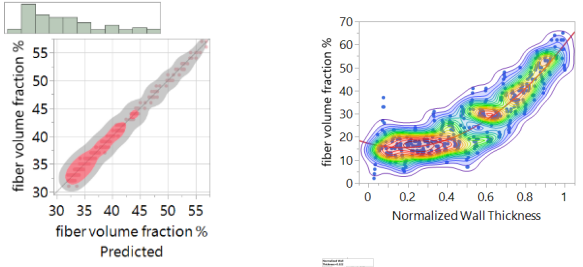
I International Scientific Conference
"CURRENT PROBLEMS OF CHEMISTRY,
MATERIALS SCIENCE AND ECOLOGY“,
Ukraine, 12.05.2021

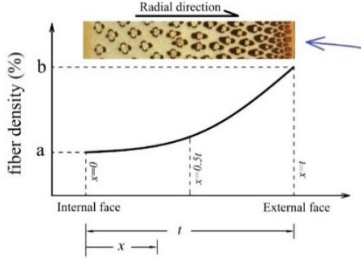
# IMAGES PROCESSING-BASED PREDICTIVE MODELING FOR FIBERS DISTRIBUTION ALONG THE RADIAL DIRECTION IN BAMBOOS *PHYLLOSTACHYS EDULIS* AND *DENDROCALAMUS GIGANTEUS*

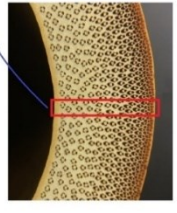

by

*A. Azadeh*
*F.C. Vitorino*
*M. Nazarkovsky*
*R. D. Toledo Filho*
*H. Savastano Jr*


```
Initial Splits(
    :Normalized Wall Thickn
    {:Normalized Wall Thickness < 0.452, {:Normalized Wall Thickness < 0.264,
    {:Normalized Wall Thickness < 0.214, {:Normalized Wall Thickness < 0.128,
    {:Normalized Wall Thickness >= 0.082}, {:Normalized Wall Thickness >=
    0.184, {}, {:Normalized Wall Thickness < 0.156, {},
    {:Normalized Wall Thickness >= 0.172}}}}, {:Normalized Wall Thickness
     >= 0.25}}, {:Normalized Wall Thickness < 0.354,
```



#DATA SCIENCE #XGBoost
#JSL #MACHINE LEARNING
#REGRESSION
#IMAGE DIGITAL PROCESSING

# Bamboo Image Processing

**Arash Azadeh**
**Postdoctoral fellow**

**USP-FZEA**

**12 May 2021**

# Objective

To provide a second order equation which represents the fiber distribution along the radial direction of bamboo section (as a Functionally Graded Material - FGM) for two bamboo species: *Phyllostachys edulis* (Moso) and *Dendrocalamus giganteus* (DG).

**KEYWORDS:** Bamboo, Image processing, Fiber distribution, Functionally Graded Material.

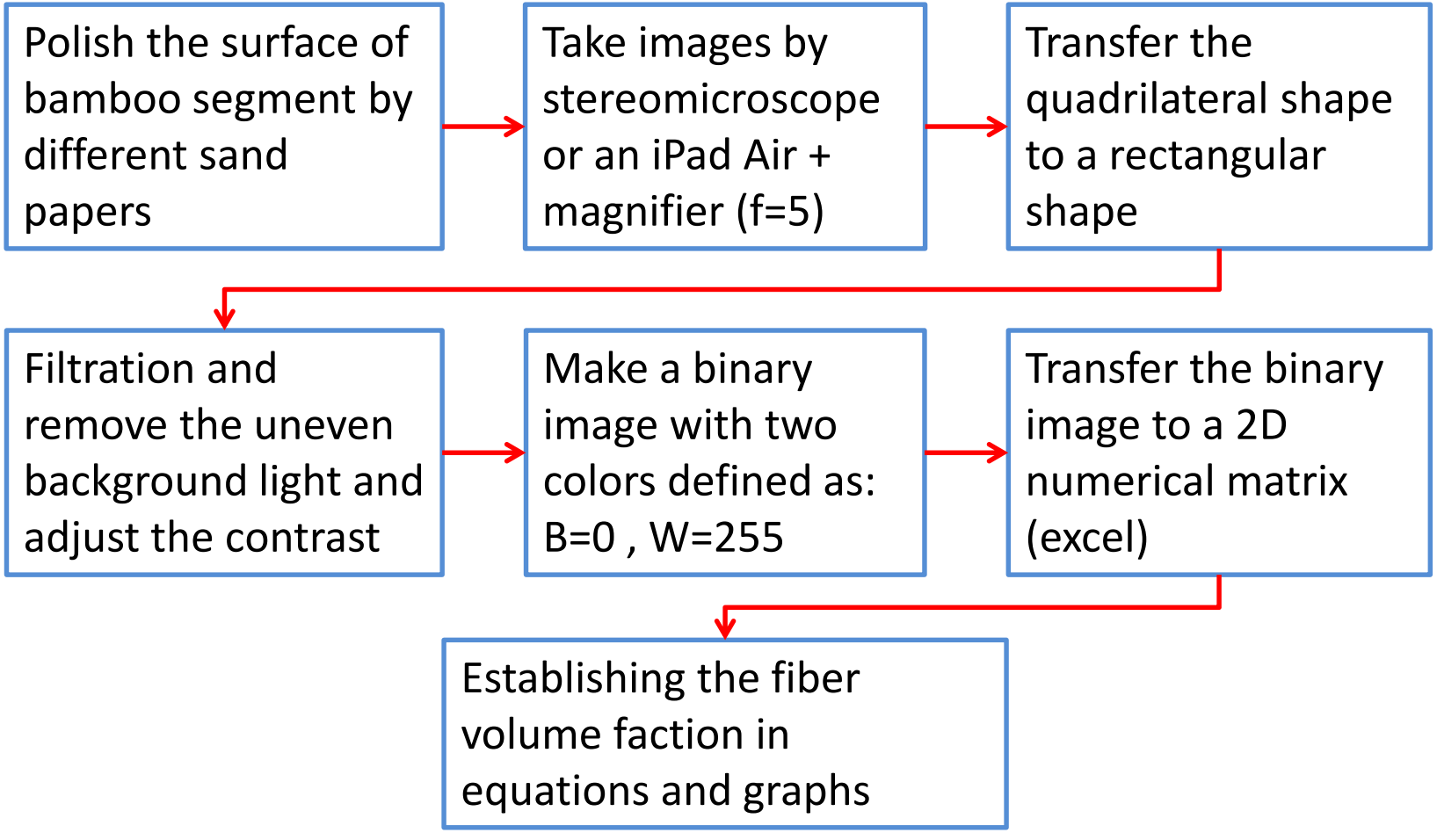
# Method

-Providing images with **optimum resolution** by using  stereomicroscope or an iPad + magnifier (f=5).

- Image **rectification** to remove the inside and outside excessive curvatures of bamboo surfaces and eliminate the unwanted warping of image due to close-up photography.

- Transferring the image data to a **numerical matrix** to measure the exact distribution of fibers.

# Steps

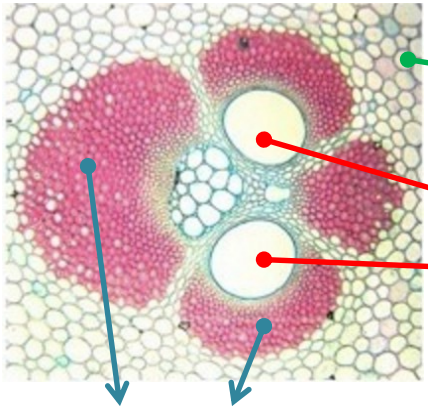

Polish the surface of bamboo segment by different sand papers
Take images by stereomicroscope or an iPad Air + magnifier (f=5)
Transfer the quadrilateral shape to a rectangular shape
Filtration and remove the uneven background light and adjust the contrast
Make a binary image with two colors defined as: B=0 , W=255
Transfer the binary image to a 2D numerical matrix (excel)
Establishing the fiber volume faction in equations and graphs

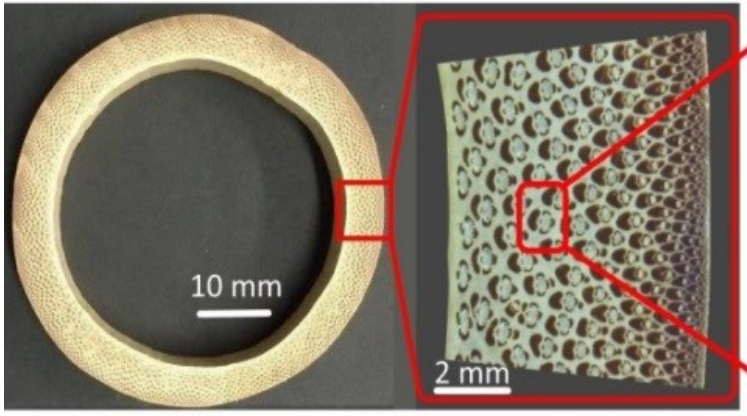
# Introduction

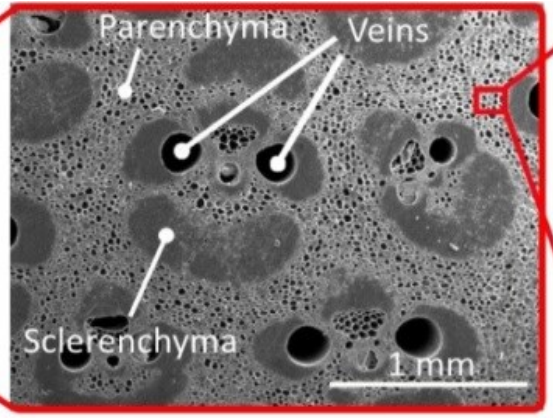
Bamboo as a three phases composite material:

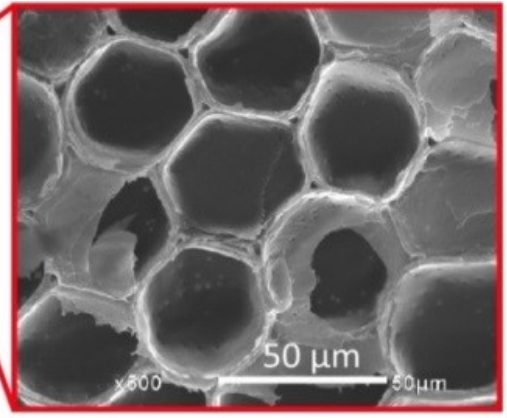
- **Sclerenchyma (Fibers)**
- **Parenchyma (Matrix)**
- Vein

Bamboo Vascular Bundle

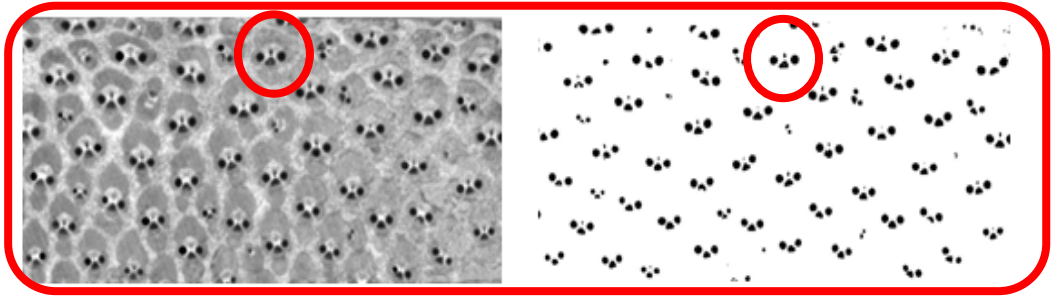


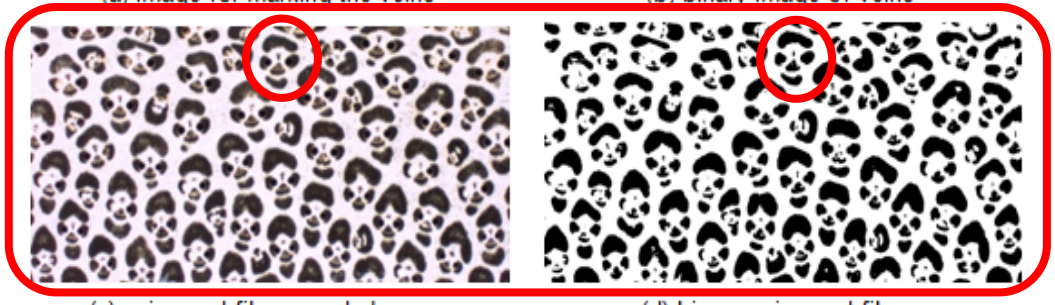



# Stereomicroscope image with appropriate lighting to separate three phases of fibers, matrix and veins



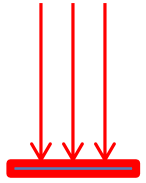
(a) image for marking the veins (b) binary image of veins



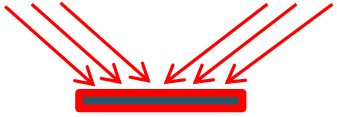
(c) veins and fibers as dark areas (d) binary veins and fibers

(e) subtraction of veins from fibers (f) pure fibers after median filtration

light source is perpendicular



light source is tilted

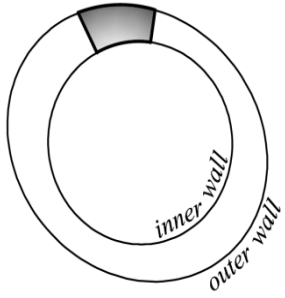


Volume fraction of bamboo components at inner and outer parts of whole section Of DG bamboo

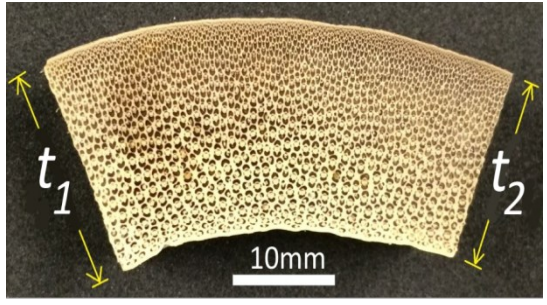
| Component | External part (%) | Internal part (%) |
|---|---|---|
| Fiber ($V_F$) | 56.1 | 30.7 |
| Matrix ($V_M$) | 42.2 | 62.9 |
| Vein ($V_V$) | 1.7 | 6.3 |

## Image rectification

- Non-uniform thickness
- Non-circular bamboo cross section

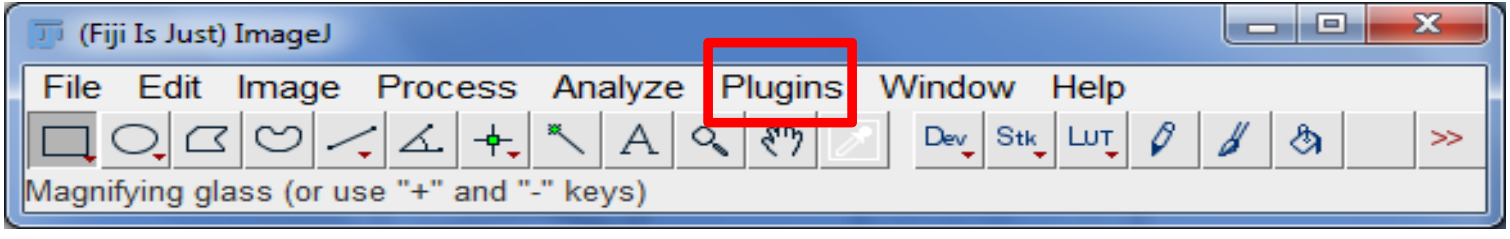



(a) Whole cross section

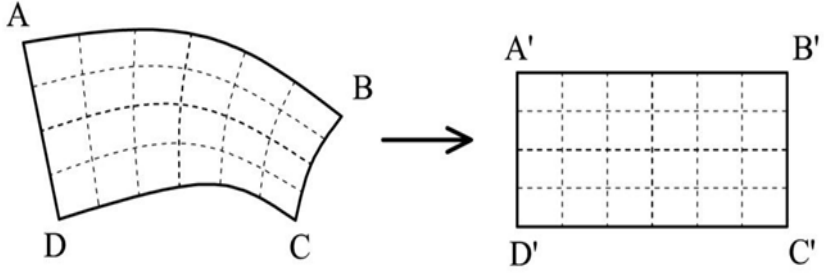



(b) Thickness variation of a bamboo segment (t1>t2)

**Software:** **"Fiji" or "ImageJ"**

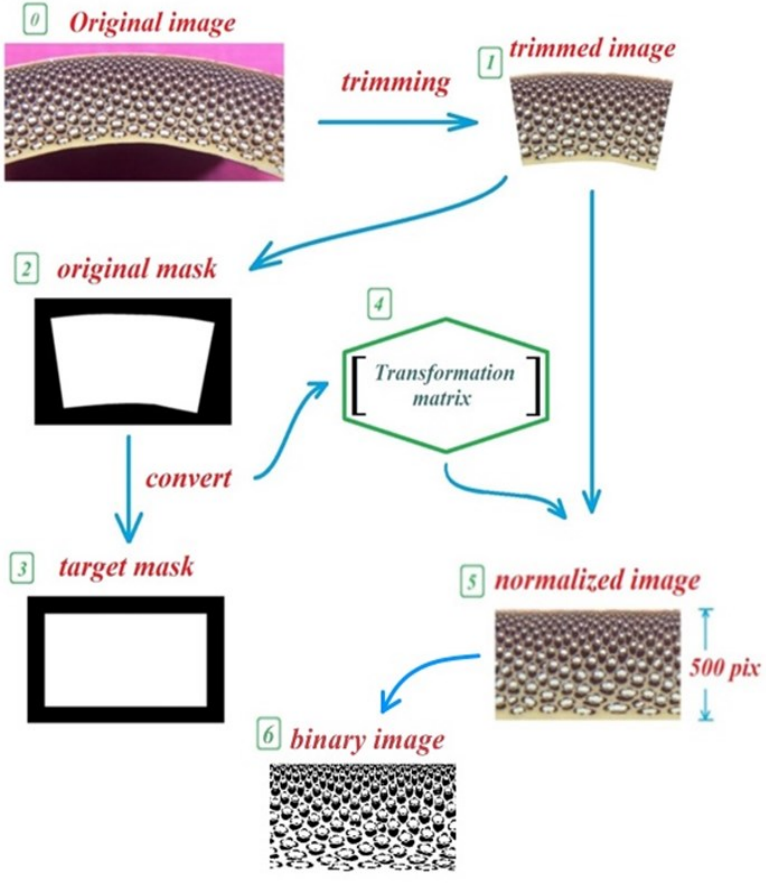


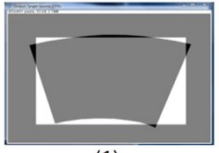

plugins>
registration>
**bUnwarpJ**




A quadrilateral section with curved sides transferred into a rectangular section

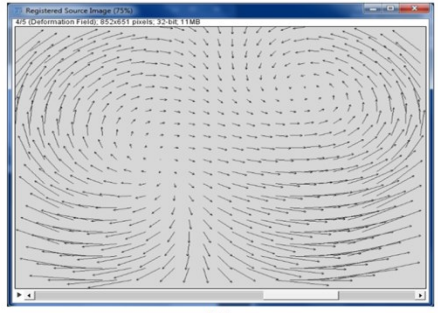
# ImageJ rectification procedure









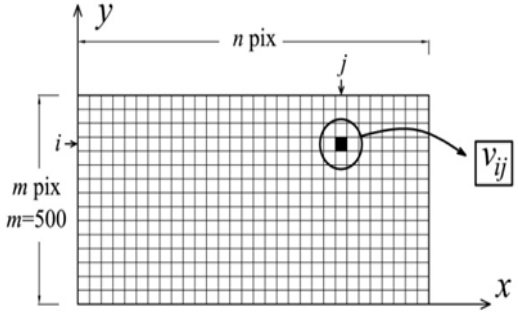
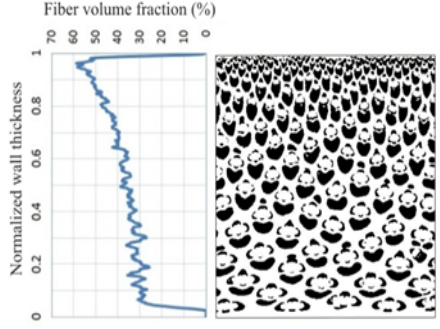
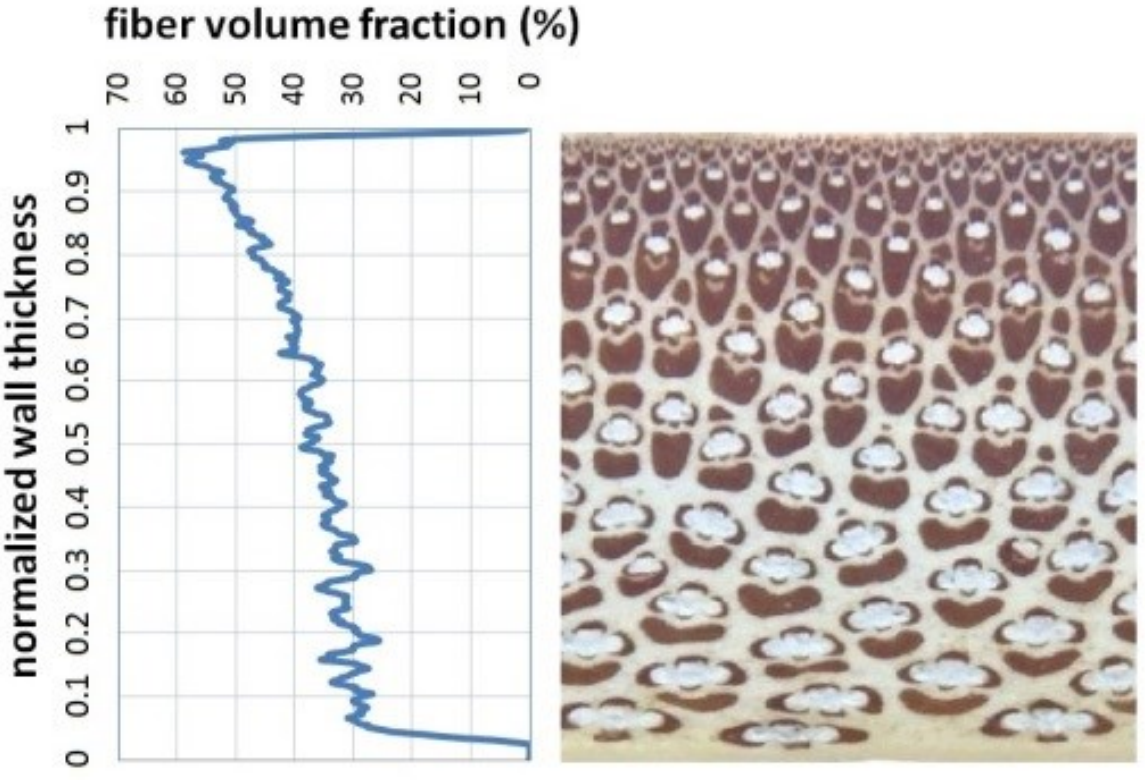




Image converted to matrix

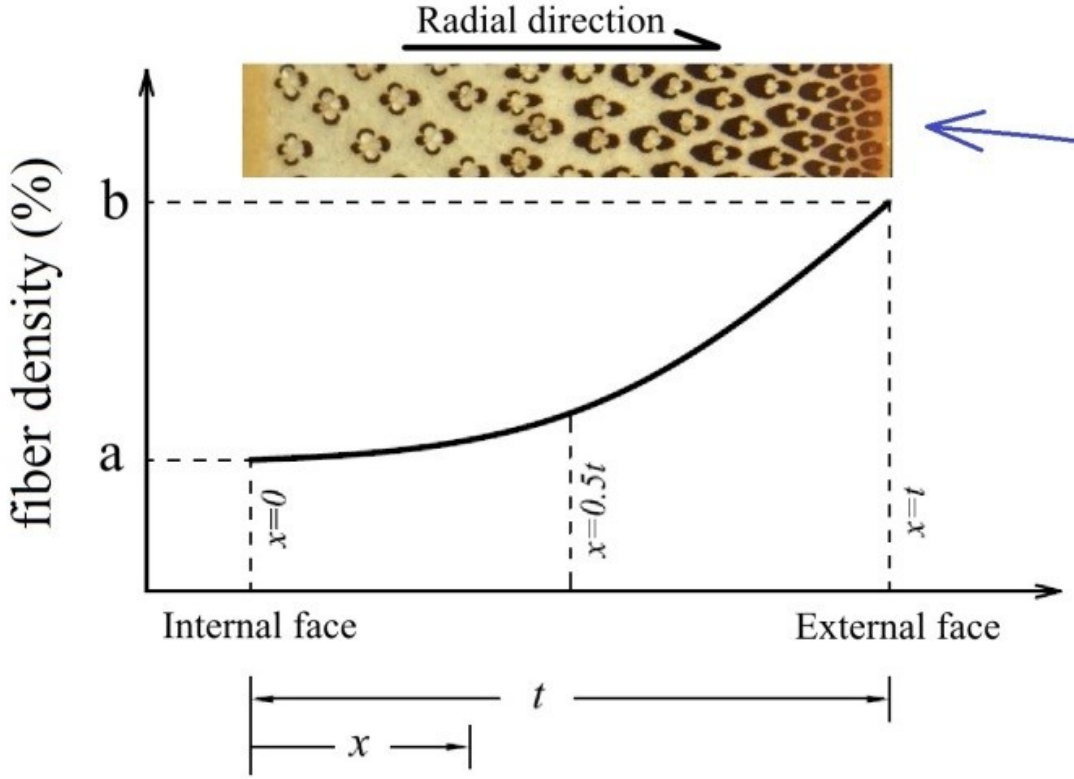


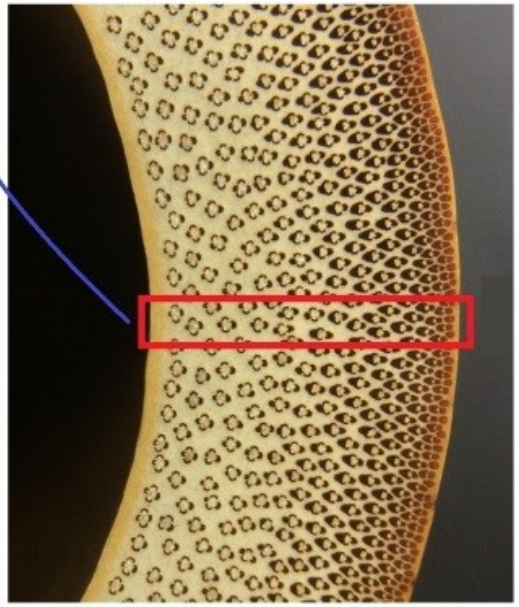

FVF in normalized wall thickness

# Fiber volume fraction at normalized wall thickness



# Bamboo is a Functionally Graded Material (FGM)



# 2 What is a Functionally Graded Material (FGM)?

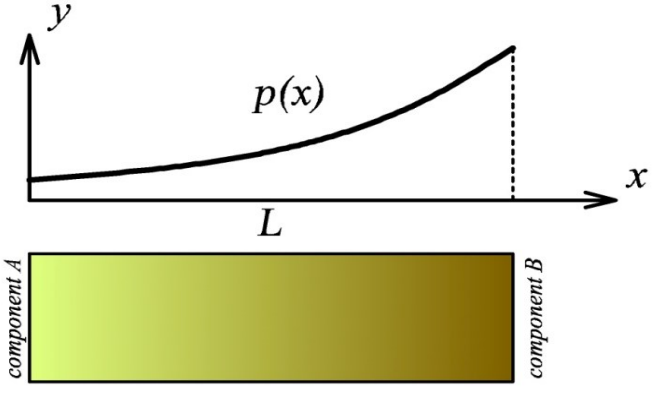
- In a functionally graded material (FGM) the properties change gradually with position.

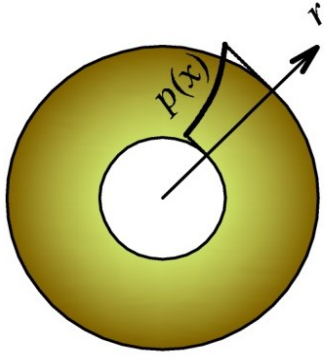
Kieback, B., A. Neubrand, and H. Riedel. "Processing techniques for functionally graded materials." *Materials Science and Engineering: A* 362, no. 1 (2003): 81-106.

- A functionally graded material (FGM) is a two component composite characterized by a compositional gradient from one component to other.

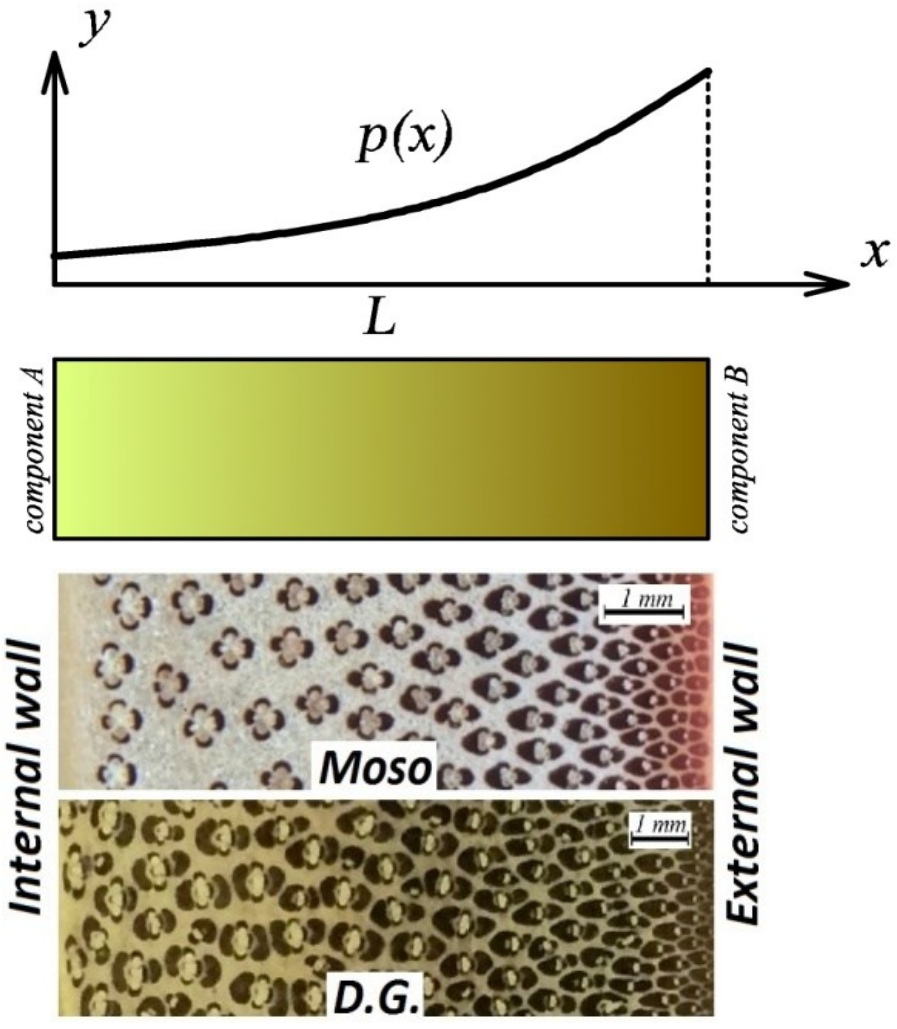
Bharti, Isha, Nishu Gupta, and K. M. Gupta. "Novel applications of functionally graded nano, optoelectronic and thermoelectric materials." *Int. J. Mater. Mech. Manuf* 1, no. 3 (2013): 221-224.

**Gradual change from one component to another alongside a direction.**



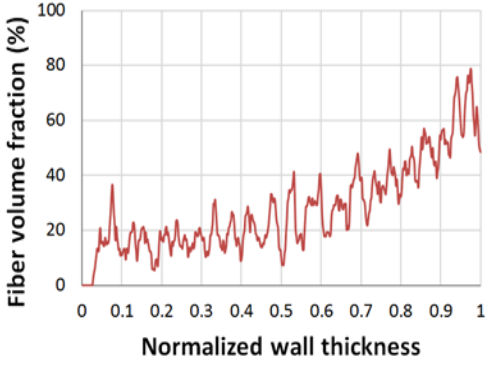



Difference between fiber distribution in *Moso* and *D.G.*

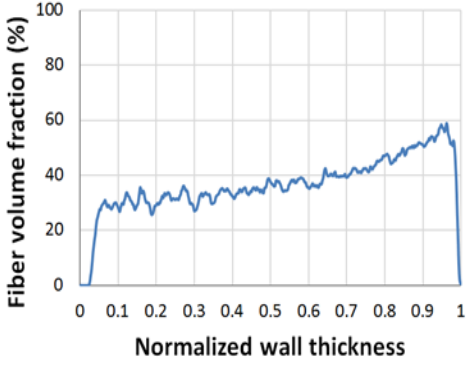



visual appearance of Moso and DG Bamboo

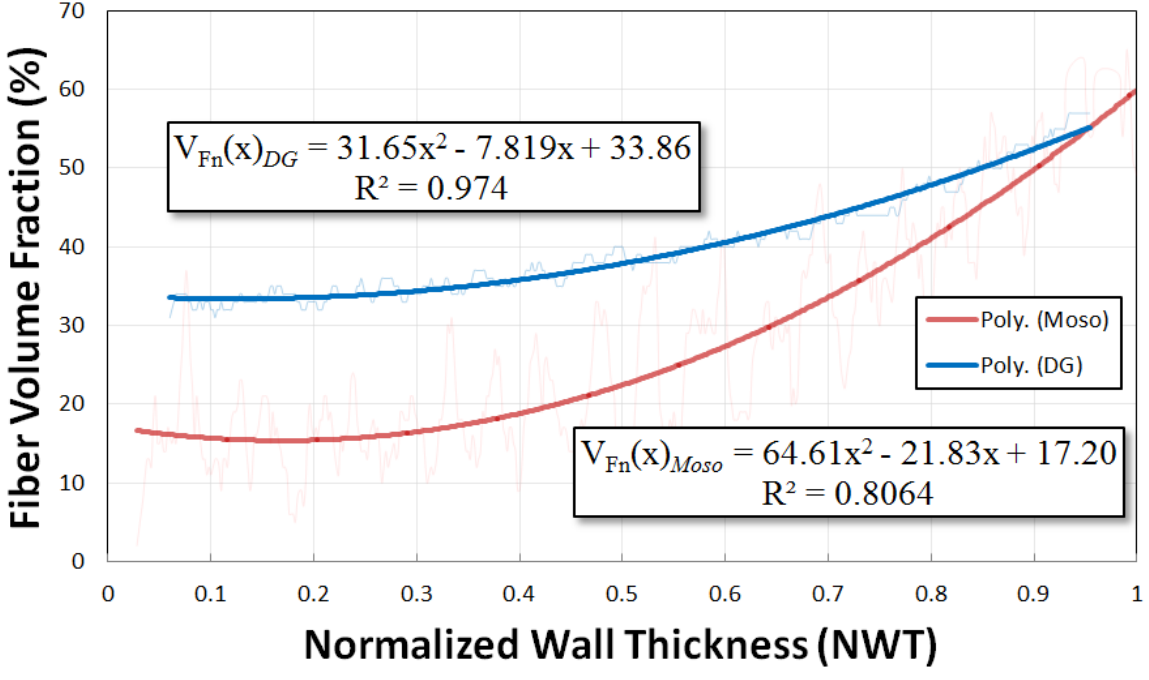
# *DG* Bamboo vs *Moso* Bamboo




(a) Moso Bamboo



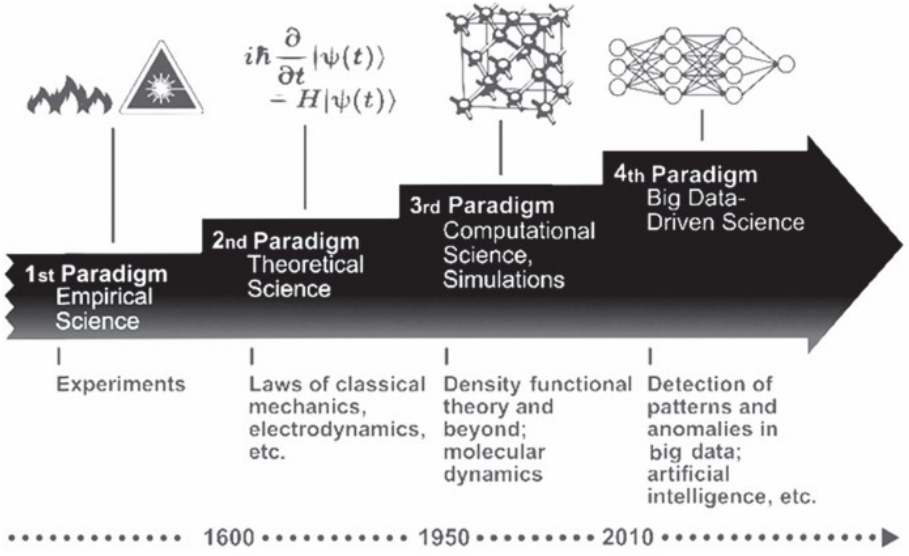

(b) DG Bamboo




curve fitting along the wall thickness for DG and Moso Bamboo

# Thanks

# Paradigm 4.Knowledge Discovery From Data




Draxl, C.; Scheffler, M. NOMAD: The FAIR Concept for Big Data-Driven Materials Science. *MRS Bull.* **2018**, *43* (9), 676–682. https://doi.org/10.1557/mrs.2018.208.

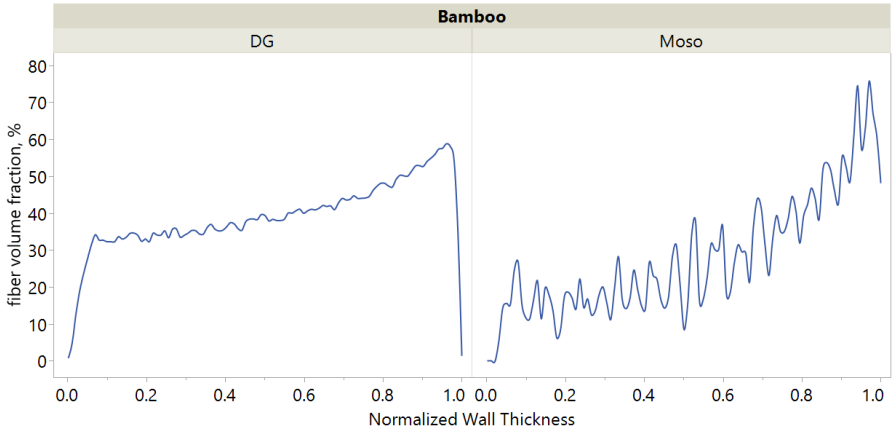
***Data science is intended to analyze, understand, and extract phenomena from data.***

Data science in chemistry can be:
– a chemical science solution of *forwarding problems* – measurement or prediction of materials' properties at unprecedented new scales;
– classification of the materials/compounds by the properties;
– the solution of inverse problems like reverse engineering of molecular and materials properties (which is not simply to "invert" physical models and run in reverse);
– a way of problem search in chemical space that is limitless.

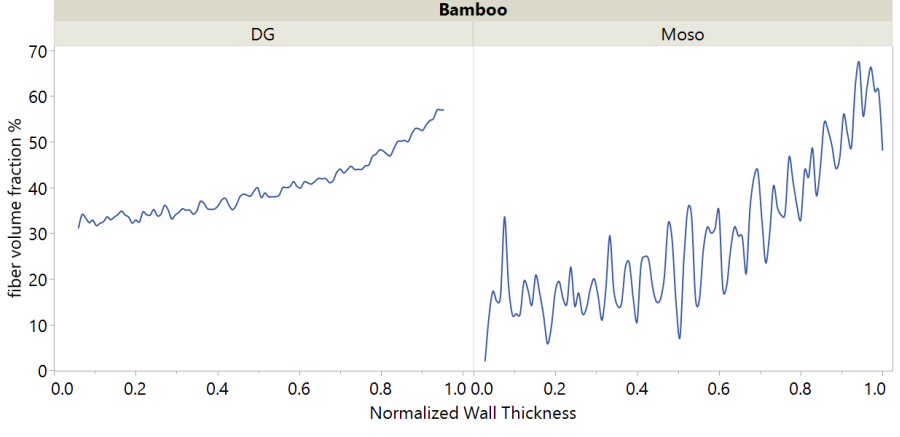
Th. Gressling "**Data Science in Chemistry**", 2021 Walter de Gruyter GmbH, Berlin/Boston

# Data Cleaning

## Initially




## After cleaning

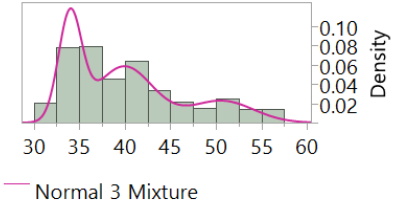



Excluded points
Moso: 30
DG: 56

# Distribution of FVF (%) for...

## DG

**fiber volume fraction %**




| Quantiles | | |
|---|---|---|
| 100.0% | maximum | 57 |
| 99.5% | | 57 |
| 97.5% | | 55 |
| 90.0% | | 50 |
| 75.0% | quartile | 44 |
| 50.0% | median | 38 |
| 25.0% | quartile | 35 |
| 10.0% | | 33 |
| 2.5% | | 32 |
| 0.5% | | 31.225 |
| 0.0% | minimum | 31 |

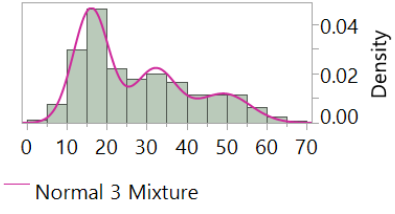
| Summary Statistics | |
|---|---|
| Mean | 40.015766 |
| Std Dev | 6.5305135 |
| Std Err Mean | 0.3099245 |
| Upper 95% Mean | 40.624871 |
| Lower 95% Mean | 39.406661 |
| N | 444 |

**Nonparametric Tolerance Intervals**

| Proportion | Lower TI | Upper TI | 1-Alpha | Actual Confidence |
|---|---|---|---|---|
| 0.900 | 32 | 54 | 0.950 | 0.9618 |

## Moso

**fiber volume fraction %**




| Quantiles | | |
|---|---|---|
| 100.0% | maximum | 65 |
| 99.5% | | 63.29 |
| 97.5% | | 57 |
| 90.0% | | 49 |
| 75.0% | quartile | 37 |
| 50.0% | median | 23 |
| 25.0% | quartile | 16 |
| 10.0% | | 12.1 |
| 2.5% | | 7.775 |
| 0.5% | | 4.355 |
| 0.0% | minimum | 2 |

| Summary Statistics | |
|---|---|
| Mean | 27.161702 |
| Std Dev | 13.930662 |
| Std Err Mean | 0.6425735 |
| Upper 95% Mean | 28.424382 |
| Lower 95% Mean | 25.899023 |
| N | 470 |

**Nonparametric Tolerance Intervals**

| Proportion | Lower TI | Upper TI | 1-Alpha | Actual Confidence |
|---|---|---|---|---|
| 0.900 | 9 | 55 | 0.950 | 0.9507 |

# Analysis of Variances

**Wilcoxon (Rank Sums)**

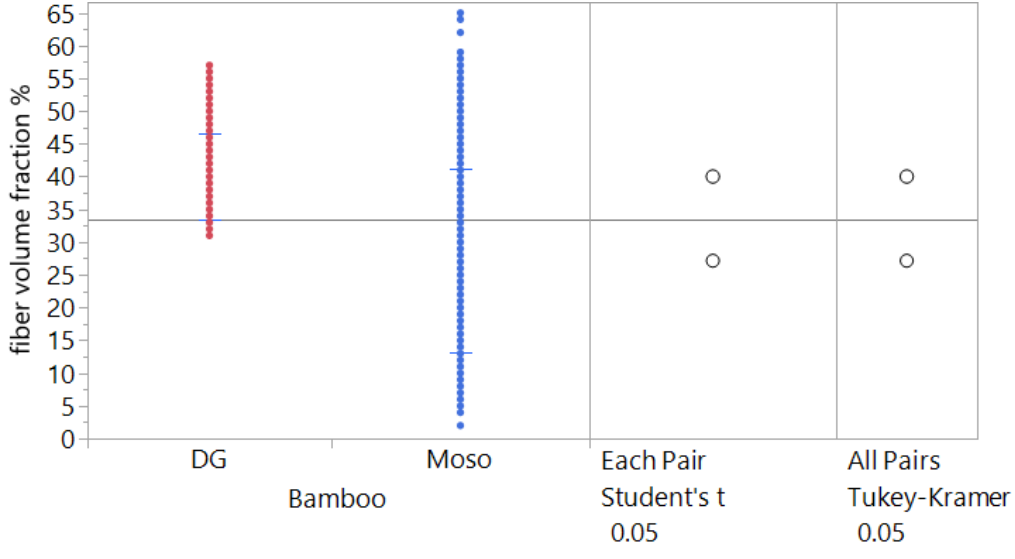
| Level | Count | Score Sum | Expected Score | Score Mean | (Mean-Mean0)/Std0 |
|---|---|---|---|---|---|
| DG | 444 | 263146 | 203130 | 592.670 | 15.053 |
| Moso | 470 | 155010 | 215025 | 329.807 | -15.053 |

**2-Sample Test, Normal Approximation**

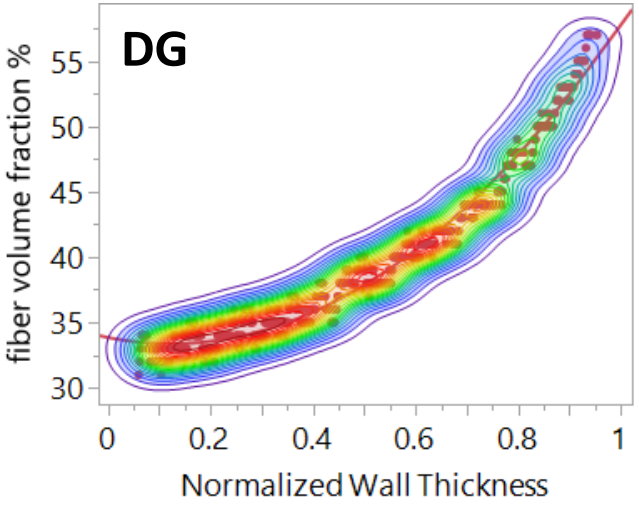
| S | Z | Prob>\|Z\| |
|---|---|---|
| 263145.5 | 15.05303 | <.0001* |

**1-Way Test, ChiSquare Approximation**

| ChiSquare | DF | Prob>ChiSq |
|---|---|---|
| 226.5976 | 1 | <.0001* |



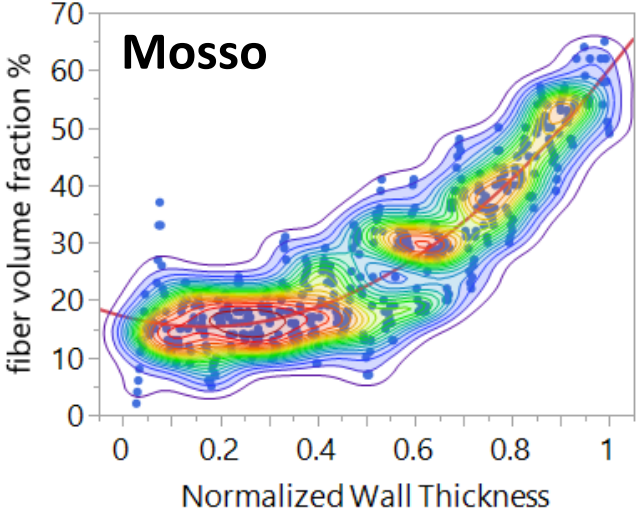
# Polynomial Fit




## Summary of Fit

| | |
|---|---|
| $R^2$ | 0.973955 |
| Root Mean Square Error | 1.056319 |
| Mean of Response | 40.01577 |
| Observations (or Sum Wgts) | 444 |

$$FVF_{DG}\,(\%) = 25.848615 + 24.026391 \cdot NWT + 31.653323 \cdot (NWT - 0.50304)^2$$




## Summary of Fit

| | |
|---|---|
| $R^2$ | 0.806359 |
| Root Mean Square Error | 6.143252 |
| Mean of Response | 27.1617 |
| Observations (or Sum Wgts) | 470 |

$$FVF_{Moso}\,(\%) = 1.176417 + 42.518244 \cdot NWT + 64.610115 \cdot (NWT - 0.49793)^2$$

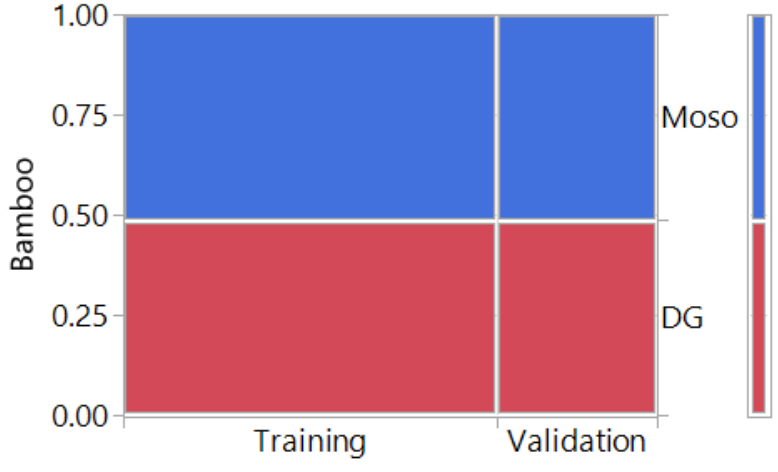
# RMSE

- The mean square error is the quotient of the sum of squares error (SSE) divided by the degrees of freedom of the error (df).
- the standard deviation of the unexplained variance.
- in the same units as the response variable.
- The degrees of freedom of the error is the degrees of freedom for the corrected total minus the degrees of freedom of the model.
- The degrees of freedom for the corrected total is N - 1. (The correction for the mean response costs one degree of freedom.)

$$RMSE = \sqrt{\frac{SSE}{df}}$$

# TRAINING AND VALIDATION SPLIT TO FIT THE MODELS

**Mosaic Plot**



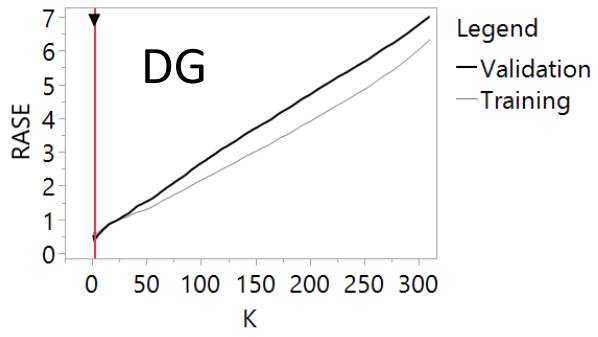

**Contingency Table**

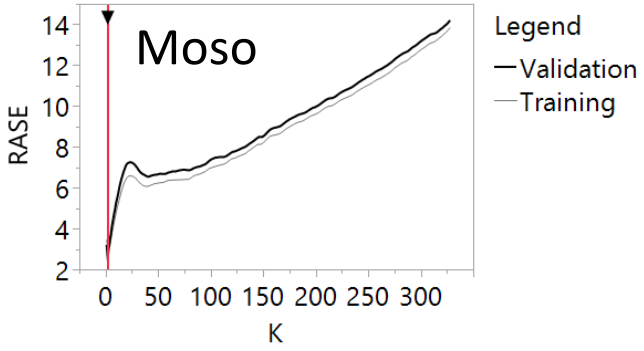
| Validation-continious7030 | Bamboo | DG | Moso | Total |
|---|---|---|---|---|
| | Count | 311 | 329 | 640 |
| | Total % | 34.03 | 36.00 | 70.02 |
| | Col % | 70.05 | 70.00 | |
| | Row % | 48.59 | 51.41 | |
| Training | | | | |
| Validation | Count | 133 | 141 | 274 |
| | Total % | 14.55 | 15.43 | 29.98 |
| | Col % | 29.95 | 30.00 | |
| | Row % | 48.54 | 51.46 | |
| Total | Count | 444 | 470 | 914 |
| | Total % | 48.58 | 51.42 | |
| | Col % | | | |
| | Row % | | | |

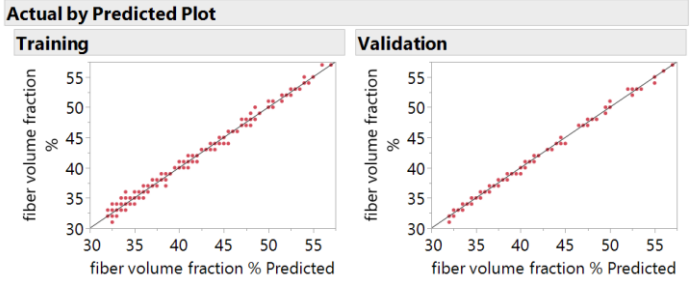
# The Models

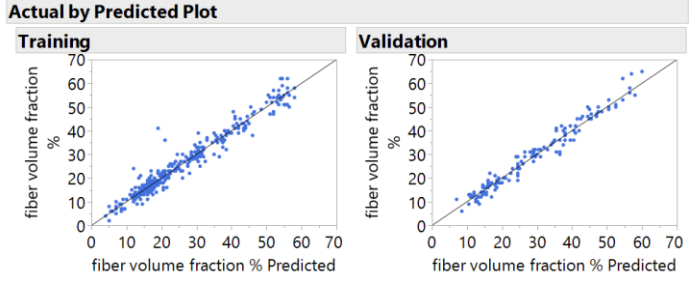
- KNN
- Decision Tree
- Bootstrap Forest
- XGBoost (5-Fold CV or Training/Validation Split)

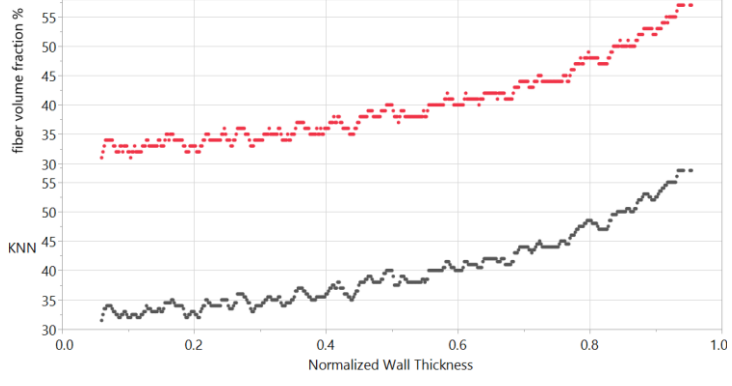
# KNN

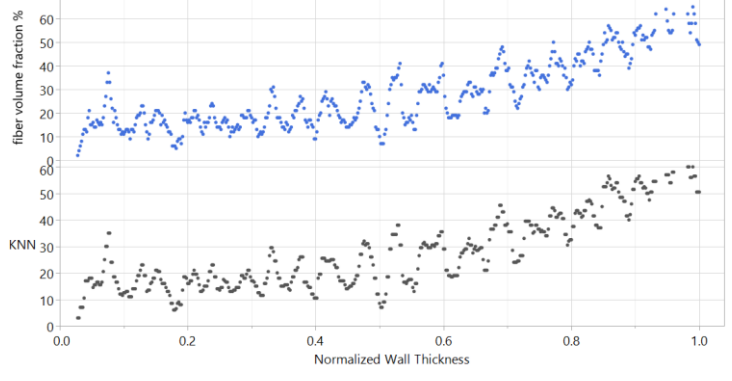


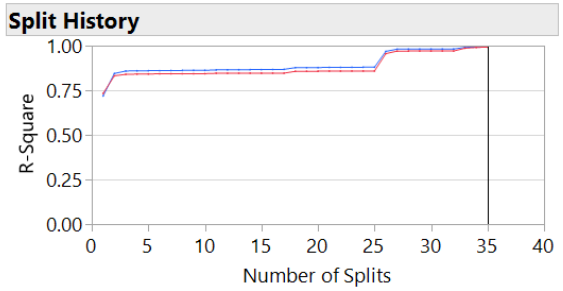



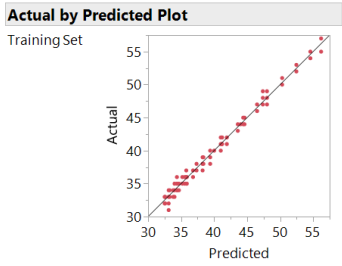



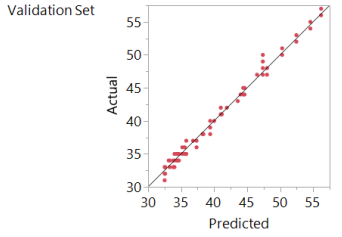

**RMSE (RASE) = 0.3902 (K=2)**

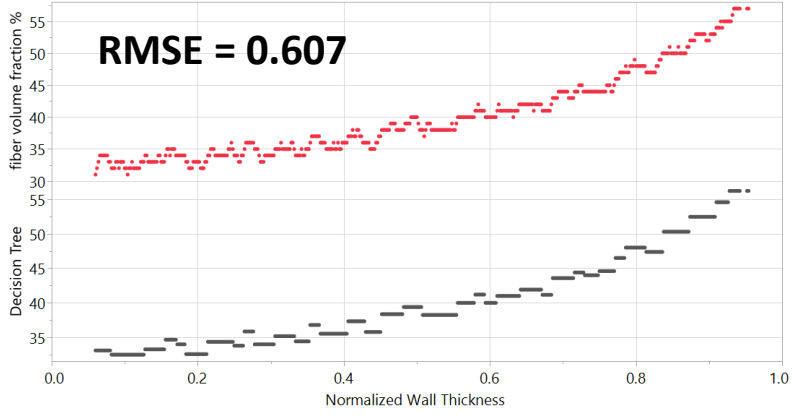
**RMSE (RASE) = 2.567 (K=2)**

# Decision Tree. DG

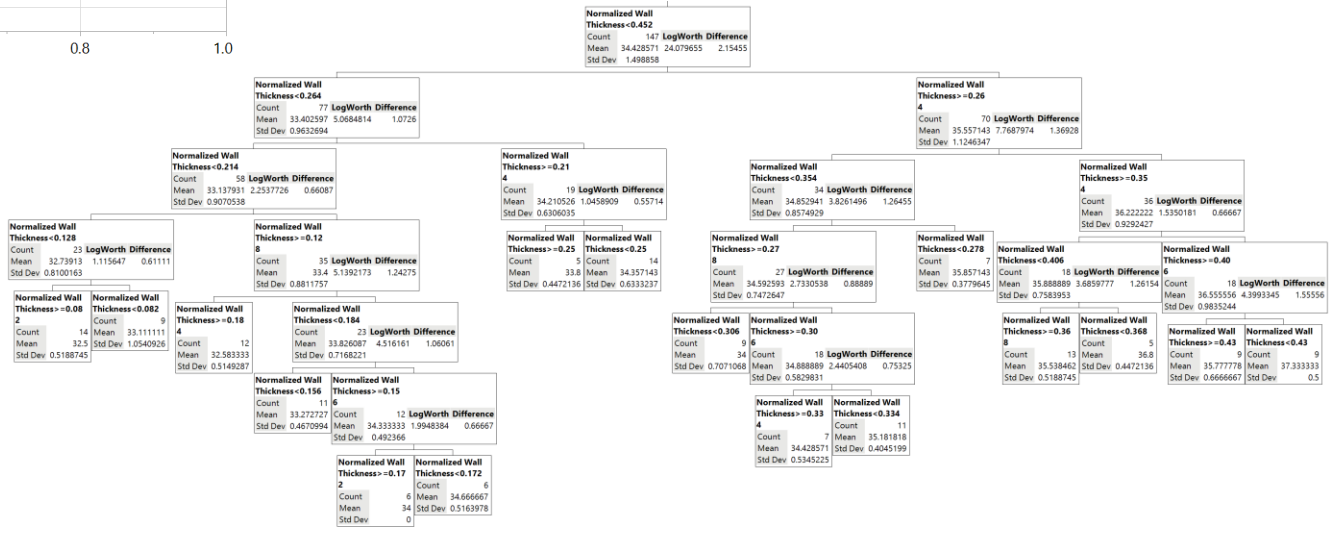


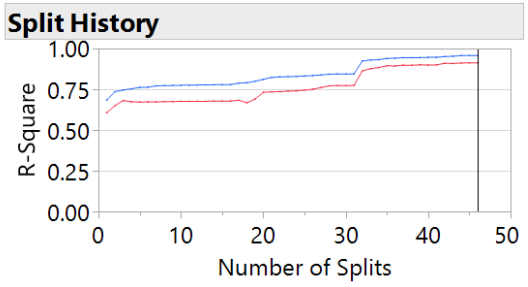



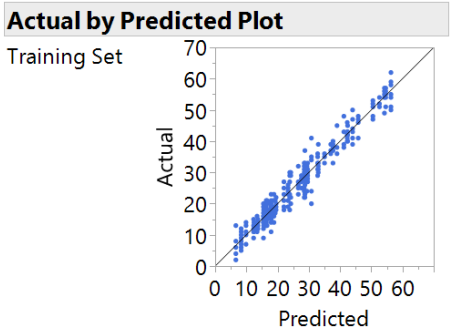

Validation Data in Red

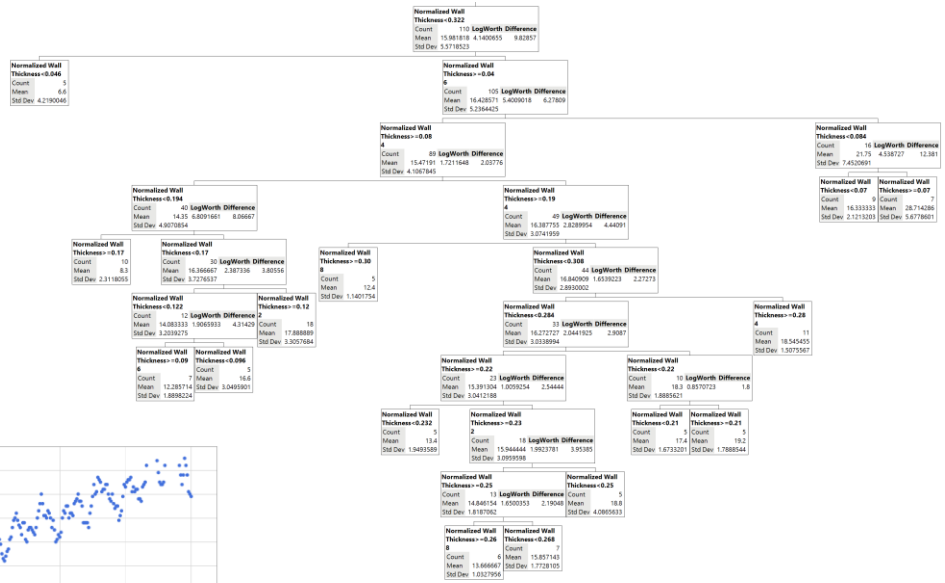


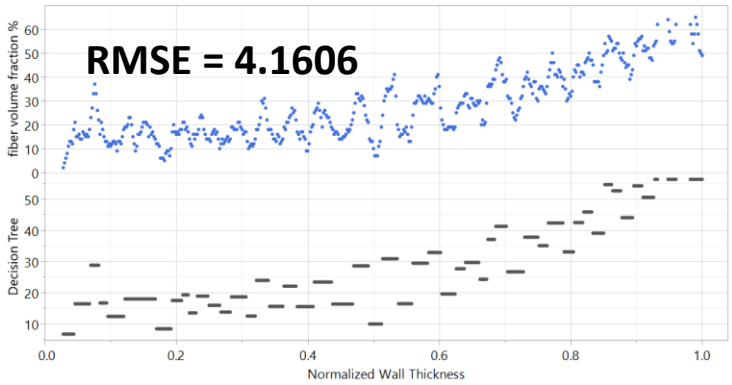
# Decision Tree. Moso



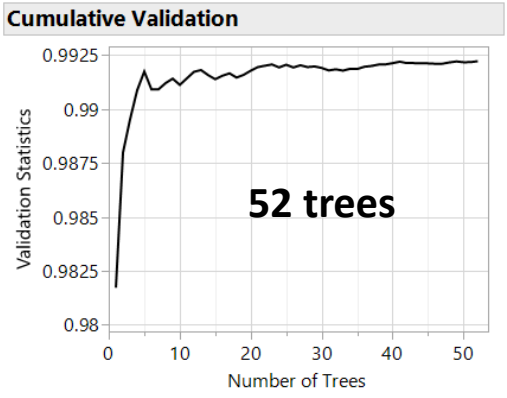

Validation Data in Red

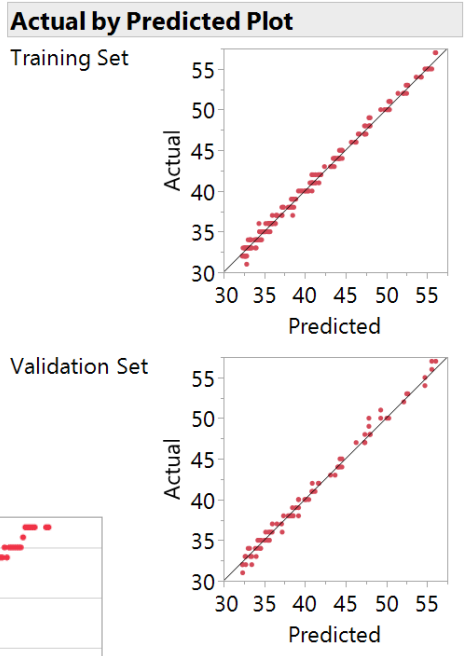


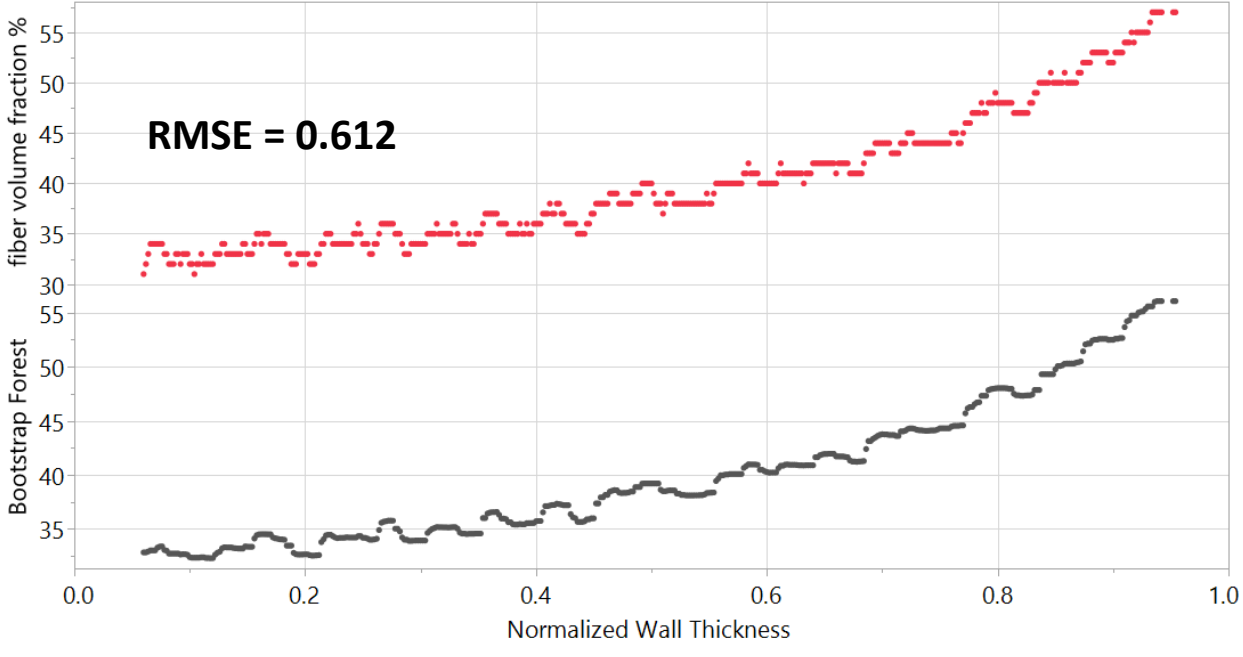





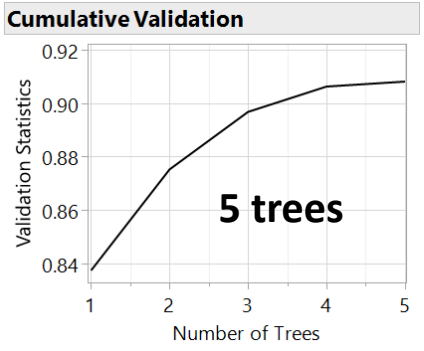
# BOOTSTRAP FOREST. DG

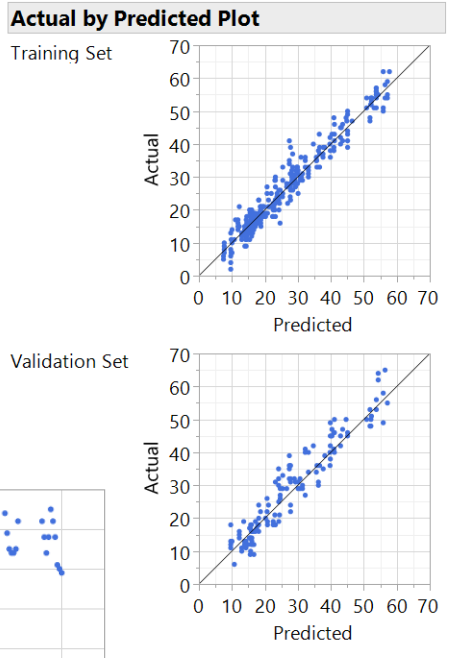


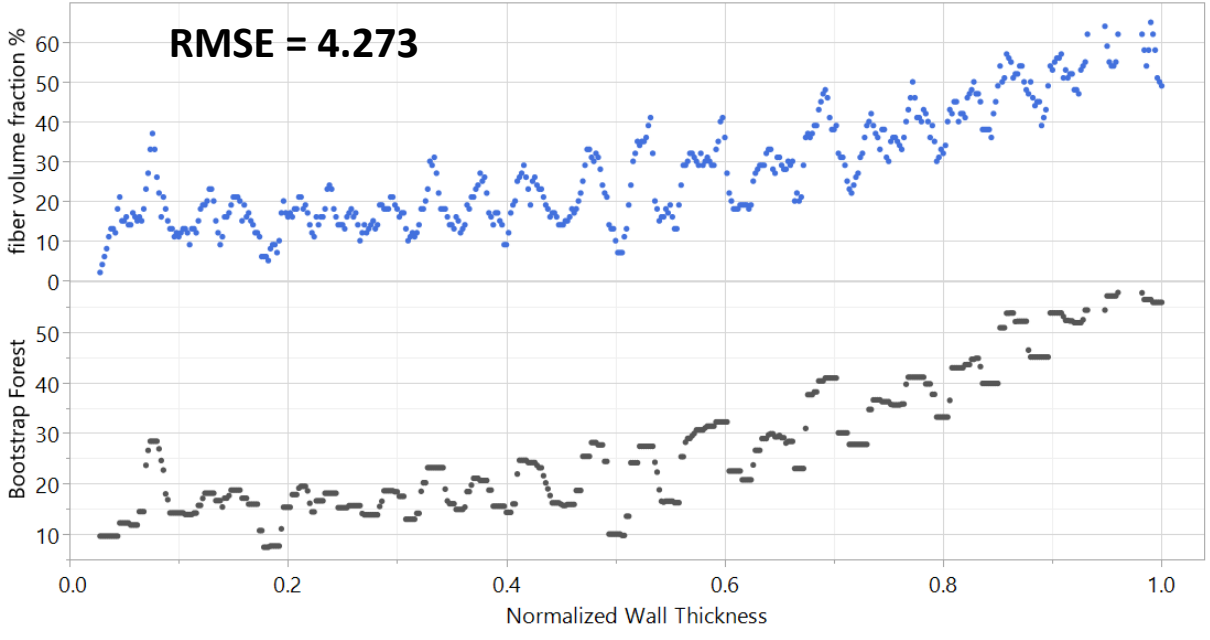



# BOOTSTRAP FOREST. Moso



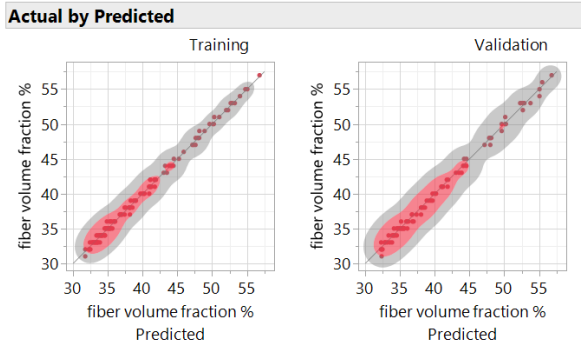



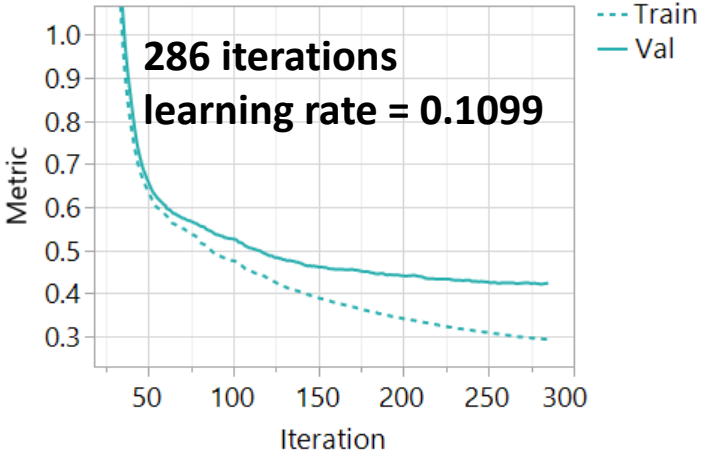



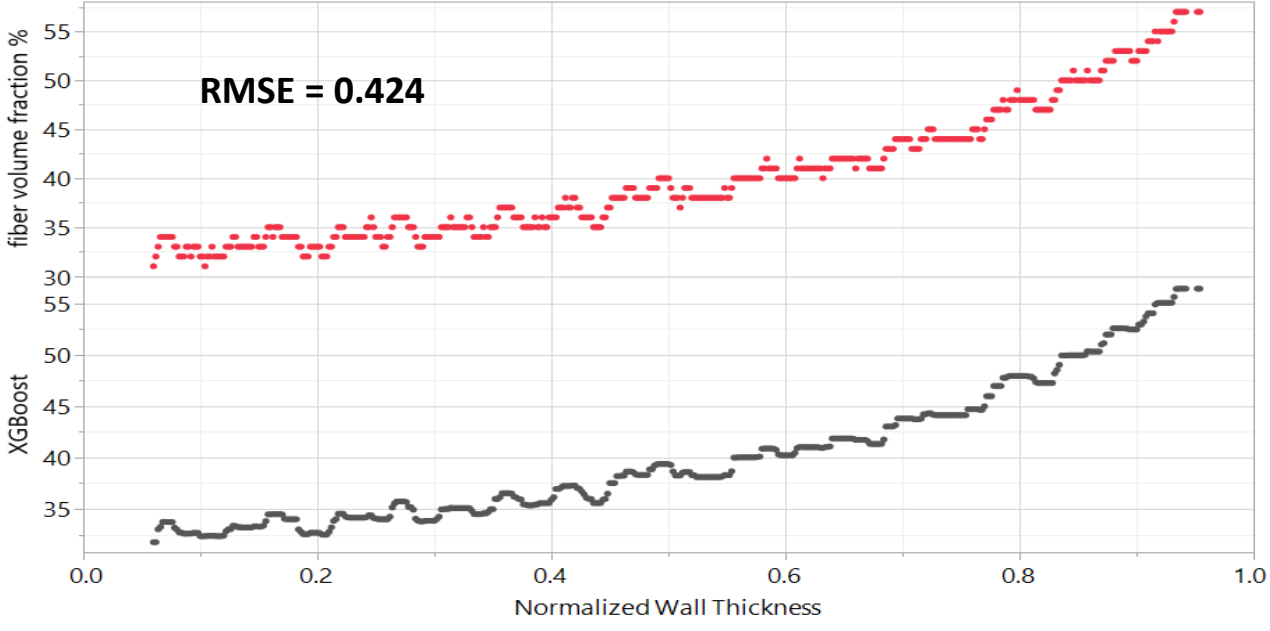
## What is XGBoost?

- XGBoost is a scalable, portable, distributed, open-source C++ library for gradient boosted tree prediction.
- The original theory and applications were developed by Leo Breiman and Jerry Friedman in the late 1990s.
- XGBoost sprang from a research project at the University of Washington around 2015 and is now sponsored by Amazon, NVIDIA, and others;
- XGBoost has grown dramatically in popularity due to successes in nearly every major Kaggle competition and others with tabular data over the past five years;
- It has also been the top performer in several published studies, including the following result from https://github.com/rhiever/sklearn-benchmarks.

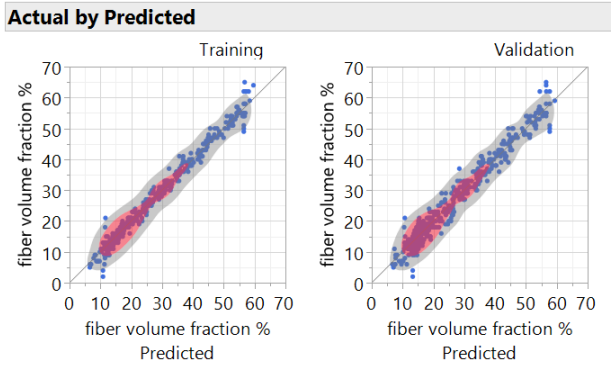
# XGBoost. DG.
# Training (70%)-Validation(30%)

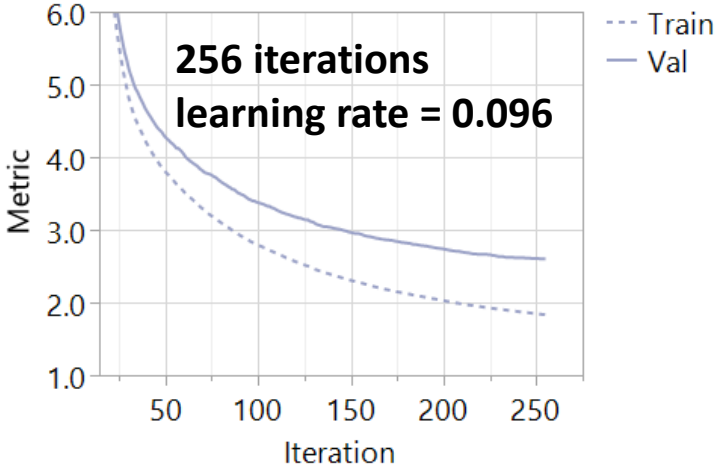


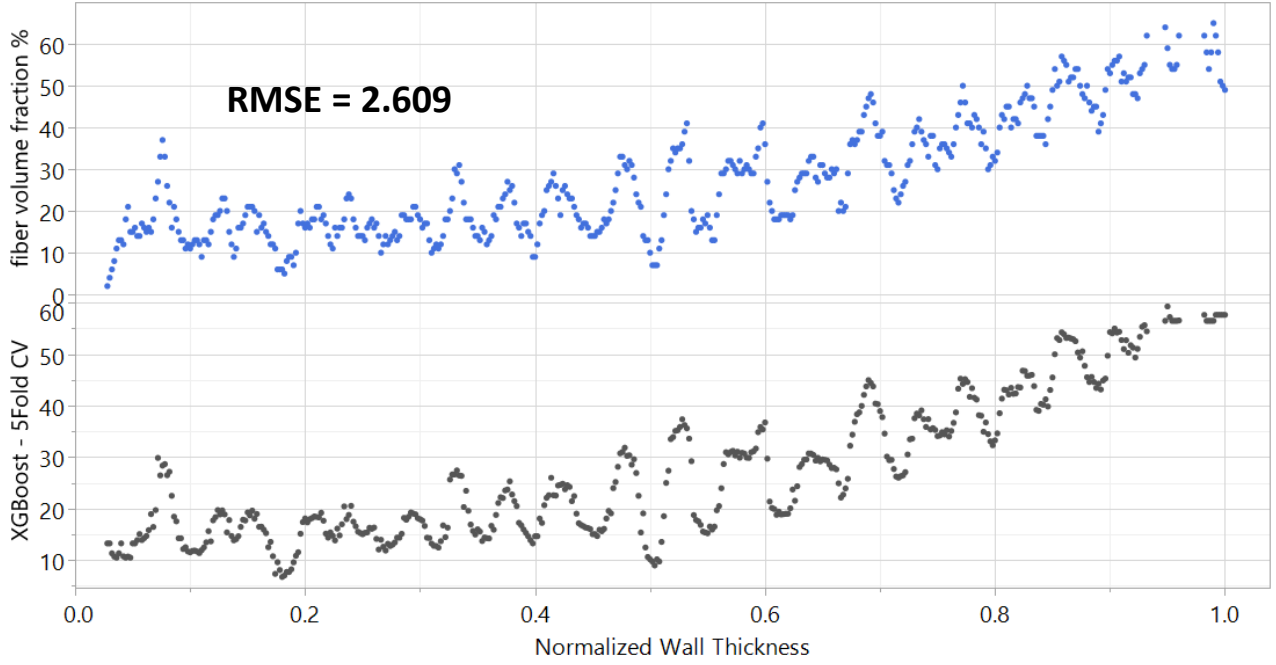



# XGBoost. Moso
# 5-Fold Cross-Validation



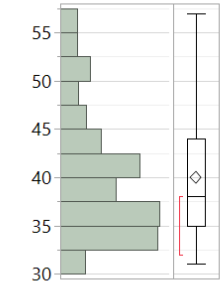



# The values distribution. Real vs Predicted. DG

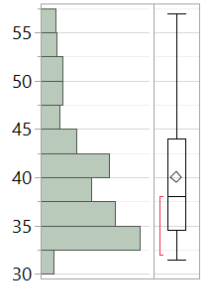
**Distributions**

**fiber volume fraction %**



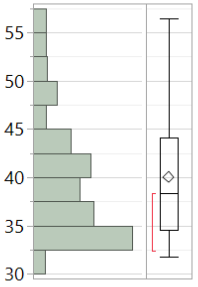

| Summary Statistics | |
|---|---|
| Mean | 40.015766 |
| Std Dev | 6.5305135 |
| Std Err Mean | 0.3099245 |
| Upper 95% Mean | 40.624871 |
| Lower 95% Mean | 39.406661 |
| N | 444 |

**KNN**

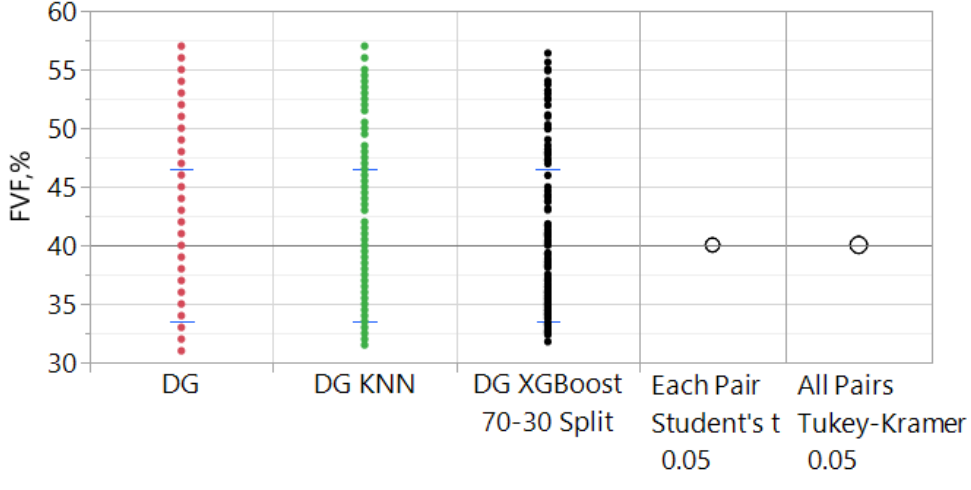



| Summary Statistics | |
|---|---|
| Mean | 40.041667 |
| Std Dev | 6.5194585 |
| Std Err Mean | 0.3093998 |
| Upper 95% Mean | 40.64974 |
| Lower 95% Mean | 39.433593 |
| N | 444 |

**XGBoost 70-30 Split**




| Summary Statistics | |
|---|---|
| Mean | 40.03405 |
| Std Dev | 6.4766136 |
| Std Err Mean | 0.3073665 |
| Upper 95% Mean | 40.638128 |
| Lower 95% Mean | 39.429973 |
| N | 444 |



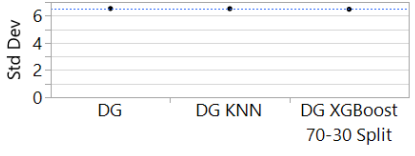

**Kruskal-Wallis Tests (Rank Sums)**

| Level | Count | Score Sum | Expected Score | Score Mean | (Mean-Mean0)/Std0 |
|---|---|---|---|---|---|
| DG | 444 | 294278 | 295926 | 662.788 | -0.249 |
| DG KNN | 444 | 295389 | 295926 | 665.291 | -0.081 |
| DG XGBoost 70-30 Split | 444 | 298111 | 295926 | 671.421 | 0.330 |

**1-Way Test, ChiSquare Approximation**

| ChiSquare | DF | Prob>ChiSq |
|---|---|---|
| 0.1185 | 2 | 0.9425 |

**Tests that the Variances are Equal**




| Level | Count | Std Dev | MeanAbsDif to Mean | MeanAbsDif to Median |
|---|---|---|---|---|
| DG | 444 | 6.530513 | 5.347932 | 5.218468 |
| DG KNN | 444 | 6.519458 | 5.347035 | 5.219595 |
| DG XGBoost 70-30 Split | 444 | 6.476614 | 5.326134 | 5.180479 |

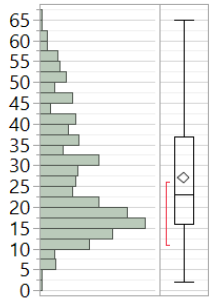
| Test | F Ratio | DFNum | DFDen | Prob > F |
|---|---|---|---|---|
| O'Brien[.5] | 0.0193 | 2 | 1329 | 0.9809 |
| Brown-Forsythe | 0.0116 | 2 | 1329 | 0.9884 |
| Levene | 0.0049 | 2 | 1329 | 0.9951 |
| Bartlett | 0.0169 | 2 | . | 0.9832 |

# The values distribution. Real vs Predicted.

## Moso

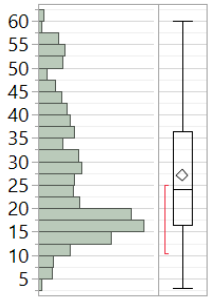
**Distributions**

**fiber volume fraction %**



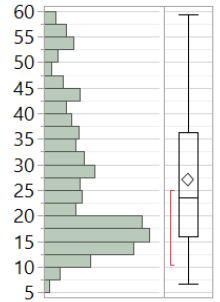
| Summary Statistics | |
|---|---|
| Mean | 27.161702 |
| Std Dev | 13.930662 |
| Std Err Mean | 0.6425735 |
| Upper 95% Mean | 28.424382 |
| Lower 95% Mean | 25.899023 |
| N | 470 |

**KNN**

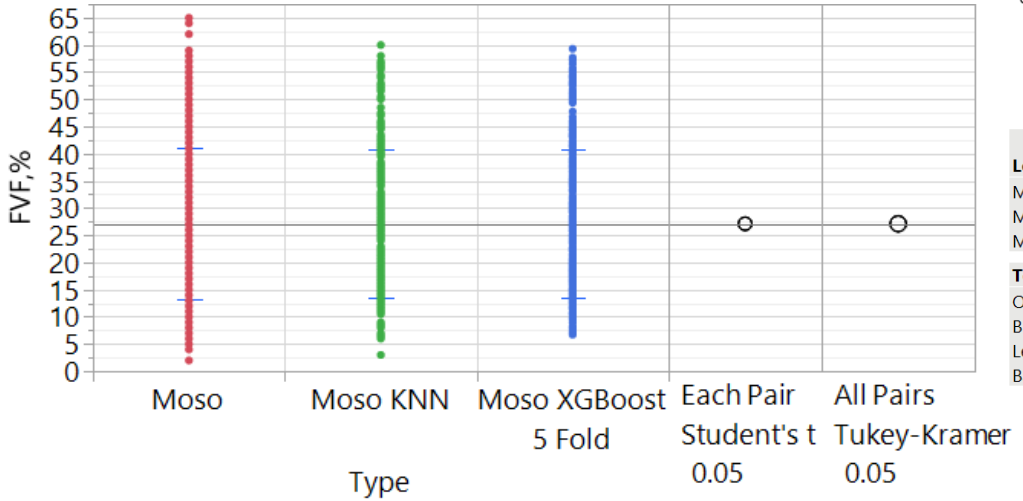


| Summary Statistics | |
|---|---|
| Mean | 27.096809 |
| Std Dev | 13.60522 |
| Std Err Mean | 0.627562 |
| Upper 95% Mean | 28.32999 |
| Lower 95% Mean | 25.863627 |
| N | 470 |

**XGBoost - 5Fold CV**



| Summary Statistics | |
|---|---|
| Mean | 27.11815 |
| Std Dev | 13.556034 |
| Std Err Mean | 0.6252932 |
| Upper 95% Mean | 28.346873 |
| Lower 95% Mean | 25.889427 |
| N | 470 |



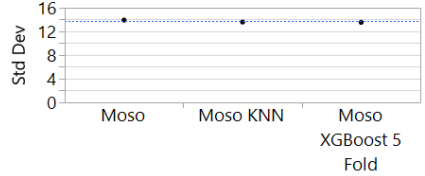

**Kruskal-Wallis Tests (Rank Sums)**

| Level | Count | Score Sum | Expected Score | Score Mean | (Mean-Mean0)/Std0 |
|---|---|---|---|---|---|
| Moso | 470 | 331394 | 331585 | 705.094 | -0.026 |
| Moso KNN | 470 | 332555 | 331585 | 707.564 | 0.135 |
| Moso XGBoost 5 Fold | 470 | 330806 | 331585 | 703.843 | -0.108 |

**1-Way Test, ChiSquare Approximation**

| ChiSquare | DF | Prob>ChiSq |
|---|---|---|
| 0.0203 | 2 | 0.9899 |

**Tests that the Variances are Equal**

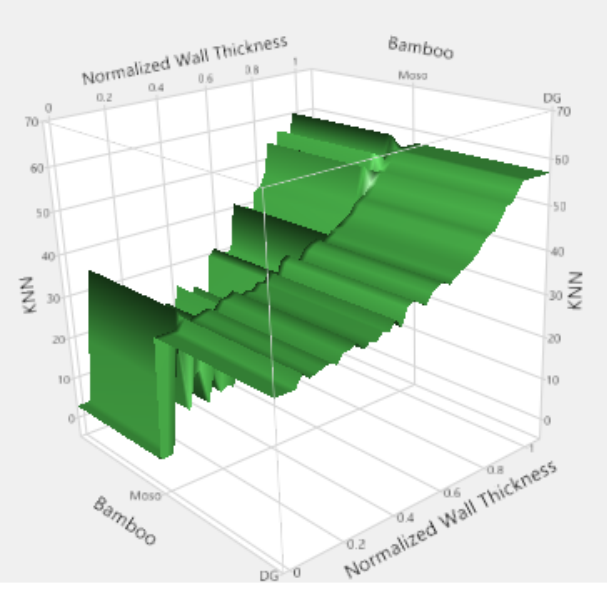



| Level | Count | Std Dev | MeanAbsDif to Mean | MeanAbsDif to Median |
|---|---|---|---|---|
| Moso | 470 | 13.93066 | 11.75393 | 11.48085 |
| Moso KNN | 470 | 13.60522 | 11.48569 | 11.21809 |
| Moso XGBoost 5 Fold | 470 | 13.55603 | 11.48107 | 11.23563 |

| Test | F Ratio | DFNum | DFDen | Prob > F |
|---|---|---|---|---|
| O'Brien[.5] | 0.2925 | 2 | 1407 | 0.7465 |
| Brown-Forsythe | 0.1395 | 2 | 1407 | 0.8698 |
| Levene | 0.2148 | 2 | 1407 | 0.8067 |
| Bartlett | 0.2065 | 2 | . | 0.8135 |

# Summary

| Model | KNN | | Decision Tree | | Bootstrap Forest | | XGBoost (5-Fold CV) | | XGBoost (70/30 split) | | Polynomial Fit | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bamboo Type | DG | Moso | DG | Moso | DG | Moso | DG | Moso | DG | Moso | DG | Moso |
| RMSE | 0.390 | 2.567 | 0.607 | 4.161 | 0.612 | 4.273 | 0.515 | 2.609 | 0.424 | 3.150 | 1.056 | 6.143 |




- **The most effective model is non-parametric KNN;**
- **To predict FVF (%) for Moso, among the parametric models, XGBoost at 5-Fold cross-validation has the lowest RMSE: 2.609**
- **To predict FVF (%) for DG, among the parametric models, XGBoost at normal 70/30 split has the lowest RMSE: 0.424**
- **Three "strongest" models and experimental data have not demonstrated significant difference in the predicted values for both types of bamboo**


researchgate.net/Michael-Nazarkovsky

michael-nazarkovsky

github.com/Nazarkovsky

orcid.org/0000-0002-7166-4184