Sorghum Field Effects: Location Correlations

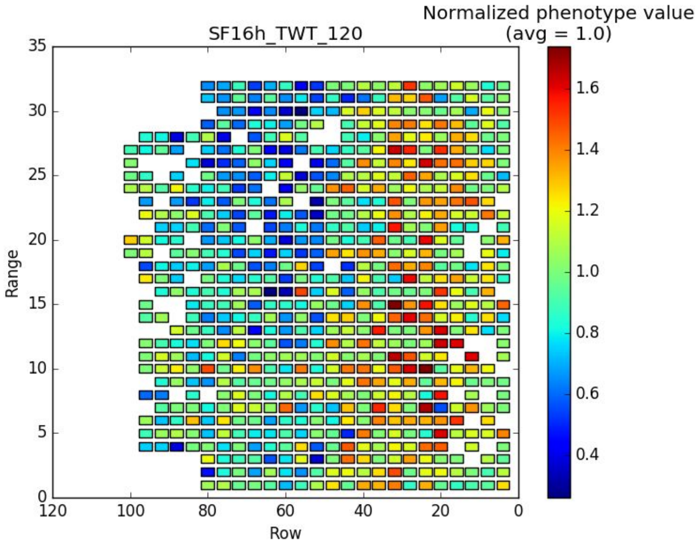
Ben Parr

bparr@cmu.edu

7/12/17

Motivation

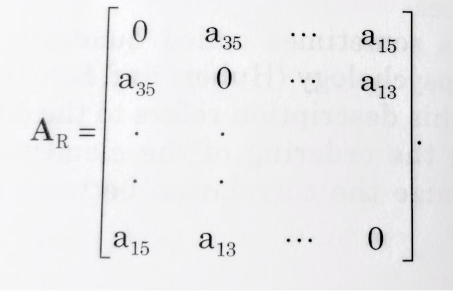
- What is the correlation between a plant's location in the field and its features?
- Is it significant?



A seeming correlation between location and a total weight of the harvest plant. Credit: Simon Heath's work

Methods

- Mantel Test using GPS locations.
  - Determined significance by randomly shuffling the GPS location of the data 10,000 times.
  - Used scikit-bio implementation.
- Computed GPS correlation for all composition, harvest, and robot features.
- Computed correlations with GPS Eastings+Northings location, as well as with Eastings only, and with Northings only.
- Finally, compared average differences between adjacent/non-adjacent plots.



$$A_R = \begin{bmatrix} 0 & a_{35} & \cdots & a_{15} \\ a_{35} & 0 & \cdots & a_{13} \\ \cdot & \cdot & & \\ \cdot & \cdot & & \\ a_{15} & a_{13} & \cdots & 0 \end{bmatrix}.$$

Example shuffling of matrix A with the new 5, 3, …, 1 order.

Results

- See Appendix for full results. Each table is sorted separately by p-value.
- All correlation coefficients are between -0.05 and 0.11
    - 2016_08_05-08_vegetation_index has largest absolute correlation (0.1080).
- East/West location has broader effect than North/South location.
    - 18 features have a p-value < 0.05 for Eastings Only.
    - Only 8 features have a p-value < 0.05 for Northings Only.

Results (by feature type)

- Composition features:
  - The Cellulose, Dry Matter and NFC prioritized features have low p-values.
- Harvest features:
  - The Total Weight feature has the lowest p-value.
- Robot features:
  - Vegetation Index, Light Interception and Laser Plant Height features have low-p-values.

Discussion

- How to correct for field effects?
  - When evaluating plants for breeding, need to normalize for field effects because the most productive plant might not have the highest biofuel output if it happened to grow in a poorer part of the field.
  - When predicting harvest features, consider synthesizing new features to improve predictive performance. Otherwise the assumption is the machine learning algorithm can take the field effects into account given the GPS locations.
- Does a max correlation of 0.11 warrant a correction?
  - Average differences between adjacent plots versus the average difference between non-adjacent plots could give some insight.
- Investigate robot features with low p-values?
  - For example, Light Interception has a relatively large differences for non-adjacent plots.

Appendix A: Data for GPS Eastings + Northings

Feature Label	Number of Data Points	Average Value of Data	Average Difference between Adjacent Plots	Average Difference between Non-Adjacent Plots	Correlation Coefficient	p-value
2016_08_05-08_vegetation_index	638	0.0700	0.0133	0.0158	0.1080	0.0001
2016_09_light_interception	623	31.6158	13.9660	18.2867	0.0560	0.0001
AD-ICP	698	0.6165	0.1042	0.1083	0.0456	0.0003
ADF	698	37.1873	4.6689	5.0671	0.0447	0.0006
2016_07_13_laser_plant_height	745	0.3868	0.1512	0.1678	0.0361	0.0010
2016_07_light_interception	482	10.9112	6.5473	7.8470	0.0490	0.0010
Cellulose	698	32.0288	4.7559	5.0943	0.0387	0.0015
SF16h_TWT_120	698	0.9246	0.4401	0.5342	0.0388	0.0029
aNDFom	698	62.6130	5.5525	6.0415	0.0381	0.0030
2016_08_light_interception	416	29.9000	16.3473	22.9669	0.0402	0.0030
NEL3x OARDC	698	60.8038	3.0126	3.2617	0.0387	0.0045
WSC Sugar	698	16.4234	5.7277	6.4311	0.0351	0.0062
Dry Matter	698	91.7598	0.7827	0.8344	0.0335	0.0089
NEM OARDC	698	57.5345	4.5891	5.0051	0.0361	0.0095
SF16h_WTL_120	698	0.7622	0.3898	0.4682	0.0325	0.0114
NEL3x ADF	698	71.9309	6.0475	6.3662	-0.0341	0.0125
NEG OARDC	698	32.9047	3.9882	4.3686	0.0346	0.0128
NFC	698	25.1992	6.5569	7.2648	0.0283	0.0151
SPCP	698	58.3669	33.4302	40.7025	0.0260	0.0414
Starch	698	0.6246	0.3766	0.3908	0.0253	0.0603

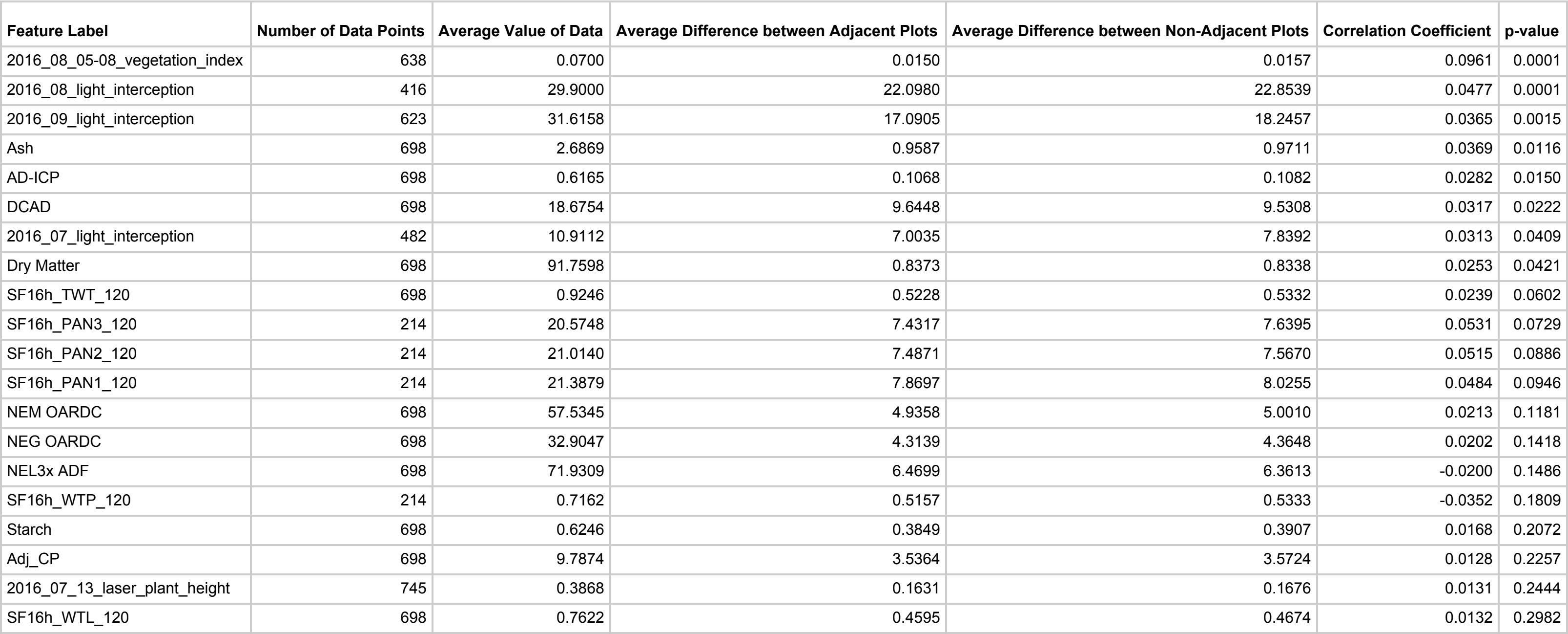
The 20 features with lowest p-values. Full results: https://goo.gl/GR5kNi#gid=1507392839

Appendix B: Data for GPS Eastings Only

Feature Label	Number of Data Points	Average Value of Data	Average Difference between Adjacent Plots	Average Difference between Non-Adjacent Plots	Correlation Coefficient	p-value
ADF	698	37.1873	4.9788	5.0635	0.0915	0.0001
aNDFom	698	62.6130	5.9088	6.0374	0.0874	0.0001
NFC	698	25.1992	7.0936	7.2585	0.0877	0.0001
SPCP	698	58.3669	39.7988	40.6285	0.0676	0.0001
WSC Sugar	698	16.4234	6.2790	6.4246	0.0916	0.0001
Cellulose	698	32.0288	4.9841	5.0917	0.0807	0.0001
2016_07_13_laser_plant_height	745	0.3868	0.1603	0.1677	0.0704	0.0001
2016_08_05-08_vegetation_index	638	0.0700	0.0156	0.0157	0.0623	0.0002
2016_09_light_interception	623	31.6158	17.4213	18.2417	0.0508	0.0003
NEL3x OARDC	698	60.8038	3.2228	3.2592	0.0559	0.0014
2016_07_light_interception	482	10.9112	7.7182	7.8272	0.0603	0.0014
AD-ICP	698	0.6165	0.1088	0.1082	0.0442	0.0022
NEL3x ADF	698	71.9309	6.3331	6.3629	-0.0488	0.0043
SF16h_HGT1_120	698	149.0917	69.1363	70.1288	0.0330	0.0089
TDN OARDC	698	65.7442	4.8354	4.8706	-0.0370	0.0169
SF16h_HGT3_120	698	148.7421	69.7097	70.8939	0.0300	0.0180
SF16h_WTL_120	698	0.7622	0.4694	0.4673	0.0354	0.0209
SF16h_HGT2_120	698	149.2063	69.5407	70.7664	0.0295	0.0233
NEM OARDC	698	57.5345	4.9809	5.0005	0.0295	0.0730
SF16h_TWT_120	698	0.9246	0.5373	0.5331	0.0261	0.0870

The 20 features with lowest p-values. Full results: https://goo.gl/GR5kNi#gid=1673187028

Appendix C: Data for GPS Northings Only



Feature Label	Number of Data Points	Average Value of Data	Average Difference between Adjacent Plots	Average Difference between Non-Adjacent Plots	Correlation Coefficient	p-value
2016_08_05-08_vegetation_index	638	0.0700	0.0150	0.0157	0.0961	0.0001
2016_08_light_interception	416	29.9000	22.0980	22.8539	0.0477	0.0001
2016_09_light_interception	623	31.6158	17.0905	18.2457	0.0365	0.0015
Ash	698	2.6869	0.9587	0.9711	0.0369	0.0116
AD-ICP	698	0.6165	0.1068	0.1082	0.0282	0.0150
DCAD	698	18.6754	9.6448	9.5308	0.0317	0.0222
2016_07_light_interception	482	10.9112	7.0035	7.8392	0.0313	0.0409
Dry Matter	698	91.7598	0.8373	0.8338	0.0253	0.0421
SF16h_TWT_120	698	0.9246	0.5228	0.5332	0.0239	0.0602
SF16h_PAN3_120	214	20.5748	7.4317	7.6395	0.0531	0.0729
SF16h_PAN2_120	214	21.0140	7.4871	7.5670	0.0515	0.0886
SF16h_PAN1_120	214	21.3879	7.8697	8.0255	0.0484	0.0946
NEM OARDC	698	57.5345	4.9358	5.0010	0.0213	0.1181
NEG OARDC	698	32.9047	4.3139	4.3648	0.0202	0.1418
NEL3x ADF	698	71.9309	6.4699	6.3613	-0.0200	0.1486
SF16h_WTP_120	214	0.7162	0.5157	0.5333	-0.0352	0.1809
Starch	698	0.6246	0.3849	0.3907	0.0168	0.2072
Adj_CP	698	9.7874	3.5364	3.5724	0.0128	0.2257
2016_07_13_laser_plant_height	745	0.3868	0.1631	0.1676	0.0131	0.2444
SF16h_WTL_120	698	0.7622	0.4595	0.4674	0.0132	0.2982

The 20 features with lowest p-values. Full results: https://goo.gl/GR5kNi#gid=1228782223