

## פיתוח מתודולוגיה למיפוי משתני צימוח בשדות כותנה בקנה מידה אזורי/ארצי

### תוך שימוש בטכנולוגיות מידע וחישה מרחוק

Development of methodology for mapping plant height and growth rate in cotton fields based on information technology and remote sensing

יפית כהן: המכון להנדסה חקלאית, וולקני. ריכוז וניהול המחקר, איפיון האפליקציה ובסיס הנתונים, ניתוח צילומים ופיתוח מודל. [yafitush@volcani.agri.gov.il](mailto:yafitush@volcani.agri.gov.il); איתן גולדשטיין, ניתוח נתונים מרחבי; אריה בוסק, מגדלי דרום יהודה, עיבוי בסיס הנתונים וניתוח.

### תקציר

בשנה שעברה, במימון של מועצת הכותנה, פותח יישומן (אפליקציה) לאיסוף נתוני גובה מחקלאים ונבנה שרת בו נאגרו הנתונים. בגלל בעיות של תוכנה וחומרה, פותח יישומן מקביל באמצעות המערכת של חברת אגריסטק. באמצעות הנתונים שנאספו הן בשנה שעברה והן השנה התקבלו למעלה מ-1500 נתונים של גובה שנמדדו באמצעות חקלאים בעשרות שדות כותנה בארץ. בגלל אילוצים שונים, רק 1009 נתונים שולבו ליצירת מודלים אמפיריים להערכת גובה באמצעות מדדים ספקטראליים. השתמשנו בשתי שיטות עיקריות: רגרסיה ליניארית ולימוד מכונה. באמצעות רגרסיה ליניארית שהתבססה על מדד SAVI הושג מודל עם מתאם גבוה וטעות של 11.5 ס"מ לסט האימות. באמצעות RF הושג מודל עם מתאם גבוה יותר וטעות של 7.9 ס"מ שזהו שיפור של כ-30%.

למודל שני יתרונות משמעותיים: תדירות גבוהה יותר בזמן ושונות מרחבית. כלומר, מצד אחד הוא מספק נתונים בתדירות של חמישה ימים ונמצא כי יחסית למדידות הידניות הוא בתדירות גבוהה יותר ומן העבר השני הוא מספק שונות של הגבהים בכל השדה בהשוואה למדידת צמחים בודדים בנקודה אחת בשדה שהיא בד"כ קרוב לשוליים.

הקוד שפותח במסגרת המחקר כולל בתוכו את כלל השלבים לקראת הגשתו לציבור החקלאים: שלב ארגון הנתונים, שלב חישוב של מדדים ספקטראליים, איחוד בסיסי הנתונים של המדידות בשטח ונתוני החישה מרחוק, שלב ניקוי הנתונים, שלב שחזור עננות (בעייתי בתצורתו הנוכחית כי הוא משחזר אחורה ולכן הנתון המשוחזר מתקבל באיחור של 5 ימים), ושל בחישוב הגובה באמצעות המודל. מה שחסר כרגע בקוד הוא היכולת להריץ את הקוד על הדימוותים ולהראות את השונות המרחבית או לחילופין לספק מפות מצב מים.

לצערנו, לא הצלחנו להשלים את ההערכה של המודל שלנו בחישוב קצב הצימוח היומי. אנחנו בתחילת התהליך אך עד לסיום המאסטר של הסטודנט לתואר שני שלנו (קרי סוף שנת 2021), אנו מתעתדים לסיים גם את התהליך הזה.

### מבוא

חישוב מנות המים להשקיה בגידולים השונים כולל כותנה נעשה ע"י הכפלת מקדם ההשקיה המתאים בהתאדות המחושבת ע"פ נוסחת פנמן-מונטית' ומקדם התיקון לפי המשוואה:  $Irrig = \alpha * Kc * ET0$  כאשר  $irrig$  זו כמות ההשקיה,  $ET0$  היא ההתאדות המחושבת (פנמן-מונטית');  $Kc$  הוא מקדם הגידול

;  $\alpha$  היא מקדם התיקון של מקדם הגידול.

כיום, החקלאים ברוב המקרים עובדים עם  $Kc$  קבועים הזמינים בשירות של שה"ם או לפי הפרסומים של מועצת הכותנה. את ההתאדות,  $ET_0$ , הם שואבים מתחנה מטאורולוגית קרובה. חישוב מנות ההשקיה לפי התאדות ומקדמי השקיה, מניחים קצב גידול רצוי ואינם מתייחסים למצבי עקת או עודף מים. כדי לבצע בקרה על ההשקיה, החקלאים מבצעים מדידות בשדה (1-2 שבוע) של מדדי צומח משתנים לפי שלב הגידול ומשווים אותם לעקומי צימוח רצויים.

**מקדם תיקון בשלב הצימוח הווגטיבי:** קצב צימוח הגבעול הראשי נמצא כמדד צמחי אמין לאפיון מצב המים של צמחי כותנה בשלב הצימוח הווגטיבי, כל עוד קצב הצמיחה גבוה מ-0.5 ס"מ ליום. מועצת הכותנה פרסמה עקום רצוי של גובה הצמח ושל קצב הצימוח לפי זן לפי תאריכים בשנה וימים מפרח בשדה. החקלאי בוחר ומסמן באמצעות מוטות במבוק מספר צמחים בקטע בשדה שאמור לייצג את החלקה כולה. פעם-פעמיים בשבוע הוא מודד את גובהם של הצמחים ומחשב את קצב הצימוח היומי. אם קצב הצימוח גבוה מהרצוי, יש להוריד ממנת המים, ולהיפך. בשלב מילוי ההלקטים, צימוח הגבעול כמעט ונפסק ולכן מדידת הקצב אינה מתאימה עוד להכוונת ההשקיה.

**מקדמי תיקון לשלב הפרודוקטיבי:** בשלב זה משתמשים במדד הנקרא: "מספר מפרקים מעל פרח צהוב" (ממפ"צ). מדד זה מייצג את היחס בין הצימוח הווגטיבי (תוספת מפרקים) לתהליך הפרודוקטיבי (התקדמות הפריחה במעלה הצמח). מספר מפרקים קטן יחסית למקובל בכל שלב מעיד על צמח מעוכב וגטטיבית. בנוסף, נמצא כי פוטנציאל המים בעלה (פמ"ע) הנמדד באמצעות תא הלחץ הוא אמצעי יעיל ואמין לבקרת ההשקיה. בדומה לגובה, גם לשני המדדים הללו, הופקו על בסיס ניסיונות רב-שנתיים עקומים רצויים. החקלאי מודד בשדה ומשווה את המדידות שלו לעקומים ומתקן את ההשקיה בהתאם.

### תיאור הבעיה

בשנים האחרונות, נצפה מעבר של החקלאים משימוש בלוחות מים קבועים לערכי התאדות מחושבת בזמן אמת בזכות עבודה משותפת של שה"ם והשירות המטאורולוגי של משרד החקלאות. עם זאת, מקדמי ההשקיה הם עדיין קבועים, ללא התייחסות לשלב הגידול בפועל. לאחרונה, עם התבססות של המחקר בתחום, התפתחו שיטות שימושיות ואף ברמה מסחרית להערכת מקדמי ההשקיה תוך התבססות על דימותי לוויין בתחום הנראה והא"א הקרוב. לעומת ההתקדמות בחישוב ההתאדות ומקדמי ההשקיה, מקדמי התיקון נעשים כאמור, על-ידי מדידות בפועל של גובה, ממפ"צ ופמ"ע. למרות הדיוק של מדידות ידניות, ההסתמכות על מספר מועט של צמחים או עלים מהווה בעיה כאשר רוצים לייצג שדות גדולים בהם יש שונות מרחבית משמעותית. טרם נעשה מחקר בארץ להערכת גובה של הכותנה וקצב צימוח באמצעים של חישה מרחוק.

בשנה החולפת, במימון של מועצת הכותנה, פותח יישומון (אפליקציה) לאיסוף נתוני גובה מחקלאים ונבנה שרת בו נאגרו הנתונים. היישומון מופיע הן את העקום של מדידות בפועל של החקלאי והן את העקומים הרצויים לפי הזן שנבחר. בצורה כזו יכול החקלאי לבקר את השקיה על ידי השוואה בין שני העקומים תוך כדי המדידה בשדה. באמצעות המערכת נאספו למעלה מ-400 נתוני גובה, מכ-60 שדות כותנה ברחבי הארץ מלמעלה מעשרה מגדלים/מדריכים. כמו כן, התקבלו נתוני גובה משדות נוספים של

מגדלים שלא עשו שימוש ביישומון. בוצע הליך של בקרת איכות הנתונים ונתונים רבים נופו בעיקר בגין חוסר ודאות לגבי המיקום של השדות מהם נאספו הנתונים. על מנת לבחון את הקשר בין גובה צמחי הכותנה ובין מדדי צימוח מדימותי הלווין החינמי Sentinel-2, השתמשנו בהפקת ממוצע מדד הצימוח לשדות הכותנה לפי תאריך קרוב באמצעות קוד בסביבת google earth engine (GEE). גם בשלב זה חלק מהנתונים נופו בעיקר בגלל עננות. שני בסיסי הנתונים חוברו והשוו והתקבל קשר ליניארי בין גובה למדד הצימוח NDVI עד לגובה של 120 ס"מ עם  $R^2=0.72$ . על בסיס המודל הזה הופקו מפות של הגובה הן סטטיות באמצעות ArcGIS והן דינאמיות באמצעות GEE באדיבותו של הראל גרינבלט המראות את השונות בתוך החלקה ( <https://harelg25.users.earthengine.app/view/cotton-height> ). הערה: כדי לראות את השונות של הגובה יש לבחור תאריכי התחלה וסוף מתאימים ולהתמקד במיקום של שדות כותנה. בצורה כזו נסגר המעגל: מחד היישומון מספק צורך מיידי של בקרת השקיה ומצד שני להציג את השונות בשדה, מה שלא ניתן להשיג במדידות של צמחים בודדים. ההישג הזה מאפשר להמשיך את הפיתוח בשני המישורים: במישור המייד, יש להוסיף ליישומון שלושה משתנים נופסים: קצב צימוח, ממפ"צ ופמ"ע. במישור העתידי: להגדיל את בסיס הנתונים על מנת לפתח מודל מדויק, אמין ויציב יותר לגובה וכן לאפשר לבחון חיתוכים לפי זנים, לפי תאריכי זריעה, ולפי אזורים.

### **מטרת המחקר**

פיתוח מתודולוגיה למיפוי גובה הצמח וקצב צימוח בשדות כותנה בקנה מידה אזורי/ארצי תוך שימוש בטכנולוגיות מידע וחישה מרחוק.

1) הרחבת היישומון לאיסוף נתונים של פרמטרים צמחיים נוספים במהלך גידול כותנה; 2) פיתוח מודל להערכת ומיפוי גובה כללי ופיתוח מודלים לפי זנים, אזורים ומועדי זריעה; 3) פיתוח יישומון שיאפשר צפייה בזמן אמת בשונות של הגובה בחלקה וכן את הערך הממוצע של החלקה לעומת המדידות הנקודתיות.

### **שיטות**

הרחבת היישומון לאיסוף נתונים של פרמטרים צמחיים במהלך גידול כותנה: בשנת המחקר הראשונה פותח יישומון עבור מכשירי אנדרואיד והועלה ל Google Play (כהן וחוב', 2020). היישומון אפשר הן לחקלאי והן למדריך לבחון האם צריך לתקן את מקדם ההשקיה ( $\alpha$ ) בהתאם למדידות גובה הצמח. בהתאם לכך, מסכי המשתמש העיקריים ביישומון היו מסך תצוגת נתוני חלקה, מסך הוספת מדידה ומסך השוואה בין גרף צימוח מומלץ ובפועל. המשתמשים לא נתבקשו להכניס את גבולות השדה במטרה לאפשר שימוש ביישומון בצורה נוחה ופשוטה מייד עם התקנת היישומון. הנתונים שהוקלדו נשמרו בבסיס נתונים מקומית במכשיר הנייד הסלולארי של המשתמש ללא אפשרות לניהול הנתונים. במקביל לאיסוף נתוני החלקה ומדידות הגובה על ידי המשתמש במכשיר הנייד, נשלחו הנתונים לבסיס נתונים מקוון שנבנה בתחנת עבודה (שרת) שהוצבה במכון להנדסה חקלאית, כדי לפתח מודל התאמה בין נתוני הגובה ובין נתוני חישה מרחוק. באמצעות היישומון נאספו בשנת 2020 נתונים מ-13 משקים (ללא שימוש הנתונים שנאספו באופן לא סדיר), 42 חלקות ו-299 נתוני גובה. לצד האפיון המוקפד של

היישומון והשרת, השימוש במערכת עורר מספר מגבלות הקשורים לשני היבטים, תוכנה וחומרה. בהיבט של היישומון, נתגלו שלוש בעיות: 1. חלק מנקודות הציון (קואורדינטות) של נתוני הגובה היו לא מדויקות, קרי, ההנחה שנקבל נקודות שהן עד 10 מ' מהמיקום שבו היה המשתמש לא היתה נכונה בכל המקרים; 2. בחלק מן המקרים, המשתמשים סימנו שהם בשדה כאשר הם לא היו בשדה; 3. הנתונים נשמרו באופן מקומי במכשיר הנייד ללא סנכרון בין משתמשים ומכשירים אחרים, קרי החקלאי לא יכול היה לצפות בנתונים שנאספו על ידי מדריך כותנה או על ידי משתמשים אחרים בגד"ש. בהיבט של החומרה, נתגלו שתי בעיות: 1) בניגוד לשרתי ענן מקוונים שזמינים 24 שעות, השרת שהוקם לא היה זמין בעת תקלות חשמל ששררו במכון להנדסה חקלאית, ולכן נתונים שהוקלדו במכשיר הנייד כאשר השרת היה מושבת לא נשמרו בבסיס הנתונים המקוון; 2) השרת לא היה מאובטח דיו ועל כן נפרץ ולא יכולנו להמשיך לעבוד איתו.

לאור המגבלות שתוארו לעיל, יצרנו קשר עם חברת אגריטסק לה יש ניסיון רב שנים בפיתוח מערכות לאיסוף נתונים, הצגתם ועיבודם. אגריטסק אינה מאפשרת למשתמשים ארעיים להשתמש במערכת אלא נעזרת בבסיס נתונים יחסי, ולכן היה צורך להגדיר מראש רשימה של משקים, חלקות ומשתמשים (חקלאים ומדריכים). בצורה זו החקלאי יכול לעקוב אחרי כל החלקות המסומנות שלו, הן ביישומון והן באתר אינטרנט על סמך נתונים שהוקלדו מכל המשתמשים שיש להם הרשאה לחלקות שלו. איור 1 מציג מספר מסכים ביישומון שפותחו על ידי אגריטסק לצורך בקרת השקיה. בשלב ראשון, במסך הבית המשתמש בוחר חלקה מתוך מפה או מתוך רשימה (בתפריט דיווח חדש), אח"כ הוא מוסיף מדידות גובה, ממפ"צ ופוטנציאל מים בעלה. היישומון מחשב אוטומטית ממוצע דגימות לחלקה ליום וקצב צימוח יומי. בסופו של תהליך החקלאי יכול להציג את טבלת הצמיחה שלו ולראות איך הוא לעומת הגרף המומלץ. בנוסף למדידות בשדה, היישומון מציג ערכי NDVI מלוויין Sentinel-2. איורים 2-5 מציגים מספר מסכים מהאתר של אגריטסק לניהול המידע, שביניהם: תצוגת נתונים בטבלה, במפה או בגרף.

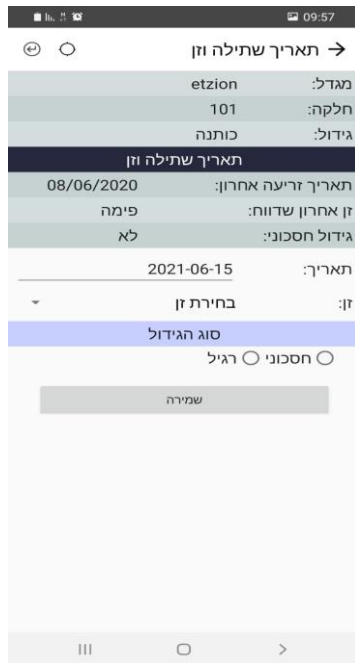
## מסך הבית



## בחירת חלקה (ממפה או מרשימה)



## עדכון פרטי החלקה



דיווח חדש (מדידות גובה, ממפ"צ, פוטנציאל מים בעלה). היישומון מחשב אוטומטית ממוצע דגימות לחלקה לתאריך דגימה וקצב צימוח.

השוואה אוטומטית של גובה הצמח המדוד וקצב צימוח מחושב לעקומים מומלצים: כחול – ערך גבוה מהרצוי, ירוק – ערך קרוב לרצוי, אדום – ערך נמוך מהרצוי

תצוגת ערכי NDVI (מינימום, מקסימום וממוצע)



קצב צימוח		גובה		תאריך
רצוי	אמיתי	רצוי	אמיתי	
-	0.32	-	145.75	27/07/20
0.58	1.5	123.2	143.5	20/07/20
0.89	1.18	115.7	133	13/07/20
0.89	0.85	115.7	130.67	13/07/20
1.22	1.54	106.9	124.75	06/07/20
1.22	1.54	106.9	124.75	06/07/20
1.54	2.54	96.8	114	29/06/20
1.54	2.77	96.8	115.67	29/06/20
1.81	1.96	85.3	96.25	22/06/20
1.81	2.46	85.3	99.75	22/06/20
1.99	3.75	72.5	82.5	15/06/20
1.99	2.25	72.5	72	15/06/20
1.99	2.11	72.5	71	15/06/20
2.04	0.61	58.4	56.25	08/06/20
-	-	42.9	52	01/06/20

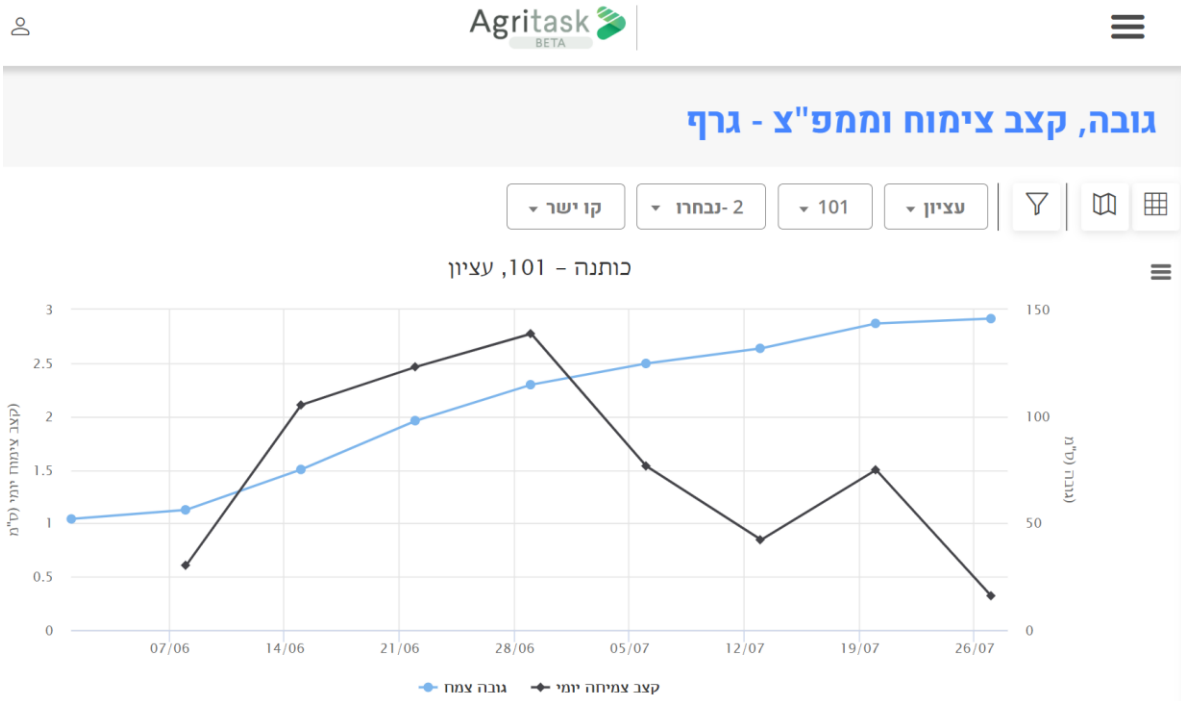
איור 1. מסכי היישומון לבקרת השקיה של אגריסטק

1-35 של 35

יצוא | יבוא | | | | | | | | | |

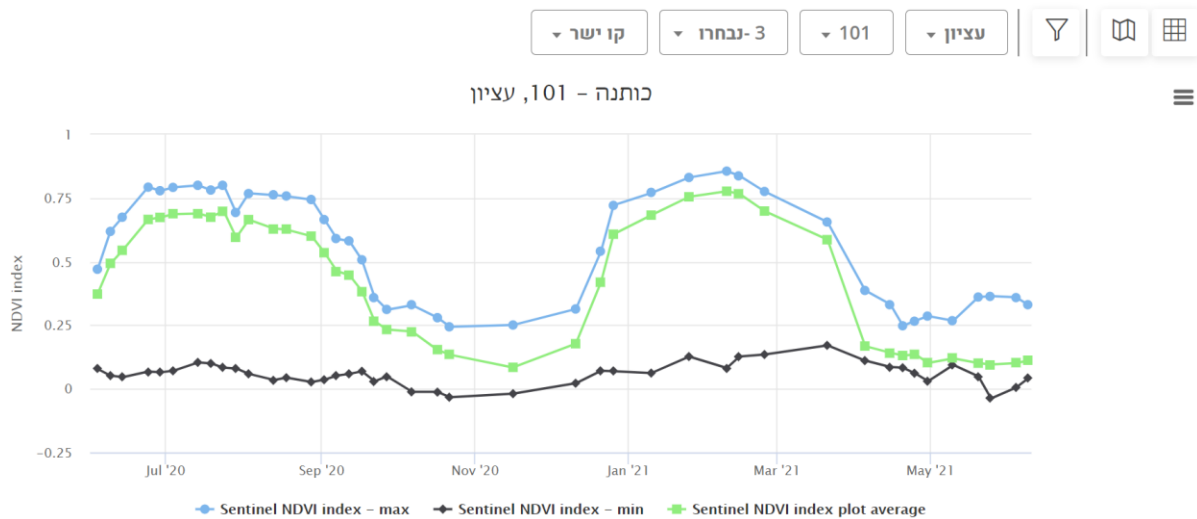
הערות	תמונות	תאור	נתונים	מדידה	גידול	אתר	מגדל	תאריך/שעה ↓	פקח
		ממפ"צ	5	ממפ"צ	כותנה - כותנה	101	עציון	07:23 27/07/2020	ניר קנטי
		קצב צימוח יומי (ס"מ)	0.32	קצב צמיחה יומי	כותנה - כותנה	101	עציון	07:23 27/07/2020	ניר קנטי
		גובה (ס"מ)	145.75	גובה צמח	כותנה - כותנה	101	עציון	07:23 27/07/2020	ניר קנטי
		ממפ"צ	6.25	ממפ"צ	כותנה - כותנה	101	עציון	07:33 20/07/2020	ניר קנטי
		קצב צימוח יומי (ס"מ)	1.5	קצב צמיחה יומי	כותנה - כותנה	101	עציון	07:33 20/07/2020	ניר קנטי
		גובה (ס"מ)	143.5	גובה צמח	כותנה - כותנה	101	עציון	07:33 20/07/2020	ניר קנטי
		ממפ"צ	7	ממפ"צ	כותנה - כותנה	101	עציון	07:31 13/07/2020	ניר קנטי
		גובה (ס"מ)	133	גובה צמח	כותנה - כותנה	101	עציון	07:31 13/07/2020	ניר קנטי
		ממפ"צ	6.5	ממפ"צ	כותנה - כותנה	101	עציון	07:26 13/07/2020	ניר קנטי
		קצב צימוח יומי (ס"מ)	0.85	קצב צמיחה יומי	כותנה - כותנה	101	עציון	07:26 13/07/2020	ניר קנטי

איור 2: דוגמא לטבלת נתוני גובה, קצב צימוח וממפ"צ לחלקה נבחרת

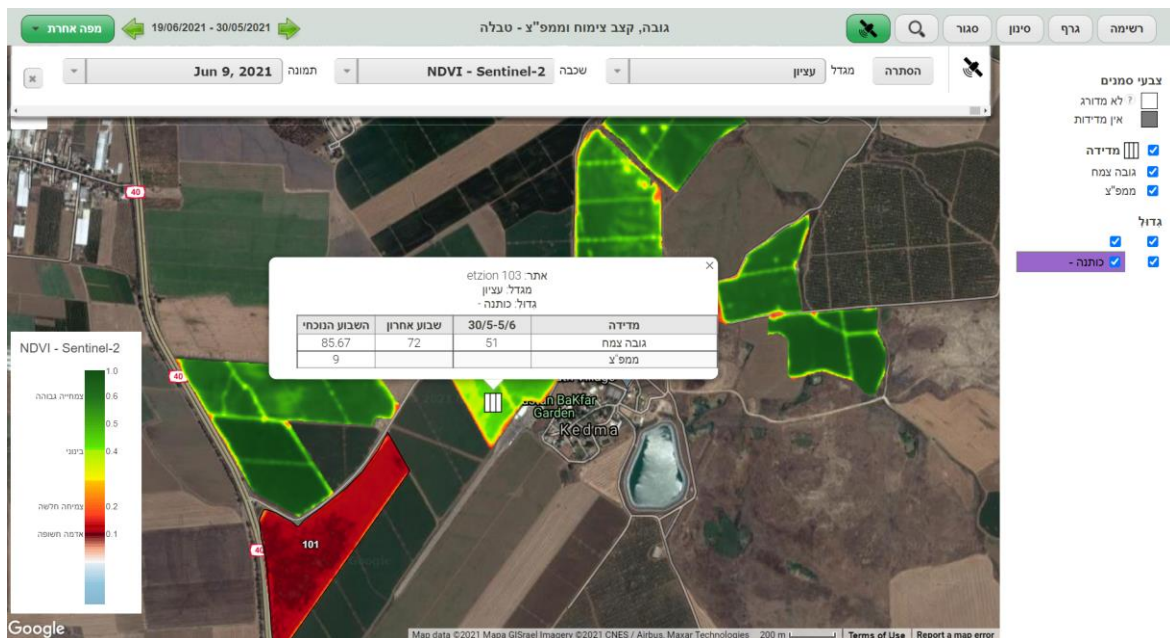


איור 3: דוגמא להצגה גרפית על ציר הזמן של גובה וקצב צימוח לחלקה נבחרת

## NDVI - גרף - סטטיסטיקות NDVI



איור 4: דוגמה להצגת NDVI על ציר הזמן לחלקה נבחרת (מיני', ממוצע ומק')



איור 5: מפת NDVI לחלקות כותנה של משק נבחר

יצירת בסיס נתוני חישה מרחוק מתוך דימותי לוויין חינמיים: בשלב זה של המחקר הוחלט להשתמש בדימותי לוויין Sentinel-2 בלבד ולא בדימותי לוויין ונוס או Planet-labs. הסיבה העיקרית לכך היא זמינות דימותי הלוויין באמצעות סביבת הניתוח Google earth engine (GEE) ללא צורך בהורדת הדימותים. באמצעות קוד הזמין [בקישור הזה](https://code.earthengine.google.com/5c468c678d49a314c08477ce7a90157f)<sup>1</sup> ניתן לקבל ערכי NDVI מבוססים על דימותי לוויין Sentinel-2 לכל תאריך בטווח נבחר לכל פוליגון בשכבה. הקלט של הקוד במקרה שלנו היה: שכבה פוליגונאלית של

<sup>1</sup> <https://code.earthengine.google.com/5c468c678d49a314c08477ce7a90157f>

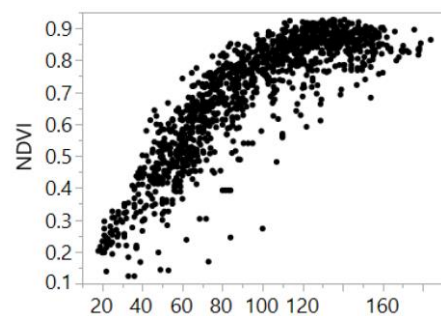
שדות הכותנה. בהקשר זה יש לציין כי כמו שניתן לשנות את השכבה הפוליגונאלית ואת התאריכים ניתן גם לשנות את סוג הלוויין ולהתאים את הערוצים כך שחישוב ה-NDVI יבוצע נכון ולא כאן המקום להאריך. בנוסף למדד NDVI חושבו בסביבת GEE 23 מדדים נוספים מארבע משפחות (טבלה 1). נספח 1 מפרט את הנוסחאות והרפרנסים של המדדים השונים.

RGB	RED EDGE		NIR		Water
CRI1	CHL_RE	IRECI	EVI	SAVI	NDWI
ARI1	ARI2	PSSRA	IPVI	ARVI	SRWI
	PSRI	CRI2	NDVI	WDVI	NDI11
	RENDVI	MCARI	DVI	PVI	
			RVI	NDI45	
				GNDVI	
2	8		11		3

טבלה 1: רשימת 24 מדדי צומח מארבע משפחות: א. שימוש בערוצי התחום הנראה בלבד; ב. שימוש בערוצי קצה-האדום; ג. שימוש בערוצי א"א קרוב ו-ד. שימוש בערוצי בליעה של מים בתחום ה-SWIR

הערכת גובה וקצב צימוח באמצעות דימותי לוויין חינוכיים: נתוני מדידות הגובה שנאספו באמצעות שני היישומונים מהשנים 2019 ו-2020 אוחדו לבסיס נתונים אחד ואליו. לצורך חישוב מודלים להערכת גובה, חיברנו את שני בסיסי הנתונים של מדידות הגובה ושל מדדי הצומח מדימותי הלוויין לפי מועדי המדידות והדימותים בסביבת אקסל ופייתון. ההתאמה בזמן לא היתה מושלמת כי דימותי הלוויין זמינים כל חמישה ימים (או בתדירות נמוכה יותר כתלות באחוז כיסוי העננים) ואילו איסוף נתוני הגובה יכול להגיע ל-1-2 פעמים בשבוע. על מנת לא לוותר על כמות גדולה של נתונים בגלל התאמת זמנים לא מושלמת הוחלט שבמידה ואין התאמה מושלמת בין יום המדידה ליום מעבר הלוויין, יילקחו דימותים בטווח של עד 5 ימים ובמידה ובטווח הזה היו שני דימותים נחשב את הערך הממוצע. בשלב הבא, הוצאו נתונים חריגים. למשל הסרת נקודות הסובלות כנראה מעננות המתבטא בנפילה מקומית בערך NDVI (נספח 2) או סדרת זמן NDVI שאינה מראה מגמה של גידול צמחי וכנראה לא מייצגים שדה כותנה (נספח 3; בעיקר בשנת 2019).

בנוסף, בגלל שידוע שהמדדים הספקטראליים בכלל וה-NDVI בפרט סובלים מרוויה כאשר כיסוי הנוף מגיע ל-100% בחרנו לנקות נתוני גובה מעל 125 ס"מ בדומה לממצא של השנה שעברה (איור 6).



איור 6: גרף פיזור של מדידות גובה מדודים (ציר X) לערכי NDVI (ציר Y)

מודלים להערכת גובה: בשנה הראשונה, פיתחנו מודל רגרסיה המבוסס על ערכי NDVI בלבד והגענו ל-RMSE של 11 ס"מ על כל הנתונים ללא חלוקה של סדרת הנתונים לכיול ואימות. בשנה זו, פיתחנו 29 מודלים מבוססי שיטות רגרסיה קלאסיות (ליניאריות, מולטי-ליניאריות, PCA ו-STEPWISE) ושיטות לימוד מכונה (ANN ו-Random Forests) כאשר כל סדרת הנתונים חולקה באופן אקראי לסדרת אימון



וסדרת אימות ביחס של 80:20, בהתאמה. מדדי הביצוע היו  $R^2$  -I (RMSE) Root Mean Square Error והם מדווחים לסדרות האימות. נספח 4 מספק הסבר קצר לכל אחת מהשיטות שהשתמשנו בהן לפיתוח מודלים. המודל הליניארי המבוסס על NDVI בלבד היווה עבורנו מעין רפרנס, כדי לבחון עד כמה מודלים אחרים יכולים לספק מדדי ביצוע טובים יותר.

הערכת קצב צימוח: להערכת קצב צימוח, חושב קצב צימוח יומי בין כל שתי מדידות גובה ולמולו קצב צימוח מחושב המבוסס על הגובה המחושב מהמודל הטוב ביותר. בחינה של הדיוק נעשתה הן באמצעות השוואה של סדרת הזמן של קצב הצימוח מערכי הגובה המדודים והמחושבים והן באמצעות חישוב של RMSE.

#### שחזור ערכי מדדי צומח ספקטראליים עקב עננות

חלק מהדימומים סובלים מעננות ויוצרים בעיה של הערכת גובה. מצד שני הורדה גורפת שלהן מקטינה את גודל בסיס הנתונים. במסגרת זו השתמשנו בשתי שיטות לשחזור ערכי מדדי צומח ספקטראליים באזורים שכוסו בעננות. על-פי כל אחת מהשיטות הוספנו את הנתונים המשוחזרים ויצרנו מודלים אמפריים שלוקחים גם את הערכים הללו בחשבון.

1. שיטת החלקת הנתונים: השיטה מתבססת על השלמת סדרת הזמן של המדדים הספקטראליים והיא מורכבת מן השלבים הבאים:

א. איתור הנקודות החשודות שסובלות מעננות באמצעות השוואת כל ערך NDVI אל מול

הערכים בשתי נקודות הזמן הסמוכות לו:  $NDVI_{t-1} > NDVI_t$  and  $NDVI_{t+1} > NDVI_t$ .

במידה שהערך נמוך משתי הנקודות הסמוכות לו הוא מסווג כסובל מעננות.

ב. החלקה של הערך באמצעות מיוצע של שתי הנקודות הסמוכות לו  $\frac{X_{t-1}+X_{t+1}}{2}$ .

ג. הבדיקה רצה בלולאה על מנת לא לפספס תרחיש של רצף של נקודות בעייתיות

2. הסרת דימומים בעל כיוון עננות גבוה מערך סף: בשיטה זו הסרנו דימומים שלמים בעלי כיוון

עננות גבוה מערך סף מסויים. בצורה כזו, הסרנו מבסיס הנתונים ליצירת המודל ערכים הסובלים

מעננות ויוצרים רעש למודל. בבדיקה של הנקודות הסובלות מעננות נמצא כי ערך המדד

הספקטראלי שלהן הוא ממוצע של שני דימומים עוקבים (זאת מכיוון שבהרבה מקרים ביום

המדדה לא היה דימום וכאמור, חושב ערך ממוצע של דימומים בפער זמן של עד חמישה ימים

לפני ואחרי המדידה). במצב כזה, אם נסיר דימומים עם עננות, נוכל לשחזר ערך ספקטראלי

אמין יותר. לדוגמה מדידת גובה שבוצעה ב-21.06 תקבל ערך מדד ספקטראלי מקביל המחושב

כממוצע של שני דימומים: 19/6 ו-24/6.



24/06/20

19/06/20

איור 7: דוגמה לדימומים עוקבים האחד סובל מעננות והשני שאינו סובל מעננות

## תוצאות

### פילוח הנתונים ליצירת מודלים

בשנה הראשונה, הצלחנו לאסוף 578 נתוני מדידות גובה אך לאחר סינון של הנתונים נותרו 243 נתונים ליצירת מודל רגרסיה ליניארית. במילים אחרות, 42% מהנתונים שנאספו התאימו ליצירת המודל. לאחר שנה נוספת של איסוף נתונים באמצעות יישומון משופר בהיבטים מסויימים, המצב התהפך כאשר 37% מהנתונים סוננו ואילו 63% נותרו ליצירת מודלים (טבלה 2).

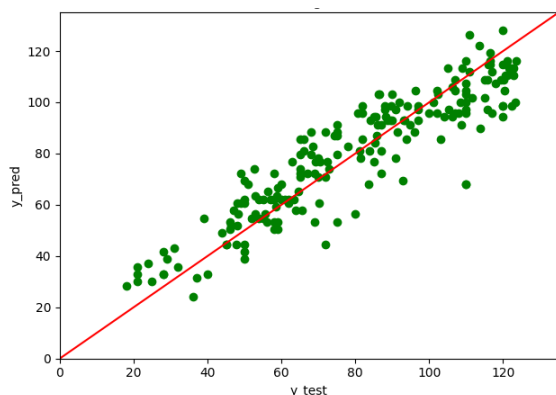
טבלה 2: פילוח הנתונים שסוננו ושותרו ליצירת מודלים

	N	Proportion (%)
Height above 125 cm	341	21%
Clouds	154	10%
Invalid Location	87	5%
Valid	1009	63%
<b>Summary</b>	<b>1591</b>	

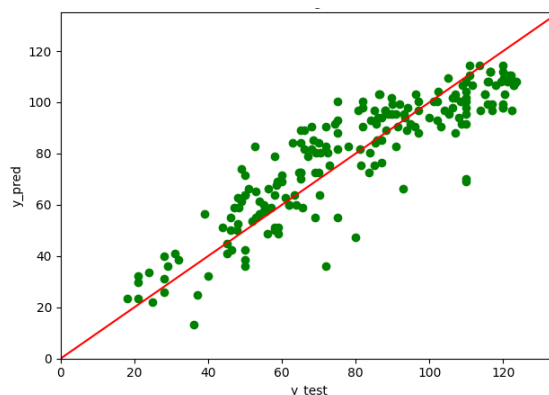
### מודלים אמפיריים להערכת גובה באמצעות מדדי צומח ספקטראליים המופקים מדימותי Sentinel-2

שיטות רגרסיה ליניארית:

איור 8 מציג דיאגרמות פיזור של סדרת האימות של ערכי גובה מדודים אל מול ערכי גובה מחושבים באמצעות רגרסיה ליניארית של NDVI ושל SAVI.



SAVI,  $R^2 = 0.83$ , RMSE=11.47 cm



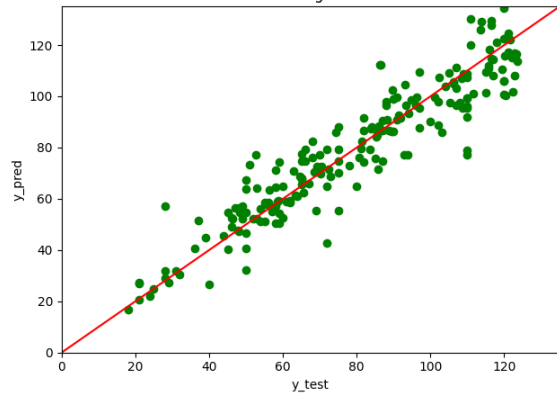
NDVI,  $R^2 = 0.79$ , RMSE=12.56 cm

איור 8: ערכי גובה מדודים (ציר Y) מול ערכי גובה מחושבים (ציר X) באמצעות רגרסיה ליניארית על בסיס NDVI ו-SAVI - סדרת האימות

ניתן לראות, כי באמצעות רגרסיה ליניארית פשוטה על בסיס מדד ספקטראלי יחיד הושגה רמת שגיאה דומה לזו שהושגה בשנה שעברה כ-11 ס"מ. אך יש לשים לב שרמת השגיאה שהושגה באמצעות הנתונים משתי העונות היא רמת השגיאה של סדרת האימות בעוד שבשנה שעברה רמת השגיאה חושבה לכלל הנתונים שהיו זמינים ללא חלוקה לסדרת כיוול ואימות ולכן מדובר בשיפור משמעותי. ניתן לייחס את השיפור לעובדה שסדרת הנתונים הדו-שנתית גדולה בערך פי ארבעה מזו של השנה הראשונה בלבד (1009 לעומת 243 נתונים). יש לשים לב שבאמצעות NDVI רמת השגיאה היא גדולה יותר בכ-10%. באמצעות רגרסיה מולטי-ליניארית עם בחירה של מדדים באמצעות Stepwise מתקבלת

תוצאה משופרת (איור 9 ואיור 10) עם  $R^2=0.87$  ו- $RMSE= 9.83$  cm. שימוש ב- PCA סיפק מודל פחות טוב עם  $R^2=0.83$  ו- $RMSE= 11.54$  cm.

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	51.374011	34.2071	1.50	0.1335
NDVI	146.87342	45.30981	3.24	0.0012*
GNDVI	-151.8754	30.4246	-4.99	<.0001*
IRECI	-70.58629	7.873858	-8.96	<.0001*
NDI45	-165.9668	45.69432	-3.63	0.0003*
DVI	304.12473	40.76513	7.46	<.0001*
RVI	4.6296752	0.518625	8.93	<.0001*
MCARI	-111.3205	32.08405	-3.47	0.0005*
NDWI	210.96335	24.00213	8.79	<.0001*
SRWI	7.5681621	2.35279	3.22	0.0013*
ARI1	11.954655	1.531464	7.81	<.0001*
CRI1	-2.718643	0.404411	-6.72	<.0001*
CHL_RE	-73.02052	36.30215	-2.01	0.0445*
NDI11	-222.1223	24.48942	-9.07	<.0001*
RE_NDVI	298.58731	82.07676	3.64	0.0003*

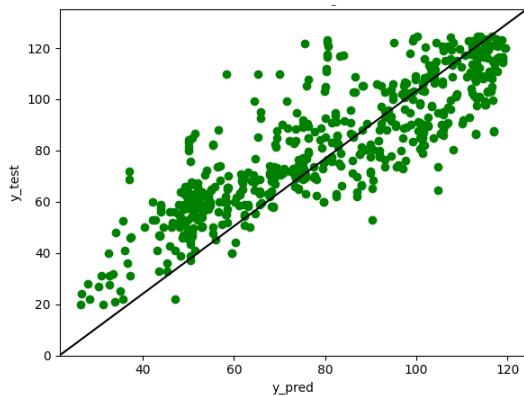


$R^2 = 0.87$ ,  $RMSE=9.83$  cm

איור 9: ערכי גובה מדודים (ציר Y) מול ערכי גובה מחושבים (ציר X) באמצעות רגרסיה מולטי-ליניארית ובחירה של מדדים באמצעות STEPWISE - סדרת האימות.

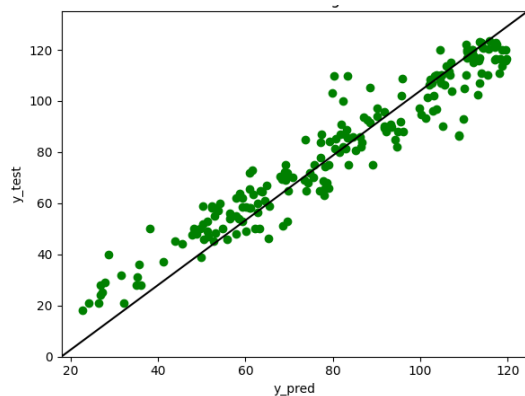
שיטות לימוד מכונה:

שימוש ב-RF על מדדים יחידים לא סיפק הערכות טובות יותר מאשר רגרסיה ליניארית. לעומת זאת, RF על כלל המדדים ועל אלה שנבחרו באמצעות STEPWISE סיפק תוצאה משופרת עם  $R^2=0.92$  ו- $RMSE= 7.9-8.1$  cm (איור 11). ביחס למודל המבוסס על NDVI זהו שיפור של 37% וביחס למודל Stepwise זהו שיפור של 22%.



$R^2 = 0.72$ ,  $RMSE=13.3$  cm

איור 12: ערכי גובה מדודים (ציר Y) מול ערכי גובה מחושבים (ציר X) באמצעות RF - סדרת האימות בחלוקה של 2019 לעומת 2020



$R^2 = 0.92$ ,  $RMSE=7.9$  cm

איור 11: ערכי גובה מדודים (ציר Y) מול ערכי גובה מחושבים (ציר X) באמצעות RF - סדרת האימות בחלוקה של 80:20.

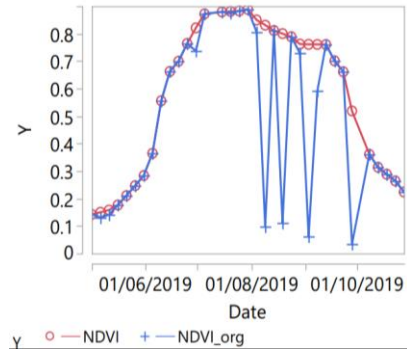
בנוסף לחלוקה לסדרות כיול ואימות של 80:20 בהתאמה, נעשה ניסיון לאמן את המודל על עונת 2019 ולאמת אותו על סדרת הנתונים של 2020 (איור 12). התוצאה פחות טובה אך היא מקבילה במידה מסויימת לתוצאה שהתקבלה ממודל המבוסס על NDVI בלבד. כך שגם במקרה הזה נראה כי RF מציג יכולות משופרות להערכת גובה באמצעות מדדים ספקטראליים. בשיטת ה-ANN התקבלות תוצאות

דומות לאלו של ה-RF. טבלה 3 מסכמת את 29 המודלים מבחינת מדדי הביצוע שלהם לסדרות האימות.

טבלה 3: סיכום המודלים שנבנו להערכת גובה מממדי צומח ספקטראליים מדימותי Sentinel-2

	method	R <sup>2</sup>	RMSE
1	Simple lm for DVI random split	0.82	11.9
2	Simple lm for NDVI random split	0.79	12.6
3	Simple lm for SAVI random split	0.83	11.5
4	Simple lm for DVI yearly split	0.73	13.1
5	Simple lm for NDVI yearly split	0.73	13.1
6	Simple LR for SAVI yearly split	0.76	12.4
7	MULTIPLE LINEAR regression all data	<b>0.85</b>	10.5
8	MULTIPLE LINEAR regression random split	0.87	9.8
9	MULTIPLE LINEAR regression yearly split	0>	50<
10	STEPWISE (backward) -all data	0.85	10.5
11	STEPWISE (backward) - random split	0.87	9.8
12	STEPWISE (backward) yearly split	0>	50<
13	PCA regression random split	0.78	12.3
14	PCA regression- all data	0.83	11.5
15	PCA regression yearly split	0.67	14.5
16	RF DVI random split	0.80	12.7
17	RF NDVI random split	0.83	11.4
18	RF SAVI random split	0.82	11.7
19	RF all VI random split	0.92	8.1
<b>20</b>	<b>RF backwards reg components random split</b>	<b>0.92</b>	<b>7.9</b>
21	RF DVI year split	0.70	13.8
22	RF NDVI yearly split	0.65	14.8
23	RF SAVI yearly split	0.68	14.3
24	RF all VI yearly split	0.73	13.4
25	RF backwards reg components yearly split	0.72	13.3
26	DNN all VI random split	0.87	10
27	DNN backwards reg components random split	0.89	9.3
28	DNN all VI yearly split	0.6	<b>15.7</b>
29	DNN backwards reg components yearly split	0>	50<

מודלי RF המשלבים בתוכם ערכים משוחזרים של רשומות הסובלות מעננות: שילוב של ערכים משוחזרים הגדיל את מספר הרשומות ל-1089. איור 13 מציג דוגמא לסדרת זמן מקורית הסובלת מעננות ועם ערכים משוחזרים. שחזור ערכים באמצעות שחזור סדרת הזמן הניב מודל עם  $R^2 = 0.89$  ו- $RMSE=8.56$  cm, ואילו ערכים משוחזרים באמצעות הסרת דימותים הניב מודל עם  $R^2 = 0.88$  ו- $RMSE=9.2$ . כלומר, למרות התוספת של הנתונים הדיוק נפגע במקצת, על-כן הוחלט להשתמש במודל ללא שחזור ערכים הסובלים מעננות.



איור 13: דוגמא לסדרת זמן מקורית (כחול) וסדרת זמן משוחזרת (אדום)

לבסוף, נבדק הדיוק של 80 הנקודות המשוחזרות באמצעות המודל RF והתקבלה שגיאה של 9.56 ס"מ. כלומר, גם אם ההחלקה לא מוסיפה לטיב המודל, היא מאפשרת הערכת גובה סבירה.

הערכת קצב צימוח: החלק הזה נמצא עדיין תחת עבודה. אנו מקווים לסיים אותו בחודשים הקרובים.

### דיון ומסקנות

הזמינות של דימותי לוויין חינוניים גדולה מאי פעם ברזולוציות ספקטרליות, מרחביות ועתיות גבוהות. ה-Sentinel 2 מספק דימותי לוויין ברזולוציה של 10-20 מטרים, כל חמישה ימים עם כ-13 ערוצים ספקטראליים, כולל ערוצים בקצה האדום וגם בא"א הבינוני הכוללים ערוצים הרגישים לתכולת מים בצמח. במחקרים רבים עד היום, בחינה של מדדים ספקטראליים בהערכה של פרמטרים צמחיים נעשתה בניסויים בקנה מידה קטן. במקביל נעשים ניסיונות של מדידות גובה בימים של המעבר של דימותי הלוויין במקומות שונים בשדה. המגבלות של שתי גישות אלו כזו נעוצות בקנה המידה הקטן ממנו נאספים הנתונים, אשר מעלים ספק ביחס לגנריות של הקשרים, כלומר, עד כמה הקשרים הללו תקפים באזורים אחרים. עם האפשרויות לאיסוף נתונים ממרחבים גדולים באמצעות טכנולוגיות מידע המאפשרות איסוף נתונים באמצעות יישומים ואיגומם בבסיס נתונים על גבי הענן, ניתן לבסס קשרים בין מדדים ספקטראליים ופרמטרים צמחיים על בסיס נתונים שנאספים משדות רבים. באופן כזה, ניתן לבחון את היעילות של המדדים על פני מרחב גדול. יותר מכך, ניתן לשלב אליהם נתונים חיצוניים כמו מדדים של קרקע חשופה, כדי לשפר את יכולת ההערכה שלהם. כלומר, ככל שנרתום חקלאים רבים יותר להשתמש באפליקציה כך תגדל האפשרות לייצר מודלים להערכת פרמטרים של צימוח המתאימים לאזורים גדולים. כאמור, בעונה הקודמת הצלחנו לקבל מודל שמאפשר הערכת גובה על בסיס NDVI עם טעות יחסית קטנה על בסיס של 250 נתונים בלבד. בעונה האחרונה הצלחנו לייצר בסיס נתונים גדול פי ארבעה. הגידול נוצר עקב ניתוח מושכל יותר של הנתונים של השנה הקודמת והן מנתונים שהצטברו בעונה השנייה. בנוסף לגידול בבסיס הנתונים, יצרנו מודלים אמפיריים הן בשיטות של רגרסיה ליניארית והן בשיטות של לימוד מכונה. המודל הליניארי על בסיס מדד ספקטראלי יחיד שופר בצורה משמעותית ביחס למודל השנה הקודמת (טעות של 11.5 ס"מ לסט האימות לעומת טעות של 11 ס"מ לכלל הנתונים בשנה הקודמת). בנוסף באמצעות מודל RF קטנה הטעות מ-11.5 ס"מ ל-7.9 ס"מ. לבסוף גם השתמשנו בשיטה פשוטה לשחזור נתונים הסובלים מעננות ולגביהם המודל הניב ערכי גובה

עם טעות של 9.6 ס"מ. יש לציין, כי חלק מהטעות נובעת מהעובדה שתאריכי המדידות בשטח ותאריכי הדימומים אינם חופפים והם יכולים להגיע להפרשים של עד 5 ימים. ככל שבסיס הנתונים יגדל עם השנים, ניתן יהיה לזקק את החלק של הטעות שנובע מההפרש הזה על-ידי סינון של נתונים שנלקחו באותו יום או בהפרשים קטנים יותר. הסיבה שלא עשינו זאת כאן היא שגם 1000 נתונים הם עדיין בסיס נתונים יחסית קטן וככל שנקטין את ההפרש מספר הנתונים יקטן משמעותית.

הצד השני של הגישה של נתוני עתק היא הרעש שיש בנתונים. גם במקרה שלנו אחוז גדול של הנתונים סלב מרעשים שונים: בעיות של מיקום, עננות וסטורציה. לכן, כמו תמיד בגישה זו, יש לשלב בכל האלגוריתמים שלב של ניקוי רעשים. הקוד שפותח במסגרת המחקר כולל בתוכו את כלל השלבים לקראת הגשתו לציבור החקלאים: שלב ארגון הנתונים, שלב חישוב של מדדים ספקטראליים, איחוד בסיסי הנתונים של המדידות בשטח ונתוני החישה מרחוק, שלב ניקוי הנתונים, שלב שחזור עננות (בעייתי בתצורתנו הנוכחית כי הוא משחזר אחורה ולכן הנתון המשוחזר מתקבל באיחור של 5 ימים), ושל בחישוב הגובה באמצעות המודל.

מה שחסר כרגע בקוד הוא היכולת להריץ את הקוד על הדימומים ולהראות את השונות המרחבית או לחילופין לספק מפות מצב מים.

לצערנו, לא הצלחנו להשלים את ההערכה של המודל שלנו בחישוב קצב הצימוח היומי. אנחנו בתחילת התהליך אך עד לסיום המאסטר של הסטודנט לתואר שני שלנו (קרי סוף שנת 2021), אנו מתעתדים לסיים גם את התהליך הזה.

בנוסף, במסגרת התואר השני, הסטודנט עובד על בניית מודלים דומים על בסיס דימומי לוויין ונוס שהוא בעל חזרתיות של יומיים ורזולוציה מרחבית של 5 מטרים. מעניין יהיה לבחון את ההבדלים בטיב המודלים.

וולקני מקים בימים אלו סביבה של הנגשת מודלים שפותחו על-ידי חוקרים בוולקני. הסביבה הזו תיקרא: ARO-TECH. בכוונתנו להגיש את הפרוייקט הזה ל- ARO-TECH שינגיש גם את היישומן לאיסוף הנתונים הצורה משופרת ומצד שני ינגיש לחקלאים מפות גובה או מפות לפי אזורים המושקים בעודף, מושקים בצורה טובה ומושקים בחוסר.

## **תודות**

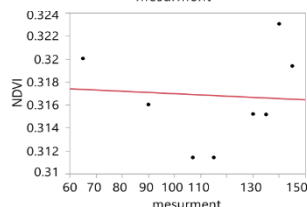
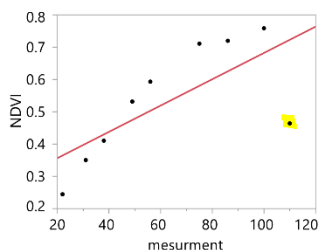
תודה גדולה לכל החקלאים שהביעו אמון ביישומונים שפותחו ועשו בהם שימוש שאפשר איסוף נתונים נרחב ופיתוח מודלים אמפיריים טובים. תודה גדולה גם לחקלאים ששלחו נתונים לטובת המודלים גם ללא שימוש ביישומונים.

נספח 1: נוסחאות של מדדי הצומח ששמשו כמשתנים להערכה של גובה הכותנה ורפרנסים

Table 1. Vegetation indices used as predictor variables for building regression and classification models of defoliation.

Acronym	Vegetation index	Equation	References
ARI 1	Anthocyanin Reflectance Index	$(1/B03) - (1/B05)$	Gitebon et al. (2001)
ARI 2	Anthocyanin Reflectance Index 2	$(B06/B03) - (B08/B05)$	Gitebon et al. (2001)
BAI	Burn Area Index	$1/(0.1 - B04)^2 + (0.06 - B08)^2$	Chuvieco, Martin & Palacios (2002)
CRI 1	Carotenoid Reflectance Index 1	$(1/B02) - (1/B03)$	Gitebon et al. (2002)
CRI 2	Carotenoid Reflectance Index 2	$(1/B02) - (1/B05)$	Gitebon et al. (2002)
CHL RED EDGE	Chlorophyll Red-Edge	$B05/B08$	Gitebon et al. (2003)
EVI	Enhanced Vegetation Index	$2.5 \times (B08 - B04)/(B08 + 6 \times B04 - 7.5 \times B02 + 1)$	Jiang, Huete, Didan & Miura (2008)
EVI2	Enhanced Vegetation Index 2	$2.5 \times (B08 - B04)/(B08 + 2.4 \times B04 + 1)$	Jiang et al. (2008)
GNDVI	Green Normalized Difference Vegetation Index	$(B08 - B03)/(B08 + B03)$	Gitebon et al. (1996)
IREDI	Inverted Red-Edge Chlorophyll Index	$(B07 - B04) \times B06/B05$	Clevers, Jong, Epema, Addink & Box (2000), Guyot & Baret (1988)
MCARI	Modified Chlorophyll Absorption in Reflectance Index	$1 - 0.2 \times (B05 - B03)/(B05 - B04)$	Daughtry (2000)
MSAVI2	Second Modified Soil Adjusted Vegetation Index	$(B08 + 1) - 0.5 \times \text{sqrt}((2 \times B08 - 1)^2 + 8 \times B04)$	Qi, Chehbouni, Huete, Kerr & Sorooshian (1994)
MTCI	MERIS Terrestrial Chlorophyll Index	$(B06 - B05)/(B05 - B04)$	Dash & Curran (2007)
NBR (NDII12)	Normalized Burn Ratio	$(B09 - B12)/(B08 + B12)$	Key et al. (2002)
NDI11	Normalized Difference Infrared Index (Band 11)	$(B08 - B11)/(B08 + B11)$	Hardisky, Klemas & Smart (1983)
NDI45	Normalized Difference Index 45	$(B05 - B04)/(B05 + B04)$	Frampton, Dash, Watmough & Milton (2013)
NDVI	Normalized Difference Vegetation Index	$(B08 - B04)/(B08 + B04)$	Rouse, Haas, Schell & Deering (1974)
NDWI	Normalized Difference Water Index	$(B03 - B08)/(B03 + B08)$	Gao (1996)
PSRI	Plant Senescence Reflectance Index - Near Infrared	$(B04 - B02)/B06$	Merzlyak, Gitebon, Chikunova & Rakitin (1999)
PSSR	Pigment Specific Simple Ratio	$B08/B04$	Blackburn (1998)
RED EDGE NDVI	Red edge NDVI	$(B08 - B06)/(B08 + B06)$	Fernández-Manoso et al. (2016)
SAVI	Soil Adjusted Vegetation Index	$1.5 \times (B08 - B04)/(B08 + B04 + 0.5)$	Huete (1988)
SZREP	Sentinel-2 Red-Edge Position	$705 + 35 \times (0.5 \times (B07 + B04) - B05)$	Guyot and Baret (1988)

Red Edge In-flection Point (REIP)	$700 + 40 \times \frac{B02 \times B03 - RE1}{RE1 - RED}$	[59]
Atmospherically Resistant Vegetation Index (ARVI)	$\frac{NIR - 2 \times RED - BLUE}{NIR + 2 \times RED - BLUE}$	[60]
Soil Adjusted Vegetation Index (SAVI)	$\frac{NIR - RED}{NIR + RED + L} \times (1 + L)$ $L = 0.5$	[61]
Modified Soil Adjusted Vegetation Index 2 (MSAVI2)	$\frac{2 \times NIR + 1 - \sqrt{(2 \times NIR + 1)^2 - 8 \times (NIR - RED)}}{2}$	[62]
Infrared Percentage Vegetation Index (IPVI)	$\frac{NIR}{NIR + RED}$	[63]
Normalized Difference Vegetation Index (NDVI)	$\frac{NIR - RED}{NIR + RED}$	[64]
Modified Soil Adjusted Vegetation Index (MSAVI)	$\frac{NIR - RED + 1}{NIR + RED + 1}$ $L = 1 - 2 \times s \times NDVI \times WDV1$ $s = 0.5$	[62]
Transformed Normalized Difference Vegetation Index (TNDVI)	$\sqrt{\frac{NIR - RED}{NIR + RED} + 0.5}$	[65]
Green Normalized Difference Vegetation Index (GNDVI)	$\frac{NIR - GREEN}{NIR + GREEN}$	[66]
Inverted Red Edge Chlorophyll Index (IRECI)	$\frac{NIR - RED}{RE1}$	[55]
Global Environmental Monitoring Index (GEMI)	$\eta \times (1 - 0.25 \times \eta) - \frac{RED - 0.125}{1 - RED}$ $\eta = \frac{2 \times (NIR - REIP) + 1.5 \times NIR + 0.5 \times RED}{NIR + RED + 0.5}$	[67]
Normalized Difference Index 45 (NDI45)	$\frac{NIR - RED}{NIR + RED}$	[68]
Perpendicular Vegetation Index (PVI)	$\sin(\alpha) \times NIR - \cos(\alpha) \times RED$ $\alpha = 45^\circ$	[69]
Difference Vegetation Index (DVI)	$NIR - RED$	[64]
Pigment Specific Simple Ratio (PSSRa)	$\frac{NIR - RED}{RED}$	[70]
Ratio Vegetation Index (RVI)	$\frac{NIR}{RED}$	[71]
Weighted Difference Vegetation Index (WDVI)	$NIR - S \times RED$ $S = 0.5$	[72]
Modified Chlorophyll Absorption Ratio Index (MCARI)	$\frac{(RED - RED - 0.2 \times (RED - GREEN)) \times RED}{RED}$	[73]
Enhanced Vegetation Index (EVI)	$\frac{2.5 \times (NIR - RED)}{NIR + 6 \times RED - 7.5 \times BLUE - 1}$	[74]
Normalized Difference Water Index (NDWI)	$\frac{NIR - SWIR}{NIR + SWIR}$	[57]
Simple Ratio Water Index (SRWI)	$\frac{NIR}{SWIR}$	[32]



נספח 2: דוגמה לפיולה מקומית בערך NDVI

נספח 3: סדרת זמן NDVI שמצביעה על כך שלא מדובר על שדה כותנה

נספח 4: הסבר קצר על השיטות של פיתוח המודלים להערכת גובה

שיטות רגרסיה קלאסיות

- i. **רגרסיה מולטי ליניארית:** היא שיטה מתמטית למציאת הפרמטרים של הקשר בין משתנה בלתי תלוי X למשתנה תלוי Y, בהנחה שהקשר ביניהם ליניארי. השיטה משמשת לניתוח מדגמים סטטיסטיים. נוסחת הרגרסיה הליניארית מחשבת את הקו הישר שעובר דרך הנקודות שבמדגם. במצב של קשר ישיר מדויק כל הנקודות במדגם ימצאו על הקו עצמו. במציאות גורמים נוספים משפיעים על המדגם והנקודות לרוב מפוזרות מסביב לקו. הקו מחושב בצורה כזאת שסכום ריבועי המרחקים של הנקודות מהקו הוא הקטן ביותר. רגרסיה ליניארית מרובה מחשבת קשר בין מספר משתנים בלתי תלויים יחד, למשתנה תלוי אחד
- ii. **PCA:** מספר המשתנים המסבירים הפונקציונליים בבעיות חיזוי הוא גדול ויכול להגיע למאות משתנים, אם לא למעלה מכך. אולם חלק גדול ממשתנים אלה מתאמים ביניהם בצורה ליניארית, מה שגורם לתופעה של יתירות במשתנים, שכן לפחות חלק מהמידע המוכלל במשתנה אחד מוכל גם במשתנה אחר המתואם אתו. שיטת PCA מתמודדת עם בעיית הקוליארייות באמצעות הגדרת

מספר קטן יותר של "משתני על" לא מתואמים שמספרם קטן משמעותית ממספר המשתנים המקוריים המשמשים בתור המשתנים המסבירים במודל החיזוי במקום המשתנים המקוריים .iii **STEPWISE** : השיטה בודקת כל משתנה בנפרד ומוצאת את המשתנים המסבירים שיש להכניס למודל בגישה של ניסוי וטעייה. גישת הרגרסיה בצעדים היא הגישה הנפוצה לבחירה של משתנים מסבירים במודל רגרסיה רב-ממדי בגישה מתחילים עם אפס משתנים במודל ומוסיפים לו משתנים מובהקים בזה אחר זה בתהליך רב-שלבי עד שמתקיימים תנאי הסיום.

### שיטות לימוד מכונה

.iv **יער אקראי (RF)**: האלגוריתם של יערות אקראיים בונה מקבץ של עצים רבים, כאשר בכל עץ בכל פיצול, הוא מגביל את המשתנים לפיהם הוא יכול לפצל ל- $m$  משתנים בלבד (מתוך  $p$  אפשריים). כמו כן, האלגוריתם מגריל תצפיות (במקום להשתמש בכל התצפיות הוא משתמש במדגם שלהן), לצורך בנייתו של עץ. אלגוריתם זה יכול לצמצם את ההשפעות של קורלציה בין משתנים, וכמו כן, הוא נותן הזדמנות למשתנים מסבירים שונים לבוא לידי ביטוי, אפילו אם הם לא בעלי העוצמה החזקה ביותר. לבסוף התוצר המתקבל הוא ממוצע החיזויים על פני כלל העצים.

.v **רשת נוירונים (ANN)**: רשת עצבית מלאכותית (ANN – Artificial Neural Network), היא מודל מתמטי חישובי שפותח בהשראת תהליכים מוחיים או קוגניטיביים המתרחשים ברשת עצבית טבעית ומשמשת במסגרת למידת מכונה. רשת מסוג זה מכילה בדרך כלל מספר רב של יחידות מידע (קלט ופלט) המקושרות זו לזו, קשרים שלעיתים קרובות עוברים דרך יחידות מידע "חבויות" (Hidden Layer). צורת הקישור בין היחידות, המכילה מידע על חוזק הקשר, מדמה את אופן חיבור הנוירונים במוח. השימוש ברשתות עצביות מלאכותיות נפוץ בעיקר במדעים קוגניטיביים, ובמערכות תוכנה שונות - בהן: מערכות רבות של אינטליגנציה מלאכותית המבצעות משימות מגוונות - זיהוי תווים, זיהוי פנים, זיהוי כתב יד, חיזוי שוק ההון, מערכת זיהוי דיבור, זיהוי תמונה, ניתוח טקסט ועוד.