Zachary Taylor

EXST 7087 – Digital Agriculture

Final Report

**Introduction**

Sugarcane (Saccharum officianarum.) is a highly important tropical crop for economics, food availability, and biofuel production. To achieve efficiency in growing this crop outside of its native range, efforts must be made to improve its resistance to disease and environmental stressors while also maintaining or improving its yields in tonnage and sucrose content. In Louisiana this effort has been chiefly led by introgression of genes from related members of the genus Saccharum that are more adapted to handle subtropical environments.(Hale et al., 2021; Todd et al., 2015) In order to combat the negative traits that are disseminated from these closely related species, a constant effort of evaluating progeny to be new parents as well as commercial cultivars is required.

In the Louisiana sugarcane breeding program it takes 12 years from the time a cross is made to the potential release of a variety to the industry(Gravois et al., 2010). During these years multiple stages of the program evaluate these varieties with differing numbers of environments as well as other aspects of experimental design such as row length, plot size, and number of replications. The stages that I primarily focus on in this study are our three trials that can be considered multi environment trials. Chronologically the first of these stages is the “On-Station Nursery” stage, which is replicated across three different environments. For this study I have combined it with the next stage, the “Off-Station Nursery” which is similarly made up of three environments leading to the “Nursery” stage being comprised of a total of six environments. Next is the “Infield” stage, which is comprised of 3 environments, this is also the first stage in which yield is determined with the use of a Weigh Wagon instead of hand counts. The final stage of the breeding program is the “Outfield” this is the test containing the largest plot sizes and the highest number of multi environment trials at ~12 locations all across the state of Louisiana.

Machine learning (ML) techniques have emerged as powerful tools for analyzing complex agricultural data and predicting plant traits based on genetic, environmental, and management factors. Among the various ML algorithms, random forests and neural networks have shown promising results in trait prediction and the identification of key factors influencing trait variation.(Kuriakose et al., 2020; Yoosefzadeh Najafabadi et al., 2023) The use of time series data and sliding window approaches can further improve model accuracy by capturing temporal patterns and changes in environmental conditions or breeding practices.

This study aims to investigate the performance of a neural network-based model for predicting multiple sugarcane traits of interest using data from three highly variable stages of the Louisiana sugarcane breeding program spanning from 1984 to 2022. The model incorporates a sliding time window approach to analyze the dataset and assess the impact of different factors, such as environment, experimental trial design, and other factors that can be attributed to the different stages.

**Dataset**

The dataset used in this study consists of breeding program data spanning from 1984 to 2022. It includes variables such as year, location, variety, stage, and traits of interest. The traits analyzed in this study are T\_SPACRE, TRS\_TON, TCA, MSTWT, POPN. The dataset was preprocessed by removing unnecessary columns and dropping rows with missing values. The data were then sorted by year, and one-hot encoding was applied to the 'VARIETY' and 'LOC' columns.

**Data Preprocessing**

A sliding time window approach was employed to analyze the dataset. Each time window spans five years, and the analysis iterates through the entire dataset, shifting the time window by one year at a time. For each window, the dataset was split into training (60%), validation (20%), and testing (20%) sets using stratified random sampling based on the 'YEAR' column. The input features were standardized using the StandardScaler from the scikit-learn library.

**Neural Network Model**

This model is a neural network implemented using TensorFlow and Keras libraries, designed to predict various traits of a given dataset. The dataset contains information on different stages (OUTFIELD, INFIELD, and NURSERY), as well as other columns such as year and variety. The model aims to predict five different traits, which are: T\_SPACRE, TRS\_TON, TCA, MSTWT, and POPN.

The neural network model is constructed using a sequential architecture consisting of 5 Dense layers. The first layer has 256 neurons with a ReLU (Rectified Linear Unit) activation function and takes input data with the shape of the training dataset's features. The subsequent layers have 128, 64, and 32 neurons respectively, all with ReLU activation functions. The final layer has a single neuron with a linear activation function to output the predicted value for the target trait.

The model is compiled with the Adam optimizer, which uses a learning rate of 0.001, and employs the mean squared error (MSE) as the loss function. The model is trained for 200 epochs with a batch size of 300. The training and validation datasets are used to optimize the model's performance during the training process. The validation dataset is useful for preventing overfitting, as it provides an unbiased evaluation of the model's performance on unseen data.

The model iterates through all possible combinations of the stages (OUTFIELD, INFIELD, and NURSERY) using the itertools library. This allows the model to learn from different scenarios and helps to identify the best stage combination for predicting each trait. For each stage combination, the model is trained on a specific window of the data, which is based on the unique years in the dataset. This sliding window approach allows the model to capture temporal patterns and trends in the data.

Once the model is trained and evaluated on the test dataset, the root mean squared error (RMSE) is calculated and stored in a DataFrame. This RMSE provides a measure of the model's predictive performance for each trait, window, and stage combination.

**Analysis**

RMSE is calculated for each grouping of window (*1-34*), stage grouping (*Outfield, Infield, Nursery, Outfield + Infield, Outfield + Nursery, Infield + Nursery, and Outfield + Infield + Nursery*), and trait (*T\_SPACRE, TRS\_TON, TCA, MSTWT, and POPN*).Once this is calculated, the variance for each grouping is also calculated by dividing the RMSE by the variance in order to acquire a coefficient of variation for each combination of the groups. Correlation matrices are also obtained for both the root mean square errors for each grouping, as well as for the variances. An absolute difference matrix is then computed between the two to look at the relationship between the error of the model and the variation within the data between the groupings within each trait. The Outfield stage performed the best, with all 5 traits of interest having the lowest RMSE value when averaged across all windows. For the variance three of the five traits (*T\_SPACRE, TCA, and POPN*) showed that the Outfield also had the lowest variation, with (*TRS\_TON and MSTWT*) having the Infield stage with the lowest variation. The RMSE to variance ratio for four of the five traits (*T\_SPACRE, TRS\_TON, TCA, and MSTWT*). had the combination of the three stages with the lowest value, with only Population having its best ratio with the Outfield.

**Conclusion**

Multiple observations can be made from these analyses. When looking at Table 4, we observe that low RMSE and low variance are directly related to one another for three of the five traits, with only TRS\_TON and MSTWT being the exceptions to this. This could potentially be explained by looking at the broad sense heritability of those traits in comparison to of the others, most specifically population. It can be hypothesized that due to the nonlinear relationship between sugar yield/stalk weight and the variation from the stages, the primary driving factor in prediction accuracy for these traits wouldn’t be as heavily influenced by the variance. The absolute difference matrices in Figure 4 the differences in the correlation between RMSE and variance within the different groupings of stage are shown. This shows the difference between the correlation of the RMSE between two groups within a trait to the correlation of variance between the same two groups for the same trait. In the difference matrix for TRS\_TON the intersection between the grouping of Outfield and Nursery with infield has a value of .66, showing a high difference between how correlated those two groupings are when comparing the rmse to when comparing the means. This can possibly be explained again by the trait not having a high correlation with the intra-stage variance and instead having its prediction accuracy being influenced by another factor.

From this study it can be observed that the Nursery stage on its own contributes the lowest prediction accuracy and the highest amount of variation. This is to be expected as it is the earliest stage of the program included in this study, that has the highest number of genotypes being evaluated. The combination of all three of the stages, while not having the lowest RMSE values, almost consistently had the best RMSE to Variance ratio. From this it can be shown that more data with more variation can help improve a model. I hypothesize that if the test set of the data for each of these stage based combinations included a more complete subset of the data that spanned all stages, that the most accuracy would come from the stage groupings from this study that had the highest RMSE to variance ratios.

**Appendix**

https://github.com/Zanthoxylem/ZT\_FinalProject

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | Table 1. Phenotypic data collected from the Louisiana sugarcane breeding program between the years of 1984 and 2021. | | | | **Variable** | **Data Type** | **Explanation** | | VARIETY | Character | Genotype name | | LOC | Character | Location/ Environment | | YEAR | Character | Year of Accession | | CROP | Numeric | Ratoon, or age of crop | | REP | Numeric | Replication | | T\_SPACRE | Numeric | Sugar Per Acre (TRS\_TON \* TCA) | | TRS\_TON | Numeric | Theoretical Recoverable Sucrose per Ton | | TCA | Numeric | Tons of Cane per acre | | MSTWT | Numeric | Stalk Weight in pounds | | POPN | Numeric | Stalk Population | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Table 2. Number of unique varieties in each 5 year window for each stage. | | | | |
| **Window** | | | **Number of Unique Varieties** | | |
| **Outfield** | **Infield** | **Nursery** |
| 1984 | 1988 | 1 | 5 | 8 | 6 |
| 1985 | 1989 | 2 | 5 | 8 | 6 |
| 1986 | 1990 | 3 | 7 | 8 | 6 |
| 1987 | 1991 | 4 | 8 | 8 | 6 |
| 1988 | 1992 | 5 | 8 | 8 | 7 |
| 1989 | 1993 | 6 | 8 | 9 | 7 |
| 1990 | 1994 | 7 | 10 | 9 | 9 |
| 1991 | 1995 | 8 | 10 | 9 | 9 |
| 1992 | 1996 | 9 | 11 | 9 | 11 |
| 1993 | 1997 | 10 | 11 | 11 | 15 |
| 1994 | 1998 | 11 | 12 | 16 | 20 |
| 1995 | 1999 | 12 | 14 | 20 | 29 |
| 1996 | 2000 | 13 | 17 | 30 | 71 |
| 1997 | 2001 | 14 | 20 | 40 | 120 |
| 1998 | 2002 | 15 | 22 | 52 | 158 |
| 1999 | 2003 | 16 | 24 | 65 | 197 |
| 2000 | 2004 | 17 | 26 | 83 | 240 |
| 2001 | 2005 | 18 | 24 | 80 | 251 |
| 2002 | 2006 | 19 | 20 | 82 | 251 |
| 2003 | 2007 | 20 | 20 | 74 | 268 |
| 2004 | 2008 | 21 | 20 | 68 | 253 |
| 2005 | 2009 | 22 | 19 | 54 | 226 |
| 2006 | 2010 | 23 | 17 | 47 | 228 |
| 2007 | 2011 | 24 | 16 | 42 | 209 |
| 2008 | 2012 | 25 | 16 | 43 | 223 |
| 2009 | 2013 | 26 | 19 | 44 | 258 |
| 2010 | 2014 | 27 | 17 | 47 | 307 |
| 2011 | 2015 | 28 | 20 | 56 | 338 |
| 2012 | 2016 | 29 | 20 | 65 | 378 |
| 2013 | 2017 | 30 | 25 | 69 | 386 |
| 2014 | 2018 | 31 | 28 | 79 | 393 |
| 2015 | 2019 | 32 | 31 | 89 | 381 |
| 2016 | 2020 | 33 | 29 | 83 | 357 |
| 2017 | 2021 | 34 | 31 | 77 | 342 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 3. Summary statistics for each trait within each stage. | | | | | |
|  | T\_SPACRE | TCA | TRS\_TON | POPN | MSTWT |
| Nursery | | | | |
| Count | 28875 | 28875 | 28875 | 28875 | 28875 |
| Mean | 7774.808 | 28.73176 | 270.4881 | 28642.55 | 2.07531 |
| Std | 2264.196 | 7.496577 | 33.9151 | 8235.352 | 0.517475 |
| Min | 199 | 0.8 | 102.1745 | 1139 | 0.37 |
| 25% | 6213.903 | 23.83352 | 249 | 23015.03 | 1.7 |
| 50% | 7700.397 | 28.68367 | 272.95 | 27581.72 | 2.03 |
| 75% | 9256.415 | 33.63669 | 294.48 | 33173 | 2.4 |
| Max | 20913.03 | 64.65815 | 478.42 | 180538.3 | 5.15 |
|  | Infield | | | | |
| Count | 5322 | 5322 | 5322 | 5322 | 5322 |
| Mean | 9019.176 | 35.12942 | 256.0969 | 32633.75 | 2.135315 |
| Std | 2522.63 | 8.860035 | 28.92587 | 9758.639 | 0.479916 |
| Min | 666.2962 | 4.94285 | 119.5579 | 8000 | 0.88 |
| 25% | 7377 | 29.4628 | 240 | 25913.2 | 1.8 |
| 50% | 8943.16 | 34.76254 | 259 | 31806.5 | 2.1 |
| 75% | 10566.77 | 40.77851 | 276 | 38414.89 | 2.45 |
| Max | 21946.48 | 85.5 | 349.54 | 78214.66 | 4.49 |
|  | Outfield | | | | |
| Count | 19680 | 19680 | 19680 | 19680 | 19680 |
| Mean | 10955.56 | 42.50204 | 257.8782 | 39752.17 | 2.169219 |
| Std | 3746.375 | 13.76151 | 32.31413 | 10837.48 | 0.534657 |
| Min | 34 | 0.2 | 87 | 2269 | 0.2 |
| 25% | 8416.621 | 33.29825 | 238 | 32670 | 1.8 |
| 50% | 10749.77 | 41.85844 | 260.6035 | 39022.5 | 2.14 |
| 75% | 13209.97 | 50.87275 | 280.4465 | 46280.63 | 2.5 |
| Max | 114560.5 | 518.7724 | 363.1243 | 112893 | 20.6 |

|  |  |  |  |
| --- | --- | --- | --- |
| Table 4. RMSE, Variance, and the Ratio of RMSE to Variance for the stage trait combinations | | | |
|  | T-SPACRE | | |
| RMSE | Variance | RMSE/ Variance Ratio |
| Infield | 2038.625 | 5640467 | 0.000371 |
| Nursery | 3102.529 | 12927503 | 0.000259 |
| Outfield | 1397.939 | 3733445 | 0.000393 |
| Infield, Nursery | 2796.242 | 11181872 | 0.000262 |
| Infield, Outfield | 1527.933 | 4625671 | 0.000335 |
| Outfield, Nursery | 2151.134 | 8887329 | 0.000278 |
| Infield, Nursery, Outfield | 2131.073 | 8875127 | 0.000254 |
|  | TRS\_TON | | |
| Infield | 24.40885 | 766.7604 | 0.034642 |
| Nursery | 27.48018 | 933.0665 | 0.032658 |
| Outfield | 22.67776 | 788.9483 | 0.029606 |
| Infield, Nursery | 25.93456 | 913.8696 | 0.029289 |
| Infield, Outfield | 22.39769 | 827.5335 | 0.027836 |
| Outfield, Nursery | 23.43519 | 918.0137 | 0.026357 |
| Infield, Nursery, Outfield | 22.87849 | 920.2945 | 0.02548 |
|  | TCA | | |
| Infield | 7.38805 | 73.03769 | 0.105049 |
| Nursery | 12.42402 | 172.3112 | 0.076675 |
| Outfield | 4.649862 | 46.34963 | 0.102286 |
| Infield, Nursery | 10.96629 | 151.6687 | 0.073786 |
| Infield, Outfield | 5.190953 | 60.07269 | 0.08763 |
| Outfield, Nursery | 8.069875 | 130.0944 | 0.070365 |
| Infield, Nursery, Outfield | 8.140519 | 128.772 | 0.066878 |
|  | POPN | | |
| Infield | 7300.049 | 77855330 | 0.000103 |
| Nursery | 7778.565 | 1.06E+08 | 7.94E-05 |
| Outfield | 5425.681 | 49381926 | 0.000113968 |
| Infield, Nursery | 7460.643 | 1.01E+08 | 7.81E-05 |
| Infield, Outfield | 5798.524 | 58465560 | 0.000101001 |
| Outfield, Nursery | 6273.594 | 93264196 | 7.88E-05 |
| Infield, Nursery, Outfield | 6371.741 | 93568034 | 7.54E-05 |
|  | MSTWT | | |
| Infield | 0.334818 | 0.190902 | 1.854365741 |
| Nursery | 0.426434 | 0.253716 | 1.71803501 |
| Outfield | 0.326764 | 0.217164 | 1.557146236 |
| Infield, Nursery | 0.380357 | 0.234813 | 1.653483855 |
| Infield, Outfield | 0.334292 | 0.216777 | 1.564312565 |
| Outfield, Nursery | 0.35672 | 0.231293 | 1.590718385 |
| Infield, Nursery, Outfield | 0.355267 | 0.229819 | 1.571517066 |
| Best performing combination of stages for each variable is highlighted | | | |

Figure 1. Mean of each trait for each of the 34 windows separated by stage.

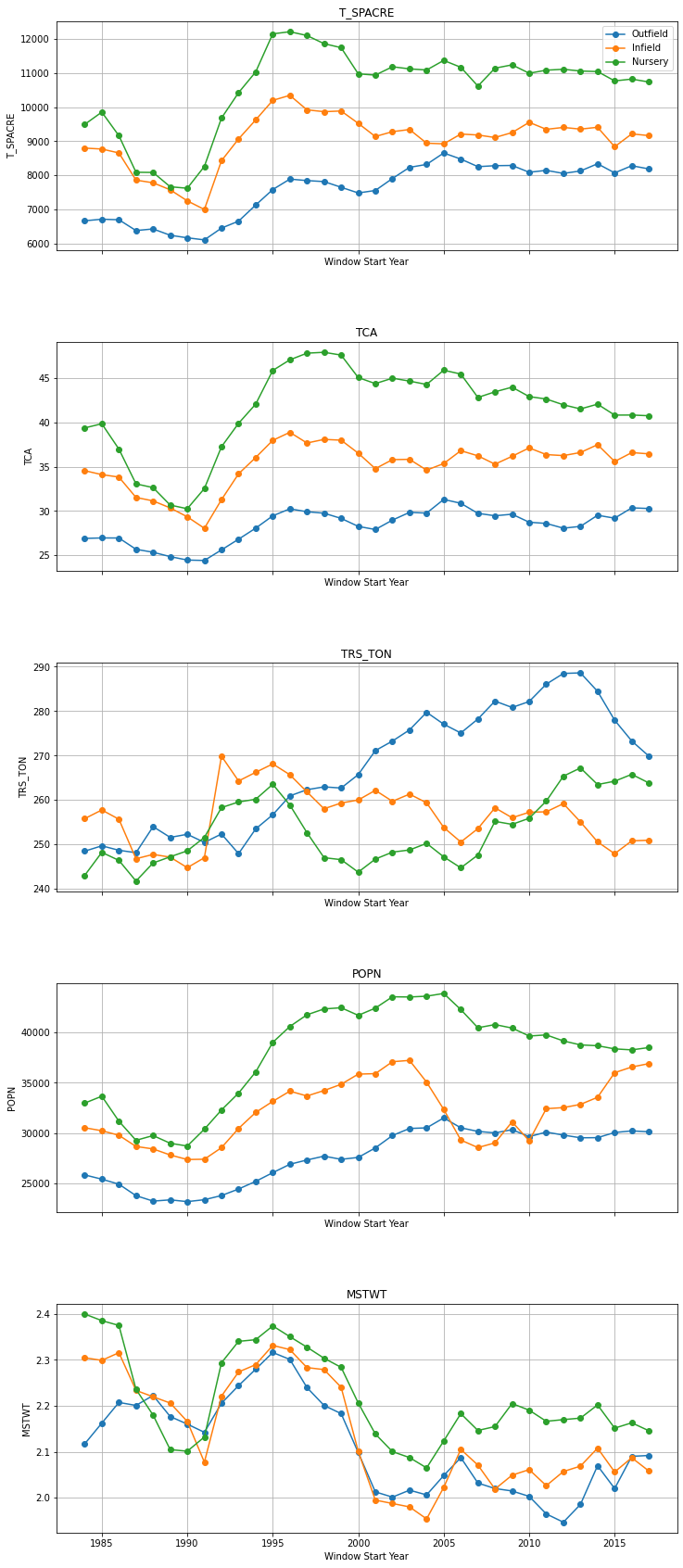
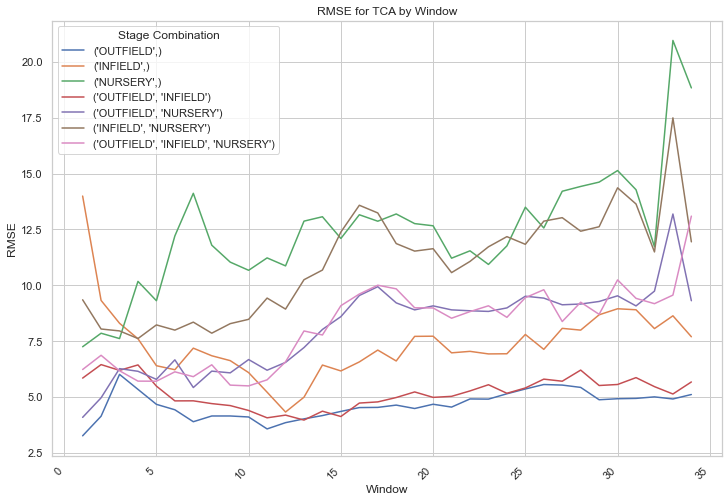
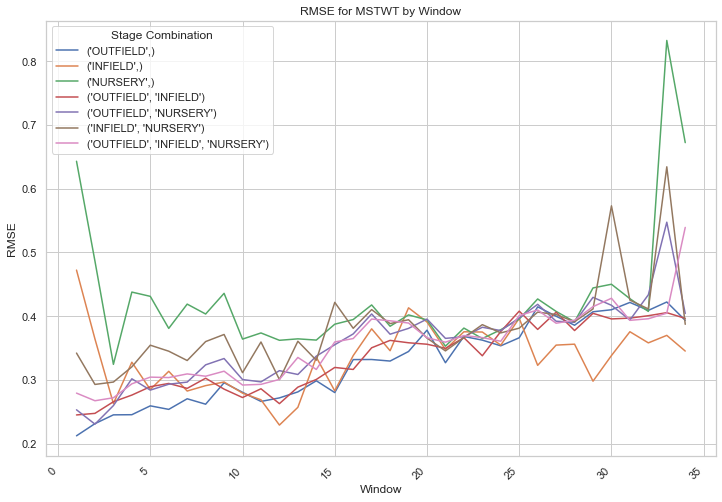


Figure 2**.** Line charts of RMSE over the windows for each of the five traits.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated



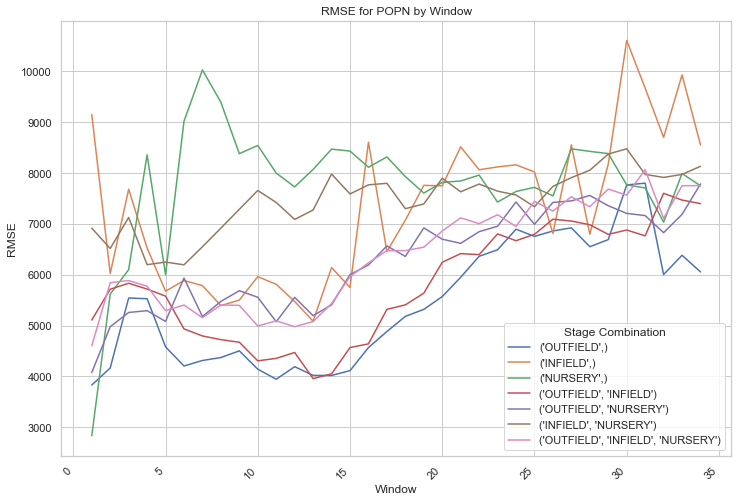


Figure 3**.** Line charts of the variance over the windows for each of the five traits.

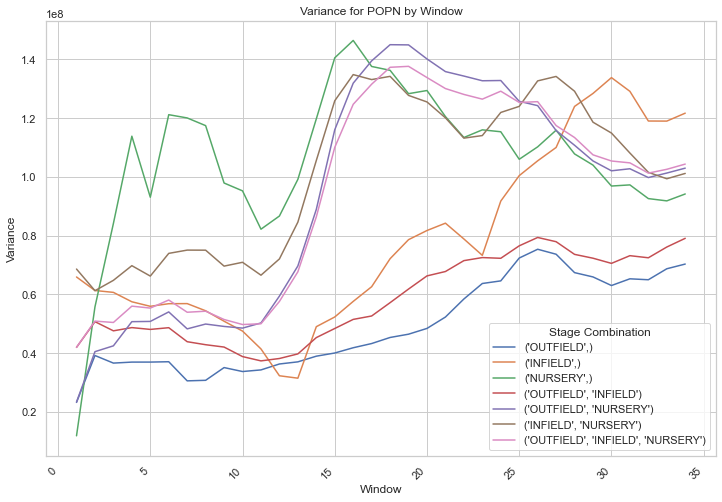
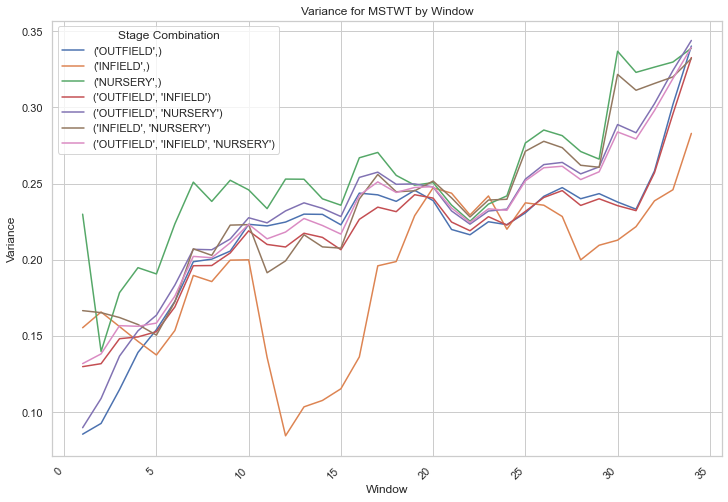
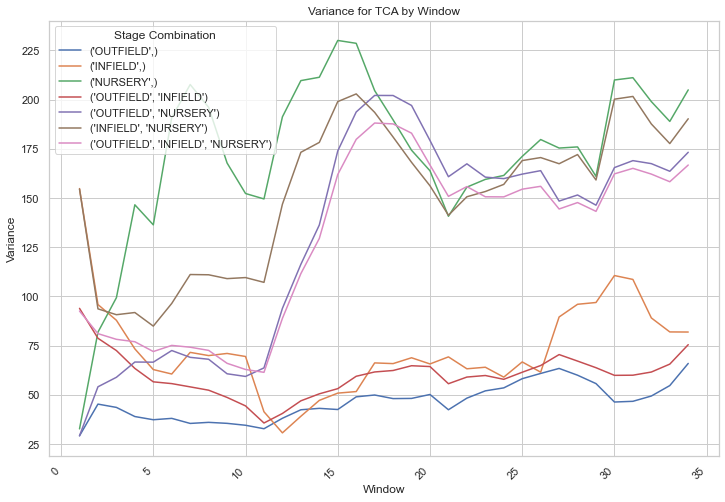
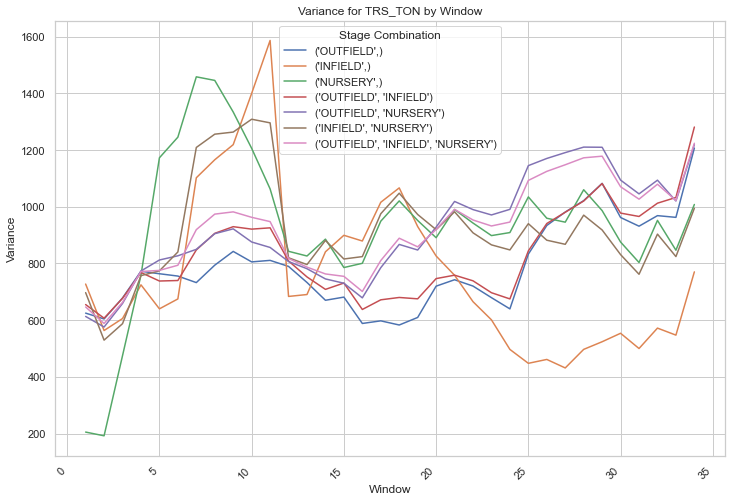
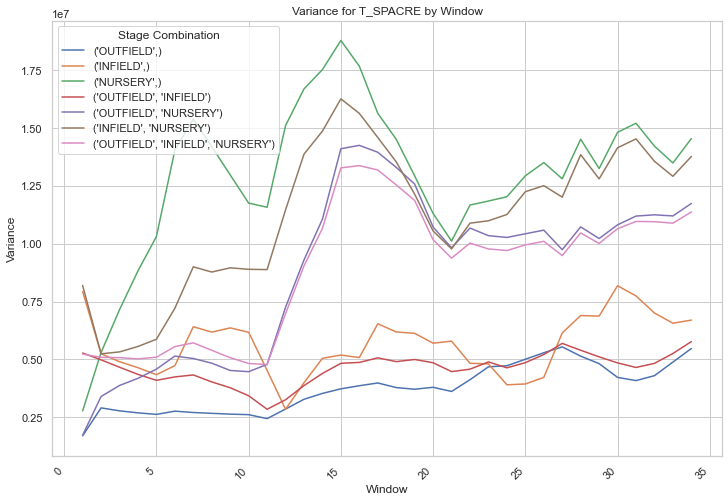
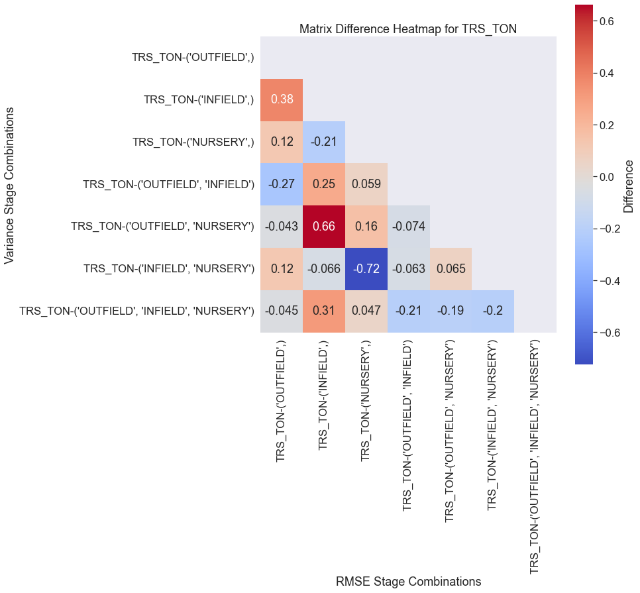
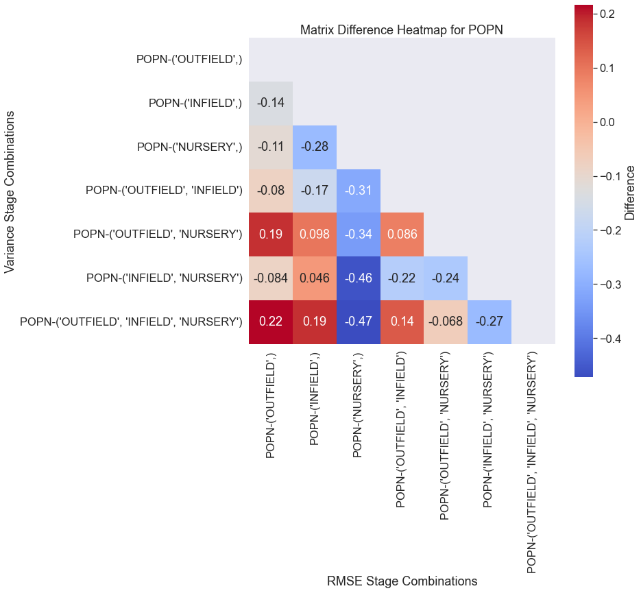
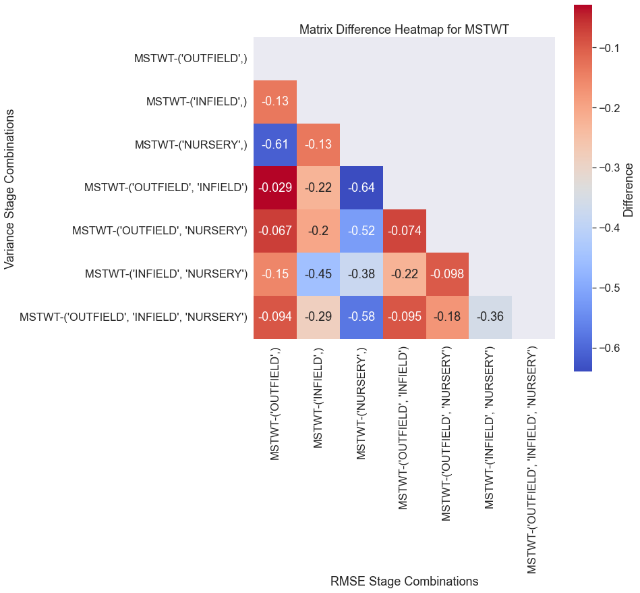
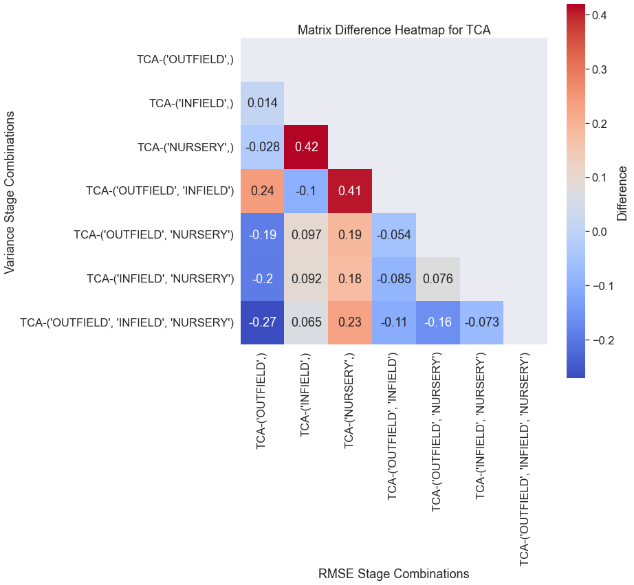


Figure 4. Absolute difference matrix from correlation matrices of RMSE and Variance

Chart, treemap chart

Description automatically generated



**References:**

Gravois, K. A., Bischoff, K. P., LaBorde, C. M., Hoy, J. W., Reagan, T. E., Pontif, M. J., Kimbeng, C. A., Hawkins, G. L., Sexton, D. R., & Fontenot, D. P. (2010). Registration of ‘L 01–283’ Sugarcane. *Journal of Plant Registrations*, *4*, 183-188. <https://doi.org/10.3198/jpr2009.10.0638crc>

Hale, A. L., Todd, J. R., Gravois, K. A., Mollov, D., Malapi-Wight, M., Momotaz, A., Laborde, C., Goenaga, R., Kimbeng, C., Solis, A., & Waguespack, H. (2021). Sugarcane Breeding Programs in the USA. *Sugar Tech*, *24*(1), 97-111. <https://doi.org/10.1007/s12355-021-01018-x>

Kuriakose, S. V., Pushker, R., & Hyde, E. M. (2020). Data-Driven Decisions for Accelerated Plant Breeding. In *Accelerated Plant Breeding, Volume 1* (pp. 89-119). <https://doi.org/10.1007/978-3-030-41866-3_4>

Todd, J., Glaz, B., Burner, D., & Kimbeng, C. (2015). Historical use of cultivars as parents in Florida and Louisiana sugarcane breeding programs. *International scholarly research notices*, *2015*.

Yoosefzadeh Najafabadi, M., Hesami, M., & Eskandari, M. (2023). Machine Learning-Assisted Approaches in Modernized Plant Breeding Programs. *Genes (Basel)*, *14*(4). <https://doi.org/10.3390/genes14040777>