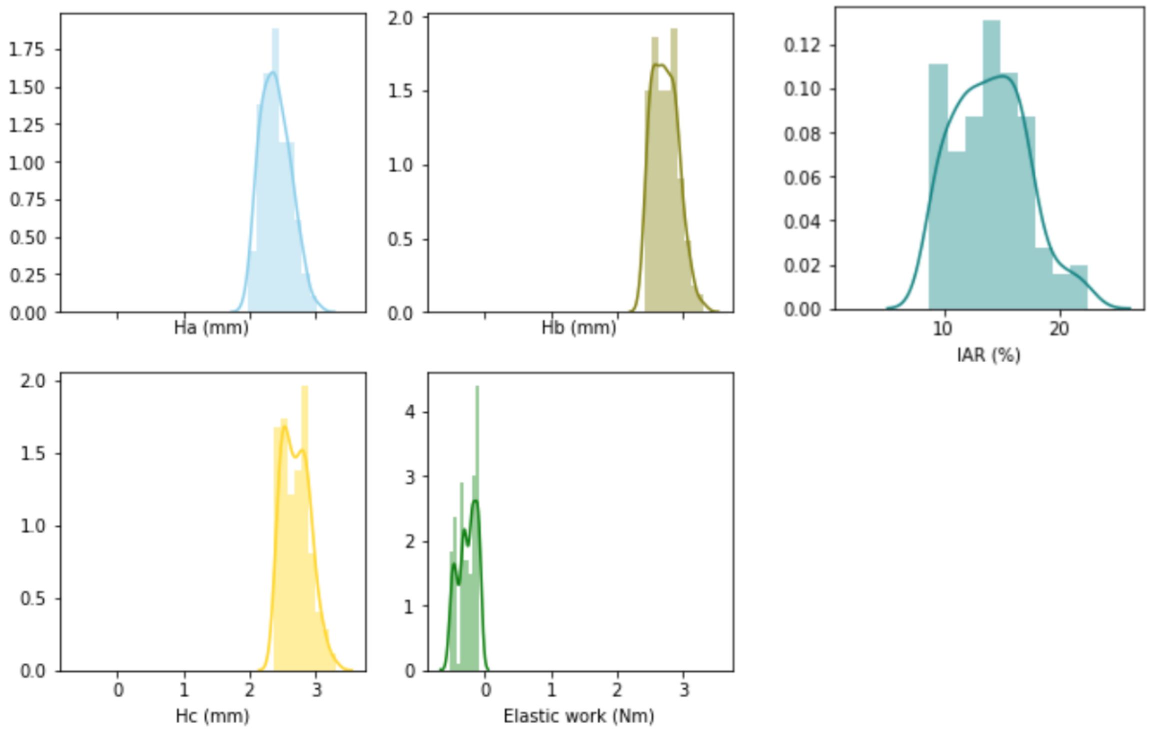
1. Introduction

Many issues in pharmaceutical manufacturing are direct consequence of tablet expansion which is caused by elastic behaviour of materials. To ensure quality of a product and facilitate industrial scale-up and manufacturing processes Quality by design principles must be employed. Prediction of elastic properties of granules if very important in pre-formulation and formulation studies. Therefore, prediction of elastic properties is very important in wet granulation scale-up process. In this experiment I set up few hypothesis testing to understand the data in better way and then, I try to evaluate different predictive models for the same data.

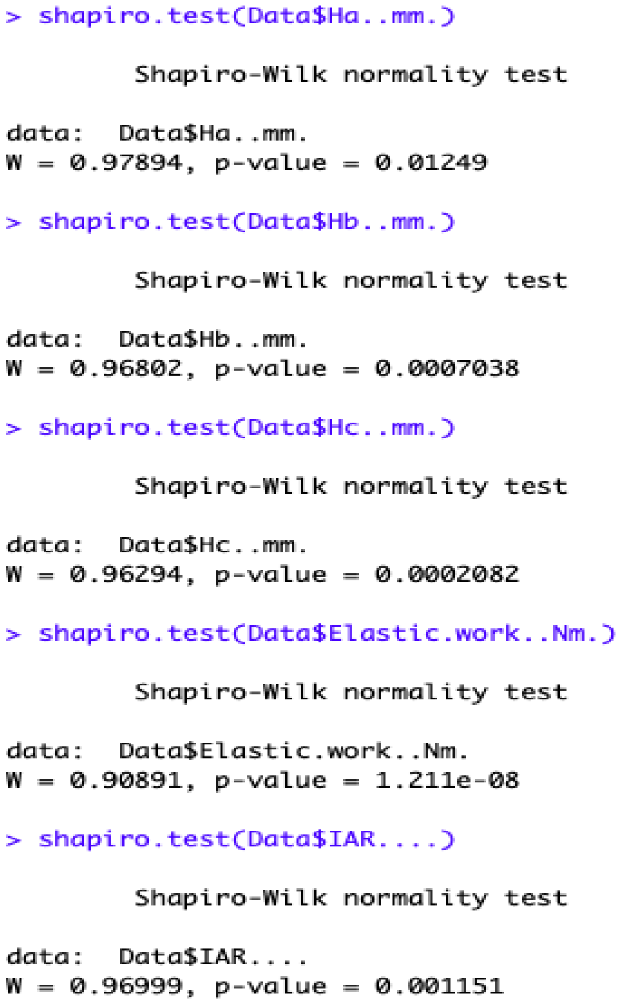
2. Statistical Data Analysis

2.1 Distribution



From the distribution of Response variables (Ha, Hb, Hc, IAR, ElasticWork), it was suspected that the sample is not normally distributed. To confirm, **Shapiro-Wilk test** had been carried out and the result mimic the same way with our initial assumption (p-value < 0.05).

Shapiro-Wilk test Result:

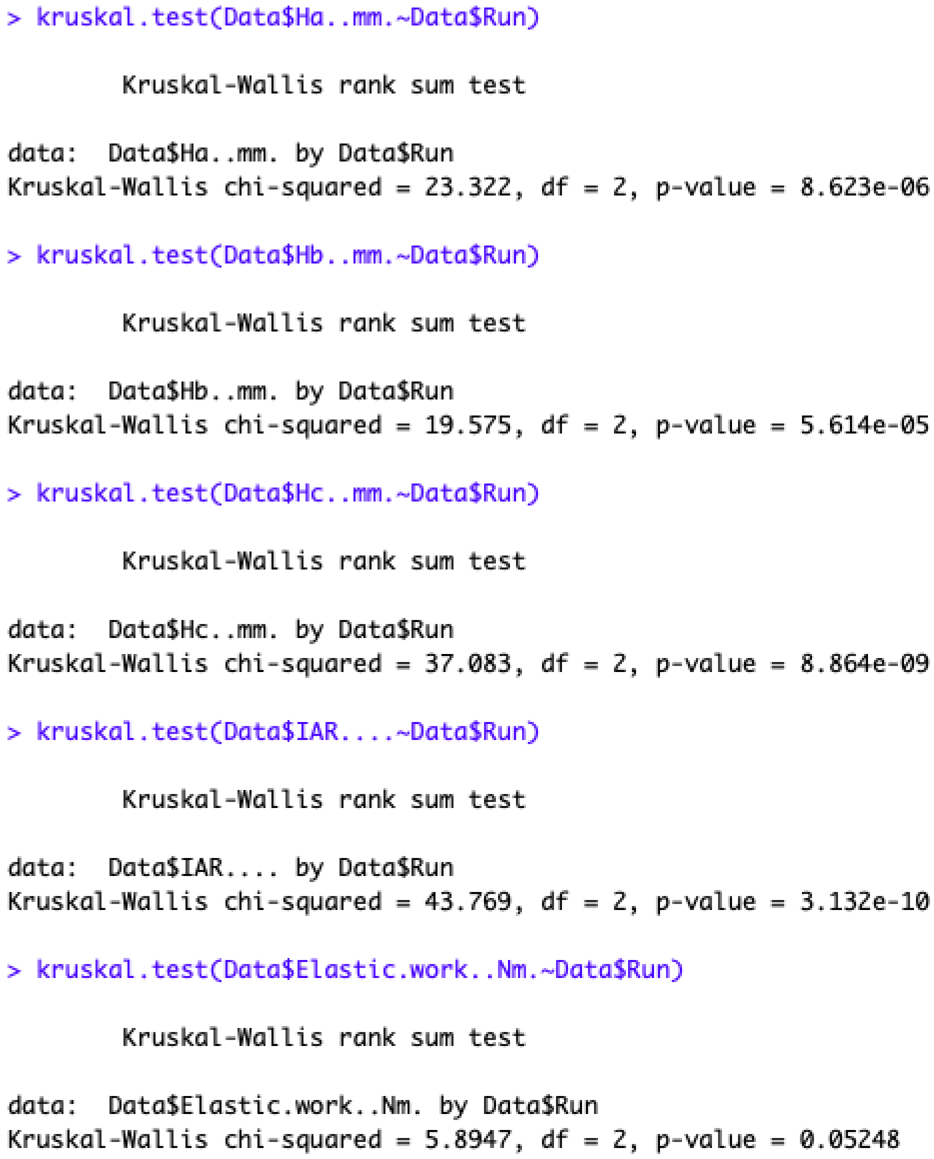


```
> shapiro.test(Data$Ha..mm.)

        Shapiro-Wilk normality test

data:  Data$Ha..mm.
W = 0.97894, p-value = 0.01249

> shapiro.test(Data$Hb..mm.)

        Shapiro-Wilk normality test

data:  Data$Hb..mm.
W = 0.96802, p-value = 0.0007038

> shapiro.test(Data$Hc..mm.)

        Shapiro-Wilk normality test

data:  Data$Hc..mm.
W = 0.96294, p-value = 0.0002082

> shapiro.test(Data$Elastic.work..Nm.)

        Shapiro-Wilk normality test

data:  Data$Elastic.work..Nm.
W = 0.90891, p-value = 1.211e-08

> shapiro.test(Data$IAR....)

        Shapiro-Wilk normality test

data:  Data$IAR....
W = 0.96999, p-value = 0.001151
```

2.2 Data Bucketing

As the response variables are not normally distributed, I may use Non Parametric Approach (Kruskal-Wallis test) to find relation between run (laboratory, commercial, pilot) and the response variables to find whether they should be treated differently.

It was found that test ran in different environment(laboratory, commercial, pilot) has made an different impact on the elastic property of the tablet, so I divided Run into buckets such as Laboratory, Commercial and pilot for our further analysis.

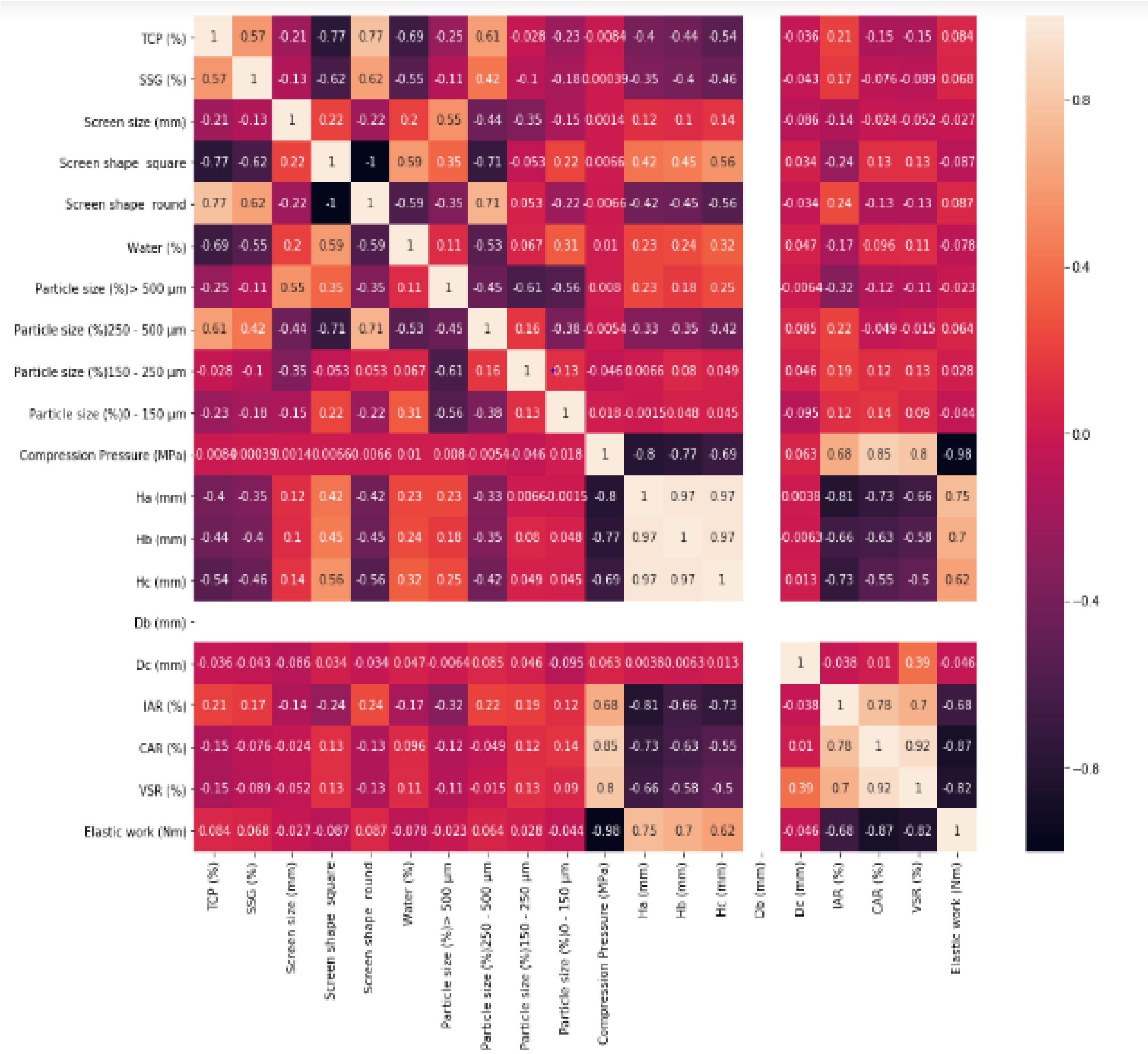


```
> kruskal.test(Data$Ha..mm.~Data$Run)

	Kruskal-Wallis rank sum test

data:  Data$Ha..mm. by Data$Run
Kruskal-Wallis chi-squared = 23.322, df = 2, p-value = 8.623e-06

> kruskal.test(Data$Hb..mm.~Data$Run)

	Kruskal-Wallis rank sum test

data:  Data$Hb..mm. by Data$Run
Kruskal-Wallis chi-squared = 19.575, df = 2, p-value = 5.614e-05

> kruskal.test(Data$Hc..mm.~Data$Run)

	Kruskal-Wallis rank sum test

data:  Data$Hc..mm. by Data$Run
Kruskal-Wallis chi-squared = 37.083, df = 2, p-value = 8.864e-09

> kruskal.test(Data$IAR....~Data$Run)

	Kruskal-Wallis rank sum test

data:  Data$IAR.... by Data$Run
Kruskal-Wallis chi-squared = 43.769, df = 2, p-value = 3.132e-10

> kruskal.test(Data$Elastic.work..Nm.~Data$Run)

	Kruskal-Wallis rank sum test

data:  Data$Elastic.work..Nm. by Data$Run
Kruskal-Wallis chi-squared = 5.8947, df = 2, p-value = 0.05248
```

2.2 Relation between Dependent and Independent Variables

Before modelling, it was required to analyse the impact of independent variables on dependent variables (Ha, Hb, Hc, IAR, ElasticWork). As part of that process I am trying to get the correlation matrix between these variables.

Correlation Matrix:



	TCP (%)	SSG (%)	Screen size (mm)	Screen shape square	Screen shape round	Water (%)	Particle size (%)> 500 µm	Particle size (%)250 - 500 µm	Particle size (%)150 - 250 µm	Particle size (%)0 - 150 µm	Compression Pressure (MPa)	Ha (mm)	Hb (mm)	Hc (mm)	Db (mm)	Dc (mm)	IAR (%)	CAR (%)	VSR (%)	Elastic work (Nm)
TCP (%)	1	0.57	-0.21	-0.77	0.77	-0.69	-0.25	0.61	-0.028	-0.23	-0.0084	-0.4	-0.44	-0.54		-0.036	0.21	-0.15	-0.15	0.084
SSG (%)	0.57	1	-0.13	-0.62	0.62	-0.55	-0.11	0.42	-0.1	-0.18	0.00039	-0.35	-0.4	-0.46		-0.043	0.17	-0.076	-0.089	0.068
Screen size (mm)	-0.21	-0.13	1	0.22	-0.22	0.2	0.55	-0.44	-0.35	-0.15	0.0014	0.12	0.1	0.14		-0.086	-0.14	-0.024	-0.052	-0.027
Screen shape square	-0.77	-0.62	0.22	1	-1	0.59	0.35	-0.71	-0.053	0.22	0.0066	0.42	0.45	0.56		0.034	-0.24	0.13	0.13	-0.087
Screen shape round	0.77	0.62	-0.22	-1	1	-0.59	-0.35	0.71	0.053	-0.22	-0.0066	-0.42	-0.45	-0.56		-0.034	0.24	-0.13	-0.13	0.087
Water (%)	-0.69	-0.55	0.2	0.59	-0.59	1	0.11	-0.53	0.067	0.31	0.01	0.23	0.24	0.32		0.047	-0.17	0.096	0.11	-0.078
Particle size (%)> 500 µm	-0.25	-0.11	0.55	0.35	-0.35	0.11	1	-0.45	-0.61	-0.56	0.008	0.23	0.18	0.25		-0.0064	-0.32	-0.12	-0.11	-0.023
Particle size (%)250 - 500 µm	0.61	0.42	-0.44	-0.71	0.71	-0.53	-0.45	1	0.16	-0.38	-0.0054	-0.33	-0.35	-0.42		0.085	0.22	-0.049	-0.015	0.064
Particle size (%)150 - 250 µm	-0.028	-0.1	-0.35	-0.053	0.053	0.067	-0.61	0.16	1	0.13	-0.046	0.0066	0.08	0.049		0.046	0.19	0.12	0.13	0.028
Particle size (%)0 - 150 µm	-0.23	-0.18	-0.15	0.22	-0.22	0.31	-0.56	-0.38	0.13	1	0.018	-0.0015	0.048	0.045		-0.095	0.12	0.14	0.09	-0.044
Compression Pressure (MPa)	-0.0084	0.00039	0.0014	0.0066	-0.0066	0.01	0.008	-0.0054	-0.046	0.018	1	-0.8	-0.77	-0.69		0.063	0.68	0.85	0.8	-0.98
Ha (mm)	-0.4	-0.35	0.12	0.42	-0.42	0.23	0.23	-0.33	0.0066	-0.0015	-0.8	1	0.97	0.97		0.0038	-0.81	-0.73	-0.66	0.75
Hb (mm)	-0.44	-0.4	0.1	0.45	-0.45	0.24	0.18	-0.35	0.08	0.048	-0.77	0.97	1	0.97		-0.0063	-0.66	-0.63	-0.58	0.7
Hc (mm)	-0.54	-0.46	0.14	0.56	-0.56	0.32	0.25	-0.42	0.049	0.045	-0.69	0.97	0.97	1		0.013	-0.73	-0.55	-0.5	0.62
Db (mm)																				
Dc (mm)	-0.036	-0.043	-0.086	0.034	-0.034	0.047	-0.0064	0.085	0.046	-0.095	0.063	0.0038	-0.0063	0.013		1	-0.038	0.01	0.39	-0.046
IAR (%)	0.21	0.17	-0.14	-0.24	0.24	-0.17	-0.32	0.22	0.19	0.12	0.68	-0.81	-0.66	-0.73		-0.038	1	0.78	0.7	-0.68
CAR (%)	-0.15	-0.076	-0.024	0.13	-0.13	0.096	-0.12	-0.049	0.12	0.14	0.85	-0.73	-0.63	-0.55		0.01	0.78	1	0.92	-0.87
VSR (%)	-0.15	-0.089	-0.052	0.13	-0.13	0.11	-0.11	-0.015	0.13	0.09	0.8	-0.66	-0.58	-0.5		0.39	0.7	0.92	1	-0.82
Elastic work (Nm)	0.084	0.068	-0.027	-0.087	0.087	-0.078	-0.023	0.064	0.028	-0.044	-0.98	0.75	0.7	0.62		-0.046	-0.68	-0.87	-0.82	1

Correlation Matrix of the Data was used for feature selection

3. Models:

I have carried out a regression analysis on the input and output variables as given below:

Input Variables:

1. TCP (%)
2. SSG (%)
3. Screen size (mm)
4. Screen shape square
5. Screen shape round
6. Water (%)
7. Particle size (%)> 500 µm
8. Particle size (%)250 - 500 µm
9. Particle size (%)150 - 250 µm
10. Particle size (%)0 - 150 µm
11. Compression Pressure (MPa)

Output Variable:

1. Ha (mm)
2. Hb (mm)
3. Hc (mm)
4. IAR (%)
5. Elastic work (Nm)

I have evaluated the following regression based algorithms for predicting response variables.

- Step-wise Liner Regression
- Linear Regression with Ridge
- Linear Regression with Lasso
- Linear Regression with ElasticNet

I have a concern of overfitting for this small dataset; so I thought to use regularisation variants of regression algorithms for my analysis.

3.1 Model for Response Ha:

3.1.1 Laboratory scale

Input Variables Considered (Best Score/Low MSE):
'TCP (%)','Compression Pressure (MPa)','Particle size (%)150 - 250 µm','Screen shape round'

Output Variable:
Ha

Model with Best Score/ Low MSE:
0.8425 / 0.03 (Linear Regression)

3.1.2 Pilot scale

Input Variables Considered (Best Score/Low MSE):
'TCP (%)','Compression Pressure (MPa)','Particle size (%)150 - 250 µm','Screen shape round'

Output Variable:
Ha

Model with Best Score/ Low MSE:
0.8881 / 0.01 (Linear Regression with Ridge)

3.1.3 Commercial scale

Input Variables Considered (Best Score/Low MSE):
'TCP (%)','Compression Pressure (MPa)','Particle size (%)150 - 250 µm','Screen shape round'

Output Variable:
Ha

Model with Best Score/ Low MSE:

0.8508 / 0.01 (Linear Regression)

3.2 Model for Response Hb:

3.2.1 Laboratory scale

Input Variables Considered (Best Score/Low MSE):
'TCP (%)','Compression Pressure (MPa)','Screen shape round','Particle size (%)150 - 250 µm'

Output Variable:
Hb

Model with Best Score/ Low MSE:
0.7873 / 0.03 (Linear Regression)

3.2.2 Pilot scale

Input Variables Considered (Best Score/Low MSE):
'TCP (%)','Compression Pressure (MPa)','Screen shape round'

Output Variable:
Hb

Model with Best Score/ Low MSE:
0.8843 / 0.004 (Linear Regression)

3.2.3 Commercial scale

Input Variables Considered (Best Score/Low MSE):
'Compression Pressure (MPa)'

Output Variable:
Hb

Model with Best Score/ Low MSE:
0.8736 / 0.003 (Linear Regression with ElasticNet)

3.3 Model for Response Hc:

3.3.1 Laboratory scale

Input Variables Considered (Best Score/Low MSE):
'TCP (%)','Compression Pressure (MPa)','Particle size (%)150 - 250 µm','Screen shape round'

Output Variable:
Hc

Model with Best Score/ Low MSE:
0.7786 / 0.03 (Linear Regression)

3.3.2 Pilot scale

Input Variables Considered (Best Score/Low MSE):

'TCP (%)','Compression Pressure (MPa)','Particle size (%)150 - 250 µm','Screen shape round'

Output Variable:
Hc

Model with Best Score/ Low MSE:
0.8461 / 0.01 (Linear Regression with Ridge)

3.3.3 Commercial scale

Input Variables Considered (Best Score/Low MSE):
'TCP (%)','Compression Pressure (MPa)','Particle size (%)150 - 250 µm','Screen shape round'

Output Variable:
Hc

Model with Best Score/ Low MSE:
0.8534 / 0.004 (Linear Regression)

3.4 Model for Response IAR:

3.4.1 Laboratory scale

Input Variables Considered (Best Score/Low MSE):
'TCP (%)','Compression Pressure (MPa)','Particle size (%)150 - 250 µm','Screen shape round'

Output Variable:
IAR

Model with Best Score/ Low MSE:
0.9052 / 1.47 (Linear Regression)

3.4.2 Pilot scale

Input Variables Considered (Best Score/Low MSE):
'TCP (%)','Compression Pressure (MPa)','Particle size (%)150 - 250 µm','Screen shape round'

Output Variable:
IAR

Model with Best Score/ Low MSE:
0.9168 / 0.63 (Linear Regression with Ridge)

3.4.3 Commercial scale

Input Variables Considered (Best Score/Low MSE):

'TCP (%)','SSG (%)','Particle size (%)> 500 µm','Compression Pressure (MPa)','Particle size (%)150 - 250 µm','Screen shape round'

Output Variable:

IAR

Model with Best Score/ Low MSE:

0.4556 / 7.55 (Linear Regression)

3.5 Model for Response Elastic Work:

3.5.1 Laboratory scale

Input Variables Considered (Best Score/Low MSE):

'Compression Pressure (MPa)'

Output Variable:

Elastic Work

Model with Best Score/ Low MSE:

0.9576 / 0.001 (Linear Regression)

3.5.2 Pilot scale

Input Variables Considered (Best Score/Low MSE):

'Compression Pressure (MPa)'

Output Variable:

Elastic Work

Model with Best Score/ Low MSE:

0.9858 / 0.0002 (Linear Regression)

3.5.3 Commercial scale

Input Variables Considered (Best Score/Low MSE):

'Compression Pressure (MPa)'

Output Variable:

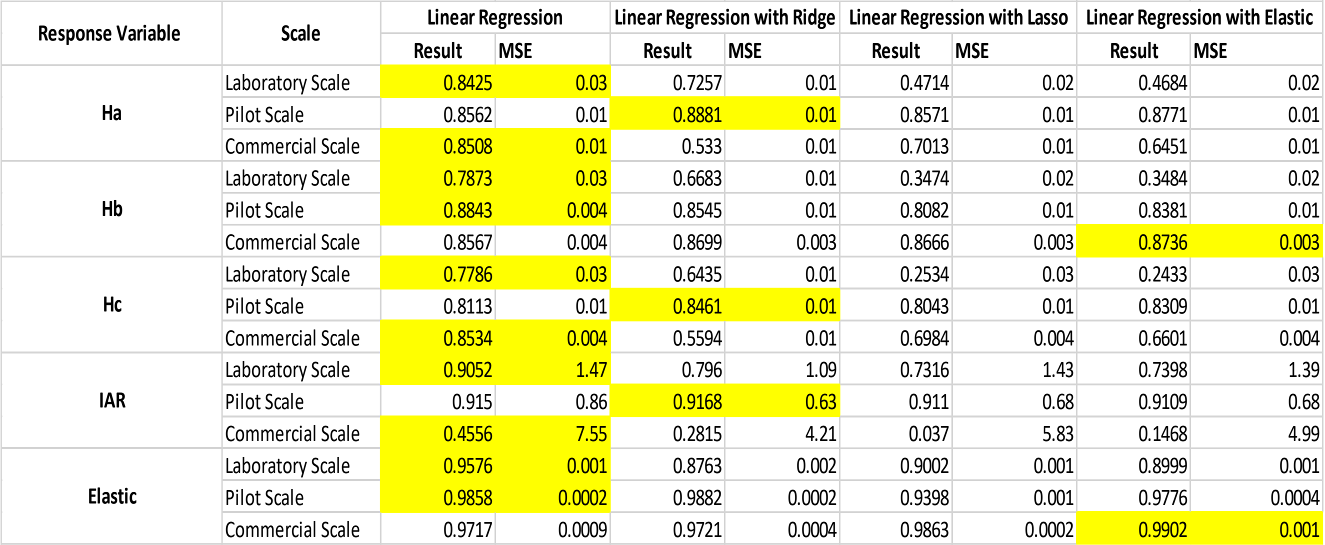
Elastic Work

Model with Best Score/ Low MSE:

0.9902 / 0.001 (Linear Regression with ElasticNet

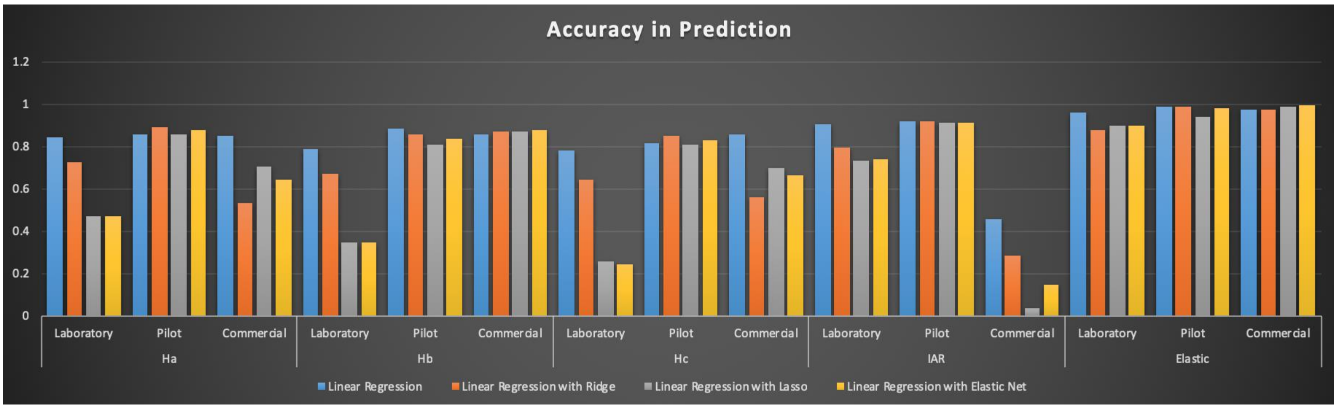
4. Results:

The table describes the Accuracy and Mean Squared Error (MSE) from all four different regression models built for four response variables. The best results are highlighted in the table.



Response Variable	Scale	Linear Regression		Linear Regression with Ridge		Linear Regression with Lasso		Linear Regression with Elastic	
		Result	MSE	Result	MSE	Result	MSE	Result	MSE
Ha	Laboratory Scale	0.8425	0.03	0.7257	0.01	0.4714	0.02	0.4684	0.02
	Pilot Scale	0.8562	0.01	0.8881	0.01	0.8571	0.01	0.8771	0.01
	Commercial Scale	0.8508	0.01	0.533	0.01	0.7013	0.01	0.6451	0.01
Hb	Laboratory Scale	0.7873	0.03	0.6683	0.01	0.3474	0.02	0.3484	0.02
	Pilot Scale	0.8843	0.004	0.8545	0.01	0.8082	0.01	0.8381	0.01
	Commercial Scale	0.8567	0.004	0.8699	0.003	0.8666	0.003	0.8736	0.003
Hc	Laboratory Scale	0.7786	0.03	0.6435	0.01	0.2534	0.03	0.2433	0.03
	Pilot Scale	0.8113	0.01	0.8461	0.01	0.8043	0.01	0.8309	0.01
	Commercial Scale	0.8534	0.004	0.5594	0.01	0.6984	0.004	0.6601	0.004
IAR	Laboratory Scale	0.9052	1.47	0.796	1.09	0.7316	1.43	0.7398	1.39
	Pilot Scale	0.915	0.86	0.9168	0.63	0.911	0.68	0.9109	0.68
	Commercial Scale	0.4556	7.55	0.2815	4.21	0.037	5.83	0.1468	4.99
Elastic	Laboratory Scale	0.9576	0.001	0.8763	0.002	0.9002	0.001	0.8999	0.001
	Pilot Scale	0.9858	0.0002	0.9882	0.0002	0.9398	0.001	0.9776	0.0004
	Commercial Scale	0.9717	0.0009	0.9721	0.0004	0.9863	0.0002	0.9902	0.001

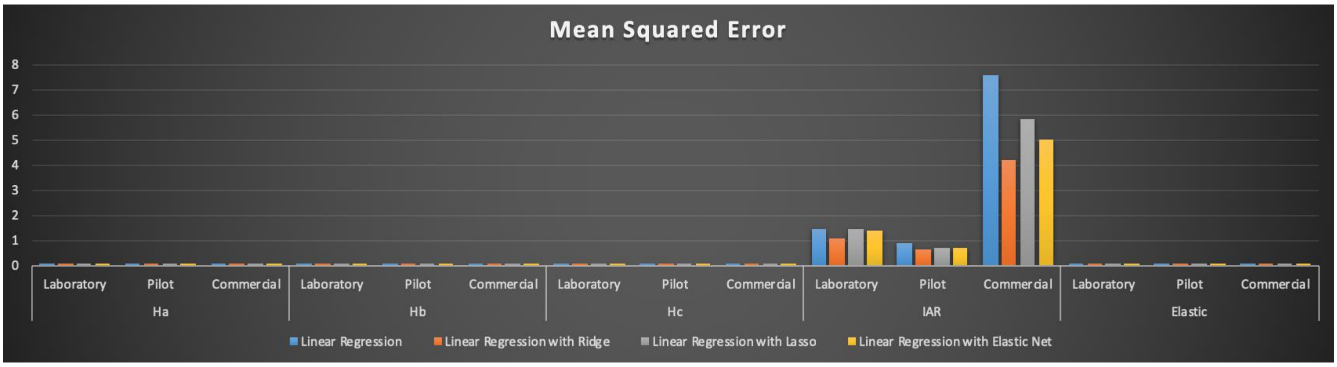
4.1 Model Accuracy

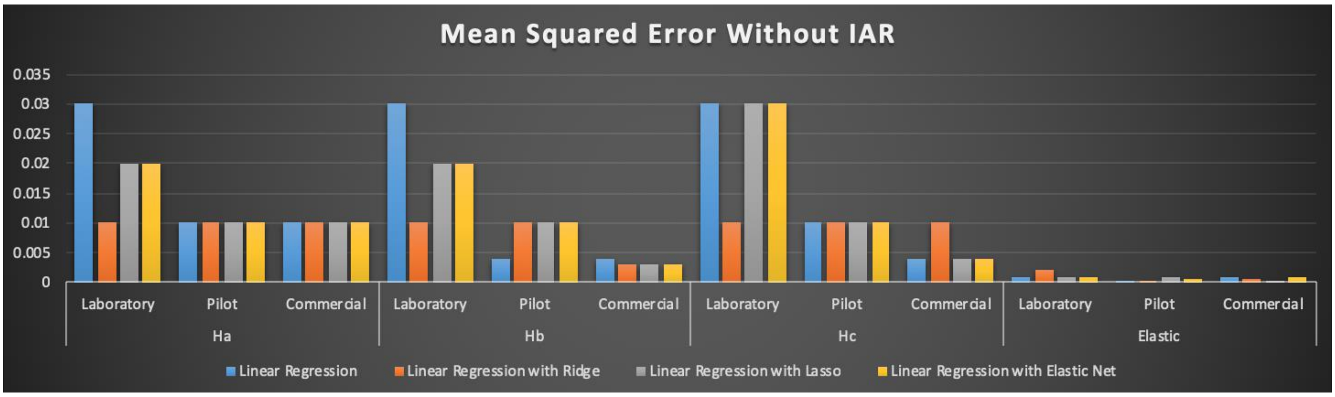


It quite evident from the above graph that, for the response variable 'Elastic Work', almost all the models predicted with similar high accuracy. However, for other response variables, stepwise regression provided a better result.

4.2 Model MSE

From the Mean Squared Analysis, it is shown that, Commercial Model Error is generally less compared to other buckets like Laboratory or Pilot. Also, with regularisation effect the models are not highly overfitted. In few occasions, Ridge Regression models gave more reliable results than others.


Mean Squared Error
Laboratory
Pilot
Commercial
Ha
Hb
Hc
IAR
Elastic
Linear Regression
Linear Regression with Ridge
Linear Regression with Lasso
Linear Regression with Elastic Net


Mean Squared Error Without IAR
Laboratory
Pilot
Commercial
Ha
Hb
Hc
Elastic
Linear Regression
Linear Regression with Ridge
Linear Regression with Lasso
Linear Regression with Elastic Net