד R כמעט שווה לו (הפרש של 0.1 מטר). מזה אפשר להסיק מסקנה, שצורת האזור לא משפיעה על שיטה של דגימה מרחבית. גם לאזור סימטרי וגם לאזור מתוח בוחרים

אותו צעד ספיראלה, כאשר המשתנים שכן משפיעים שווים. אפשר לנסות להסביר את העובדה, למרות שהיא לא טריוויאלית ולא ברורה מאליו.

גרף 9

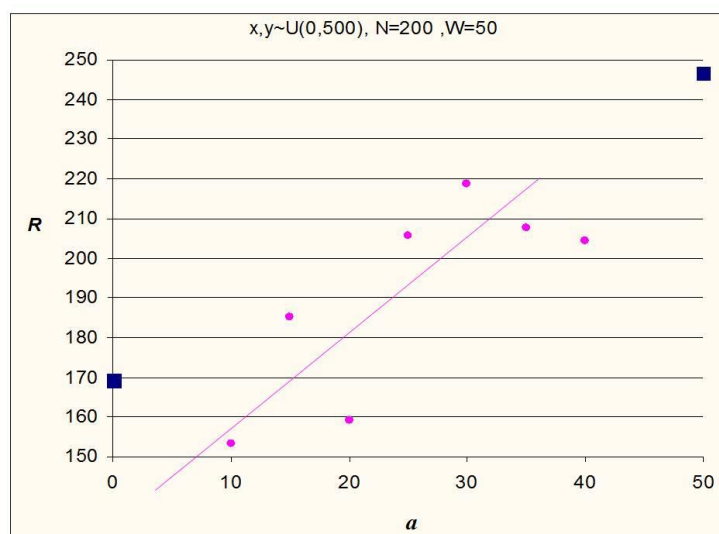


4.4 שבר דגימה

פרמטר שבר דגימה לא שייך לתכונות האזור, אבל הוא יכול להשפיע חזק על תהליך הדגימה. לכן עשינו בדיקה בסיסית גם עבור שבר דגימה. בתור תנאים קבועים בחרנו שטח ריבועי 500×500 , וצפיפות נקודות – 800 (200 נקודות על שטח זה). מצב זה דומה למוצג בגרף 3 עם היפוך שבר דגימה $W=10$. ביצענו הניסיונות גם בשתי נקודות נוספות של $W=50$ ו-5.

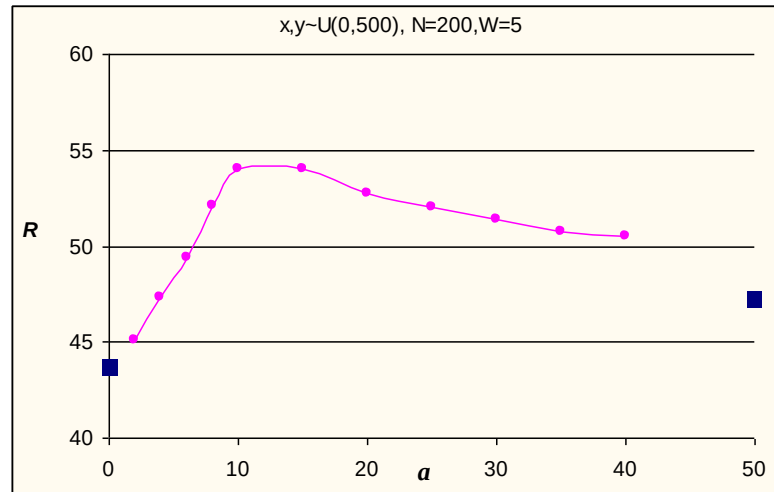
כשהיפוך שבר דגימה היה 50 (נדגמו רק 4 נקודות מ-200) פיזור נקודות בגרף 10 די גדול, ולא הודגמה נקודת מקסימום לוקאלי. המקסימום הוסג בקצה ימני של הדיאגרמה, כאשר מיון מתבצע לפי זווית, שאופייני עבור מדגמים קטנים.

גרף 10



כאשר היפוך שבר דגימה W הוא 5 (נדגמו 40 נקודות מתוך 200) המצב הוא שונה. בגרף 11 רואים תלות מובהקת בין a ל- R וקיים מקסימום לוקאלי באזור 10 – 15 של פרמטר a . לאחר מכן נצפית ירידה איטית של עקומת התלות, וערך בקצה הימני (מיון לפי זווית) הרבה יותר נמוך מהמקסימום. במקרה של $W=10$ המקסימום נמצא באזור ימני יותר – 20-25. בעזרת גרפים 3 ו-11 אפשר לבחור "נקודת פשרה" עבור שני המצבים – $a=15$, שבה מדד הפריסה קרוב למקסימום של שניהם.

גרף 11



בסיכום הסעיף ניתן לומר ששבר הדגימה משפיע על תהליך הדגימה וצריך ללמוד את ההשפעה במשימות קונקרטיות.

4.5 מדגם מקרי פשוט בהשוואה עם דגימה מרחבית

שאלה טבעית היא – בכמה שיטת הדגימה המוצעת משפרת תוצאת הדגימה שנעשה על ידי דגימה מקרית פשוטה. את ההשוואה נעשה על ידי אותו פרמטר R . להשוואה בחרנו כמה נקוות אופטימליות המוצגות למעלה עם צפיפות שונה ושטח שונה.

במצב המתואר בגרף 2 ($N=100$; $P=400$; $x, y \sim U(0, 500)$) בנקודה אופטימלית ($a=30$) לפי דגימה מרחבית משיגים ערך המדד ממוצע ~ 125 עם סטיית תקן 12.5. לעומת זאת חישוב לפי ממ"פ נותן ערך המדד ממוצע – 92 וסטיית התקן – 17.3.

במצב המתואר בגרף 3 ($N=200$; $P=800$; $x, y \sim U(0, 500)$) בנקודה אופטימלית ($a=25$) לפי דגימה מרחבית משיגים ערך המדד ממוצע ~ 82 עם סטיית תקן 6. לעומת זאת חישוב לפי ממ"פ נותן ערך המדד ממוצע – 62 וסטיית התקן – 8.

במצב המתואר בגרף 7 ($N=800$; $P=800$; $x,y \sim U(0,1000)$) בנקודה אופטימלית ($a=20$) לפי דגימה מרחבית משיגים ערך המדד ממוצע ~ 78 עם סטיית תקן 2.7. לעומת זאת חישוב לפי ממ"פ נותן ערך המדד ממוצע – 59 וסטיית התקן – 3.6.

אפשר לראות שבכל המקרים שנבדקו יחס בין ערך המדד שהתקבל בעזרת דגימה מרחבית גדול יותר בערך בשליש מערך שמתקבל לפי ממ"פ. האמירה הזאת נכונה גם עבור סטיית התקן של המדד.

5. סיכום ומסקנות

1. בעבודה מוצעת שיטה לדגימה מרחבית של מבנים בתוך אזור סטטיסטי. השיטה ממקסמת פריסת מבנים במדגם ונותנת לכל מבנה ולכל פרט סיכוי שווה להידגם.
2. לצורך מדידת פריסת המבנים נבחר מדד (R) אשר מחושב כממוצע של המרחקים המינימליים בין כל מבנה למבנים אחרים באזור. המדד הזה (א) לוקח בחשבון מיקום של כל מבנה ביחס למבנים האחרים, ו-(ב) מאפשר לשמור על פריסה ממוצעת של כל המבנים במדגם.
3. הוגדרו גורמים עיקריים אשר יכולים להשפיע על שיטת הדגימה: צפיפות מבנים באזור, שטח האזור, פרמטר צורת אזור, שבר דגימה. לכל פרמטר נמצא תחום השתנות שלו לפי נתוני GIS בארץ.
4. לכל הגורמים בתחום השתנותם נאמד ערך אופטימלי של פרמטר שיטת הדגימה (פרמטר ליניארי של ספירלה). האמידה נעשתה בעזרת סימולציות של התפלגות מבנים אחידה בתוך אזור מלבני (הוראה שפרמטר צורה ממוצע של אזורים סטטיסטיים בארץ מתאים לצורה ריבועית).
5. נעשתה השוואה של מדד פריסת מבנים R בדגימת מבנים מרחבית אופטימלית לדגימה מקרית פשוטה. בכל המקרים שנבדקו יחס בין ערך המדד שהתקבל בעזרת דגימה מרחבית גדול יותר בערך בשליש מערך שמתקבל לפי ממ"פ. במקרה של דגימה מרחבית סטיית התקן של המדד קטנה יותר גם בערך בשליש לעומת ממ"פ.
6. תוצאות האמידה של פרמטר שיטה ניתן ליישם בתהליך דגימה מרחבית של מבנים באזורים סטטיסטיים.

רשימת מקורות

- Clark, P.J. and Evans, F.C., (1954). Distance to Nearest Neighbor as a Measure of Spatial Relationships in Population. *Ecology*, Vol. 35; Issue 4: 445-453
- Bostoen, K. and Chalabi, Z., (2006). Optimization of household survey sampling without sample frames. *International Journal of Epidemiology* ;35:751–755
- Serfling R.E. and Sherman I.L., (1975). *Attribute Sampling Methods*. US Department of Health and Human Services, Public Health Service: Washington, D.C., Publication No. 1230.

Abstract

Dwellings which are adjacent to each other are usually populated by people who belong to the same socio-economic level. The geographical factor is therefore usually an important sampling variable in order to represent all the layers of the population in the area. In this work we will present a unique sampling method which is based on a spiral that evolves in the space where the dwellings are located, while maintaining a uniform sampling probability and a representative sample in space. We will also present a minimax type index that characterizes the maximum of the minimal average distance between each sampled building to its nearest neighbor. The index will characterize, in some way which will be defined in the paper, the optimal disposition of the sampled units over the space. We will also relate to the shape of the space where the dwellings are located in the sampling frame. We will define a unit-less index to the shape of an area and present results showing that the optimal spatial sampling doesn't strongly depend on the shape of the area in its main range of the statistical areas. The efficiency of the method will be demonstrated under different geographical factors by simulations and compared with the results which would have been obtained by simple random sampling.

Keywords: Spatial sampling, Archimedes spiral, Minimax.

The Central Bureau of Statistics (CBS) encourages research based on CBS data, such as this work. Works of research of this sort are not official publications of the CBS, and therefore the opinions and conclusions expressed in these publications are those of the authors and do not necessarily represent those of the CBS.

**Published by the Central Bureau of Statistics, 66 Kanfei Nesharim St.,
Corner Bachi St., P.O.B 34525, Jerusalem 91342, Israel**

Tel. 972-2-6592666; Fax: 972-2-6521340

Internet Site: www.cbs.gov.il

E-Mail: info@cbs.gov.il

WORKING PAPER SERIES

No. 107

Spatial Sampling

Theodor Itskov and Alexandra Katzenelenbogen

March 2018