הערכת יבול בשומשום

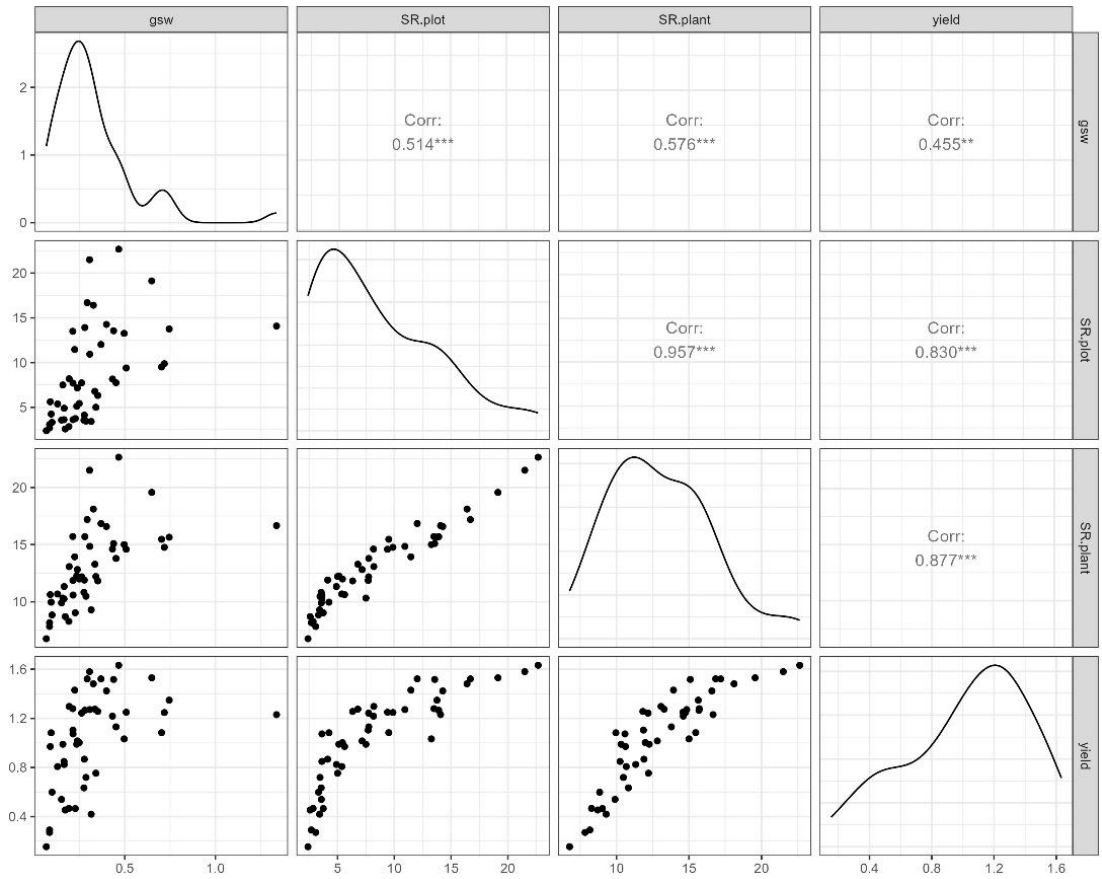
מיכל מנס

מבוא

שומשום הוא גידול וותיק מאוד שמגודל כבר אלפי שנים בחבל הלבנט. לזרעי השומשום חשיבות רבה בתזונה של בני האדם, במיוחד עבור אלו שלא צורכים מזון מהחי – בשל תכולת הברזל הגבוהה וויטמינים מקבוצת B. כגידול קיץ עם שורש מעובה ומעמיק, חשיבותו גדולה גם כחלק ממחזור הגידולים פה בארץ.
אך גידול השומשום הוא יתום מבחינה מחקרית, הוא מגודל כיום בעיקר במדינות עולם שלישי כשתהליך הגידול הוא ידני לחלוטין וללא מיכון כלל – דבר שמביא ליבול נמוך בהרבה מהפוטנציאל. בשנים האחרונות נעשים מאמצים מחקריים גדולים – גנטיים ואגרוטכניים – על מנת למכן את תהליך הזריעה והקציר, כדי שישוב לגדול כאן בישראל, כמו גם בשאר המדינות המפותחות.
כחלק ממאמצים אלו, במחקר שלי נלמדות כמויות ההשקיה המיטביות לשומשום, והדרכים המיטביות למדוד ולאפיין ביעילות את מצב הצמח תוך מיקוד ביבול הסופי שיניב השדה.
לשם כך ערכנו ניסוי שדה בקיץ שעבר ובו גידלנו שני זני שומשום ב6 רמות השקיה (4 חזרות*2 זנים*6 טיפולים = 48 חלקות), ומדדנו לאורך עונת הגידול מגוון מדדים. חלק מהמדדים נלקחו באופן ידני (לדוגמא גובה צמח, לחץ מים בעלה, מוליכות פיוניות, יבול..), חלק מהמדדים נלקחו באמצעים של חישה מרחוק, בצילום ע"י רחפן עם מצלמה תרמית והיפר-ספקטרלית.
מטרת הפרוייקט הנוכחי היא להעריך את תכונת היבול באמצעות תכונות מדודות (ידני\ חישה מרחוק) אחרות, תוך כדי מציאת התכונות המשמעותיות שתורמות למטרה זו. אם נשיג את המטרה – נוכל למדוד מספר מדדים מצומצם ולחסוך משמעותית בכח אדם ומשאבים, מבלי להתפשר על דיוק בתחזית היבול.
למטרה זו – כל התכונות איתן נעבוד הן כמותיות רציפות, והנתונים שננתח הם מיום שיא הגידול – 69 יום לאחר זריעה. נשתמש בריגרסיה רבת משתנים, שכן אנו רוצים להסביר את תכונת היבול באמצעות מספר תכונות מסוימות, את הרכב התכונות המיטבי למודל נבחר בשיטת AIC.

שיטות ותוצאות

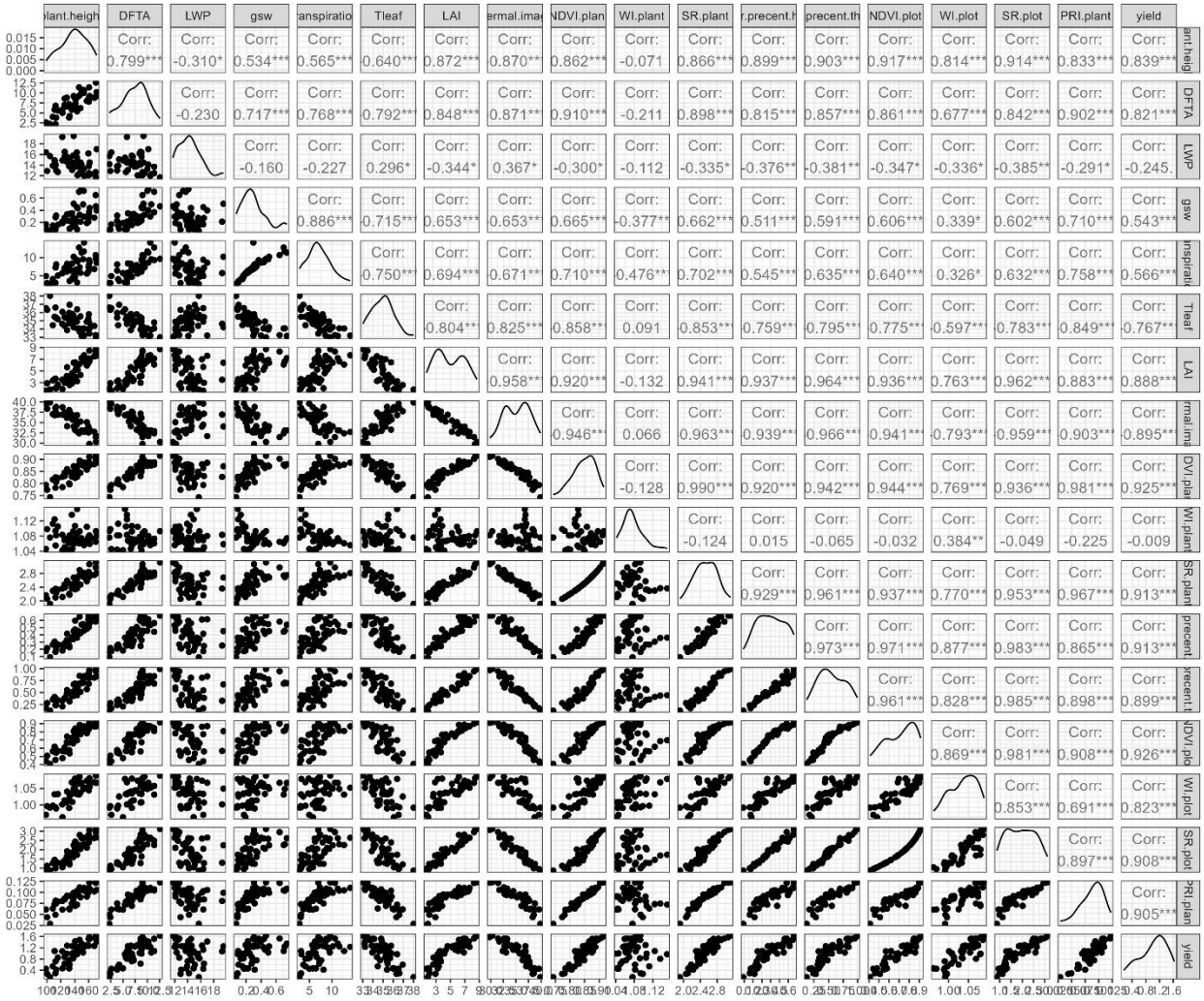
ריגרסיה רבת משתנים היא שיטה באמצעותה מעריכים תכונה מסויימת (במקרה שלנו- יבול) באמצעות תכונות מדודות אחרות. ברצונינו למצוא תמהיל תכונות שמצליח להעריך יבול בצורה טובה, ולא רק להסביר את השונות הקיימת, ולכן שיטה זו מתאימה להשגת המטרה.
בבסיס השיטה עומדות שלוש הנחות מודל – קשר לינארי בין המשתנים, התפלגות נורמאלית, קורלציה לא גבוהה. נתחיל אם כן ראשית בלבסס שהנחות מודל אלה מתקיימות.
נתרשם מהנתונים באופן ויזואלי בעזרת החבילה PerformanceAnalytics שמאפשרת לנו לראות במשולש העליון את מקדם פירסון והמובהקות שלו, במשולש התחתון את פיזור הנתונים, ובאלכסון רואים את צורת ההתפלגות של הנתונים. נתבונן על הנתונים בצורתם הטבעית ללא נירמול.
נתמקד במדדים שנראים בעייתיים מבחינת הנחות המודל (מוליכות פיוניות, SR.plot, SR.plant – שני האחרונים הם אינדקסים שחושבו מתוך הנתונים הספקטרליים, לפי הספרות – קורלטיביים לתכולת כלורופיל ואחוז כיסוי).



גרף 1: התפלגות תכונות שמראות חריגות בקשר להנחת המודל ללינאריות

נשים לב לנקודה חריגה גבוהה במוליכות פיוניות ש"מפריעה" ללינאריות, נשמיט אותה (הקוד מחפש את האינדקס עבורו המוליכות פיוניות מקסימלית, ומכניס במקום הערך הנ"ל את הערך הממוצע למוליכות פיוניות. מטרת החילוף בממוצע במקום בNA היא למנוע איבוד נתונים של חלקה שלמה).
בנוסף ניתן לראות שהנתונים הספקטרליים SR.plot ו - SR.plant דרושים לטרנספורמציה לוגריתמית כדי שיראו קשר לינארי, ולכן נבצע גם את זה.

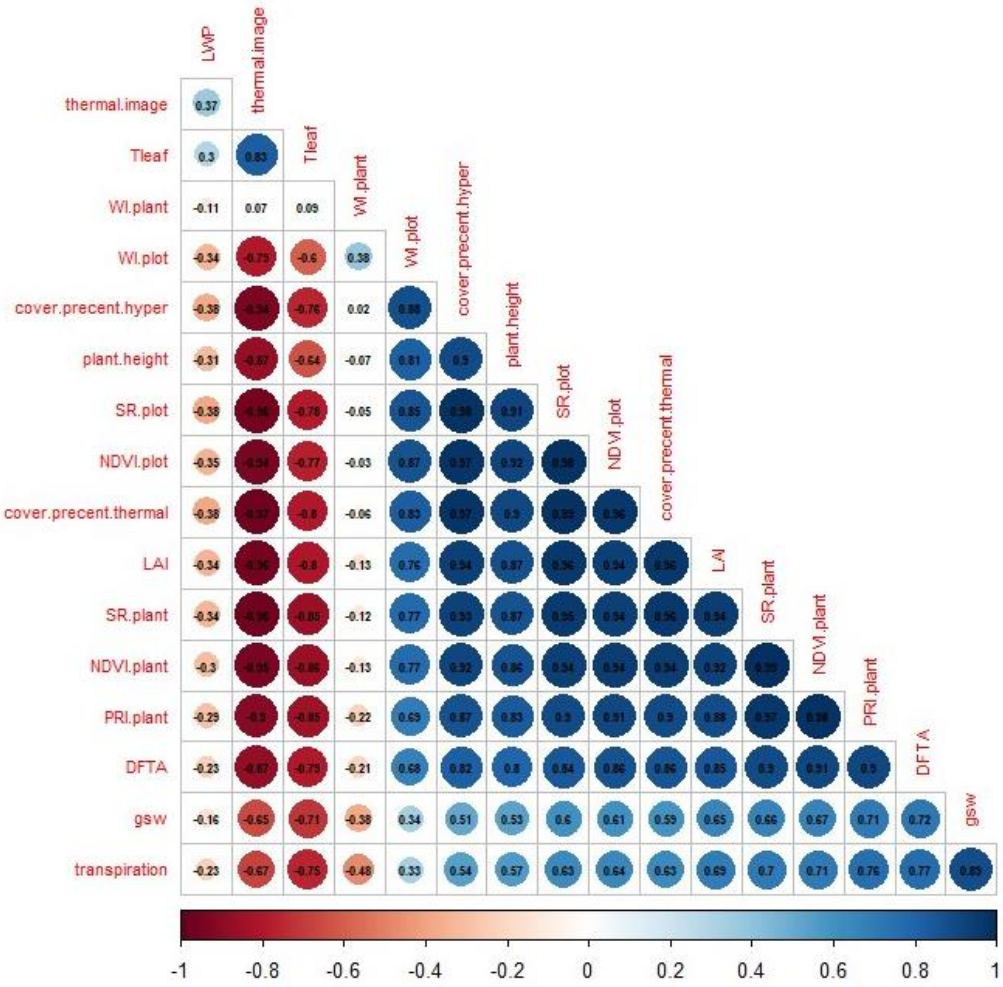
וכעת ניתן להתרשם שוב מהנתונים המטופלים, ולראות שההתפלגויות נורמאליות, והקשרים הם לינאריים.



גרף 2: התפלגות כל התכונות – גרף פיזור, צורת התפלגות, קורלציה ומובהקותה.

וכעת נתמקד בחוזק הקשר הלינארי, שאמנם ניתן לראות בגרף שלעיל, אך נוכל להמחיש בצורה וויזואלית יותר. נשתמש בספריית corrplot, וכך כשנכניס את כל הנתונים נוכל לראות את חוזק הקשר הלינארי צבוע בצבע מתאים (לפי מפת הצבעים המופיעה בתחתית הגרף) וגם מבוטא בגודל העיגול. בצורה זו, עיגולים גדולים יותר וצבועים בצבע חזק יותר מייצגים קורלציה גבוהה יותר – שניתן גם לראות את ערכה המספרי בכיתוב שחור.

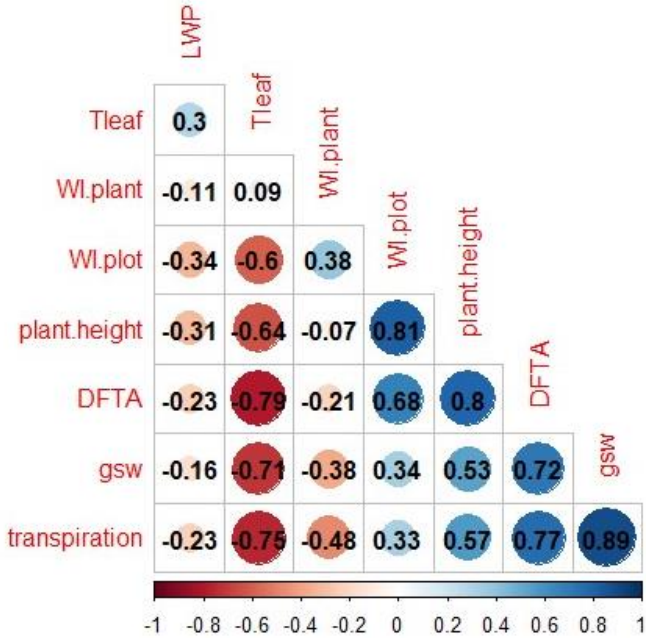
אם נתבונן על הקורלציות, נוכל להבחין במספר תכונות קורלטיביות במיוחד (גרף 3).
כאמור, למודל הרגרסיה הלינארית עלינו להכניס תכונות בעלות קורלציה נמוכה, ולכן נבחר רק את מחצית התכונות הפחות קורלטיביות. לשם כך – נחשב את מקדם פירסון של כל תכונה עם שאר התכונות, ונסכום את הערך המוחלט של המקדמים עבור כל תכונה כזו. כך, תכונה שיש לה קורלציה חזקה עם הרבה תכונות – תקבל ערך גבוה יותר מתכונה שיש לה קשר חזק עם מעט תכונות, או כזו שיש לה קשר חלש עם הרבה תכונות. נבחר 8 מהתכונות שקיבלו את "הציון" הכי נמוך (קרי, פחות קורלטיביות). בפועל – הקוד עושה את החישוב הנל ומסדר לפי הסדר, ואני בוחרת ב8 התכונות הראשונות.

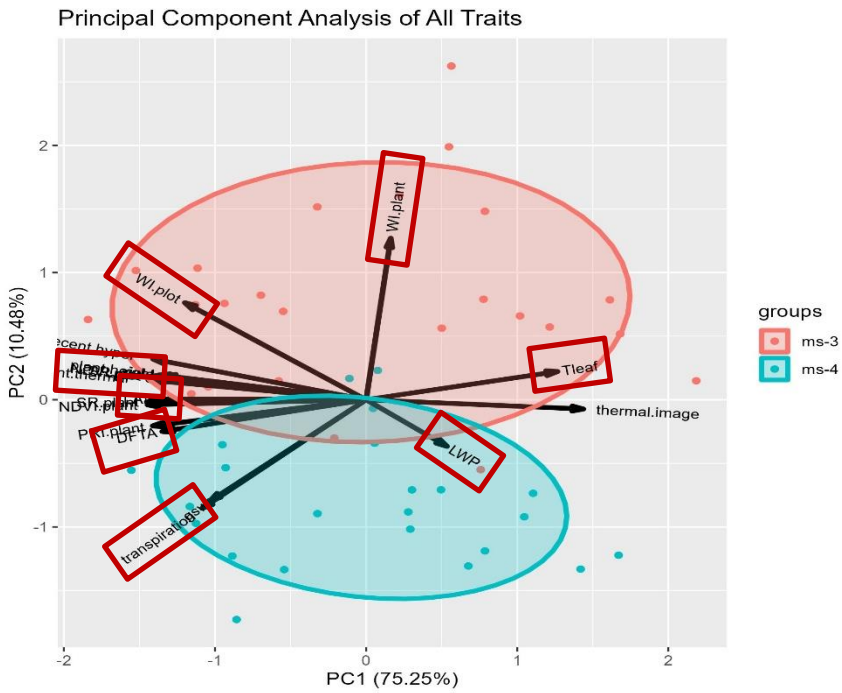


גרף 3: מפת קורלציות של כל התכונות

לפי שיטה זו, התכונות שנבחרו הן אלו:
מבין הנתונים הספקטרליים נבחרו WI.plot ו- WI.plant (מדובר בwater index, לפי הספרות קורלטיבי לתכולת מים בצמח). יתר התכונות שנבחרו נמדדו ידנית (תוצאות מדידות תא לחץ, מוליכות פיוניות, טמפרטורה עלוותית, טרנספירציה מדודה, גובה צמח, מרחק קודקוד הצמח מפרח פתוח ראשון).

נרצה לוודא שהתכונות שבחרנו מייצגות את השונות שקיימת לנו בשדה בצורה טובה. לשם כך ניתן בעזרת ספריית ggbiplot לעשות אנליזת PCA עם כל התכונות ההתחלתיות, כאשר בהדגשת הוקטורים העצמיים של התכונות שבחרנו לפי קורלטיביות נמוכה אנו רואים שמירב השונות מיוצגת על ידם. הצבעים השונים של הנקודות מייצגים שני זנים שנבחנו בניסוי, האליפסות מקיפות נקודות שנמצאות בטווח של 2 סטיות תקן מנתוני הזנים.
ליד גרף זה מוצגת גם מפת קורלציות ממוקדת של המדדים שנבחרו, ניתן לראות שרוב הקורלציות נמוכות, ודבר זה יכול להניח את דעתינו בקשר להנחת המודל השלישית.





גרף 4: אנליזת PCA עם וקטורים נבחרים מודגשים

גרף 5: מפת קורלציות של תכונות נבחרות

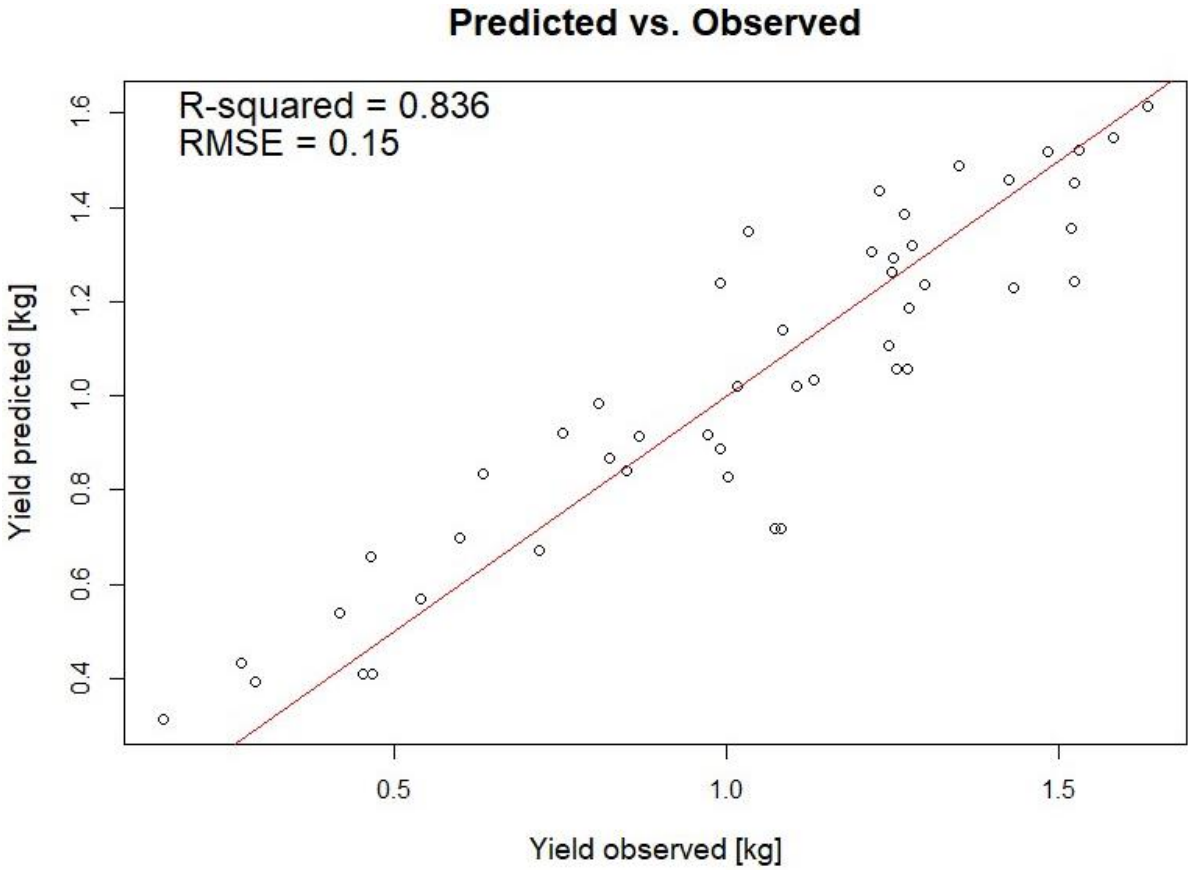
כעת, משהנחות המודל מתקיימות, ניתן להכניס את 8 התכונות למודל ריגרסיה רבת משתנים, ולבצע בחירת משתנים באופן הדרגתי לפי אופטימיזציה של AIC. המודל מבצע צעדים הלוך וחזור כאשר בכל צעד הוא מוסיף או מוריד משתנה אחר, ובודק כיצד הפעולה השפיעה על מדד הAIC. בסוף התהליך מתקבלות מספר תכונות שהשילוב שלהן מעריך ברמת הדיוק הגבוהה ביותר את היבול.

```
Call:
lm(formula = target_variable ~ WI.plant + WI.plot + Tleaf, data = least_correlated_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.31707 -0.10566 -0.01087  0.08531  0.36379 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.39292    1.62349   0.242 0.809889    
WI.plant    -4.48838    1.16961  -3.838 0.000394 ***
WI.plot      8.76324    1.02469   8.552 6.59e-11 ***
Tleaf       -0.10355    0.02885  -3.589 0.000829 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1562 on 44 degrees of freedom
Multiple R-squared:  0.8466,	Adjusted R-squared:  0.8361 
F-statistic: 80.93 on 3 and 44 DF,  p-value: < 2.2e-16
```

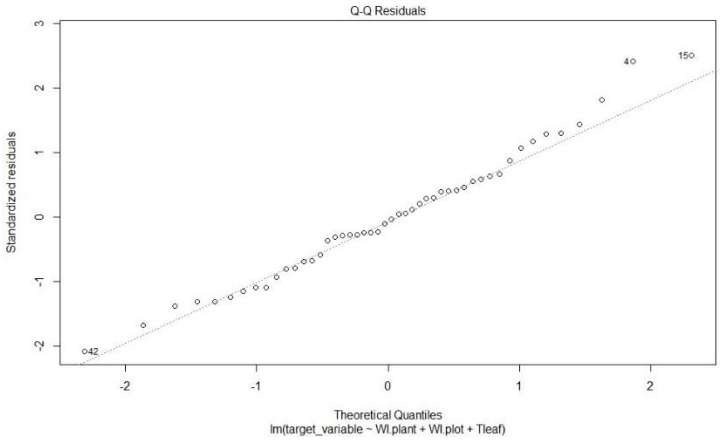
גרף 6: תוצאות מודל רגרסיה רבת משתנים.

המודל מצליח לאמוד את היבול בדיוק של 83.6%, ברמת מובהקות גבוהה, באמצעות שלוש תכונות – טמפרטורת עלה מדודה, ושני אינדקסים ספקטרליים.
ניתן להמחיש זאת באמצעות הגרף של יבול חזוי אל מול יבול מדוד, בו גם ניתן לראות שהטעות של המודל קטנה

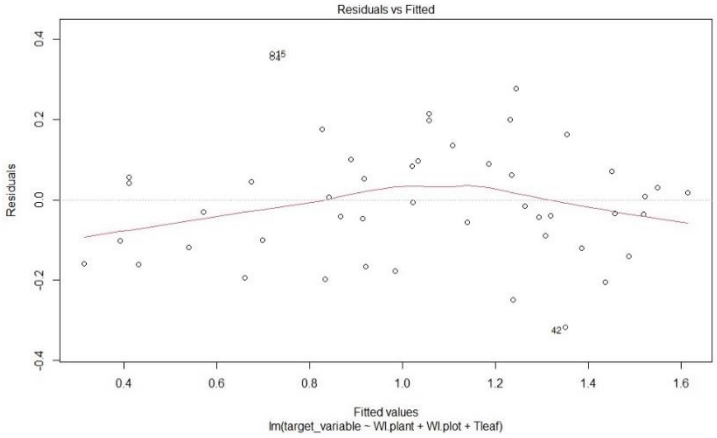
יחסית לטווח הערכים. בנוסף ניתן לראות בגרפים 8-9 שהנתונים מתפלגים נורמאלית, והשאריות מפוזרות אקראית – מדד נוסף לעמידה של המודל בהנחות הבסיס.



גרף 7: ערכי יבול בק"ג חזויים מול מדודים לפי מודל הרגרסיה.



גרף 8: עקומת התפלגות נורמאלית של נתוני המודל



גרף 9: התפלגות השאריות של נתוני המודל

דיון ומסקנות

ריגרסיה לינארית מרובת משתנים מאפשרת לנו לא רק לתאר תופעה קיימת אלא גם להסביר ולמדל אותה.
בבואנו לבצע הערכת יבול בשומשום, גידול יתום שלא ניתן להישען בו על ידע קודם בספרות, מוטב לבדוק כיוונים רבים ככל האפשר. נמדדו 8 תכונות מורפולוגיות ופיזיולוגיות שבגידולים אחרים מקובל להעריך באמצעותן את היבול. איסוף המדדים האלה הוא סיזיפי ולא פרקטי בסדרי גודל של שדות מסחריים.
הכוח הטמון בטכנולוגיה של חישה מרחוק - שמאפשרת לנו הצצה אל תכונות שניתן לגלות עוד לפני שרואים אותן בעין - מרחיב לנו את יריעת האפשרויות פי כמה וכמה ואף ניתן ליישום בשדות מסחריים.

אך החכמה היא גם לדעת לבחור בצורה נבונה את התכונות.
מבחינה זו, החישה מרחוק – אליה וקוץ בה, שכן האינדקסים הספקטרליים שחישבנו למטרות עבודה זו היו קורלטיביים מאוד אחד לשני. מתוך נקודת פתיחה של מחצית תכונות מדודות ידנית ומחצית בחישה מרחוק, כעת רק רבע מהתכונות שנכניס למודל הריגרסיה הן של חישה מרחוק. צוואר הבקבוק המשמעותי הזה הוא מגבלה במודל הסטטיסטי שבחרנו.
(בעצם ישנן שיטות מתקדמות רבות לחילוץ נתונים ספקטרליים שיכולים לעזור להתגבר על האתגר שציינו.)

בסופו של דבר, גם לאחר צוואר הבקבוק, באנליזת הPCA ניתן לראות שמירב השונות מוסברת על ידי כלל התכונות שמדדנו, ואף ניתן לראות שהרכיב השני מוסבר בעיקר ע"י הזן. וכשמתמקדים בתכונות שנבחרו לכניסה למודל אנו רואים שהן מתפרשות לכל הכיוונים, קרי מייצגות לנו את השונות בצורה מספקת.
העובדה הזו מאשרת לנו ששיטת המיון השאירה מחד תכונות שמתיישבות עם הנחת המודל השלישית, ומאידך לא וויתרה על תכונות משמעותיות מדי.

את טיב הערכת מודל ניתן לבחון לפי ערך adjusted R squered הגבוה, וכן על פי המובהקות הגבוהה.
מעניין לראות שהמודל הכניס את שתי התכונות החישתיות לתוך המודל, וזה יכול להצביע לנו על חשיבות החישה מרחוק בהגברת הדיוק של הערכת היבול. ניתן לראות זאת כהצלחה במובנים של הגברת היעילות בכוח אדם.
ממעוף הרחפן, ובעזרת תכונה קרקעית מדודה של טמפרטורת עלווה, אפשר להעריך את היבול שיתקבל בשדה.

אוכלוסית העולם ההולכת וגדלה מציבה בפני קהילת המדענים משימה חשובה מאין כמותה – תזונה בריאה ומספקת לכולם. חקלאים נדרשים לגדל יותר מזון בפחות שטח, ועל כן עלינו לשאוף לייעול מקסימלי הן בכח עבודה והן בשטחי הגידול. בניסויים שאנחנו עורכים אנו משקיעים זמן ומשאבים רבים על מנת לבוא בפתח תגליות מדעיות שיסייעו לנו בהשגת משימות אלו – להשיג יותר, בפחות השקעה.
העובדה שהצלחנו להעריך את היבול בחלקות השומשום בניסוי בדיוק גבוה של 83%, וזאת באמצעות 3 תכונות מדודות בלבד – רק אחת מהן נלקחת באופן ידני, כשלעצמה מוכיחה שהמשימה ניתנת לביצוע.