2023

Supervised Machine Learning to Predict Farinograph Stability of Spring Wheat Flour

Brendan Kienlen

SUPERVISED MACHINE LEARNING TO PREDICT FARINOGRAPH STABILITY OF SPRING WHEAT FLOUR

BY

BRENDAN KIENLEN

A thesis submitted in partial fulfillment of the requirements for the

Master of Science

Major in Plant Science

South Dakota State University

2023

THESIS ACCEPTANCE PAGE
Brendan Kienlen

This thesis is approved as a creditable and independent investigation by a candidate for the master's degree and is acceptable for meeting the thesis requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Karl Glover
Advisor Date

David Wright
Department Head Date

Nicole Lounsbery, PhD
Director, Graduate School Date

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Karl Glover for accepting me as a graduate student and pushing me past the boundaries of what I thought myself to be capable of. He has been a great mentor and I will not forget what he's done for me.

I would like to thank the rest of my committee, Dr. Melanie Caffe and Dr. Chris Saunders. Dr. Caffe provided crucial plant breeding knowledge in her graduate course, and this thesis would not have been possible without the data science courses and guidance provided by Dr. Saunders.

I would also like to thank all my coworkers at the Seedhouse, especially those within the Spring Wheat Breeding Program. Help is never far away in the Seedhouse, and usually comes paired with a good laugh. To my coworkers in the Spring Wheat Breeding Program, to say I wouldn't have been able to do this without all of you would be an understatement.

Lastly I'd like to thank my family, especially my parents, who always listen to my crazy ideas and encourage me to keep pushing. I'd be lost without you guys.

CONTENTS

ABBREVIATIONS

AACC: American Association of Cereal Chemists

AFLP: Amplified fragment length polymorphism

AGP: Arabinogalactin-peptides

AYT: Advanced Yield Trials

ELISA: Enzyme-linked immunosorbent assay

GWAS: Genome-wide association study

GBS: Genotype-by-Sequencing

HRSW: Hard Red Spring Wheat

HRWW: Hard Red Winter Wheat

HMW-GS: High molecular weight glutenin

LMW-GS: Low molecular weight glutenin

MTI: Mixing Tolerance Index

NCI: Northern Crops Institute

NDSU: North Dakota State University

NSP: Non-starch polysaccharides

PCR: Polymerase Chain Reaction

QTL: Quantitative Trait Loci

RAPD: Random amplification of polymorphic DNA

RMSE: Root Mean Square Error

SNP: Single Nucleotide Polymorphism

SSR: Simple Sequence Repeats

SRWW: Soft Red Winter Wheat

WE-AX: Water-extractable arabinoxlyans

WU-AX: Water-unextractable arabinoxylans

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

SUPERVISED MACHINE LEARNING TO PREDICT FARINOGRAPH STABILITY OF SPRING WHEAT FLOUR

Hard red spring wheat (*Triticum aestivum* L., HRSW) flour is mostly used to produce wheat-based foods where dough strength is a major quality component. Maintaining adequate levels of dough strength is a key objective in the development of HRSW cultivars. In commercial settings, dough strength is often measured using Farinograph stability. Due to resource constraints within breeding programs, flour quality is often measured by other methods, such as the Mixograph and Glutomatic. The objective of this research was to determine whether data from the Mixograph and Glutomatic could be used to properly predict Farinograph stability. Farinograph and Mixograph data spanning 6 years and 7 locations were used for the analysis. The Farinograph stability data provides a check for the accuracy of the predictive models, which were developed with the Mixograph and Glutomatic data. Two regression-based models were ran, using parameters obtained from the Mixograph and Glutomatic tests: Random Forest, and XGBoost. The Random Forest model had an average of 162 seconds difference between a predicted time and an observed time, while the XG Boost model had an average of 196 seconds difference between predicted and observed. Observations with low stability times often had predictions greater than their true times, while those with high stability times frequently had predictions which were less than their true times. Although this was the case, selecting materials in a breeding program with predictions of greater than 1200 seconds would most often lead to satisfactory results.

Chapter 1: Literature Review

1.1 Introduction

Common wheat (*Triticum aestivum L.*) is one of the 'big three' cereal crops, with over 600 million tons being harvest annually. Wheat also has an unrivalled range of cultivation, from 67º N in Scandinavia and Russia to 45º S in Argentina, while also growing at higher elevations in the tropical and sub-tropical regions (Shewry, 2009). Also referred to as 'Bread Wheat', common wheat is a hexaploid species (AABBDD) which allows its high genomic plasticity and capacity for broad adaptation. As a result, nearly 95% of global wheat production is of the hexaploid variety, with the remaining 5% being contributed mostly to durum wheat (Mastrangelo & Cattavelli , 2021). Durum wheat is primarily used in the production of pastas. Grain from common wheat is used both for animal feed and production of products for human consumption, which are derived from its flour. The most obvious of these is bread, which has religious significance in Judaism, Christianity, and the Islamic faiths (Shewry, 2009).

Wheat, produced in nearly every part of the United States, is the third largest U.S. crop in terms of both value and acreage, behind corn and soybeans (Vocke & Ali, 2014). Unlike most other crops, however, wheat has distinct varieties that are produced in different regions or over different seasons. The result is wide variation in the costs of wheat production across growing areas, inherent in the diversity of inputs and production practices. These costs can affect the competitiveness of wheat with other crops in each region and the profitability of planting wheat (Vocke & Ali, 2014).

1.2 Origin of Domesticated Wheat

The hexaploid wheat (AABBDD) contains three different genomes each derived from different diploid species *viz., Triticum urartu* (AA genome), *Aegilops speltoides* (controversial, BB genome), and *Aegilops tauschii* (DD genome) (Feldman et al., 1995). According to the archeological records, wheat originated in Southeast Turkey. Initially, the progenitor species containing AA and BB subgenomes were discovered (Aaronsohn, 1910) and these were hybridized followed by a doubling of chromosomes which resulted in tetraploid fertile wheat, *T. turgidum* (AABB) (Gohar et al., 2022). Then the *T. turgidum*, wild emmer, was domesticated in the Fertile Crescent. Afterward, *T. turgidum* hybridized with a diploid species A. *tauschii* which resulted in the formation of hexaploid wheat (AABBDD; McFadden, 1944). Wheat has two copies of each of its three genomes, with seven chromosomes belonging to each respective genome, bringing the total chromosome number to 42.

Initially, wheat was spread to Greece, Cyprus, India, and Egypt followed by other countries around the world (Cooper, 2015). The wheat grain in primitively cultivated species was long, thin, and small in size. The first naturally mutated traits were non-brittle rachis and naked grain that were responsible for the domestication of wheat (Pourkheirandish et al., 2018). Selections made by early farmers were on the basis of phenotypic traits such as grain size, color, and non-shattering spikes (Eckardt, 2010).

Before using cereals to make bread, they were used to make porridges which is believed to be the first form of cereals being used as a human food source (Beldrok et al., 2000). Sumerians were the first to bake unleavened bread sometime around 6000 B.C. (Beldrok et al., 2000). The Egyptians were the ones who began using yeast they created

from brewing beer in their bread around 3000 B.C. as well as developing a bread oven that could bake multiple loaves at one time (Beldrok et al., 2000).

1.3 Wheat Classification

Wheat production in the United States is largely made up of winter & spring wheat (hexaploid), and durum wheat (tetraploid). Winter wheat production represents approximately 70 percent of total U.S. wheat production, on average. Spring wheat typically constitutes about 25 percent of total U.S. wheat production, and durum wheat makes up the remainder (Bond & Liefert, 2019). Regional climatic differences across the United States account for much of the variation in the class of wheat grown, each with its own production practices and associated costs (Vocke & Ali, 2014). Northern wheat producers, for example, generally chose to plant spring wheat varieties that are harvested in the fall because winter wheat, which is planted in the fall for summer harvest, can often be killed by the cold during winter dormancy (Vocke & Ali, 2014).

Hard red spring wheat (HRSW) and hard red winter wheat (HRWW) are both known for producing high protein levels in their flour, which can be blended with lower protein wheat to produce loaf bread as well as specialty breads (Vocke & Ali, 2014). Soft red winter wheat (SRWW) is used in the production of cakes, cookies, and crackers (Vocke & Ali, 2014). White wheat is used for noodles, crackers, cereals, cookies, white crusted bread, and other wheat products which do not require the high-protein flour provided by the hard red wheats (Bond & Leifert, 2019). Durum wheat is used in the production of pasta and represents 3-5% of domestic wheat production (Bond & Leifert, 2019; Vocke & Ali, 2014).

1.4 Wheat Grain Quality

Wheat quality usually refers to the processing quality, which is mainly dependent on the content characteristics of storage proteins in wheat grains, and directly determines market price and end-use value (Peng et al. 2022). Wheat processing quality is represented by the physical and chemical characteristics of the dough, which make it possible to process wheat into a variety of food products, some of which were mentioned previously (Peng et al., 2022). Dough properties are mainly determined by the gluten proteins, glutenin, and gliadin (Peng et al., 2022). Glutenins can be subdivided into high molecular weight glutenin subunits (HMW-GS) and low molecular weight glutenin subunits (LMW-GS) (Shewry et al., 2002).

HMW-GS is the main factor determining gluten elasticity, which is encoded by *Glu-1* genes which are located on chromosome 1 of each of the A, B, and D genomes (Peng et al., 2022). Dough with strong gluten, such as those produced by HRSW and HRWW varieties, has high ductility resistance and can maintain stability over comparatively long periods of mixing (Peng et al., 2022). The dough retains gas produced during fermentation in discrete cells evenly distributed in the dough, whereas a dough with lower gluten strength can cause the excessive expansion of gas cells during baking, resulting in the collapse of cell walls and aggregation of cells, resulting in a rough bread texture (Don et al., 2006). This illustrates the importance of strong gluten in the goals of wheat breeding programs.

1.5 Wheat Flour

Protein content in wheat grains typically ranges from 10-18%, and protein content is generally considered to be positively correlated with wheat processing quality, particularly dough strength (Peng et al., 2022). From a structural standpoint, wheat grains can be divided into embryo, endosperm, and seed coat. The desirable portion involved in flour production is the endosperm, which contains starch and protein (Zhang, 2020). When the wheat is milled into flour, the seed coat and embryo portions of the grains are removed as much as possible and only the endosperm is transformed into flour (Zhang, 2020). Therefore, the main components of wheat flour are starch and protein. As a result, the quality components of flour are largely determined by the quality of starch and protein (Zhang, 2020).

1.5.1 Starch

Wheat flour consists mostly of starch (63-72%) and is present in the endosperm. It consists of the glucose polymers, amylose, and amylopectin (Van Der Borght et al., 2005). Amylose is essentially linear in structure, while amylopectin is highly branched and consists of chains. Typical levels of amylose and amylopectin are 25-28% and 72-75%, respectively (Van Der Borght et al., 2005).

When wheat flour is used to make bread, steamed buns, and steamed bread, yeast is added to make the dough expand to improve the taste quality. During milling a small, but significant proportion of starch granules in flour are physically damaged. The level of starch damage varies with the severity of grinding and the hardness of the wheat (Van der Borght et al., 2005). In the process of dough fermentation, the damaged starch in wheat

flour is more likely to be broken down into different sugars by yeast and further improve the quantity of yeast present. This also results in a production of carbon dioxide, which will be confined by the swelling starch and gluten after water absorption and finally make the dough expand (Zhang, 2020). The “bubbles” seen in a slice of bread are a result of the trapped carbon dioxide, and the end-product should possess a fluffy texture and taste.

1.5.2 Protein

Protein is the second most important component in wheat flour, and like starch it mainly exists in the endosperm of wheat grains (Zhang, 2020). Wheat grain contains about 12% proteins which can be divided into two main groups: gluten and non-gluten proteins.

Non gluten proteins (15-20%) consist of albumins which are soluble in water, and globulins, which are insoluble in water but soluble in dilute salt solutions. Gluten proteins, the main storage proteins in wheat, represent 80-85% of total wheat protein, and can be divided into two groups, gliadins and glutenins (Van Der Borght et al., 2005). Upon hydration and mixing, they form a strong, cohesive viscoelastic network that allows the wheat flour dough to retain the gases produced by yeast, referenced above (Van Der Borgh et al., 2005). In the process of dough kneading, the gluten proteins repeat the process of polymerization and depolymerization, covalent bonding in the dough increases, and the gluten network gradually becomes strong. However, if dough kneading is too long, the number of covalent bond breaks will increase, and the dough quality declines.

During mixing, gliadins behave as a viscous liquid, which imparts extensibility and cohesiveness to the dough. Flours with high gliadin content will produce a weak, sticky, inelastic dough (Wrigley et al., 2006). The glutenins present in wheat dough are responsible for the dough being rubbery and elastic (Barak et al., 2014). Flours containing a high glutenin content are strong, tough, elastic, and have non-adhesive gluten proteins (Wrigley et al., 2006).

Glutenins have a strong negative correlation with peak viscosity, breakdown viscosity, and pasting temperature while gliadins have a positive correlation with breakdown viscosity, setback, and final viscosity (Barak et al., 2014). Therefore, a correct balance of viscoelastic properties is crucial in quality aspects. If dough is too viscous, it will not maintain its desired final shape. If the dough lacks in elastic properties, it is difficult to form into the desired shape of the final product (Barak et al., 2014).

1.5.3 Non-Starch Polysaccharides

Wheat contains polysaccharides other than starch. non-starch polysaccharides. (NSP) are present in the cell walls of the endosperm and bran tissues. These NSPs are composed of arabinoxylans, B-glucans, cellulose and arabinogalactan-peptides (Van Der Borght et al., 2005). Wheat arabinogalactan-peptides (AGP) consist of large polysaccharide moieties (92%-94%) covalently linked to a 15 amino acid peptide (6-8%) (Van Der Borght et al., 2005).Arabinoxylans represent only about 2% of flour weight (Garofalo et al., 2011). Arabinoxylan amount, structure, and physiochemical properties vary widely among wheat varieties, including its molecular weight, distribution, branching pattern, water extractability, and interactions with other cell wall components such as lignin or cellulose (Garofalo et al., 2011). Arabinoxylans can further be

characterized by their water extractability, being either extractable or unextractable. Non-starch polysaccharides are also closely related to gluten forming proteins and have effects on their properties, thereby affecting the functional and dough rheological properties and by extension, bread quality (Garogalo et al., 2011).

Water extractable arabinoxylans (WE-AX) yield highly viscous solutions, while water unextractable arabinoxlyans (WU-AX) have a strong tendency to absorb water and swell (Van Der Borght et al., 2005). Endogenous arabinoxylans have a negative effect on dough handling properties, which may result from either the immobilization of some water necessary for complete hydration of gluten proteins, interference of interactions between arabinoxylans and gluten proteins, or both (Van Der Borght et al., 2005).

1.5.4 Lipids

Although lipids are a minor component in wheat flour (2-2.5%) they are considered to have significant impacts on flour and dough functionality by interacting with gluten proteins and starch, thereby stabilizing gas bubble cells in breadmaking (Gonzalez-Thuillier et al., 2015). Lipids in wheat grains display large structural diversity and comprise neutral and polar components.

Neutral components are comprised of acylglycerols and free fatty acids, while polar components are constituted of glycolipids and phospholipids (Gonzalez-Thuillier et al., 2015). Triglycerides, which are neutral lipids, are the most abundant lipids present in wheat flour accounting for approximately 95% of the total lipids. Phospholids, which are a vital component of cell membranes and maintain cellular integrity, represent 3-4% of the lipid profile in wheat grains (Ferrer et al., 2015).

1.5.5 Byproducts

During the milling process, byproducts of wheat flour are produced. These include bran, shorts, middlings, and germ. Bran is the outer layer of the wheat kernel that contains high levels of fiber, as well as vitamins and minerals. Wheat bran contains about 70% of fibers in wheat grains. This fiber is composed of insoluble fibers such as cellulose and lignin, as well as soluble fibers such as pectin and beta-glucan (FAO, 2019).

Shorts and middlings are the intermediate products between the bran and the flour. Shorts and middlings contain high amounts of protein, vitamins, and minerals such as iron, zinc, and magnesium. Due to their high nutritional value, shorts and middlings are often used as a dietary supplement in animal feed (FAO, 2019).

Germ is the embryonic part of the wheat kernel that contains high levels of vitamins, minerals, and unsaturated fatty acids. Wheat germ is also a rich source of vitamin E, which is an antioxidant that plays a crucial role in protecting cells from oxidative damage. Germ is commonly used in the production of health supplements and is also used as a flavoring agent in some food products (Wei et al., 2015).

1.6 Wheat Flour Quality Measurements

Several methods of analysis for measuring the dough quality of wheat flour exist. The following sections will cover three: Farinograph, Mixograph, and Glutomatic. Mixograph and Glutomatic analyses are conducted within the South Dakota State University HRSW breeding program as a means of determining quality potential of early-generation populations and breeding lines , while the Farinograph analyses are conducted

on more advanced breeding lines at the Northern Crops Institute at North Dakota State University in Fargo, North Dakota.

1.6.1 Farinograph

The Brabender Farinograph is an instrument used to analyze rheological qualities of wheat flour during mixing and the creation of bread dough. Flour samples are placed into a bowl (50g for small bowl, 300g for large bowl procedures) then water is added. Once water is added, the bowl rotates around a mixer while a computer continually measures resistance of the dough against the mixer throughout the duration of the test. This resistance is measured in Brabender units, which are plotted on a graph called the farinogram. Results of Farinograph analyses are widely used to predict functionality of flour and determine its suitability for end-use consumers such as millers and producers of baked goods.

The Farinograph measures several different rheological behaviors of wheat flour dough: arrival time, peak time, mixing tolerance index, departure time, stability, and water absorption capacity. Arrival time can be defined as the amount of time between water uptake and the time it takes the Farinograph curve to reach 500 Brabender units (BU) (approved method 54-21.02; AACC, 2011).

Peak time is indicative of elapsed time between water uptake and maximum dough strength, which is indicated as the highest point of the curve on the farinogram. Peak time can be interpreted as the optimized amount of mixing time to produce a flour with high structural integrity (Reese et al., 2007).

Mixing tolerance index (MTI) is the ability of the dough to recover its structure over time during the mixing process. MTI is measured by recording the resistance of the dough to mixing after it has been subjected to an external force. A dough with good mixing tolerance is able to recover its structure and maintain its quality during baking, or inversely, the dough will break down and soften (Diosi et al., 2015).

Departure time occurs when the top of the curve falls below 500 BU on the farinogram (Reese et al., 2007). This is the point in the mixing process in which the dough's structural integrity starts to deteriorate, commonly perceived as over-mixing. From here the baking quality of the dough steadily worsens.

Stability is arrival time minus departure time, or the amount of time in which the curve of the farinogram resides above the 500 BU line (Reese et al., 2007). Stability is widely used as an indicator of overall flour protein quality and is of particular interest to end-use consumers.

The Farinograph procedure, as previously stated, consists of either large bowl (300g) or small bowl (50g) flour samples. The following is the American Association of Cereal Chemists (AACC) approved operating procedure for the Farinograph analysis (approved method 54-21.02; AACC, 2011):

Procedure

1) Adjust the Farinograph thermostat, a temperature of 30 ± 0.2 ° needs to be maintained. Check the temperature of the circulating water, check that the water is circulating freely through the hose and bowl jackets, and confirm that the flow pattern matches the equipment manual.

2) Adjust the position of the base plate to be horizontal, then fix the four foot-screws with their locknuts.

3) Check that the chart paper is exactly horizontal. Two small plates on spring-loaded hinges are the guides for the paper and can be adjusted.

4) To clean, at the end of each test, while the machine is running, add dry flour ro the bowl to make a dough with a consistence of 800-900BU within 1 minute of mixing with the test dough. Stop the machine, unscrew the bowl, discard the dough, and scrape the bowl with a plastic spatula. Clean the bowl with a damp cloth and wipe completely dry.

Constant flour weight procedure for large and small bowls

1) Sensitivity: There are four sensitivity settings, two choices of position linkage between balance levels, and two choices of weights (400 and 1000). Chose the correct sensitivity for the bowl size.

2) Zero position of the scalehead pointer: Adjust the scalehead pointer to the zero position when the instrument is running at 63±2 rmp with the mixer empty.

3) Adjustment of bandwidth: The damping device should be adjusted, after the oil in the damping chamber has been at temperature for 1 hour or more, and after the damping piston has been moved up and down several times. Raise the dynamometer until the scalehead pointer indicates 1,000BU. Measure the amount of time it takes for it to go from 1,000BU to 100 BU (should be 1±0.2s).

Large bowl procedure

1) Turn the thermostat and circulating pump on 1 hour before use
2) Determine the flour moisture content and keep the flour in moisture proof containers
3) Place 300 ±0.1g of flour (14% moisture basis) in the bowl
4) Fill the large burette with room temperature water, making sure the tip is full, and that the automatic zero adjustment is functioning.
5) Set the pin-point to 9 minutes, turn the machine on to the 63rpm setting and then run for 1 minute until the zero minute line is reach. Then begin to add water to the right front corner of the bowl from the burette to the expected absorption of the flour. When the dough begins to form, scrape the sides of the bowl with a plastic scraper, working counterclockwise. Cover the bowl with the plexiglass cover to prevent any evaporation. If the mixing curve will be higher than 500BU add more water, this will be used to estimate the absorption for the next attempt.
6) The first titration rarely has a curve with the maximum resistance centered at 500BU, in the next titration adjust the absorption up or down until it is within 20 of 500BU.
7) In the final titration, all the water within 25s of opening the burette. Let the instrument run until an adequate curve is produced for evaluation. Then, lift the pen from the paper, and clean the bowl.

8) Report the absorption values to the nearest 0.1% and calculate the absorption on a 14% moisture basis using the following equation:

$$Absorption\% = \frac{(x + y - 300)}{3}$$

X=mL of water needed to produce a curve with maximum consistency centered on the 500BY line

Y= g of flour used

Small bowl procedure

The same method is used except that 50±0.1g of flour is added. Titration is conducted with a small burette instead of a large one. The absorption rate is calculated with the following equation:

$$\text{Absorption}\% = 2(\text{x} + \text{y} - 50)$$

X= mL of water needed to produce a curve with maximum consistency centered on the 500BY line

Y= g of flour used

Interpretation of the Farinogram is derived from the Farinograph curves, an example of which is shown in (Figure 1).

1.6.2 Mixograph

In a similar fashion to the Farinograph, the Mixograph records dough and gluten properties of wheat flour by measuring the resistance of a dough to mixing (Wheat Marketing Center Inc, 2004). Mixograph analysis measurements include, but are not

limited to, water absorption, peak time, peak width, peak value, and peak right value. The Mixograph curve provides information related to gluten strength, optimum dough development time, and mixing tolerance (Wheat Marketing Center Inc, 2004).

The peak time illustrates the dough development time which begins when the recorder is started and ends when the dough has reached its maximum consistency (Wheat Marketing Center Inc, 2004). The Mixograph mixing tolerance is the resistance of a dough to breaking down during continues mixing, this measurement is expressed as a score relative to a control (Wheat Marketing Center Inc, 2004). Weak gluten flours have shorter peak times and less of a mixing tolerance than in strong gluten flours (Wheat Marketing Center Inc, 2004). While the Mixograph exhibits similar qualities in the nature of its measurements, it should be noted that it does not provide a direct measurement for dough stability, as the Farinograph does. This is a crucial detail pertaining to the objective of this research.

The following is the AACC approved operating procedure for the Mixograph Analysis (approved method 54.40.02; AACC, 1999):

1) The moisture content of the flour should be determined, then weigh the flour samples (10 or 35g on a 14% moisture basis) to 0.01g. The flour should be kept in moisture proof containers.

2) Room temperature needs to be maintained at 25±1° for 24 hours a day. The equipment, flour, and water should be at room temperature. The mixing bowl can be soaked with water between samples but should be dried before the next use.

3) After long idle periods, two or three mixograms of standard flour should precede the other recordings.

4) Transfer weighed flour to the dry mixograph bowl, this can be aided with a camel-hair brush.

5) With a tongue depressor or spatula, move the flower between two bowl pins to create a triangular shaped hole in the middle.

6) Before starting the mixogram be sure the ink is running freely from the pen

7) The mixogram should be started on a major arc and run for a fixed time (typically 8-10 minutes).

8) Place the bowl in position on the Mixograph, dispense the water from an automatic pipet, lower the mixing head, and start recording the mixogram.

The dough absorption is calculated using the following equation (14% moisture basis):

$$y = 1.5x + 43.6$$

X= percent of flour protein content

Y= Percent absorption of water

1.6.3 **Glutomatic**

The Glutomatic analysis test provides information on the quantity and estimates the quality of gluten in wheat or flour samples. Gluten is responsible for the elasticity and extensibility characteristics of flour dough. Wet gluten reflects protein content and is a common flour specification required by end-users in the food industry (Wheat Marketing Center, 2004). Wet gluten is produced by washing the wheat flour in a salt solution, which removes starch and other solubles, leaving only the saturated gluten (Wheat Marketing Center, 2004). The wet gluten is then centrifuged and forced through a sieve, with the percentage of gluten left on the sieve being measured as Gluten Index (GI). GI is an indicator of the strength of the gluten within the flour sample (Wheat Marketing Center, 2004).

The following is the AACC approved operating procedure for the Glutomatic analysis (approved method 54-40.02; AACC, 2000)

1) Place the 88-μm polyester screen in the washing chamber. On top of the screen place the plastic chamber wall with the cylindrical insertion tool inside. Wash from top to bottom to remove any leftover debris.

2) Add wash liquid to the washing chamber to wet the polyester screen. Hit the screen three times on your hand covered with a cloth to remove excess water. Add 10±0.01g of well mixed flour onto the screen that contains a film of liquid to prevent the flour from falling through.

3) Add 4.8ml of wash solution from a dispenser while holding the chamber at about a 30° angle. Shake the chamber gently in circular motions to spread the liquid over the sample.

4) Assemble the washing chamber onto the Glutomatic and start it for a 20s dough mixing and 5min gluten washing cycle. The wash liquid flow rate should be 5-56ml/min.

5) At the end of the cycle, remove the gluten from the chamber without tearing to place it in a centrifuge.

Wet gluten content and gluten index

1) Place the wet gluten from wash chamber into a separate gluten index cassette in a centrifuge.

2) Centrifuge for 30sec at 6000±5rpm for 1min.

3) Remove the gluten from the cassette. With a spatula, remove the gluten that has passed through the sieve. Weigh the gluten to the nearest 0.01g. Leave the gluten on the scale.

4) With tweezers remove the gluten that is remaining on the top of the sieve and weigh for the total wet gluten.

Dry gluten content and water binding in wet gluten

1) Take the total amount of wet gluten, place it in the center of a lower heating surface or a dryer.

2) Close the dryer and dry at 150° for 4min.

3) With tweezers, remove the dry gluten and weigh to the nearest 0.01g.

The calculations for total wet gluten, gluten index, dry gluten, and water binding in wet gluten are as follows:

$$Wet\ gluten\ content\%\ (14\%\ moisture\ basis) = \frac{total\ wet\ gluten(g) * 860}{100 - \%sample\ moisture}$$

$$Gluten\ index = \frac{wet\ gluten\ remaining\ on\ sieve(g) * 100}{total\ wet\ gluten(g)}$$

$$Dry\ gluten\ content\%\ (14\%\ moisture\ basis) = \frac{total\ dry\ gluten(g) * 860}{100 - \%sample\ moisture}$$

$$Water\ binding\ capacity\ (water\ bound\ in\ wet\ gluten)\% = wet\ gluten\ content\% - dry\ gluten\ content\%$$

1.7 Machine Learning

Machine Learning and data analytics are interdependent and interrelated fields of study that focus on deriving decisive insights. Machine learning models are used to learn patterns in data in either one of two ways, supervised or unsupervised learning (Doshi & Varghese, 2022). In a supervised learning model, the data analyst has access to a set of predicting features, measured on many observations, as well as access to the measured response variable for these observations. The goal is then to predict the response using the set of predictors and observations (James et al., 2021). In unsupervised learning, we do not have access to the response variable, and only have access to the predictors and the observations.

The models used in this research were Random Forest and a boosted tree model called XG Boost. Random Forest and XG Boost are both classified as decision tree models, which are simple and useful tools for interpretation, and can be used in either a

regression or a classification setting (James et al., 2021). The models were trained using supervised learning.

1.7.1 Random Forest

Random Forest provides an improvement over other decision tree models by decorrelating the trees (James et al., 2021). For example, when bagging models build their decision trees, a random sample of the predictors is chosen each time a split in the tree is considered. In Random Forest models, a random sample of *m* predictors is chosen as split candidates from the full set of *p* predictors. The split in the tree is only allowed to use one of those *m* predictors (James et al., 2021). A fresh set of *m* predictors is taken at each split in the tree, and usually $m = \sqrt{p}$, that is, the number of predictors used at a split in the tree is approximately equal to the square root of the total number of predictors in the model (James et al., 2021).

1.7.2 XG Boost

Boosted trees are grown sequentially, meaning each tree is grown using information form previously grown trees, thus each tree is fit on a modified version of the original data set. Boosting models learn slowly, and avoid the issue of using a single large decision tree which may result in overfitting (James et al., 2021). Essentially, Boosting models train many weak models, with each iteration using the residual errors of the pevious models to fit the next model. The final prediciton is a weighted sum of all of the tree predictions. In XG Boost, trees are built in parallel instead of being built sequentially.

1.8 Environment

The quality of wheat depends not only on its genetic potential for particular quality characteristics, but also on its ability to realize this potential in actual production and under different environmental conditions (Horvat et al., 2012). Wheat quality properties usually are influenced by the interaction of genotype and environment, however, the magnitude of the interaction effects are often smaller compared with genotype and environment main effects (Horvat et al., 2012). However, both genotype and environment, and their interacting effects, have an influence on the relationship of flour protein composition to loaf volume (Guttieri et al., 2000).

Temperature and fertilization contribute to environmental effects on end-use quality. Temperature effects, particularly high temperatures during grain filling, can significanly elevate protein content while lowering the performative functionaly of grain protein, ultimately changing the rheological properties of flour (Guttieri et al., 2000). Nitrogen fertilization also strongly influences the quantity of protein in wheat flour, but the magnitude of the influence of increased fertilization varies between cultivars (Guttieri et al., 2000). Other factors, such as amount of precipitation, distribution of that precipitation, and the duration of grain fill also have an influence on the end-use quality of wheat flour (Guttieri et al., 2000).

1.9 Breeding for Quality Improvement

Early selection processes in wheat domestication likely had little to do with quality aspects pertaining to milling, but more likely selection was based upon the tolerance of individual plants to the biotic and abiotic stressors of the time and region (Kiszonas & Morris, 2017). This selection was entirely visual and assessed plants' tolerance to disease, drought, shattering resistance, etc.

The basis for more modern methods of crop improvement mostly developed starting in the 18th century and continuing through the 19th and 20th centuries. In the 18th century, Joseph Goottlieb Kölreuter established proof of heredity in plant species (Kiszonas & Morris, 2017). The 18th century saw the development of Darwin's "Origin of Species" and the concepts of natural selection, as well as Gregor Mendel's work establishing the Laws of Heredity (Kiszonas & Morris, 2017). Looking at wheat in particular, Thomas Andrew Knight is credited as the first individual known to have made a cross with wheat. In 1859, Louis de Vilmorin advocated the selection of individual plants – a technique now known as pure line selection – to ensure genetic purity (Kiszonas & Morris, 2017).

These discoveries and practices had a large impact on many areas, including plant breeding, but did not directly address the need to improve the end-use quality of wheat lines. The McDougall Brothers of London examined that the baking quality of wheats from around the world had a direct relationship to the gluten content of the flour produced by a given wheat line (Richardson, 1884). Hybridization of wheat lines, advances in flour milling, industrial mechanization as well as construction of transportation infrastructure all had a profound influence on wheat improvement in the early 20th century (Kiszonas & Morris, 2017).

In 1907, Thomas Burr Osborne established four major delineations of wheat proteins based on solubility: albumens (water soluble), globulins(10% salt soluble), prolamin (gliadins), and glutenin (dilute acid or alkalai) (Osborne, 1907). The ability to separate gluten into gliadin and glutnenin created a new avenue of selecting for gluten quality (Kiszonas & Morris, 2017).

In 1959, early work with gluten began with moving-boundary electrophoresis, which exposed high levels of variation across wheat varieties with regards to both glutenins and gliadins (Wrigley & Bietz, 1988). Advancements in electrophoresis methodology, as well as methods such as liquid chromatography, mass spectrometry, and the enzyme-linked immunosorbent assay (ELISA) provided further insights for the analysis of gluten & end-use quality (Kiszonas & Morris, 2017).

In 1977, Sanger et al. developed the 'Sanger Method' of DNA sequencing (Sanger et al., 1977). This layed the ground work for many new breeding capabalities, especially once it was paired with the polymerase chain reaction (PCR) (Mullis et al., 1987). PCR systems allowed for the detection of quantitative trait loci (QTLs) via random amplification of polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), simple sequence repeats (SSR) and single nucleotide polymorphisms (SNPs) (Kiszonas & Morris, 2017). The QTL's detected through these early marker systems allowed for the selection of wheat lines which had QTL's which the breeder might deem as beneficial or productive. The continual use and improvement of genotyping and marker technologies have resulted in even greater breeding tools, such as genotype-by-sequencing (GBS) and genome-wide association studies (GWAS), which can be quite expensive but typically provides breeders with a great deal of information in regards to quality (Kiszonas & Morris, 2017).

References

AACC Approved Methods of Analysis, 11th Ed. Methods 38-12.02, 54-21.02 & 54-40.02. Approved Methods of Flour Quality Analyses. Cereals & Grains Association, St. Paul, MN, U.S.A. (2011).

Aaronsohn, A. (1910). Agricultural and botanical explorations in Palestine. Washington, DC: US Government Printing Office.

Barak, S., Mudgil, D., & Khatkar, B.S. (2014). Influence of Gliadin and Glutenin Fractions on Rheological, Pasting, and Textural Properties of Dough. *International Journal of Food Properties*, 17:1428–1438, 2014

Beldrok, B., Mesdag, J., & Dinner, D. A. (2000). Bread-making quality of wheat: a century of breeding in Europe. *Springer Science & Business Media.*

Bond, J. K., & Liefert, O. (2019). Wheat sector at a glance. Usda-Ers. https://www.ers.usda.gov/topics/crops/wheat/wheat-sector-at-a-glance/#classes

Cooper, R. (2015). Re-discovering ancient wheat varieties as functional foods. *J. Tradit. Complement. Med.* 5, 138–143. doi:10.1016/j.jtcme.2015.02.004

Diósi, G., Móré, M., & Sipos, P. (2015). Role of the farinograph test in the wheat flour quality determination. Acta Universitatis Sapientiae, Alimentaria, 8(1), 104–110. doi:10.1515/ausal-2015-0010

Don, C., Mann, G., Bekes, F., & Hamer, R.J. (2006). HMW-GS affect the properties of glutenin particles in GMP and thus flour quality. *J. Cereal Sci.* 44, 127–136. doi: 10.1016/j.jcs.2006.02.005

Doshi, M., Varghese, A. (2022) Smart Agriculture Using Renewable Energy and AI-powered IoT. AI, Edge and IoT Smart Agriculture.

Eckardt, N. A. (2010). Evolution of domesticated bread wheat. American Society of Plant Biologists.

Feldman, M., Lupton, F., & Miller, T. (1995). "Wheats," in Evolution of Crop Plants. Editors J. Smartt, and N. w. Simmonds. 2nd (London: Longman Scientific), 184–192.

Ferrer, E., Alegria, A., Farre, R., Abian, J.,, Rossello, C., & Pons, A. (2015) Lipid composition of wheat flours with different breadmaking potential. Food Chemistry, 172, 300-306.

Food and Agriculture Organization of the United Nations (FAO). (2019). Small scale wheat milling.

Garófalo, L., Vazquez, D., Ferreira, F., & Soule, S. (2011). Wheat flour non-starch polysaccharides and their effect on dough rheological properties. Industrial Crops and Products, 34(2), 1327–1331. doi:10.1016/j.indcrop.2010.12.003

Gohar, S., Sajjad, M., Zulfiqar, S., Liu, J., Wu, J., & Rahman, M., (2022). Domestication of newly evolved hexaploid wheat – a journey of wild grass to cultivated wheat. Frontier Genetics. doi:10.3389/fgene.2022.1022931

González-Thuillier, I., Salt, L., Chope, G., Penson, S., Skeggs, P., & Tosi, P., (2015). Distribution of Lipids in the Grain of Wheat (cv. Hereward) Determined

by Lipidomic Analysis of Milling and Pearling Fractions. Journal of Agricultural and Food Chemistry, 63(49), 10705–10716. doi:10.1021/acs.jafc.5b05289

Guttieri, M. J., Ahmad, R., Stark, J. C., & Souza, E. (2000). End-Use Quality of Six Hard Red Spring Wheat Cultivars at Different Irrigation Levels. Crop Science, 40(3), 631. doi:10.2135/cropsci2000.403631x

Horvat, D., Drezner, G., Dvojkovic, K., Simic, G., Magdic, D., & Spanic, V. (2012) End-Use Quality of Wheat Cultivars In Different Environments. Agricultural Institute of Osijek

James, G., Witten, D., Hastie, J., Tibshirani, R. (2021) An Introduction to Statistical Learning with Applications in R, 2nd Edition.

Kiszonas, A. M., & Morris, C. F. (2017). *Wheat Breeding for Quality: A Historical Review. Cereal Chemistry Journal.* doi:10.1094/cchem-05-17-0103-fi

Mastrangelo, A.M., & Cattivelli, L. What Makes Bread and Durum Wheat Different? Trends in Plant Science, 26(7), 677–684. doi:10.1016/j.tplants.2021.01.004

McFadden, E. (1944). The artificial synthesis of Triticum spelta. Rec. Genet. Soc. Am. 13, 26–27.

Mullis, K.B., Erlich, H.A., Arnheim, N., Horn, G.T., Saiki, R.K., & Scharf, S.J. (1987) Process for amplifying, detecting and/or cloning nucleic acid sequences. U.S. Patent 4,683,195. July 28, 1987.

Osborne, T.B. (1907) The Proteins of the Wheat Kernel. Carnegie Inst. Washington Publ.: Washington, D.C.

Peng, Y., Zhao, Y., Zitong, Y., Zeng, J., Xu, D., Dong, J., & Ma, W. (2022). Wheat Quality Formation and Its Regulatory Mechanism. Frontiers in Plant Science. Volume 13. https://doi.org/10.3389/fpls.2022.834654

Pourkheirandish, M., Dai, F., Sakuma, S., Kanamori, H., Distelfeld, A., Willcox, G., et al. (2018). On the origin of the non-brittle rachis trait of domesticated einkorn wheat. Frontier Plant Science. 8, 2031. doi:10.3389/fpls.2017.02031

Reese, C.L., Clay, D.E., Beck, D, & Englund, R. (2007) Is Protein Enough for Assessing Wheat Flour Quality? Western Nutrient Management Conference. Salt Lake City, UT 7. 85-90

Richardson, B. W. (1884). Our book shelf. Nature, 148.

Sanger, F., Nicklen, S., & Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. 74:5463-5467

Shewry, P.R., Halford, N.G., Belton, P.S., & Tatham, A.S. (2002). The structures and properties of gluten: an elastic protein from wheat grain. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 357, 133–142. doi: 10.1098/rstb.2001.1024

Shewry, P.R. (2009) Wheat. Journal of Experimental Botany, Vol. 60, No. 6, pp. 1537–1553, 2009 doi:10.1093/jxb/erp058

Van Der Borght, A., Goesaert, H., Veraverbeke, W. S., & Delcour, J. A. (2005). Fractionation of wheat and wheat flour into starch and gluten: Overview of the main processes and the factors involved. Journal of Cereal Science, 41(3), 221–237. https://doi.org/10.1016/j.jcs.2004.09.008

Vocke, Gary, & Mir Ali. U.S. Wheat Production Practices, Costs, and Yields: Variations Across Regions, EIB-116. U.S. Department of Agriculture, Economic Research Service, August 2013.

Wei, Y., Wang, Y., Wan, Z., Wang, T., Yang, L., & Chen, Z. (2015) Extraction, characterization, and utilization of wheat germ oil: A review. Journal of Food Science, 80(4), R893-R902.

Wheat Marketing Center Inc. (2004). Wheat and Flour Testing Methods: A Guide to Understanding Wheat and Flour Quality. In Wheat Marketing Center Ink, North American Export Grain Association.

Wrigley, C. W., Békés, F., & Bushuk, W. (2006). Chapter 1 Gluten: A Balance of Gliadin and Glutenin. In Gliadin and Glutenin: The Unique Balance of Wheat Quality (Issue March). https://doi.org/10.1094/9781891127519.002

Wrigley, C.W. & Bietz, J.A. 1988. Proteins and Amino Acids. Wheat, Chemistry & Technology. Y. Pomeranz, ed. AACCI: St. Paul, MN.

Zhang, Ang. (2020) Effect of wheat flour with different quality in the process of making flour products. *Int. J. Metrol. Qual. Eng.*, Volume 11, 2020 https://doi.org/10.1051/ijmqe/2020005

Chapter 2: Supervised Learning to Predict Farinograph Stability

SUPERVISED MACHINE LEARNING TO PREDICT FARINOGRAPH STABILITY OF SPRING WHEAT FLOUR

Hard red spring wheat (*Triticum aestivum* L., HRSW) flour is mostly used to produce wheat-based foods where dough strength is a major quality component. Maintaining adequate levels of dough strength is a key objective in the development of HRSW cultivars. In commercial settings, dough strength is often measured using Farinograph stability. Due to resource constraints within breeding programs, flour quality is often measured by other methods, such as the Mixograph and Glutomatic. The objective of this research was to determine whether data from the Mixograph and Glutomatic could be used to properly predict Farinograph stability. Farinograph and Mixograph data spanning 6 years and 7 locations were used for the analysis. The Farinograph stability data provides a check for the accuracy of the predictive models, which were developed with the Mixograph and Glutomatic data. Two regression-based models were ran, using parameters obtained from the Mixograph and Glutomatic tests: Random Forest, and XGBoost. The Random Forest model had an average of 162 seconds difference between a predicted time and an observed time, while the XG Boost model had an average of 196 seconds difference between predicted and observed. Observations with low stability times had predictions greater than their true times, while observations with high stability times often had predictions which were less than their true times. Although this was the case, selecting materials in a breeding program with predictions of greater than 1200 seconds would most often lead to satisfactory results.

2.1 Introduction

Common wheat (*Triticum aestivum L.*) is one of the 'big three' cereal crops, with over 600 million tons being harvested annually (Shewry, 2009). Wheat also has an unrivalled range of cultivation, from 67º N in Scandinavia and Russia to 45º S in Argentina, while also growing at higher elevations in the tropical and sub-tropical regions (Shewry, 2009). Also referred to as 'Bread Wheat', common wheat is a hexaploid species (AABBDD) which confers its high genomic plasticity and capacity for broad adaptation. As a result, nearly 95% of global wheat production is of the hexaploid variety, with the remaining 5% being contributed mostly to durum wheat (Mastrangelo & Cattavelli , 2021). Durum wheat is primarily used in the production of pastas. Grain from common wheat is used both for animal feed and production of products for human consumption, which are derived from its flour. The most obvious of these is bread, which has religious significance in Judaism, Christianity, and the Islamic faiths (Shewry, 2009).

Wheat, produced in nearly every part of the United States, is the third largest U.S. crop in terms of both value and acreage, behind corn and soybeans (Vocke & Ali, 2014). Unlike most other crops, however, wheat has distinct varieties that are produced in different regions or over different seasons. The result is wide variation in the costs of wheat production across growing areas, inherent in the diversity of inputs and production practices. These costs can affect the competitiveness of wheat with other crops in each region and the profitability of planting wheat (Vocke & Ali, 2014).

End-use quality is a major point of emphasis within breeding programs. End-use consumers want flour that will have high stability over a long period of mixing, if the flour does not have a favorable level of stability it will begin to breakdown prematurely

and the resulting quality of the baked good (often bread) will suffer as a result. Because of this, many different analyses have been employed to determine the quality of wheat flour prior to being used in a baking setting. Methods of analysis which are of particular interest to this research include the Brabender Farinograph (Brabender GmbH & Co. KG, South Hackensack, New Jersey) USA, the Mixograph (National Manufacturing Company, Lincoln, Nebraska, USA) and the Glutomatic (Perkin-Elmer, Inc., Waltham, MA, USA). Flour samples which attain a Farinograph stability time of 15 minutes (900 seconds) or greater are considered sufficient for end-use products. The goal of this research was to determine if machine learning models could be used to accurately predict the stability time of spring wheat research lines, providing breeders with an additional tool in their selective index.

2.2 Materials & Methods

2.2.1 Grain Production and Flour Sample Preparation

All Farinograph, Mixograph, and Glutomatic tests were carried out with flour samples prepared from grain produced in Advanced Yield Trial (AYT) field plots of the SDSU Spring Wheat Breeding Program. Each AYT was composed of 48 entries, 12 of which were checks. Each trial was conducted as a randomized complete block design composed of three replications. After plot harvest, samples to be used for flour analysis of each entry were obtained by creating composite samples through mixing grain from each rep into a single container. Grain samples were then tempered for 18 to 24 hours to obtain a moisture content of ~15%. After tempering, flour was produced by passing composite grain samples through a Quadromatic Junior grain mill (Brabender GmbH & Co. KG (South Hackensack, New Jersey).

2.2.2 Flour Quality Analyses

Flour protein and moisture content were determined with a Near-Infrared Reflectance (NIR) Systems 6500 Monochromators (Foss, Laurel, MD, USA). Each of these measurements were used to determine appropriate amounts of water to be added to each sample for Mixograph and Farinograph analyses. Glutomatic and Mixograph tests conducted as a part of this research were carried out on all 48 AYT samples produced at each of three selected locations each year according to AACC Approved Methods 38-12.02 and 54-40.02 (AACC, 2011) on the SDSU campus. Farinograph analyses (AACC Approved Method 54-21.02; AACC, 2011) were generally performed on a subset of 33 AYT flour samples that were previously subjected to Glutomatic and Mixograph tests.

An exception was that instead of a subset, all 48 flour samples from the three growing locations were subjected to Farinograph analysis in 2017. These tests were performed by personnel at the Northern Crops Institute in Fargo, ND. Upon completion, numerical data were provided for analysis after merging into a single dataset that also contained Glutomatic and Mixograph observations.

2.2.3 Data Analysis

2.2.3.1 Exploratory Data Analysis

All data analysis was performed in R Statistical Software (R-4.2.3; R Studio – 2023.03.0+386). The data used to train the statistical models was taken from years 2015-2020 (Table 1) across 7 growing locations (18 environments), where 95 unique genotypes were included. Some genotypes were included in all six growing seasons while some others were included in only a single year. In total, this resulted in 639 observations to be used in the training of the statistical models. The test set, which contained data from 2021, spans 3 locations (Table 1) and contains 99 observations from 33 separate genotypes. Although the Farinograph data was contained within the same data file as the Mixograph and Glutomatic data, it was not used for model training or testing. An exception, however, was that of Farinograph stability, expressed in seconds, as it was the response variable for the prediction models. Summary statistics of the data set are presented in Table 2.

Parameters MTW, GI, NIR Moisture, TW, and SKCS HI were found to be statistically significant when modeled with the entire dataset. A new linear model (Table

3) was then constructed to examine the effects of these parameters, with some previously significant parameters becoming statistically insignificant in the smaller model. Pairwise correlation plots of the recorded stability and predictive parameters can be found in Figure 1.

2.2.3.2 Model Building

Both the Random Forest and XG Boost models were built using a loop function that served as a cross validation. Random Forest and XG Boost were chosen due to the nature of the data set, in conjunction with how the models operate. Both models are also well known for their predictive performance, most notably XG Boost's performance in Kagle Competitions, a competition for predictive modeling.

Random Forest provides an improvement over other decision tree models by decorrelating the trees (James et al., 2021). For example, when bagging models build their decision trees, a random sample of the predictors is chosen each time a split in the tree is considered. In Random Forest models, a random sample of *m* predictors is chosen as split candidates from the full set of *p* predictors. The split in the tree is only allowed to use one of those *m* predictors (James et al., 2021). A fresh set of *m* predictors is taken at each split in the tree, and usually $m = \sqrt{p}$, that is, the number of predictors used at a split in the tree is approximately equal to the square root of the total number of predictors in the model (James et al., 2021).

Regarding XG Boost, oosted trees are grown sequentially, meaning each tree is grown using information form previously grown trees, thus each tree is fit on a modified version of the original data set. Boosting models learn slowly, and avoid the issue of

using a single large decision tree which may result in overfitting (James et al., 2021). Essentially, Boosting models train many weak models, with each iteration using the residual errors of the pevious models to fit the next model. The final prediciton is a weighted sum of all of the tree predictions. In XG Boost, trees are built in parallel instead of being built sequentially.

Five years of training data would be used to train an iteration of the model, which would then be used to predict the stability of the remaining training year's observations. This process was repeated until every year in the training set had received the "left out" treatment. Figures 2-7 show the training process of the Random Forest model as it is tested over each training year. Figures 8-13 show the training process of the XG Boost model as it is tested on the training years. Tables 4 & 5 show the RMSE of the training models, which are similar, although the RMSE for the XG Boost training model is higher in every year except 2019. Both models contain the same predictive parameters and response variable:

$$Stability = MTW + GI + NIR\ Moisture + TW + SKCS\ HI$$

2.3 Results

Both models exhibit similar visual results (Figures 14 & 15). The vertical line in Figures 14 & 15 represents 900 seconds, or the threshold for acceptable stability. The diagonal line represents a 1:1 slope and provides a visual bearing of predictive accuracy for both individual observations and the model as a whole.

We can effectively analyze 4 quadrants on the graphs when considering the intersection of the two slopes. The upper left quadrant represents observations which were recorded below 900 seconds, but their predicted stability was inflated by the model. The lower left quadrant represents observations whose stability were also recorded below 900 seconds, although these observations were understated in their prediction of stability. That is, the observed stability was higher than the predicted stability. The upper right quadrant represents observations whose stability was observed to be over 900 seconds, but the prediction of stability was higher than that which was actually observed. The final quadrant, the lower right, represents observations which were observed to have stability times above 900 seconds, but the predicted times were less than the times observed.

Both the Random Forest model and the XG Boost model place most observations in the lower right quadrant, meaning the predicted stability was less than what was actually observed. Interestingly, the inverse also appears to be true for both models. The upper left quadrant, which represents observations whose stability was less than 900 seconds but over stated by the models, has more observations placed within it when compared to the lower left quadrant. This is especially true for the Random Forest model, though the XG Boost model displays a more balanced distribution of predictions below 900 seconds.

Table 6 provides the RMSE of the test models, with the Random Forest model (12.733) displaying a lower RMSE than the XG Boost model (14.027). Tables 7 & 8 show summary statistics for the residuals of the Random Forest and XG Boost models, respectively. The mean represents the average difference between the observed stability time of the *i'th* observation and the predicted stability time of the *i'th* observation. The

mean of the Random Forest model residuals was 162.13, meaning we can expect about 2 minutes and 42 seconds of error when considering predicted vs observed stability for a random observation. The mean of the XG Boost model residuals was 196.77, meaning we can expect about 3 minutes and 17 seconds of error between predicted vs observed stability for a random observation.

2.4 Discussion

Dough stability has a direct impact on the end-use quality of baked goods. This should incentivize wheat breeders to make Farinograph stability a major focal point for consideration and selection within breeding programs. As important as this trait is to wheat growers and the entire wheat industry, it is difficult to measure directly in a breeding program where only small quantities of flour are available in early generations, where selection for stability could have its greatest impact.

The Farinograph stability measurement for a given flour sample is indicative of the dough's rheological characteristics over mixing-time and can be used as a selective measure by breeders. However, the end-use quality of wheat depends not only on its genetic potential for particular quality characteristics, but also on its ability to realize this potential in actual production and under different environmental conditions (Horvat et al., 2012). Wheat quality properties usually are influenced by the interaction of genotype and environment, however, the magnitude of the interaction effects are often smaller compared with genotype and environment main effects (Horvat et al., 2012). However, both genotype and environment, and their interacting effects, have an influence on the relationship of flour protein composition to loaf volume (Guttieri et al., 2000).

The models constructed in this research did not consider the year, genotype, or location as factors. The premise behind this was the hope to build a predictive model which could accurately predict dough stability time using only data collected from laboratory flour analyses. The Random Forest and XG Boost testing models displayed signals of accurate predictability, especially when examining the predictions on observations whose recorderd stability times were in the 700-1300 second range. However, the models did struggle with observations whose recorded stability times were toward the lower and higher end of the dataset's range. The average discrepancy between an observed stability time and a Random Forest prediction was over two and a half minutes, while the average discrepancy between observed stability and an XG Boost prediction was just over 3 minutes. Figures 2-7 and 8-13 show the variability that the year in which an observation is collected can have a major impact on the predictable nature of the obesrvation. It has also been shown that genetic potential has an impact on the rheological characteristics of wheat flour (Horvath et al., 2012). It is likely that the models would improve if conditions allowed them to consider other factors, such as the year, genotype, environment, and weather data; as well as any interaction effects between these factors. In their current state, the Random Forest model appears to be superior to the XG Boost model.

Given enough data and training years, it is possible that these models could improve such that the Farinograph analysis of more "mature" lines in a breeding program could be neglected, so long as there is Farinograph data from previous years. This would allow the breeder to use cheaper methods of flour quality analysis, while still performing a sort of "wellness check" on the stability of check lines or varieties nearing commericial

release. This would also free up time and financial resources for the younger heterogenous populations and newly-derived lines to be tested via the Farinograph analysis, providing the breeder with an additional tool in their selective index. These operative possibilites should incentivize further research into the predictablity of dough stability within research varieties.

2.5 Conclusion

As long as wheat grain is being used for the production of baked goods, flour stability will be a point of emphasis for wheat breeding programs. It has been previously shown that flour stability is dependent on genotype, environment, and the interactions between the two. The Farinograph is a dependable and industry-trusted method of measuring dough stability, but may be cost prohibitive for some breeding programs with consideration to time, labor, and amount of sample required (sample preperation requiring its own time and labor as well). Less costly methods of dough analysis such as the Mixograph and Glutomatic have been shown to provide data which can be useful for predicting dough stability (Schumate, 2020).

The Random Forest model had an average prediction error of roughly 2 minutes and 40 seconds, while the XG Boost model had an average prediction error of just over 3 minutes. However, these models were restricted in regards to their predictive parameters, and would likely improve given more data and model building parameters such as year, genotype, location, and environmental effects, as well as any interactions between them. The models shown in this research are quite accurate within the interquartile range of the observed stability times. While the models did struggle with observations outside of this range, it can still supply wheat breeding programs with additional insights and

opportunities, and should therefore incentivize further efforts of achieving a model with an even higher level of predictive accuracy.

References

AACC Approved Methods of Analysis, 11th Ed. Method 54-21.02. Rheological Behavior of Flour by Farinograph: Constant Flour Weight Procedure. Cereals & Grains Association, St. Paul, MN, U.S.A. (2011).

AACC Approved Methods of Analysis, 11th Ed. Method 54-40.02. Physical Dough Tests: Mixograph Method. Approved 1999. Cereals & Grains Association, St. Paul, MN, U.S.A. (2011).

AACC Approved Methods of Analysis, 11th Ed. Method 38-12.02. Gluten: Wet Gluten, Dry Gluten, Water-Binding Capacity, and Gluten Index Approved 2000. (AACC Approved Methods of Analysis, 11th Ed. Method 54-40.02. Physical Dough Tests: Mixograph Method. Cere. (2011).

Aaronsohn, A. (1910). Agricultural and botanical explorations in Palestine. Washington, DC: US Government Printing Office.

Barak, S., Mudgil, D., & Khatkar, B.S. (2014). Influence of Gliadin and Glutenin Fractions on Rheological, Pasting, and Textural Properties of Dough. *International Journal of Food Properties*, 17:1428–1438, 2014

Beldrok, B., Mesdag, J., & Dinner, D. A. (2000). Bread-making quality of wheat: a century of breeding in Europe. *Springer Science & Business Media.*

Bond, J. K., & Liefert, O. (2019). Wheat sector at a glance. Usda-Ers. https://www.ers.usda.gov/topics/crops/wheat/wheat-sector-at-a-glance/#classes

Cooper, R. (2015). Re-discovering ancient wheat varieties as functional foods. *J. Tradit. Complement. Med.* 5, 138–143. doi:10.1016/j.jtcme.2015.02.004

Diósi, G., Móré, M., & Sipos, P. (2015). Role of the farinograph test in the wheat flour quality determination. Acta Universitatis Sapientiae, Alimentaria, 8(1), 104–110. doi:10.1515/ausal-2015-0010

Don, C., Mann, G., Bekes, F., & Hamer, R.J. (2006). HMW-GS affect the properties of glutenin particles in GMP and thus flour quality. *J. Cereal Sci.* 44, 127–136. doi: 10.1016/j.jcs.2006.02.005

Doshi, M., Varghese, A. (2022) Smart Agriculture Using Renewable Energy and AI-powered IoT. AI, Edge and IoT Smart Agriculture.

Eckardt, N. A. (2010). Evolution of domesticated bread wheat. American Society of Plant Biologists.

Feldman, M., Lupton, F., & Miller, T. (1995). "Wheats," in Evolution of Crop Plants. Editors J. Smartt, and N. w. Simmonds. 2nd (London: Longman Scientific), 184–192.

Ferrer, E., Alegria, A., Farre, R., Abian, J.,, Rossello, C., & Pons, A. (2015) Lipid composition of wheat flours with different breadmaking potential. Food Chemistry, 172, 300-306.

Food and Agriculture Organization of the United Nations (FAO). (2019). Small scale wheat milling.

Garófalo, L., Vazquez, D., Ferreira, F., & Soule, S. (2011). Wheat flour non-starch polysaccharides and their effect on dough rheological properties. Industrial Crops and Products, 34(2), 1327–1331. doi:10.1016/j.indcrop.2010.12.003

Gohar, S., Sajjad, M., Zulfiqar, S., Liu, J., Wu, J., & Rahman, M., (2022). Domestication of newly evolved hexaploid wheat – a journey of wild grass to cultivated wheat. Frontier Genetics. doi:10.3389/fgene.2022.1022931

González-Thuillier, I., Salt, L., Chope, G., Penson, S., Skeggs, P., & Tosi, P., (2015). Distribution of Lipids in the Grain of Wheat (cv. Hereward) Determined by Lipidomic Analysis of Milling and Pearling Fractions. Journal of Agricultural and Food Chemistry, 63(49), 10705–10716. doi:10.1021/acs.jafc.5b05289

Guttieri, M. J., Ahmad, R., Stark, J. C., & Souza, E. (2000). End-Use Quality of Six Hard Red Spring Wheat Cultivars at Different Irrigation Levels. Crop Science, 40(3), 631. doi:10.2135/cropsci2000.403631x

Horvat, D., Drezner, G., Dvojkovic, K., Simic, G., Magdic, D., & Spanic, V. (2012) End-Use Quality of Wheat Cultivars In Different Environments. Agricultural Institute of Osijek

James, G., Witten, D., Hastie, J., Tibshirani, R. (2021) An Introduction to Statistical Learning with Applications in R, 2nd Edition.

Kiszonas, A. M., & Morris, C. F. (2017). *Wheat Breeding for Quality: A Historical Review. Cereal Chemistry Journal.* doi:10.1094/cchem-05-17-0103-fi

Mastrangelo, A.M., & Cattivelli, L. What Makes Bread and Durum Wheat Different? Trends in Plant Science, 26(7), 677–684. doi:10.1016/j.tplants.2021.01.004

McFadden, E. (1944). The artificial synthesis of Triticum spelta. Rec. Genet. Soc. Am. 13, 26–27.

Mullis, K.B., Erlich, H.A., Arnheim, N., Horn, G.T., Saiki, R.K., & Scharf, S.J. (1987) Process for amplifying, detecting and/or cloning nucleic acid sequences. U.S. Patent 4,683,195. July 28, 1987.

Osborne, T.B. (1907) The Proteins of the Wheat Kernel. Carnegie Inst. Washington Publ.: Washington, D.C.

Peng, Y., Zhao, Y., Zitong, Y., Zeng, J., Xu, D., Dong, J., & Ma, W. (2022). Wheat Quality Formation and Its Regulatory Mechanism. Frontiers in Plant Science. Volume 13. https://doi.org/10.3389/fpls.2022.834654

Pourkheirandish, M., Dai, F., Sakuma, S., Kanamori, H., Distelfeld, A., Willcox, G., et al. (2018). On the origin of the non-brittle rachis trait of domesticated einkorn wheat. Frontier Plant Science. 8, 2031. doi:10.3389/fpls.2017.02031

Reese, C.L., Clay, D.E., Beck, D, & Englund, R. (2007) Is Protein Enough for Assessing Wheat Flour Quality? Western Nutrient Management Conference. Salt Lake City, UT 7. 85-90

Richardson, B. W. (1884). Our book shelf. Nature, 148.

Sanger, F., Nicklen, S., & Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. 74:5463-5467

Shewry, P.R., Halford, N.G., Belton, P.S., & Tatham, A.S. (2002). The structures and properties of gluten: an elastic protein from wheat grain. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 357, 133–142. doi: 10.1098/rstb.2001.1024

Shewry, P.R. (2009) Wheat. Journal of Experimental Botany, Vol. 60, No. 6, pp. 1537–1553, 2009 doi:10.1093/jxb/erp058

Schumate, B. (2020) Predicting Farinograph Staibility of Wheat Flour with Mixograph and Glutomatic Tests. SDSU Spring Wheat Breeding Program, South Dakota State University

Van Der Borght, A., Goesaert, H., Veraverbeke, W. S., & Delcour, J. A. (2005). Fractionation of wheat and wheat flour into starch and gluten: Overview of the main processes and the factors involved. Journal of Cereal Science, 41(3), 221–237. https://doi.org/10.1016/j.jcs.2004.09.008

Vocke, Gary, & Mir Ali. U.S. Wheat Production Practices, Costs, and Yields: Variations Across Regions, EIB-116. U.S. Department of Agriculture, Economic Research Service, August 2013.

Wei, Y., Wang, Y., Wan, Z., Wang, T., Yang, L., & Chen, Z. (2015) Extraction, characterization, and utilization of wheat germ oil: A review. Journal of Food Science, 80(4), R893-R902.

Wheat Marketing Center Inc. (2004). Wheat and Flour Testing Methods: A Guide to Understanding Wheat and Flour Quality. In Wheat Marketing Center Ink, North American Export Grain Association.

Wrigley, C. W., Békés, F., & Bushuk, W. (2006). Chapter 1 Gluten: A Balance of Gliadin and Glutenin. In Gliadin and Glutenin: The Unique Balance of Wheat Quality (Issue March). https://doi.org/10.1094/9781891127519.002

Wrigley, C.W. & Bietz, J.A. 1988. Proteins and Amino Acids. Wheat, Chemistry & Technology. Y. Pomeranz, ed. AACCI: St. Paul, MN.

Zhang, Ang. (2020) Effect of wheat flour with different quality in the process of making flour products. *Int. J. Metrol. Qual. Eng.*, Volume 11, 2020 https://doi.org/10.1051/ijmqe/2020005

Table 1. Years and locations for all 738 AYT samples.

Year		Location	
2015	Brookings, SD	Selby, SD	Watertown, SD
2016	Brookings, SD	Selby, SD	Groton, SD
2017	Letcher, SD	Watertown, SD	Groton, SD
2018	Brookings, SD	Miller, SD	Groton, SD
2019	Groton, SD	Selby, SD	Watertown, SD
2020	Agar, SD	Brookings, SD	Watertown, SD
2021	Brookings, SD	Faulkton, SD	Selby, SD

Table 2. Summary statistics of the explored data set.

Parameter	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Sec	180	750.2	994	1048.9	1305.2	2820
WZ	58.1	64	66.2	66.48	68.9	75.7
NIR Moisture	7.78	12.23	12.7	12.67	13.37	14.41
NIR Protein	10.7	13.45	14.14	14.24	15	18.53
NIR Ash	0.36	0.41	0.43	0.4357	0.45	0.55
Dry	1.022	1.316	1.404	1.421	1.514	2.54
Good	1.87	3.326	3.571	3.572	3.821	4.864
GI	46.49	88.07	94.53	91.22	97.55	99.78
WG	27.12	36.18	38.94	39.42	42.36	62.27
DG	10.22	13.16	14.04	14.21	15.14	25.4
WB	16.52	22.94	25	25.21	27.35	36.87
TW	53.84	59.36	60.64	60.43	61.68	65.28
K Protein	12.49	14.66	15.31	15.41	16.1	19.1
K Ash	1.168	1.4551	1.573	1.57	1.69	1.98
Bran	21.44	26.42	29.08	39.51	54.3	63.7
Shorts	6.961	10.942	14.855	16.32	20.5	36.1
SKCS HI	29.72	65.35	70.11	70.33	75.74	89.63
SKCS WT	22.45	29.05	31.01	31.03	32.96	40.66
SKCS DM	2.319	2.628	2.708	2.711	2.797	3.071
MLT	0.53	3.06	3.945	4.389	5.277	15.34
MLV	33.24	43.57	46.48	46.83	49.7	59.07
MLS	-0.18	6.091	9.317	9.538	12.642	33.745
MLW	12.25	22.65	26.91	27.19	31.81	47.19
MLI	11.54	101.2	131.48	143.82	176.11	422.46
MPT	1.53	4.06	4.945	5.389	6.277	16.34
MPV	39.87	48.27	51.74	52.68	56.77	72.42
MPW	9.897	20.397	24.366	24.53	28.788	41.103
MPI	56.62	153.64	183.91	194.41	225.38	465.5
MRT	2.53	5.06	5.945	6.389	7.277	17.34
MRV	38.08	45.85	48.94	49.5	52.76	64.83
MRS	-12.259	-5.511	-3.814	-4.024	-2.198	0.677
MRW	3.453	14.489	18.288	17.777	21.164	31.623
MRI	103.4	206.1	236.2	245.8	276.9	508.7
MTT	8.08	8.33	8.83	9.458	9.99	19.34
MTV	28.29	39.44	41.88	42.02	44.57	55.75
MTS	-7.605	-1.368	-1.0295	-1.0763	-0.7295	0.355
MTW	1.457	7.216	10.175	10.626	13.462	25.388
MTI	262.4	346	373.9	382.4	408.5	623.9
MTXV	33.86	42.94	45.75	46.08	48.85	59.47

MTXS	-5.911	-3.778	-2.977	-2.96	-2.172	-0.256
MTXW	2.442	10.674	14.159	14.034	17.247	27.538
MTXI	141	252.7	282.5	290.1	321.5	520.5

Table 3. Summary results of the linear model.

	Estimate	Std. Error	T-Value	P-Value	DF	F-Statistic	R^2	P-Value (Model)
Intercept	-584.126	464,490	-1.258	0.209	5	124.7	0.502	2.2E-16
MTW	46.058	3.290	13.992	2.00E-16	608			
GI	14.361	1.513	9.491	2.00E-16				
NIRMoisture	-25.999	13.298	-1.955	0.051				
TW	3.963	7.103	0.558	0.577				
SKCS_HI	-1.310	1.553	-0.843	0.399				

Table 4. Root Mean Square Error of the Random Forest training model.

Year	RMSE
2015	312.411
2016	261.732
2017	324.05
2018	245.451
2019	356.613
2020	400.968

Table 5. Root Mean Square Error of the XG Boost training model.

Year	RMSE
2015	349.493
2016	278.254
2017	427.99
2018	289.936
2019	345.146
2020	465.616

Table 6. Root Mean Squared Error of the Random Forest & XG Boost testing models.

	Test RMSE
Random Forest	12.733
XG Boost	14.027

Table 7. Summary statistics of the predictive models on the 2021 testing data, in terms of seconds.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
Random Forest	-634.94	-27.17	153.14	162.13	325.01	906.46
XG Boost	-687.37	22.26	163.67	196.77	363.82	1008.12

Figure 1: Pairwise correlation plots, histograms, and scatterplots of Stability & statistically significant parameters

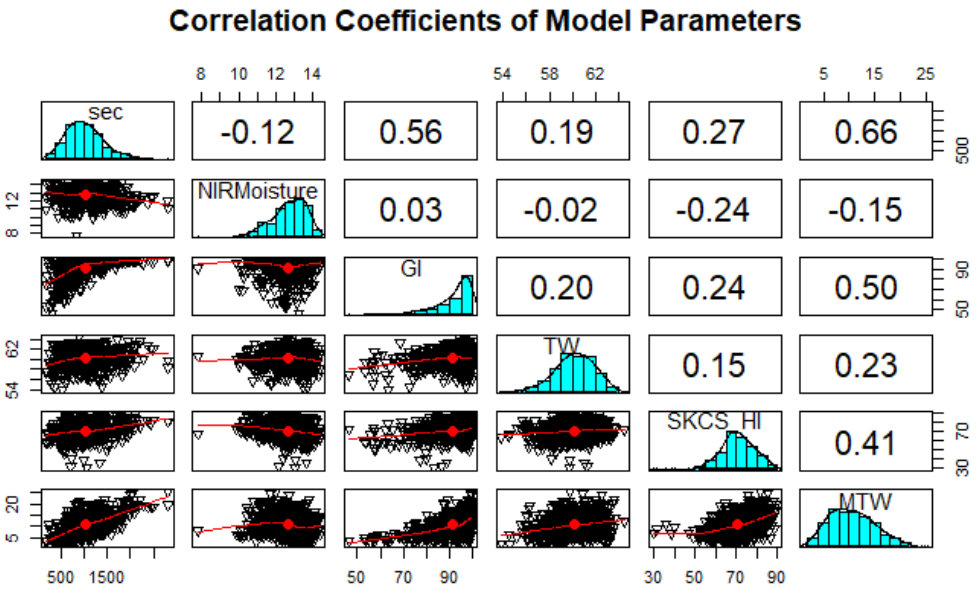

Figure 2. Training of Random Forest model on year 2015.

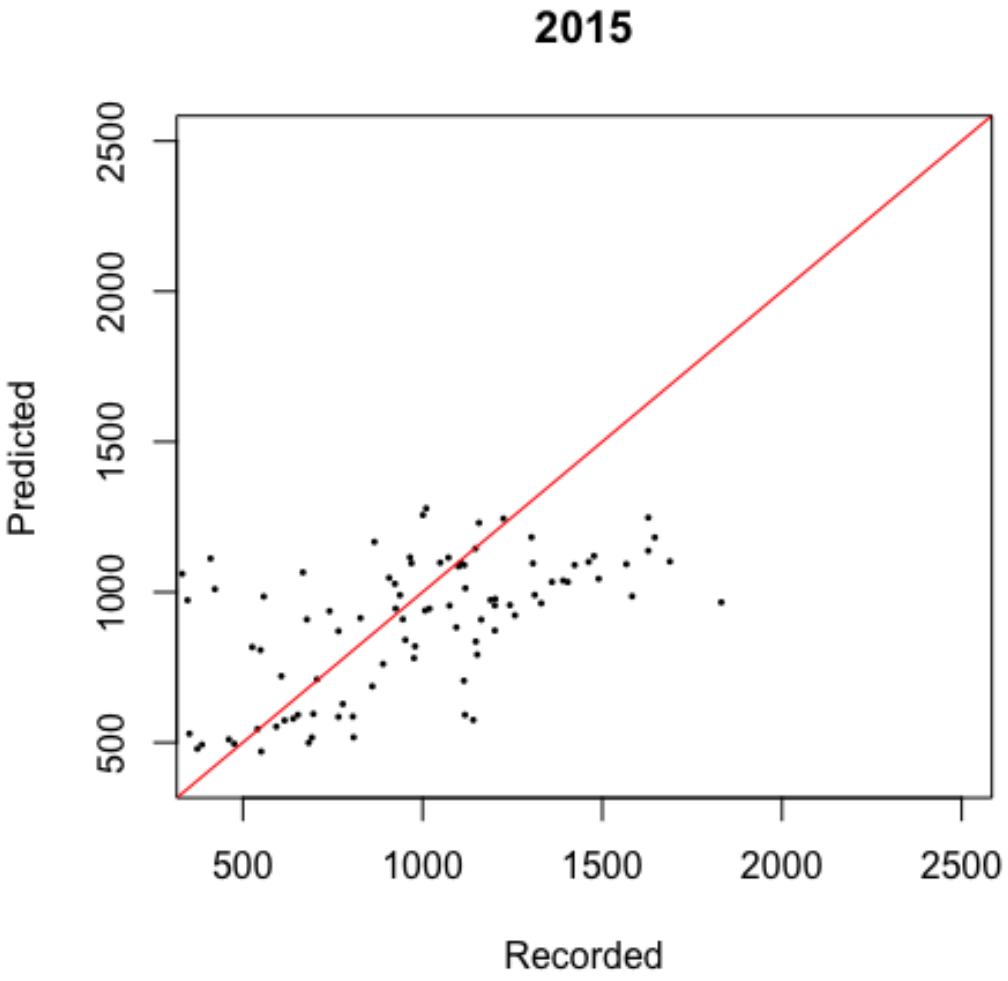

Figure 3. Training of Random Forest model on year 2016.

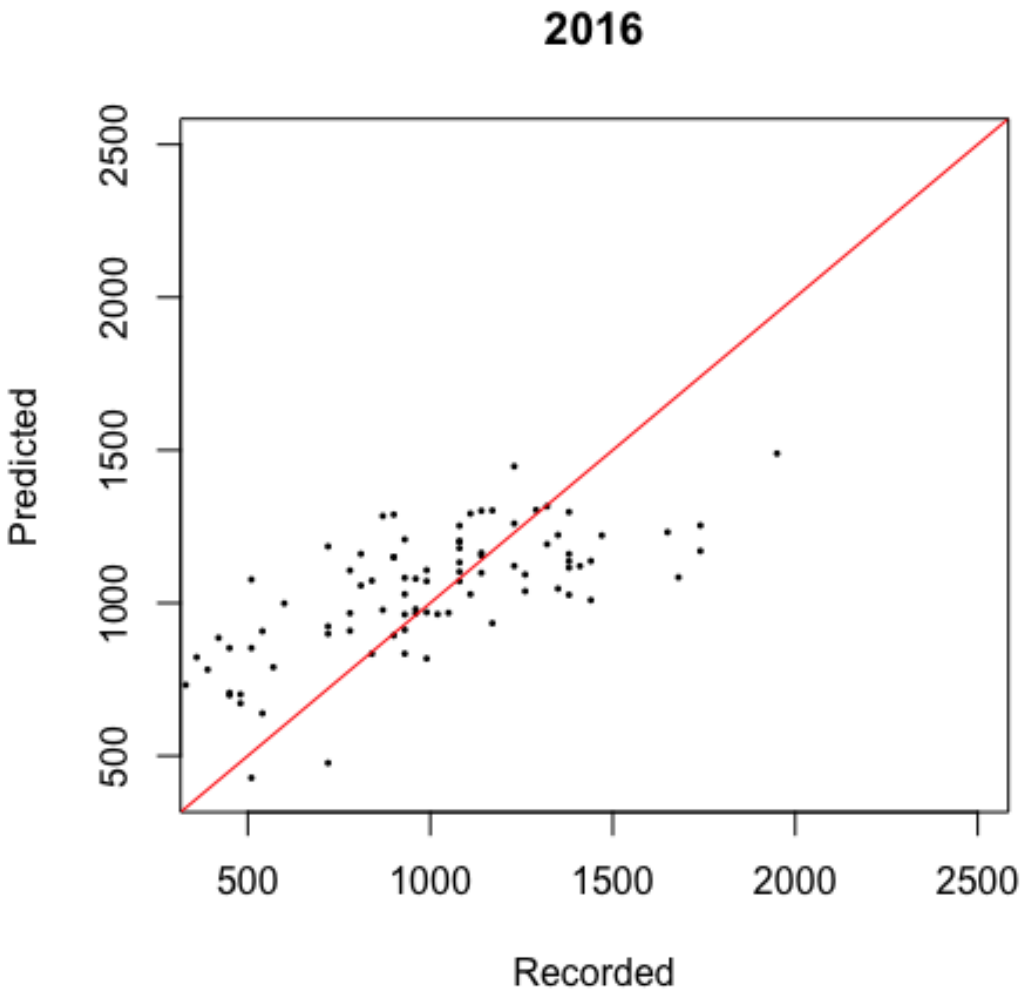

Figure 4. Training of Random Forest model on year 2017.

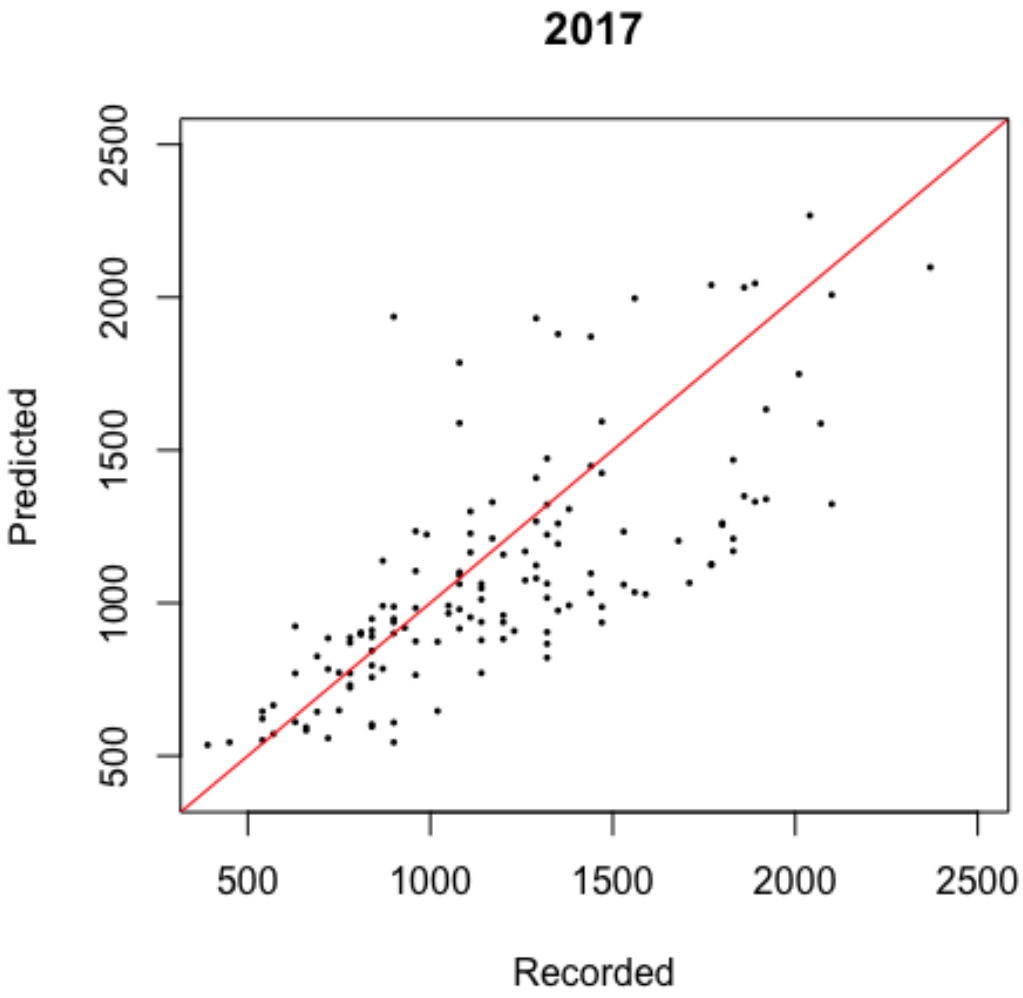

Figure 5. Training of Random Forest model on year 2018.

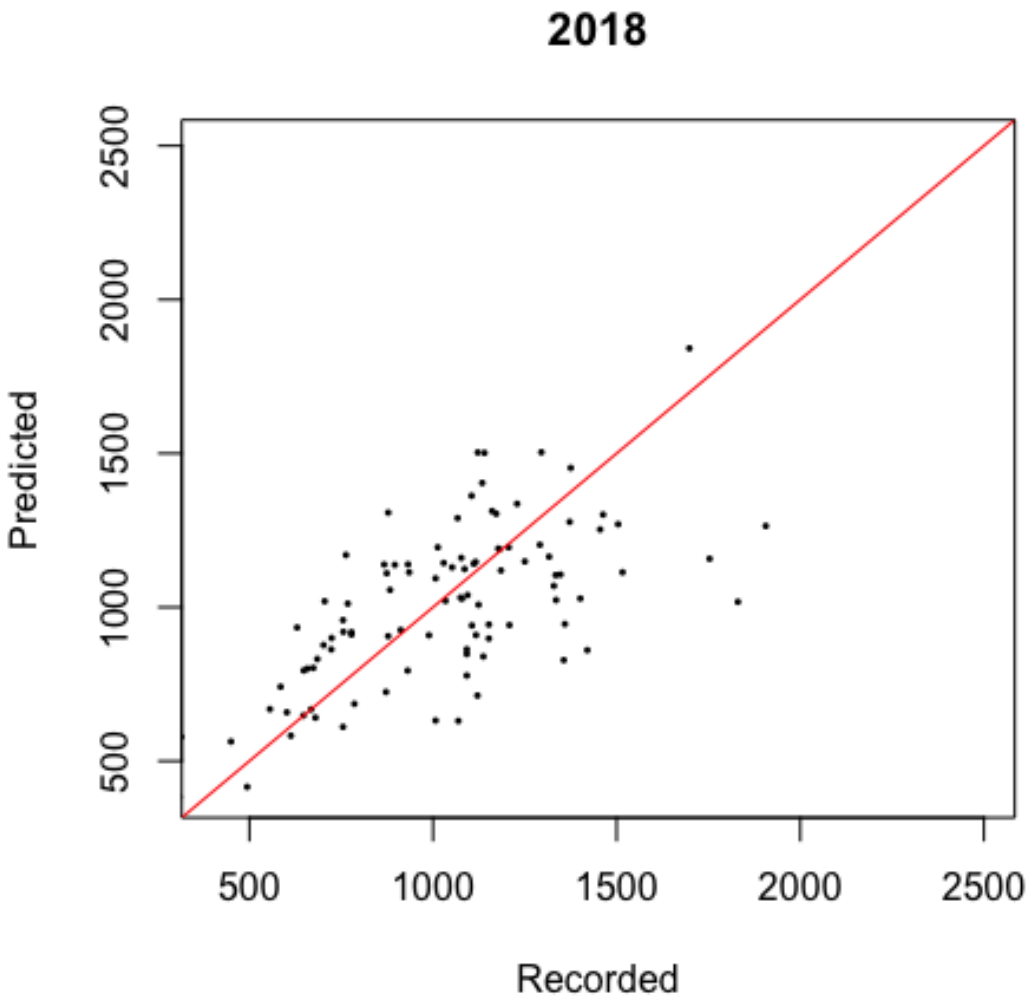

Figure 6. Training of Random Forest model on year 2019.

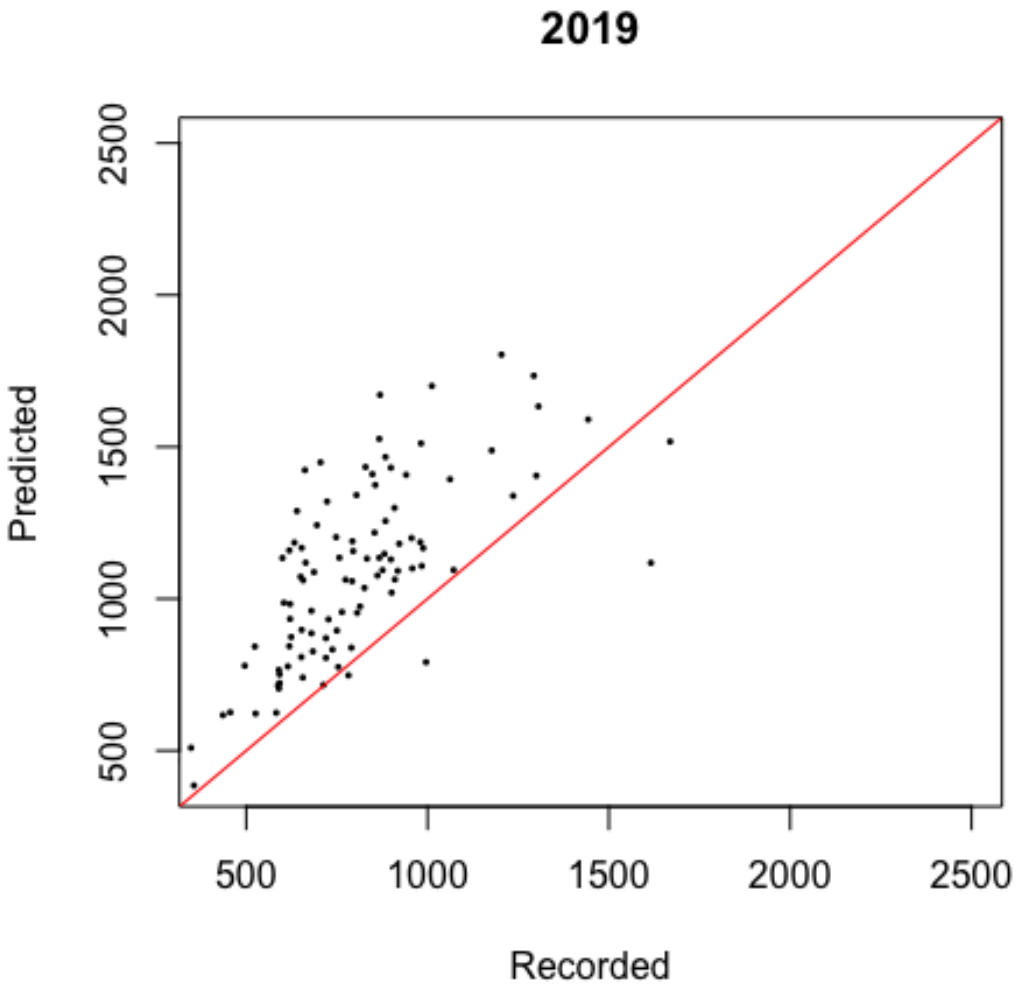

Figure 7. Training of Random Forest model on year 2020.

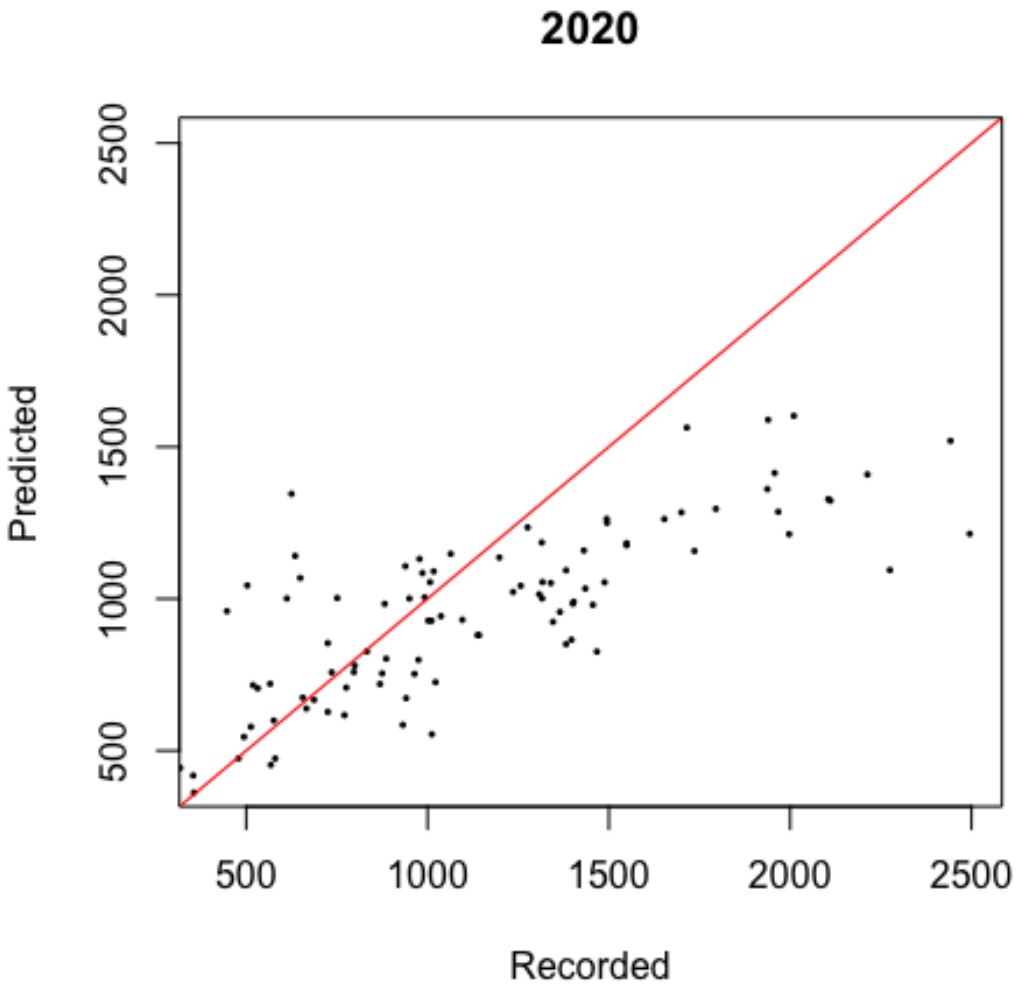

Figure 8. Training of the XG Boost model on year 2015.

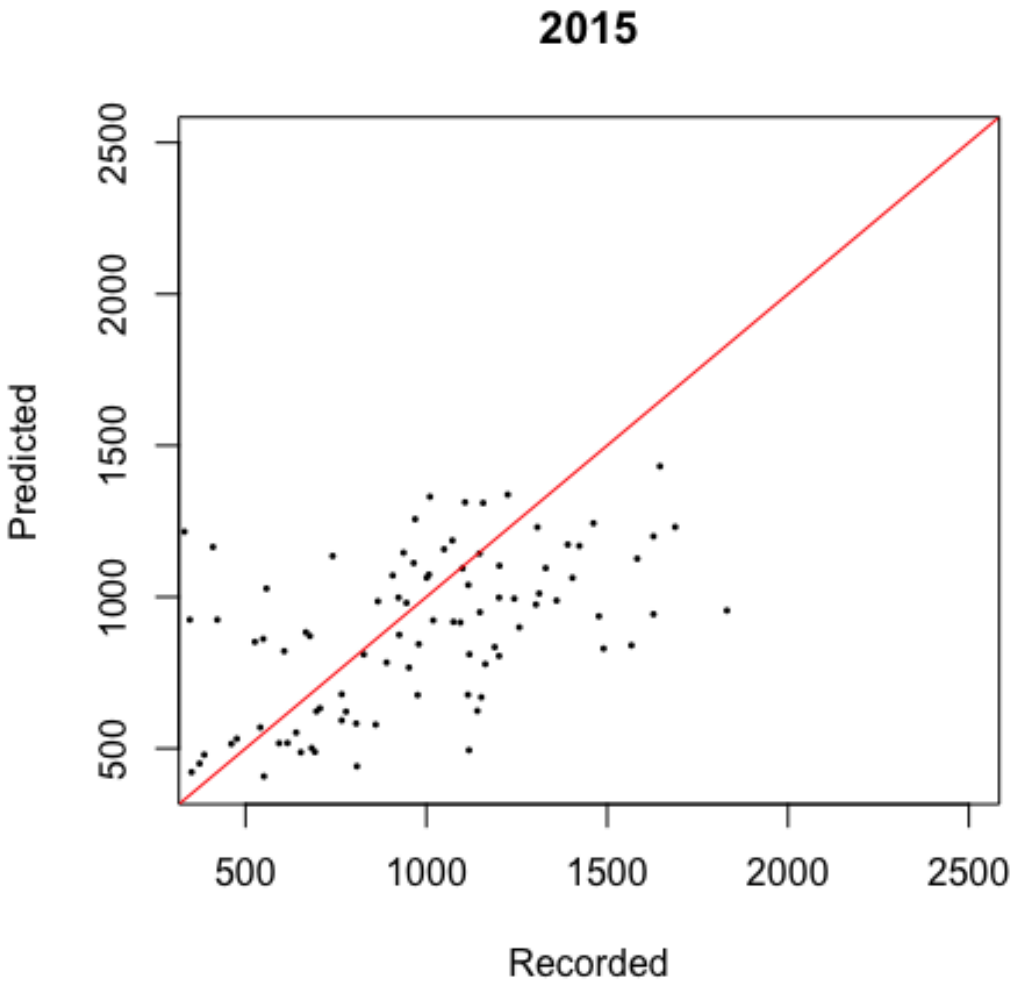

Figure 9. Training of the XG Boost model on year 2016.

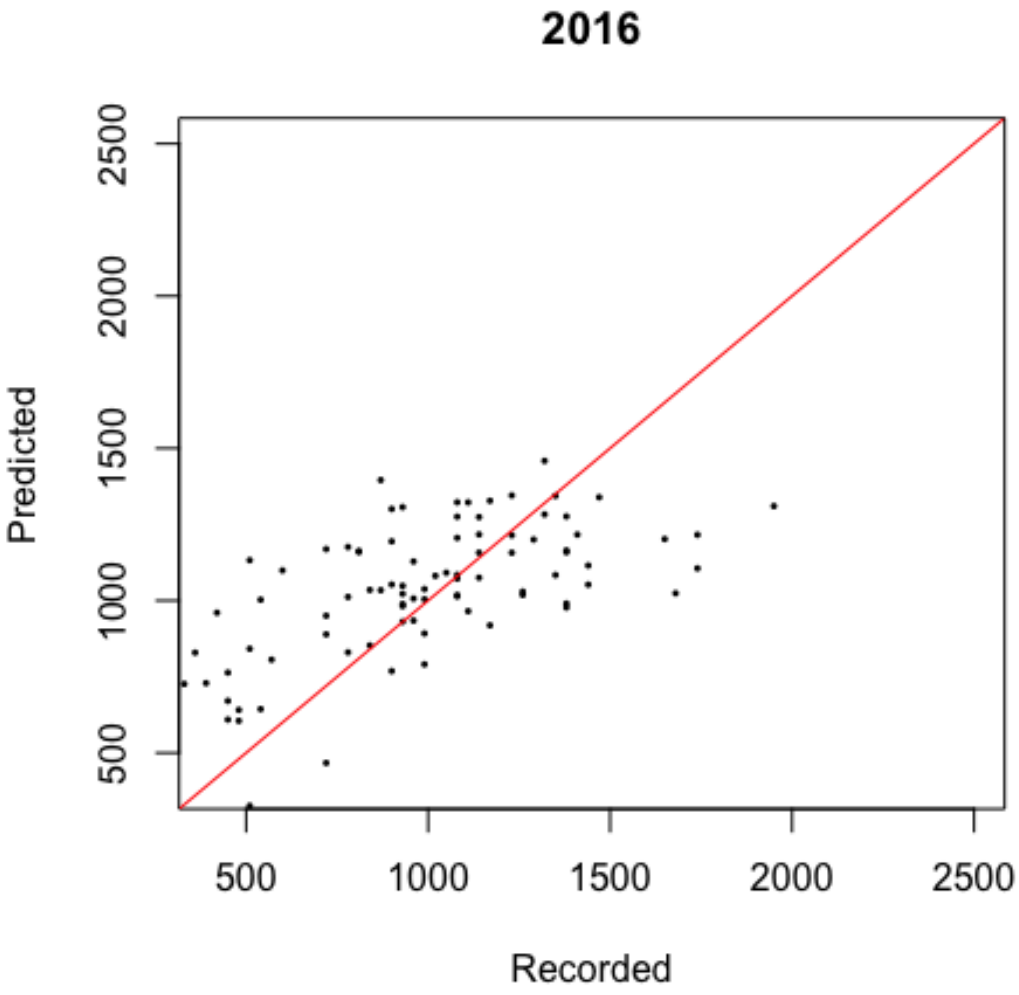

Figure 10. Training of the XG Boost model on year 2017.

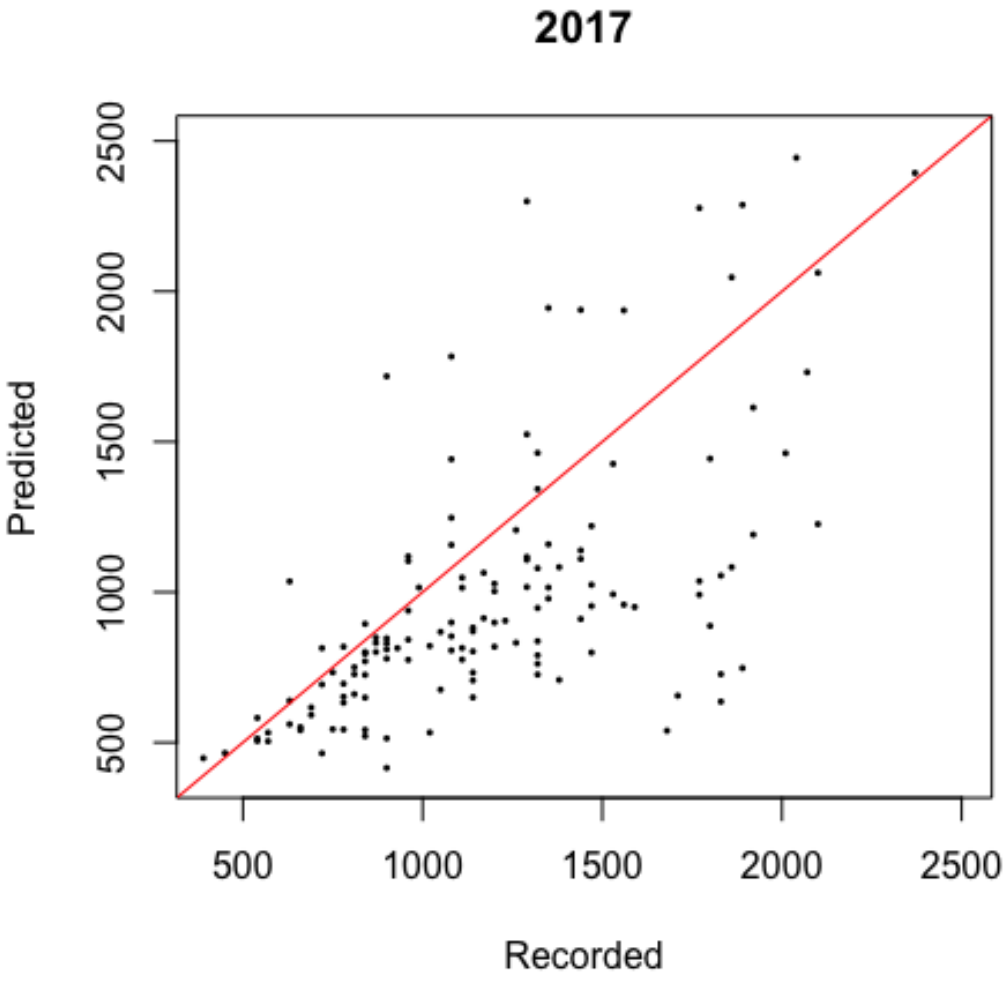

Figure 11. Training of the XG Boost model on year 2018.

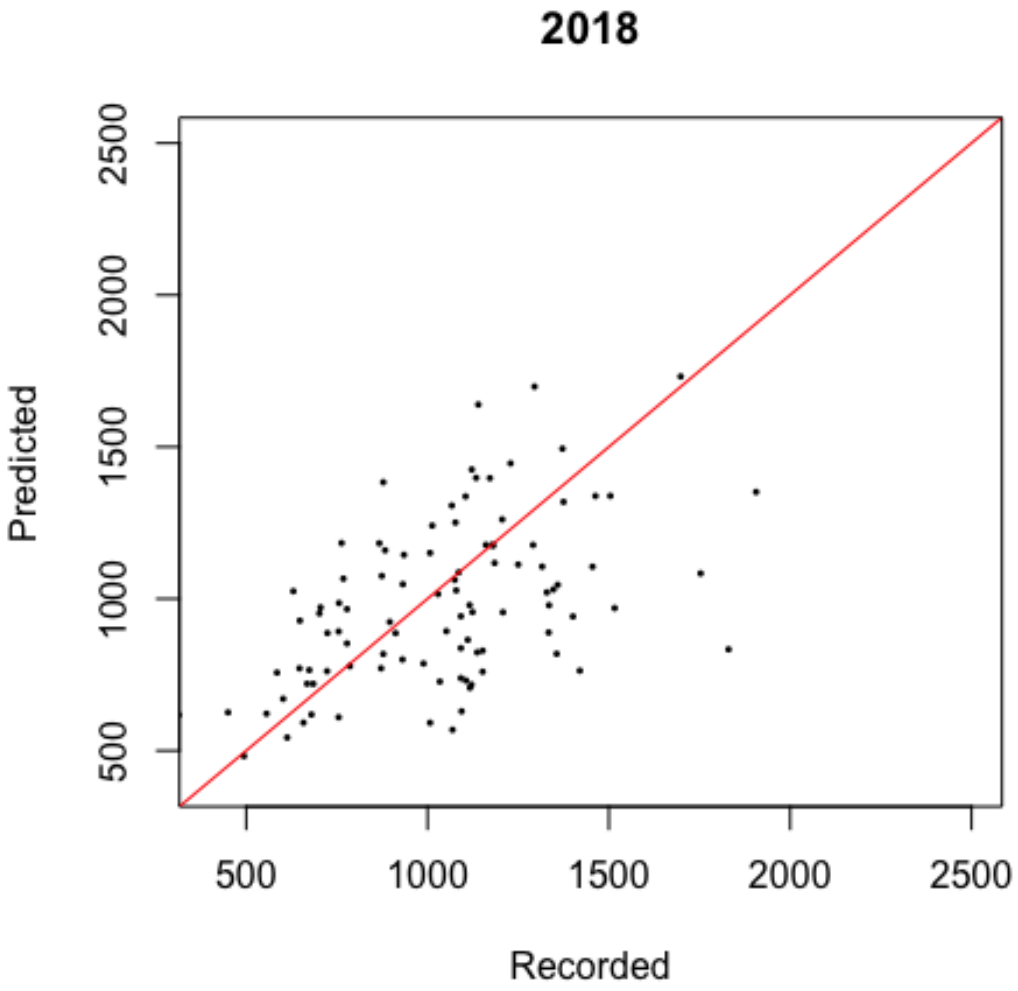

Figure 12. Training of the XG Boost model on year 2019.

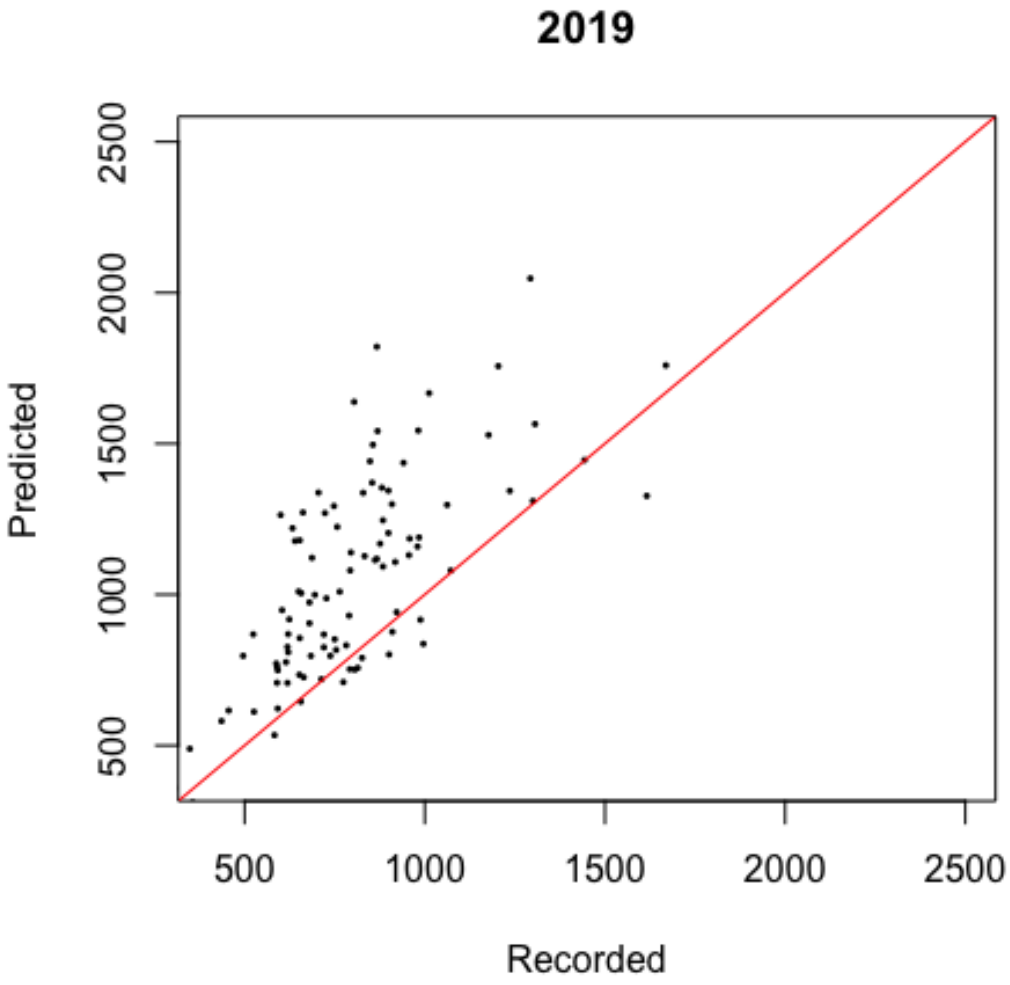

Figure 13. Training of the XG Boost model on year 2020.

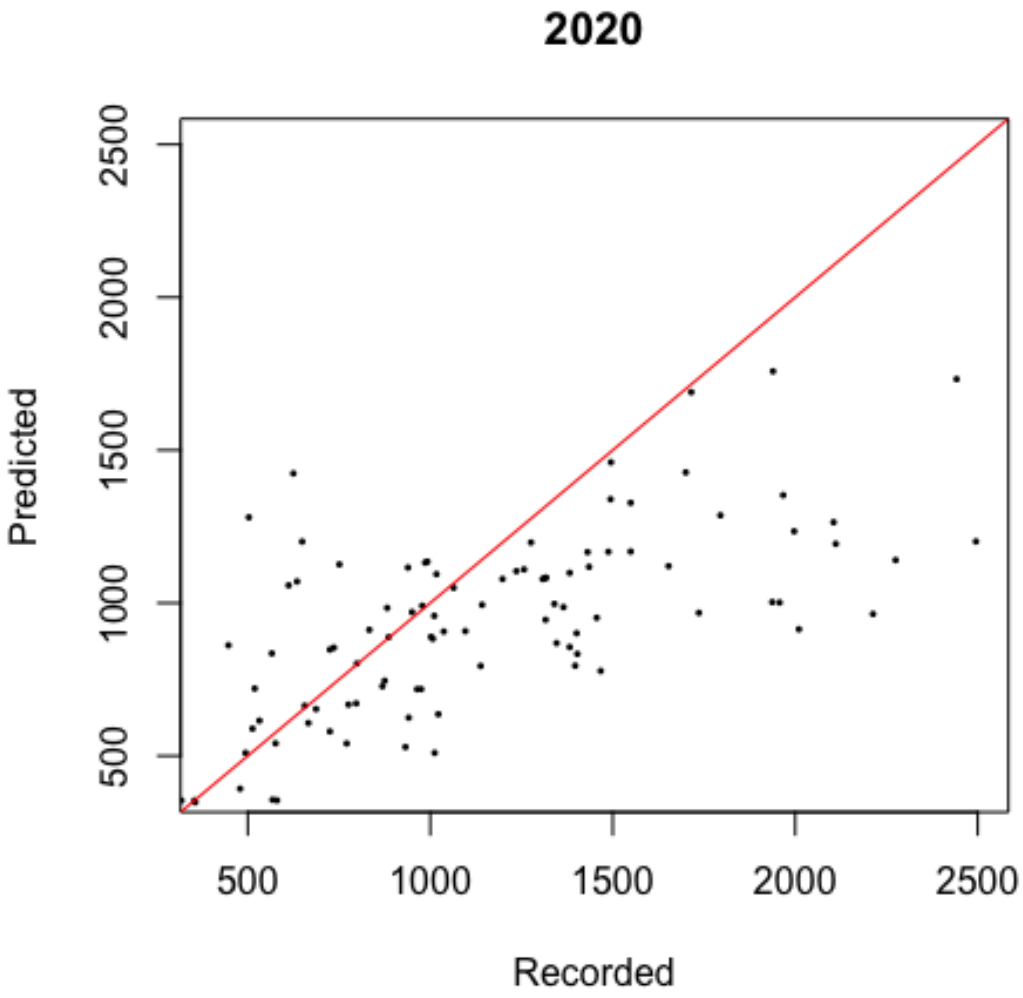

Figure 14. Test results of the Random Forest model on the 2021 data.

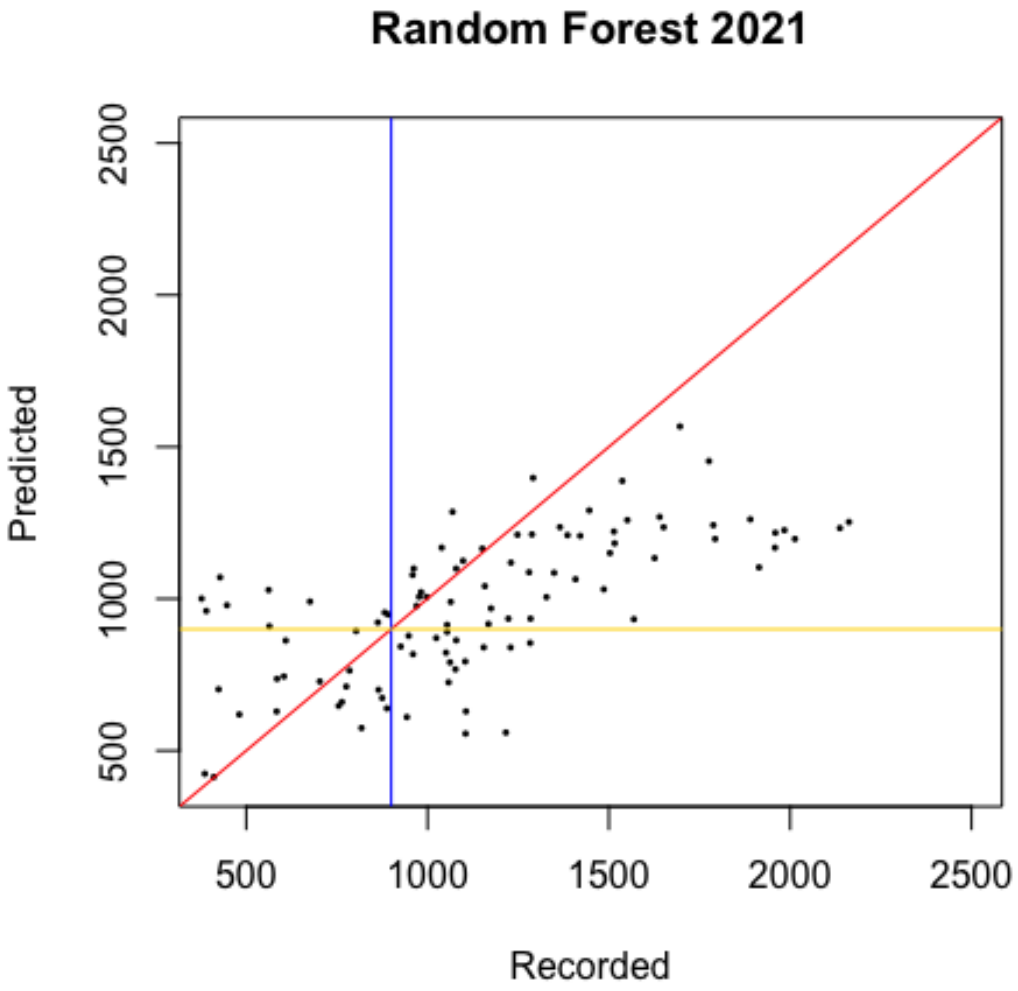

Figure 15. Test results of the XG Boost model on the 2021 data.