חקירת המאפיינים הסטטיסטיים של אומדן צפיפות הקרנל הכפול

הרולד שיפ

תַקצִיר

מניע: ישנן דוגמאות רבות באפידמיולוגיה של נתונים המתרחשים כנקודות שמופיעות במרחב. דוגמה אחת היא כתובת הבית של אנשים בשכונה, והאם אנשים חשופים פיתחו מצב רפואי מסוים במהלך פרק זמן מסוים. נתונים אלה מאפשרים לאדם לחשב את שיעור השכיחות של מצב זה. לאחרונה, טכניקה המכונה צפיפות הקרנל הכפול שימשה להערכת פונקציות $\lambda(.,.) =: \lambda$. הפונקציה λ היא היחס בין שתי פונקציות אחרות λ_I ו- λ_P. כל אחת מהפונקציות לעיל λ_I ו- λ_P מוערכת בעצמה על ידי החלקת קבוצה של תצפיות נקודה לתפקוד רציף וחלק, תוך שימוש בהערכת עוצמת הקרנל, וערך λ בכל נקודה הוא האומדן של λ_I בנקודה זו חלקי האומדן של λ_P באותה נקודה.

החוקרים נוטים לאחרונה להשתמש בצפיפות הקרנל הכפול בשל אובדן מידע עקב איחוד נתונים באיזורים גיאוגרפיים כמו ערים או שכונות. איחוד זה נחשב נחוץ הן לצמצום השפעות השונות על הטעויות היחסיות של התוצאות והן לשמירה על הפרטיות של אנשים שחלו במחלה. השימוש בצפיפות הקרנל הכפול מאפשר גם לקשר בין נתוני מיקום למקורות נתונים אחרים, כגון רמות רווחה, זיהום אוויר ותכונות אחרות שאינן זמינות ליחידים.

עד כמה שידוע לנו, המאפיינים הסטטיסטיים של צפיפות הקרנל הכפולה עדיין לא מובנים במלואם. ברצוננו לדעת האם אומדני הסיכון או שיעור ההיארעות המתקבלים באמצעות צפיפות הקרנל הכפול מדויקים או מטעים בתנאים שונים. דאגה מיוחדת היא נקודות שיא של פונקציית הסיכון האמיתית. נקודות אלה מצביעות על נקודה שסביבה הסיכון להידבקות במחלה הוא הגבוה ביותר. בעוד דיוק האומדן של גודל הסיכון בנקודות שיא אלה הוא חשוב, זה קריטי כי אומדן המיקום יהיה מדויק, כך שלא ליצור כל האסוציאציות מטעה עם נקודות עניין אחרות בקרבת מקום. מחקר זה בוחן חלק מהמאפיינים הסטטיסטיים של צפיפות הקרנל הכפול. בפרט, אנו מנתחים באופן אמפירי כיצד גורמים שונים של האוכלוסייה והתפלגות השכיחות משפיעים על הדיוק של צפיפות הקרנל הכפול.

תיאוריה ושיטה: סוג המחקר שאנו מעוניינים בו מדבר על אירועים של מחלה כרונית. המדד הנפוץ של תדירות בשימוש בספרות נקרא שיעור ההיארעות. עבור אוכלוסייה מסוימת, תקופת זמן ומערכת של תקריות, אנו יכולים לחשב את שיעור השכיחות הכולל על ידי לקיחת המספר הכולל של האירועים וחלוקת האוכלוסייה כולה. עם זאת, אנו עשויים להשוות את שיעור ההיארעות במקומות שונים בתוך שטח. לשם כך, אנו מחשבים את שיעור השכיחות לפי נקודה.

היחידה הבסיסית של המחקר שלנו היא הניסוי, סדרה של סימולציות מונטה קרלו שהורצו עם קונפיגורציה ראשונית זהה ועם מדידות זהות. כל ניסוי מופעל עם קבוצה קבועה של פרמטרים. אנו מפעילים סדרה של ניסויים, משתנה

פרמטר אחד בכל פעם, או במקרים מסוימים שני פרמטרים במקביל, כדי לבחון את ההשפעה של פרמטר זה על הדיוק של צפיפות הקרנל הכפול.

כדי לחשב את צפיפות הקרנל הכפול, יש לקבוע פרמטר הידוע כרוחב הפס, ודיוק צפיפות הקרנל הכפול תלוי בה במידה רבה. בכל ניסוי השתמשנו בשתי טכניקות, כלל האצבע של סילברמן ו-Least Squares Cross Validation כדי לבחור את רוחב הפס. אנו משווים את התוצאות של שתי הטכניקות הללו.

על מנת לתאר את הדיוק של צפיפות הקרנל הכפול כשיטה להערכת פונקצית הסיכון האמיתית λ אנו משתמשים במספר מידות דיוק. בפרט, עבור כל ניסוי אנו מודדים mean integrated squared error, mean integrated absolute error, שגיאה ה-סופרימום, הטית השיא, סחיפת השיא, הטית centroid וסחיפת ה-centroid.

תרומתו של מחקר זה תהיה לענות על שאלות המחקר הבאות:

- כיצד הדיוק של צפיפות הקרנל הכפול מושפע מ:
  - משך המחקר,
  - פונקציות שיעור שונים,
  - גודל האוכלוסייה,
  - והתפלגויות האוכלוסייה השונות.
- כאשר מסתכלים על הגורמים לעיל, כמה מדויקת היא צפיפות הקרנל כפול:
  - באופן גלובאלי, על אזור המחקר בכללותו,
  - גודל בנקודות השיא,
  - המיקום של נקודות השיא.

תוצאות: השגנו תוצאות המראות כי הגדלת התוחלת של מקרים N, כגון מה יקרה בעת הגדלת משך המחקר, מקטין את רוחב הפס שנבחר בגורם של $N^{-1/6}$. הוא גם מקטין את טעות האומדן של צפיפות הקרנל הכפול במובן היחסי של טעויות גלובאליות לפי גורם של $N^{-3/4}$, וגם לגודל נקודה ולטעויות שיא. עם זאת, במקרה זה הטעויות המוחלטות גדלו עם תוחלת התקריות. כמו כן, נצפה כי העליה ב - σ המרווח של פונקציית הסיכון (קצב) גורמת גם לשגיאת אומדן מופחתת הן במונחים מוחלטים והן במונחים יחסיים של $\sigma^{-1.4}$. כמו כן, ראינו כי הגדלת תוחלת האירועים במקביל לגודל האוכלוסיה צמצמה את טעות האומדן, בשיעור של כ-$N^{-2/7}$, וכי אוכלוסייה לא אחידה הגדילה מעט את טעות האומדן, למעט כאשר התפלגות האוכלוסייה הייתה מאוד צרה.

מסקנה: בדוגמאות שבדקנו, מצאנו כי סטטיסטית, צפיפות הקרנל הכפול יכולה לתת קירוב טוב לסיכון אמיתי או לפונקציית שכיחות. זה נכון במונחים של הדיוק הכללי של השיעור בכל נקודה נתונה, כאשר השגיאה הריבועית הממוצעת המשולבת הממוצעת הייתה כמעט תמיד פחות מ -5% מן האמת, למעט במקרים עם וריאציה קיצונית של צפיפות האוכלוסין. זה היה גם מעריך טוב עבור המיקום של השיא, שבו בממוצע זה היה בין 5-7% בגודל של אזור המחקר עבור אוכלוסיות אחידות ו 15% עבור אוכלוסיות עם נקודת שיא בתוכן. על פי רוב, צפיפות הקרנל הכפול המעיטה בהערכת גודל השיא, במיוחד כאשר משתמשים באימות צולב כדי לבחור את רוחב הפס. התוצאות שלנו לא הראו הבדל גדול בין תוכניות סילברמן ובין תוכנית באימות צולב לאימות רוחב פס, למעט בקביעת מיקום השיא. שתי התכניות בחרו רוחב פס בין 5% ל -20% מגודל אזור המחקר. מאחר שהדיוק משתפר עם מספר התצפיות, זה

נבון להגדיל את מסגרת הזמן של המחקר כאשר מספר התצפיות בשנה מסוימת קטן מ 50. בניסויים שלנו זה היה שיעור שכיחות של בין 0.1% ל 0.5% . במקרים בהם צפיפות האוכלוסין משתנה במידה רבה על פני שטח מחקר, צפיפות הקרנל הכפול פחות מדויקת מאשר באוכלוסיות שונות.

מגבלות: מחקר זה משתמש בנתונים מדומים שנדגמו מפונקציית "שיעור אמת" ידועה כדי למדוד את הדיוק של אומדן צפיפות הקרנל הכפול. עם זאת, פונקציות האוכלוסייה ושיעורי השכיחות שלנו, המאפשרות לנו לחשב את האוכלוסייה ואת האירועים בנקודות במרחב, אינם מייצגים לחלוטין כל אוכלוסייה או תקריות בפועל. משמעות הדבר היא שהתוצאות שלנו מתארות את הרגישות של צפיפות הקרנל הכפול לגורמים שמעניינים אותנו, אך אינן מספקות ערבויות לגבי הדיוק של צפיפות הקרנל הכפול בכל מחקר ספציפי שבו נעשה שימוש.

מגבלה נוספת של מחקר זה היא השווינו רק שתי טכניקות מבחר רוחב פס. ישנן מספר טכניקות אשר לא שקלנו, כולל בוררים רוחב פס אדפטיבית אשר עשוי לתת תוצאות טובות יותר במיוחד תחת חלוקות האוכלוסייה משתנה מאוד.

כמו כן, בשל בעיות חישוביות, אנו מניחים כי יש לנו מספיק נתונים על מנת להעריך באופן מדויק את צפיפות האוכלוסין ולכן השתמשנו בפונקציה קבועה עבור זה במכנה של צפיפות הליבה הכפולה. משמעות הדבר היא כי אנו עובדים תחת תרחיש אידיאלי שבו אין אי ודאות בהערכת צפיפות האוכלוסייה.