Example PReMiuM profile regression analysis with variety trial public data

We carried out a test real-data analysis of the PReMiuM R package using public maize variety trial data from state trials. It was computationally demanding to retrieve the weather data from NOAA and we thus used only the monthly average of a few weather variables for this preliminary analysis (NCEP Reanalysis Derived data provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, http://www.esrl.noaa.gov/psd/*)*. We limited our analysis to the ten hybrid varieties that were planted in the largest number of fields (from 12 to 20 farms within each state variety trial). There were no hybrids that were present in all states, so these data are unbalanced. For 2014 data Stapleton lab data analysts used the RNCEP package to retrieve weather data monthly averages for six randomly chosen NOAA weather variates [see http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html for details on these variables] that were retrieved based on the latitude-longitude coordinates given by the variety trial metadata. We used trial data from three states (URLs for data download are http://vt.cropsci.illinois.edu/corn.html, http://www.agronomy.k-state.edu/services/crop-performance-tests/corn/2014-corn-performance-test.html and https://varietytesting.missouri.edu/archive/2014-Corn-Complete.pdf). We thus had 124 rows with six columns of weather data. The input data files and R code are available at (https://github.com/AEStapleton/premium_followup).

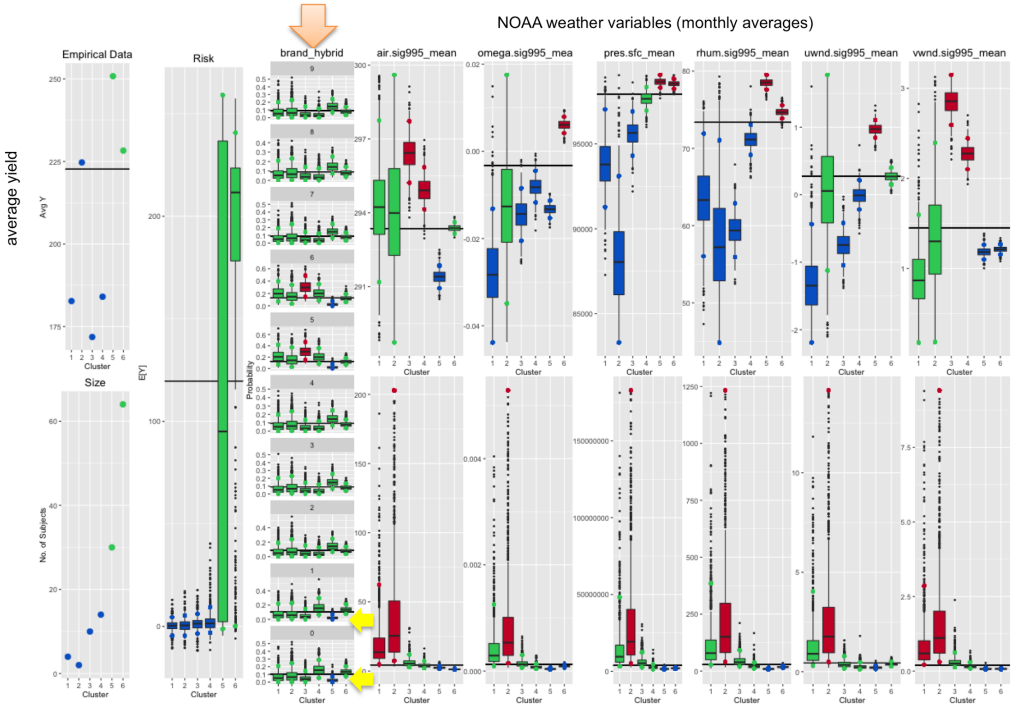

Fig. 2 PReMiuM profile box plots of analysis outputs. Summary plot outputs of PReMiuM analysis, with the PReMiuM clusters shown on the x axis of each graph. Weather variates are,

from left to right, air temperature (in Kelvin), wind (omega), air pressure, relative humidity, crosswind velocity, and vertical wind velocity. Blue color indicates less than the overall mean, green color indicates that the value is near the overall mean, and red color indicates that the values are higher than the overall mean. Hybrids are listed by number in the plot headers, as
AgriGold A6499STXRIB with 0
AgriGold A6533VT3PRIB with 1
Beck-XL 5939AMXT with 2
Beck-XL 6365AMX with 3
Burrus-6T54 3000GT with 4
DEKALB-DKC53-78 SSRIB with 5
MAT CHECK-LATE with 6
Power Plus-5C17 AMXT with 7
Power Plus-6P75 AMXT with 8
Stone-6378RIB with 9

Clusters 1 and 2 include very small number of data points, and have correspondingly high standard deviations. Cluster 6 included fields in Kansas, Missouri and Illinois (results not shown), with weather variables that were near the overall average. Cluster 5 included weather conditions related to high yield, such as higher humidity and lower air temperature (Fig. 2). We note that variety by environment crossover interactions were not completely explained by this small data analysis; this is visible in Fig 2, as most hybrids are equally likely to be in all clusters (orange arrow, mid-left hybrid subplot graphs, all green bars in all clusters) except for the Agrigold hybrids 0 and 1 (yellow arrows, with less probability of being in cluster 5) and the hybrids that were only present in Kansas fields (hybrids 5 and 6). This result would be expected in public trials, as commercial breeders would place optimized varieties into such trials. As an example of this phenomenon, the maturity check hybrid in the Kansas trails has much more variable yields across farms than the commercial Dekalb hybrid (Fig. 3b). Our small-scale variety trial data analysis can be used for describing the specific ranges of weather components in the groups of field environments. For example, the high-yielding environments in cluster 5 can be described as low-VPD environments (Gholipoor et al., 2013), and the cluster 3 and 4 fields would be high-VPD, water-limited environments.

We next conducted model fits to find the effect sizes for hybrid (genotype) factor and hybrid*environment factor, in order to calculate the genotype/genotype*environment ratio. An effective environment-grouping method would have a smaller hybrid*cluster ratio than the hybrid*farm model fit ratio (Chenu, 2015). In our analysis of the 2014 public variety trial data, we observed environment-group ratio differences and the expected decrease in the effect size ratio (Table 1). As shown graphically in Fig. 2, there is still significant variance in the interaction term after clustering with this limited set of NOAA monthly-mean variables (Table 1 last P-value row).

Table 1 Comparison of Model Factor Fits

Model term	P value	G/GxE effect size ratio
hybrid	P<0.0001	1.3
hybrid*farm	P<0.0001	
hybrid	P<0.0001	0.77
hybrid*PReMiuMcluster	P<0.0001	

Grey row background indicates the model fit using the farm identifier; white row background indicates model fit using the PReMiuM cluster variable. Generalized regression models were fit with a gamma distribution using JMP Pro v12 (SAS, Inc)..

When we visualized the data ranges stratified by PReMiuM cluster (Fig. 3a) we saw separation into shared (brown) and several smaller, more farm-specific groups, suggesting that farms within these clusters could be considered as grouped test regions characterized by specific component weather variables. As an example of the use of this type of information, a Stone breeder might wish to use pink-cluster farms to select for varieties with high yield potential but limited market area, while farms in the brown cluster might be useful for developing widely adapted germplasm. We can also see variety-specific G x E patterns, such as AgriGold A6499TXRIB placement in one cluster across all farms (ie low G x E for yield) while AgriGold A6533VT3PRIB appears in two clusters (Fig. 3a) and thus has more G x E; we note that both AgriGold varieties were planted on all farms. This difference between these two hybrids is not visible in the overall plot of hybrid by yield (Fig. 3b). A more detailed analysis of weather and farm characteristics would be of interest in the Kansas locations, as both the hybrids that were planted on all farms in Kansas (DEKALB and MAT-CHECK-LATE) grouped into four different environment clusters. This may reflect irrigation or other management practice difference in some of these fields.

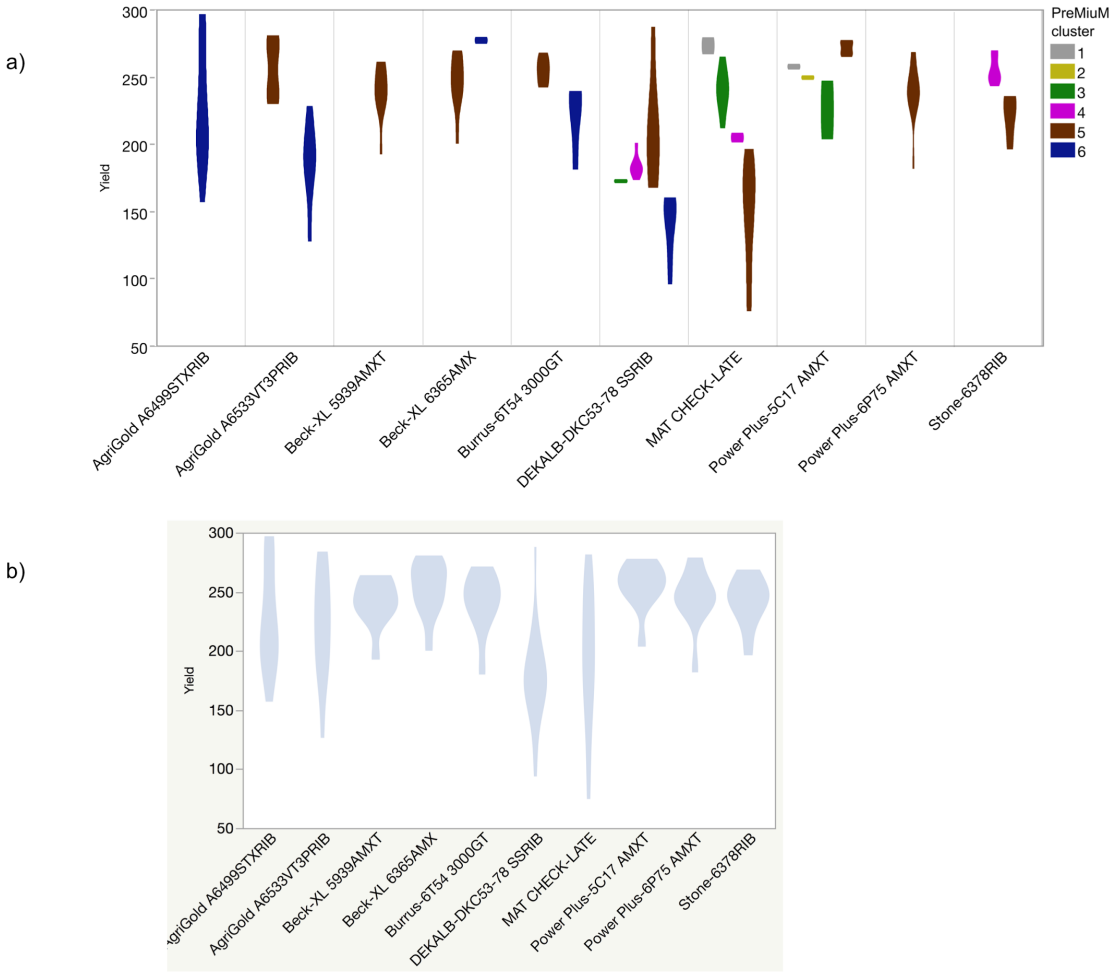

Figure 3 Yield Comparisons Across Public Variety Trials a) Violin plot of PReMiuM cluster assignment by hybrid, showing range of variation and distribution of measured yield values. b) Violin plot of yield by hybrid name; the pale blue cloud indicates the range and distribution of measured yield values.

Public variety trial data can only be used for environment grouping for one year at a time, as the hybrids are typically only planted in one year (and we cannot use relatedness information from these commercial varieties to examine SNPs shared across years). In a larger-scale real data analysis we would be especially interested in weather covariate clusters that have variety covariates (or better yet, particular SNP sets/genotype covariates) that appear in only one cluster of weather variates even though they were planted in all fields. [In other words, a cluster pattern like we observed for hybrid 5 in Fig. 2 but from a balanced design rather than missing data.] We would also want to fit SNPs that result in this type of SNP-specific cluster pattern, in order to better understand the allele-envirotype relationship. This type of outcome would 'explain' the environmental covariates that are interacting with that genotype or allele. It is known that within-year environment groups are unstable (Navabi et al., 2006), so public variety trial data are likely to be limited in envirotyping power.