Modelling 'Concord' Berry Weight Dynamics

Golnaz Badr[1], Terence R. Bates[1]

[1]Cornell Lake Erie Research and Extension Lab, Portland, NY 14769

A paper from the Proceedings of the
14th International Conference on Precision Agriculture
June 24 – June 27, 2018
Montreal, Quebec, Canada

Abstract. The growth and development of Concord (*Vitis labruscana* Bailey) depends on internal and external factors. As a result, both vegetative and reproductive cycles of Concord vary based on growing season and vine status. Fresh berry weight also fluctuates depending on the growing season and location of the vineyard. Knowledge of berry weight dynamics across growing season is essential to accurately predict final yield at harvest based on early season crop estimates. The main objective of this study was to develop the state of the art methodology to precisely estimate Concord fresh berry weight.

The experiment was conducted from 2011 to 2014 at nine vineyards distributed along the Lake Erie American Viticulture Area (AVA). Data collection on Fresh Berry Weight (FBW) was carried out for each vineyard starting two weeks pre-veraison until harvest. The Percent of final FBW was computed for each vineyard across individual growing seasons using FBW and FBW at 100 days after bloom. The weather data including daily Growing Degree Days (GDD) were obtained from Cornell University Network for Environment and Weather Applications (NEWA). A Machine Learning (ML) Randomforest (RF) algorithm was adopted to model the dynamics of percent of final FBW for each individual vineyard. The model performance was evaluated by comparing the observed and predicted FBW for the test subsets. Several statistical metrics such as Mean Error (ME), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), correlation coefficient (r), and model efficiency (EF) were computed and reported to compare the efficiency of the model for each vineyard.

The results of the model evaluation indicated that for all the years and the sites combined ME was 0.6 % and MAE was 6 % while RMSE was 1.3%, r was 0.9, and EF was 92%. The results also indicated that the ME, MAE, RMSE, and r varied depending on the vineyard location as well as the vineyard management status. The highest MAE and RMSE associated to a vineyard that was not well-maintained. This study was able to apply RF technique to successfully capture the dynamics of Concord FBW across multiple regions and growing seasons.

Keywords. *Concord, Berry Weight, Modelling, Machine Learning, Random Forest, Yield Component, Crop Estimation.*

Introduction

Concord (*Vitis labruscana* Bailey) growth and development and eventually its yield is affected by various internal and external factors. Concord grape crop yield is estimated based on destructive or non-destructive sampling schemes to help growers, juice processors, and decision makers to better design their work plans for the necessary order of tasks that needs to happen during the growing season to ensure a marketable crop yield.

Grapevine yield components are the factors in grapevine reproduction that when multiplied together they compose the yield obtained from a single vine or an entire vineyard (Coombe and Dry, 2001). The yield components for a single vine includes total number of buds per vine, total number of shoots per bud, total number of clusters per shoot, total number of berries per cluster and berry weight (Equation 1):

$$Yield = \frac{buds}{vine} \times \frac{shoots}{bud} \times \frac{clusters}{shoot} \times \frac{berries}{cluster} \times berry\ weight\ , \quad (1)$$

The total yield for an entire vineyard is calculated by summing up calculated yield (using Equation 1) for all the individual vines (Keller, 2010). The vineyard row by vine spacing as well as the trellis and training design controls the total number of vines in a vineyard and also the vine size (Keller, 2010). There are various destructive and non-destructive methods to estimate yield components early season and these methods are used to come up with accurate predictions of final yield at harvest. However***, for accurate prediction of final crop yield at harvest based on early season crop estimations, knowledge of berry weight dynamics plays a vital role***. In other words, accurate berry weight predictions are necessary for conversion of early season yield estimates to final yield at harvest predictions. Among all the yield components, berry weight prediction is the most labor intensive and costly practice. Therefore, any technique that improves the current state of the art on berry weight measurement is of interest both for growers and Concord industry.

Few studies have focused on predicting berry weight for wine grapes (*Vinifera* spp) and the authors of this current research are not aware of any studies that have focused on berry weight predictions for Concord grapevines. Fernandez Martiez et al. (2011) used data mining and artificial intelligence techniques to predict variations in 'Tempranillo' grape berry weight during ripening process. They reported that the non-parametric models behaved the best for prediction of the variables and among all the models "Gaussian Processes" had the highest accuracy with Root Mean Square Error (RMSE)= 0.0939 and Mean Absolute Error (MAE) = 0.0748.

Triolo et al. (2017) studied the simultaneous effect of major factors influencing berry mass. The model was based on vine water status, nitrogen status, berry weight, berry seed mass, and seed number. These factors were measured from veraison to harvest in 'Cabernet franc' vineyards. They conducted a multiple linear regression and reported that all the inputs had a significant effect on berry weight but vine water status represented the most important factor among all.

In the Lake Erie American Viticultural Area (AVA), the common practice for Concord yield estimation has been based on the assumption that on 30 Days After Bloom (DAB) berry weight usually

corresponds to 50% of final fresh berry weight at harvest (Bates et al. 2018). Pool et al. (1993) used Growing Degree Days (GDD) accumulation as a proxy for berry development and indicated that GDD accumulation to about 611° C (10 °C base) usually coincided with the developmental stage that berries were about 50% of their final fresh weight.

Machine learning

Machine Learning (ML) is defined as the field of study that assigns computers the ability to learn without being explicitly programmed (Samuel, 1963). Statistics and ML began their interface in the 1980s (Ratner, 2012) when ML researchers became familiar with the classical problems of statisticians mainly on predicting outcome of continuous or categorical variables. ML technically refers to "the computational process of automatically inferring and generalizing a learning model from sample data" (Dua and Du, 2011). To effectively describe the dependences among data, these learning models use statistical functions (Jain et al., 2000). In addition, the correlation and causalities between input and output is also described by the learning models (Jain et al., 2000).

ML has been used to predict yield for various agricultural crops such as: wheat (Newlands et al., 2014; Pantazi et al., 2016, Johnson et al., 2016, Veenadhri et al., 2014), barley (Johnson et al., 2016), Canola (Johnson et al., 2016), Cotton (Papageorgiou et al., 2011), Maize (Gonzalez-Sanchez et al., 2014; Veenadhri et al., 2014), Soybean(Veenadhri et al., 2014), Rice (Gandhi et al., 2016; Veenadhri et al., 2014), Apples (Papageorgiou et al., 2013), Pepper (Gonzalez-Sanchez et al., 2014), Common bean(Gonzalez-Sanchez et al., 2014), Chickpea (Gonzalez-Sanchez et al., 2014), Potato (Gonzalez-Sanchez et al., 2014), and Tomato (Gonzalez-Sanchez et al., 2014).

Supervised ML is defined as when the ML algorithm is trained by a meaningful sample data to develop a model (Dua and Du, 2011). Among various supervised ML methods Random Forest (RF) is regarded the most popular bagging ensemble classifier.

Random forest

Random forest (RF) was initially proposed by Breiman (2001) as an improvement of previous methods such as bagging of classification trees (Brieman, 1996), this is due to the fact that RF added an additional layer of randomness to the bagging process. This technique altered the way a regression or classification tree is constructed through building each individual tree using a different bootstrap sample of data. RF technique nodes are split using the best among a subset of predictors randomly chosen at that node. RF has lots of known advantages over other statistical techniques (Breiman, 2001): a) user friendly due to low number of parameters; b) better accuracy; c) faster procedure compared to bagging or boosting; d) useful internal estimation of error and, e) robust against overfitting.

The RF algorithm is developed as a package for R programming language by Liaw and Wiener (2002). The RF algorithm initially draws n_{tree} bootstrap samples from the original data, in the next step for each of the bootstrap samples the algorithm grows an unpruned regression tree with the following modification: randomly samples m_{try} of the predictors at each node and choose the best split among variables. Subsequently, the RF algorithm predicts new data by aggregating the predictions of the n_{tree} trees which for regression is done by averaging. The error estimate rate is obtained using training data, where at each bootstrap iteration, the data that is not in the bootstrap sample is predicted using the tree grown with the bootstrap sample. This is called "out of bag" or OOB data. Subsequently the OOB predictions are aggregated and the error rate is calculated and is called OOB estimate of error rate.

The main objective of this study was to develop state of the art methodology to accurately predict final fresh berry weight for Concord grapes at different stages during growing season.

Materials and Methods

Vineyards description

Experimental plots were established in Concord vineyards across Lake Erie AVA (Figure 1). A total of nine Concord experimental plots were selected based on their geographical location in the Lake Erie AVA: a) three vineyard plots in the lake zone; b) three vineyard plots in the bench zone, c) three vineyard plots in the escarpment zone (Figure 1). At each site, from 2011-2014, weekly fruit samples were collected from selected vines starting two-weeks pre-veraison to harvest and measured for berry weight. The final yield harvest was also measured and reported for the selected vines. The row-by-vine spacing was 2.6 × 2.4 m in the experimental vineyards and commercial standards for floor, nutrient, pest, and disease management were adopted (Jordan et al., 1980). Pre-and post-emergence herbicides were used for no-till weed management to maintain a 1.2 m weed-free zone under the vines, and the row centers were treated with one glyphosate application at bloom. Around budbreak a single application of ammonium nitrate fertilizer was surface broadcasted across the block at a rate of 56 kg/ha of actual N. The NY and Pennsylvania pest management guidelines for grapes (Weigle, 2006) were adopted for choosing proper fungicide and insecticide, as well as their application rates.

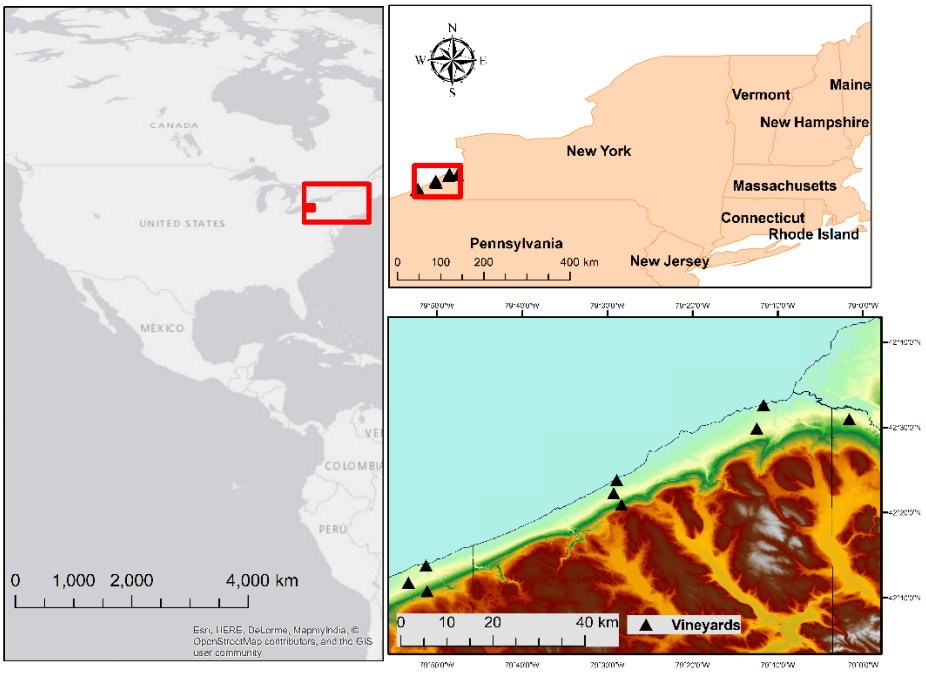

Figure 1: Location of the experimental plots that were established in the Lake Erie American Viticulture Area (AVA).

Digital Elevation Model (DEM) was used to obtain the elevation information for each experimental vineyard (NED, 2018). Percent slope rise and aspect was calculated for each experimental vineyard using ArcGIS spatial analysis tool (ArcGIS10.5.1 ESRI, Redlands, CA 2018). The soil type and drainage class were extracted from a google earth streaming interface that has been developed by UC Davis Soil Resource Lab (SoilWeb Earth, 2018). The google earth interface streams the United States Department of Agriculture- National Cooperative Soil Survey (USDA-NCSS) Soil Survey Geographic database (SSURGO, 2018) and the State Soil Geographic (STATSGO, 2018) soil survey products depending on the scale of the maps the source of the data changes within the interface. Among the nine sites, site 6 had the highest elevation and site 1 had the lowest elevation. Site 9 had the overall steepest slope (8.5%) and site 7 was the flattest one (0.2%) (Table 1). The majority of soil type in the region are silt loam and the majority of vineyards were moderately well-

drained to well-drained (Table 1).

Table 1: summary of experimental vineyard properties.

Vineyard ID	Long.	Lat.	Elevation (m)	Slope (%)	Aspect	Soil type	NEWA Weather Station	Zone	Drainage Class
1	-79.19	42.54	185.7	1.4	N	Niagara silt loam	Silver Creek, NY	Lake-East	Somewhat poorly drained
2	-79.20	42.49	234.0	3.2	NW	Chenango gravelly loam	Sheridan, NY	Bench-East	Well-drained
3	-79.03	42.51	257.5	1.6	E	Chenango gravelly loam	Versailles, NY	Escarpment-East	Well-drained
4	-79.48	42.39	192.4	1.2	W	Elnora Fine sandy loam	Portland, NY	Lake-Central	Moderately well-drained
5	-79.48	42.37	234.4	0.9	N	Pompton silt loam	Portland, NY	Bench-Central	Moderately well-drained
6	-79.47	42.35	334.9	6.4	N	Hornell silt loam	Portland Escarpment, NY	Escarpment-Central	Somewhat poorly drained
7	-79.85	42.23	220.1	0.2	W	Harborcreek-tyner complex	Northeast lab PA	Lake-West	Somewhat excessively drained
8	-79.88	42.19	230.6	0.6	NW	Pompton silt loam	Harborcreek, PA	Bench-West	Moderately well-drained
9	-79.85	42.18	323.0	8.5	NW	Mardin silt loam	Harborcreek, PA	Escarpment-West	Moderately well-drained

Weather data

Daily weather data for each experimental plot was obtained from Network for Environment and Weather Applications (NEWA, 2017). For each site a representative X and Y coordinate (geographical latitude and longitude) was selected and the closest weather station to the experimental plot was selected using the coordinate location of the plots. In case, the weather station was relatively new and there was a lack data records for the whole period of study, then the weather data was obtained from the second closest station. Daily GDD accumulation for each station was obtained for the whole period of study.

The daily GDD for NEWA weather stations is calculated using T_{min} and T_{max} with a 10 °C base temperature (Equation 2). GDD is a measure of grapevine thermal time, based on the assumption that the growth and development of grapevines linearly increase with any increase in the mean temperature. GDD calculation was based on Winkler et al. (1974).

$$GDD = \sum_{i=1}^{n}(T_i - T_b), \qquad (2)$$

where T_i is the mean daily air temperature starting from bloom date and T_b is the base temperature for grapevines (10 °C). Base temperature is the temperature at which vines resume their growth and development in spring. In this study no heat units accumulated when the average temperature was

below the base temperature. For the purpose of this study, the bloom date was used as the start point for GDD accumulation (*i*= bloom date) for each growing season. This was mainly done based on the assumption that the GDD accumulation prior to bloom date might not be directly affecting berry weight. The bloom date was recorded at site 5 from 2011-2014 and historical phenology records were also available for site 5 (Table 2); therefore, the bloom date in all the other 8 experimental vineyards were approximated by the bloom date at site 5.

Table 2: Phenology dates recorded at site 5 and historical phenology dates based on sites 5 records. This dates were used for GDD calculations across all the other sites.

Phenology stage	30-year average	2011	2012	2013	2014
Budbreak	4-May	10-May	25-Apr	3-May	12-May
Bloom	14-Jun	11-Jun	5-Jun	10-Jun	16-Jun
Veraison	22-Aug	20-Aug	10-Aug	19-Aug	25-Aug

Model input data

A data set was compiled based on daily GDD_{bloom} accumulation for all the nine sites for 2011-2014. The approximate bloom date (Julian date) for site 5 was used as the start date for *GDD* accumulation.

Percent final fresh berry weight

In this study, the recorded berry weights at each experimental vineyard were converted into percent of final berry weight assuming that the berries were ready for harvest at 100 DAB (Figure 2). This was done by diving the recorded berry weight for each individual day by the recorded berry weight at 100 DAB.

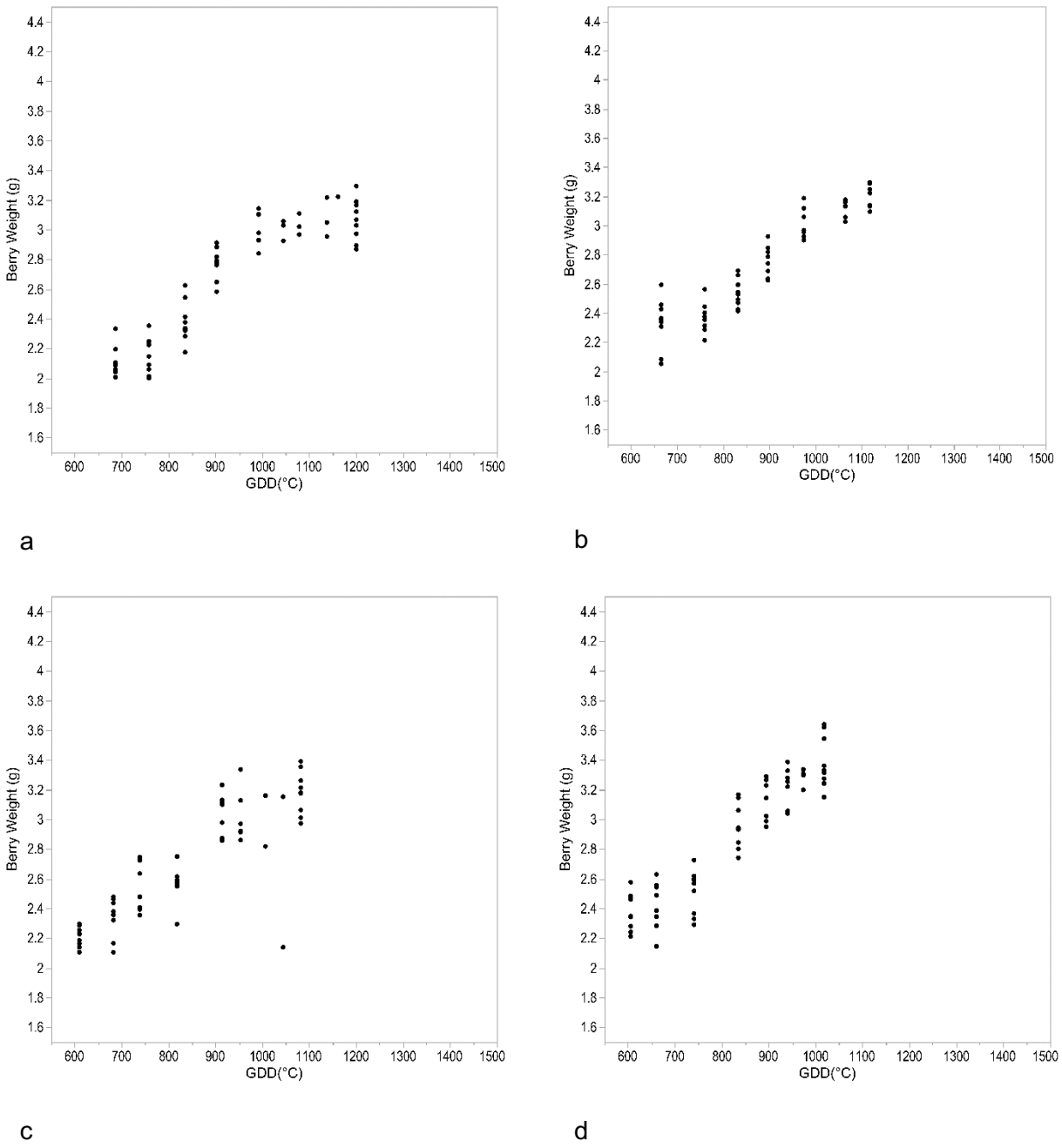

Figure 2: Berry weight (g) dynamics based on GDD for site 5: a) 2011; b) 2012; c) 2013; d) 2014.

The compiled data set for each individual experimental vineyard was then divided to test and train data-sets using a 0.75 split ratio and it was later used in the algorithm development process. The train dataset was used to train the mathematical algorithm and the test dataset was used for evaluation of the results.

Random forest algorithm implementation

The Random forest (RF) algorithm was developed using "randomForest" package in R programming language (RF, 2017). The type of RF was set to regression since the variables were continuous. The algorithm was developed using 1000 trees with a node size of 5. The number of variables randomly sampled as candidates at each split (ntry) was set to 3. The initial algorithm was developed using

percent of final berry weight as dependent variable and the independent variable was set to GDD_{bloom}. The percent of final fresh berry weight for each experimental vineyard was predicted and used for evaluation of model performance.

The mean of square of residuals for RF algorithm was computed using the following equation (Liaw and Wiener, 2002; Equation 3.):

$$MSE_{OOB} = n^{-1} \sum_1^n \{y_i - \hat{y}_i^{OOB}\}^2, \quad (3)$$

where $\hat{y}_i^{OOB}$ is the average of "Out of bag" or OOB predictions for the i^{th} observation. And percent of variance explained by RF algorithm was computed using the following equation (Liaw and Wiener, 2002; Equation 4.):

$$1 - \frac{MSE_{OOB}}{\hat{\sigma}_y^2}, \quad (4)$$

where $\hat{\sigma}_y^2$ is computed with n as divisor.

Model performance

The predicted and observed percentage of final berry weight were compared for individual experimental vineyards and the error was reported. Mean error (ME), Mean Absolute Error (MAE) (Equation 5), and Root Mean Square Error (RMSE)(Equation 6), Correlation Coefficient (r), and model efficiency (EF; Greenwood et al., 1985; Equation 7) were computed and reported for individual experimental vineyards.

$$MAE = \frac{1}{n} \sum_{i=1}^n | O_i - \bar{O} |, \quad (5)$$

$$RMSE = \left(\sum_{i=1}^n (O_i - \bar{O})^2 / n \right)^{1/2}, \quad (6)$$

$$EF = 1 - \left[\frac{\sum_{i=1}^n (S_i - O_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2} \right], \quad (7)$$

where O_i is the observed value, S_i is the simulated value, $\bar{O}$ is the mean observed value, and n is the number of observations.

The overall procedure of model development can, therefore, be divided into five major steps including data collection, data compile, model development, model performance evaluation, and model implementation (Figure 3). The data collection step covers the sampling, data entry, and obtaining data from various sources. Data compile mainly includes organizing the important variables for individual experimental vineyards and defining the training and testing schemes. Model development includes defining the algorithm properties and assigning the dependent and independent variables. The model predictions are computed and recorded at the model development stage. In addition, the percent of variance explained and mean square error of RF algorithm is computed at the model development stage. Model performance evaluation using the error measures such as ME, MAE, RMSE, and r usually occurs at the model evaluation stage and the overall error is computed at this stage. Finally, if the model performs reasonably well based on its error measures then the model can be implemented as a decision support tool that can potentially be adopted by researcher, growers, and extension educators (Figure 3).

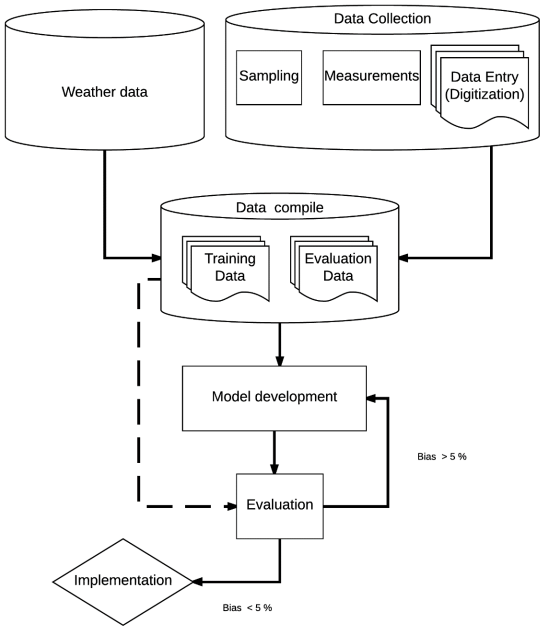

Figure 3: Overall research method.

Results and Discussion

In this study RF algorithm was employed to develop a model that has the ability to predict the dynamics of berry weight throughout the growing season for nine different experimental vineyards across multiple growing seasons. The prediction results were reported as percentage of final fresh berry weight for individual experimental vineyards. The model was initialized with GDD_{bloom} as independent variable. To satisfy the model parsimony rules, this study only focused on GDD_{bloom} as the independent variable to predict the percent of final fresh berry weight. Having fewer variables as model inputs, makes the implementation of the model and its future application more user friendly. In addition, fewer inputs, can reduce the chance of introducing error into the model.

Percent of final fresh berry weight for individual experimental vineyards were computed and statistics for each individual site was computed (See Figure 4 as an example). The results indicated that, on average highest percent of final fresh berry weight was obtained for site 4 (86±11%) and lowest was for site 6 (78±12 %) among all the experimental vineyards for all the years. This can be partially explained by the vineyard properties as vineyard is located in the escarpment zone with the highest elevation that translates to a colder region with a lower overall GDD_{bloom} accumulation and lower overall air temperature. In addition, site 6 soil drainage class is defined as “some-what poorly drained” which might be also a contributing factor to the lower overall development of berries in that site (Table 1). On the other hand, site 4 is located near the Lake Erie with one of the lowest elevations which potentially translated to relatively milder temperatures and higher GDD_{bloom} accumulations that can partially explain the higher growth and development rates as indicated by percent of final fresh berry weight.

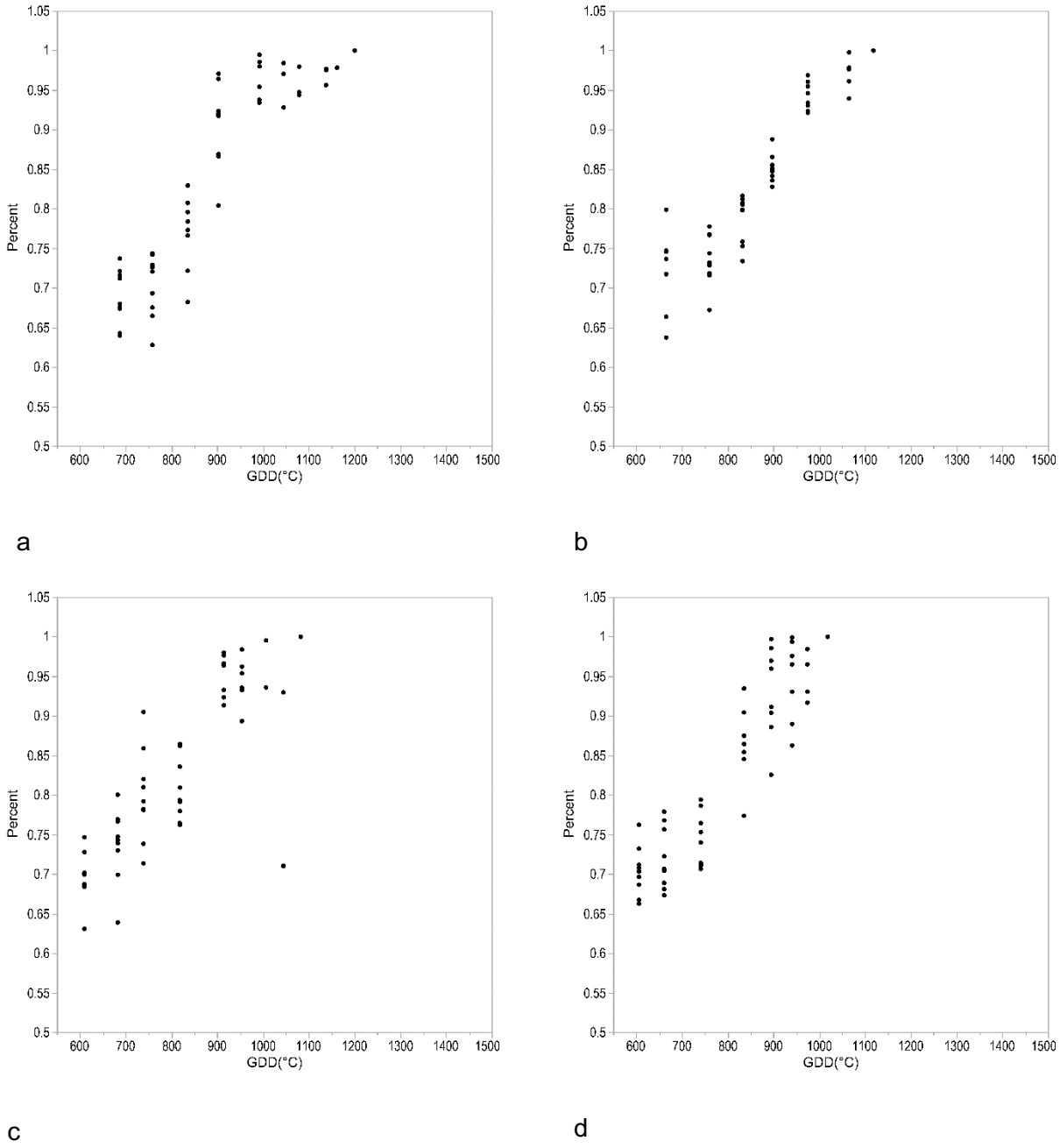

Figure 4: Percent of final fresh berry weight progress with increasing GDD for site 5: a) 2011; b) 2012; c) 2013; d) 2014.

Model performance

The results of comparing the predicted and observed percent of final fresh berry weight (Table 3) indicated that ME was on average 0.06 %, with vineyard 4 having the highest ME (2.1 %) and vineyard 1 having the lowest ME (0.02%). MAE was on average 6 % with vineyard 4 having the highest bias (10%) and vineyard 1 having the lowest MAE (3%). The RMSE on average was 1.4 % and the highest

RMSE was calculated for vineyard 4 (4.7%) and the lowest RMSE was calculated for vineyards 1 and 5(0.4 %) (Table 3). On average these models were able to explain 92 % of the variance in the dependent variable (EF), the models for vineyard 5 had the highest EF (97 %) and the vineyard 4 had the lowest EF (79%) (Table 3). Overall percent of variation explained by RF algorithm was 80.91 % with vineyard 4 having the lowest variance explained by the model (70%) and vineyard 1 indicated the highest variance explained by the RF algorithm (87.73%). The mean squared of error for the RF algorithm was 0.0062 and the site 4 has the highest MSE for RF (0.02) while site 1 had the lowest MSE for RF at 0.0002 (Table 3).

Table 3: Summary of model performance evaluation results. Mean Error (ME), Mean Absolute Error (MAE), RMSE (Root Mean Square Error), correlation coefficient (r), Efficiency (EF), variation explained by Random Forest mode (Var. RF), Mean Square Error for the Random Forest model (MSE RF)

Vineyard ID	ME	MAE	RMSE	r	EF	Var. RF (%)	MSE RF
1	0.0002	0.03	0.004	0.94	0.96	87.73	0.0002
2	0.0090	0.08	0.020	0.93	0.96	86.97	0.0080
3	0.0090	0.08	0.013	0.93	0.98	80.19	0.0120
4	0.0210	0.10	0.047	0.84	0.79	70.03	0.0200
5	0.0020	0.04	0.004	0.94	0.97	84.31	0.0030
6	0.0040	0.05	0.007	0.88	0.93	80.08	0.0030
7	0.0060	0.05	0.011	0.85	0.86	76.70	0.0040
8	0.0020	0.05	0.012	0.88	0.90	75.36	0.0040
9	0.0020	0.04	0.006	0.91	0.95	86.85	0.0020
Average	0.0061	0.06	0.014	0.90	0.92	80.91	0.0062

The correlation coefficient between the observed and estimated percentage of final berry weight was on average 0.92 where vineyard 4 had the lowest correlation among all the sites (0.84) and vineyards 1 and 5 had the highest correlation at 0.94 (Table 3; Figure 5). The overall trend in the model performance evaluation indicated that despite having relatively high accuracy, site 4 constantly is flagged as the experimental vineyard with highest bias compared to the other experimental vineyards. This behavior can be partially attributed to the NEWA weather station data that was used for vineyard 4, as it might not well-represent the micro-climate in the vineyard due to the fact that vineyard 4 was located about 2.8 km away from the Portland, NY weather station.

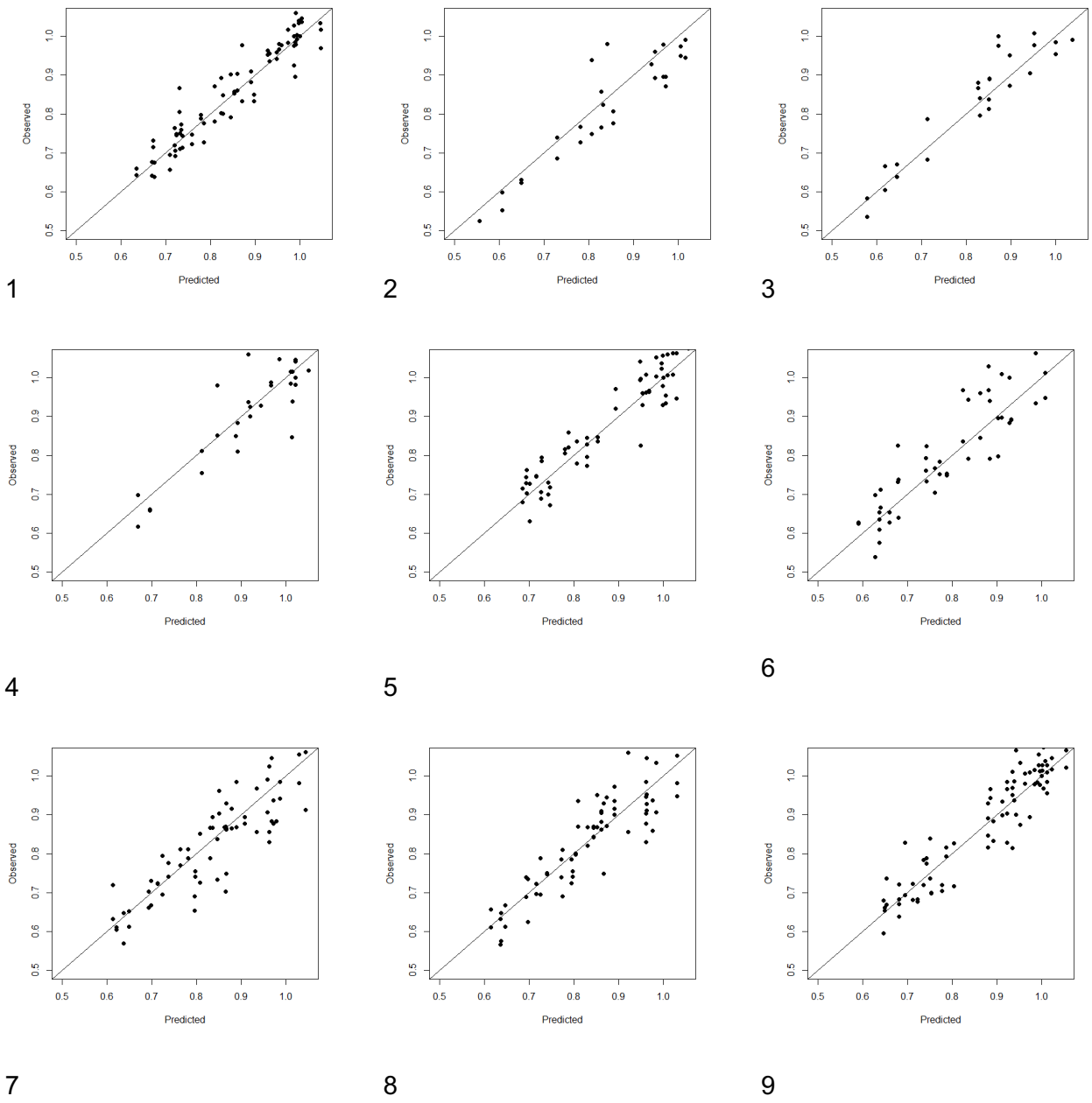

Figure 5: Comparison of predicted and observed percent of fresh berry weight for all the nine sites (site1-9).

The overall variation in model performance and the associated bias is also impacted by the cultural practices and management techniques in various experimental vineyards. Based on the data sampled across all these experimental vineyards, it is inferred that differences in topography, soil, micro-climate, closeness to waterbody (in this case Lake Erie), and cultural practices are contributing factors to the overall variation in Concord berry weight dynamics. Hence, any developed model that should be able to robustly predict the berry weight dynamics has to be least sensitive to changes in the micro-scale biophysical environment surrounding Concord grapevines. Such a model would have the ability to predict the berry weight accurately at a regional scale and can later be trained to predict the berry weight dynamics at a continental scale. This study was able to develop such a robust model that performed well at a regional scale. Future studies should focus on implementing the results of this study into a decision support tool possibly by using user friendly web applications that can be used by decision makers, researchers, and growers to make timely and better informed decisions

earlier during the growing season.

Summary

A four-year study from 2011-2014 was designed to measure berry weight at weekly intervals starting at two weeks pre-veraison until harvest. The measurements were carried out at nine different Concord experimental vineyards across Lake Erie grape AVA and the data was compiled to a train and test data-set including GDD_{bloom} and percent of final fresh berry weight. A RF algorithm was developed and trained to predict the percentage of final berry weight based on GDD_{bloom} in each experimental vineyard. The results indicated that the model had an overall ME of 0.6 % and MAE was on average 6% with a range from 3 to 10 %. RMSE was on average 1.4 % and the correlation coefficient between predicted and observed percent of final fresh berry weights were 0.9 and the model was overall 92 % efficient in explaining the variation in the data. The results are promising as the overall bias is low. Therefore, it is recommended that this model also get tested in other geographical regions and with other grape varieties to come up with a general berry weight modeling scheme that has been tested across multiple climate zones.

Acknowledgements

The authors would like to thank Cornell Lake Erie Research and Extension lab personnel for their effort in collecting the berry weight data. This research was funded by: The Lake Erie Processor Fund, the New York wine and Grape Foundation, and USDA-NIFA-SCRI Project # 2015-51181-24393.

References

ArcGIS.10.5.1. ESRI (Environmental Systems Research Institute, Relends, CA). (2018). Software.

Bates, T., Dresser, J., Eckstorm, R., Badr, G., Betts, T., & Taylor, J. (2018). Variable rate mechanical crop adjustment for crop load balance in 'concord' vineyards. In IEEE proceedings on IOT, May 2018, Tuscany, Italy.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24 (2),123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Coombe, B.G., & Dry, P.R. (2001). *Viticulture* Volume 2: Practices (7th printing), Winetitles, Adelaide, Australia.

Dua, S. & Du, X. (2011). *Data Mining and Machine Learning in Cybersecurity*. Auerbach Publications. Taylor & Francis group.

Fernandez Martinez, R., Ascacibar, F. J. M.-P., Espinoza, A. V. P., & Lorza, R. L. (2011). Predictive modelling in grape berry weight during maturation process: comparison of data mining, statistical and artificial intelligence techniques. *Spanish Journal of Agricultural Research*, *9*(4), 1156–1167.

Gandhi, N., Armstrong, L. J., Petkar, O., & Tripathy, A. K. (2016). Rice crop yield prediction in India using support vector machines. In *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)* (pp. 1–5).

Gonzalez-Sanchez, A., Frausto-Solis, J., & Ojeda-Bustamante, W. (2014). Predictive ability of machine learning methods for massive crop yield prediction. *Spanish Journal of Agricultural Research*, *12*(2), 313–328.

Greenwood, D.J., Neeteson, J.J., & Daycott, A. (1985). Response of potatoes t N fertilizer dynamic model. *Plant Soil*, 85, 185–203.

Jain, A.K., Duin, R.P.W., & Mao, J. (2000). *Statistical pattern recognition: A review*. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (1), 4–37.

Johnson, M. D., Hsieh, W. W., Cannon, A. J., Davidson, A., & Bédard, F. (2016). Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. *Agricultural and Forest Meteorology*, *218–219*, 74–84.

Jordan T.D., Pool R.M., Zabadal T.J., & Tomkins J.P. (1980). *Cultural practices for commercial vineyards*. Cornell Extension Misc. Bull. 111, pp. 29-30. NY State College of Agriculture and Life Sciences, Cornell University, Ithaca.

Keller, M. (2010). Chapter 6 - Developmental Physiology. In *The Science of Grapevines* (pp. 169–225). San Diego: Academic Press.

Liaw, A. & Wiener, M. (2002). Classification and regression by randomForest. *R News,* 2, 18–22.

NED. (2018). National Elevation Dataset. https://lta.cr.usgs.gov/NED. Accessed 27 June 2018.

NEWA. (2017). Network for Environment and Weather Applications. http://newa.cornell.edu/. Accessed 11 April 2018.

Newlands, N. K., Zamar, D. S., Kouadio, L. A., Zhang, Y., Chipanshi, A., Potgieter, A., et al. (2014). An integrated, probabilistic model for improved seasonal forecasting of agricultural crop yield under environmental uncertainty. *Frontiers in Environmental Science*, 2.

Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L., & Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture*, *121*, 57–65.

Papageorgiou, E. I., Markinos, A. T., & Gemtos, T. A. (2011). Fuzzy cognitive map based approach for predicting yield in cotton crop production as a basis for decision support system in precision agriculture application. *Applied Soft Computing*, *11*(4), 3643–3657.

Papageorgiou, E. I., Aggelopoulou, K. D., Gemtos, T. A., & Nanos, G. D. (2013). Yield prediction in apples using Fuzzy Cognitive Map learning approach. *Computers and Electronics in Agriculture*, *91*, 19–29.

Pool R.M., Dunst R.E., Crowe D.C., Hubbard H., Howard G.E. & DeGolier G. (1993). Predicting and controlling crop on machine or minimal pruned grapevines. *In: Proceedings of the Second Nelson J. Shaulis Grape Symposium: Pruning Mechanization and Crop Control*. Pool R.M. (Ed.), NY State Agricultural Experiment Station, pp. 31-45.

Ratner, B. (2012). *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data*, Second Edition. Auerbach Publications.

RF. (2017). Ranfomforest. https://cran.r-project.org/web/packages/randomForest/randomForest.pdf. Accessed 12 April 2018.

Samuel, A. (1963). Some studies in machine learning using the game of checkers, In Feigenbaum, E., and Feldman, J., Eds., *Computers and Thought* (pp. 14–36). McGraw-Hill, New York.

SoilWeb Earth. (2018). UC Davis Soil Resource Lab. https://casoilresource.lawr.ucdavis.edu/soilweb-apps/. Accessed 27 April 2018.

SSURGO (2018). Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. Soil Survey Geographic (SSURGO). https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/?cid=nrcs142p2_053627. Accessed 27 April 2018.

STATGO (2018). Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. U.S. General Soil Map. (STATSGO2).https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/geo/?cid=nrcs142p2_053629. Accessed 27 April 2018.

Triolo, R., Roby, J. P., Plaia, A., Hilbert, G., Buscemi, S., Di Lorenzo, R., & van Leeuwen, C. (2018). Hierarchy of Factors Impacting Grape Berry Mass: Separation of Direct and Indirect Effects on Major Berry Metabolites. *American Journal of Enology and Viticulture*, *69*(2), 103–112.

Veenadhari, S., Misra, B., & Singh, C. (2014). Machine learning approach for forecasting crop yield based on climatic

parameters. In *2014 International Conference on Computer Communication and Informatics* (pp. 1–5).

Weigle, TH. (2006). *New York and Pennsylvania Pest Management Guidelines for Grapes*. Cornell University, Ithaca.

Winkler, A.J., Cook, J.A., Kliewer, W. M., Lider, L. A., & Cerruiti, L. (1974). *General Viticulture*. 4th edition. University of California Press, Berkley.