Concrete Compressive Strength Data Analysis

RMSC 4002 TERM PROJECT

TIAN Ling 1155002110
XIAO Shiyu 1155002082
XING Yue 1155014329
ZHOU Zhihao 1155014412
He Zixuan 1155014579
2013 FALL

Abstract

In this paper, we try to figure out an efficient classification rule to identify the concrete strength level based on different ratio of its raw materials. First, outliers are detected and removed from original dataset. Second, PCA is then conducted for better understanding of the correlations among 8 explanatory variables as well as the relationship between concrete strength level and 8 raw materials. Finally, with "Age" treated as dummy variables, three classification methods – Multinomial Logit (MNL), Classification Tree (Ctree) and Artificial Neural Network (ANN) are applied for concrete strength level classification and comparison.

1. Introduction

Concrete is the most important and widely utilized construction material in the world because of its extremely good durability over other materials. Among the factors contributing to concrete's durability, compressive strength is indispensable. Concrete majorly consists of cement, water, aggregates, non-chemical and chemical admixture.

Cement and water are two primary ingredients. In general, lower water to cement ratio will result in higher strength, which might be the dominating factor that controls concrete compressive strength (Yeh, 1998). However, supplementary materials might also contribute to the compressive strength. For instance, fly ash can cause negative influence to concrete strength at the early stage of concrete (Topcu & Sarıdemir, 2008). Model with only linear predictors perform poorly because the ingredients and cement properties might associate nonlinearly (Erdal, 2013).

Addition to ingredients mixture, time or age will also cause discrepancy in concrete compressive strength. There is a benchmark age—28 days—as the standard for people to examine the strength and other qualities in practice (Alilou & Teshnehlab, 2010).

The compressive strength range of usable concrete starts from 17 MPa, and the higher compressive strength is, the more structures the concrete can be applied to. For example, residential buildings require compressive strength between 17 MPa and 34 MPa, while special buildings, such as skyscraper, require 28 MPa and more (Cement Organization, n.d.). Levels of compressive strength utilized in the prediction models are customized to the dataset in order to derive a meaningful prediction, which also coincides with its practical application.

2. Data Preparation

2.1. Outlier Detection

There are totally 1030 observations, among which certain ingredient combinations' compressive strength are recorded at different ages. As the experiment results, blast furnace slag, fly ash and super plasticizer are controlled variables, which means some observations lack one, or two, or even tree of the mentioned ingredients. This data structure paralyzes Mahalanobis distance computation if outliers are detected on the whole dataset basis. Consequently, we separated data into eight groups, and the result is shown on Table 1 in Appendix.

2.2. Principal Component Analysis

In order to model the concrete compressive strength level, it might be necessary to reduce the number of variables without loss of large amount of information. Principle Component Analysis (PCA) is a traditional approach to reduce the dimension of original dataset and preserve as much information as possible.

The correlation relationship among these 8 variables is detected using covariance matrix. The last variable "Age" has relatively weak correlation with all other 7 variables. As introduced before, the former 7 variables – components of raw material to produce concrete are intrinsic factors. Meanwhile, age, which denotes the duration before measurement, is an extrinsic factor to affect the compressive strength. For simplicity, the intrinsic factors and extrinsic factor are separated, and only the internal ones are used to conduct principal component analysis.

There are 7 principal components in intrinsic factors. The scree plot (See Appendix Figure 2) can be used to identify the adequate number of components we should use. Since the total variation is relatively small, the first 5 components, which have explained approximate 97% of all variances, are good enough to represent all the intrinsic factors. The transformed five new components are independent with each other from the random pattern of the scatterplot of principal component scores (See Appendix Figure 3).

The first component, as we can see in the loading table (See Table 2), has the largest positive loading in Water and largest negative loading in Super plasticizer. As super plasticizer is known as high range water reducer, we can treat these two as positively related to water. Moreover, with negative loadings in Fly ash and Fine aggregate, the first component can be regarded as one of the most important factor in concrete production – water-cement ratio. For the second and forth components, the significant variables are Coarse aggregate and Fine aggregate, respectively. We simply interpret the former component as large materials and the latter one as small materials in physical size. The third component is dominated by the variable Cement, which is the most important material of concrete. Meanwhile, the coefficient of cement and other variables (except the insignificant coefficient of super plasticizer) are opposite, which implies that the third component can be also constructed as Contrast of Cement and Mortar. For the fifth component, two variables of Aggregate possess positive coefficients while Water and Fly ash show negative coefficients. Since aggregate in concrete production is use to form the concrete as skeleton, however, water and fly ash have the opposite function that is to increase its plasticity. In consequence, the final component represents the shaping ability of concrete.

Table 1: Loadings for PCA analysis (first 5 components)

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Cement	0.0874	-0.0298	0.8245	-0.0774	-0.2410
Blast_Furnace_Slag	0.2250	-0.6289	-0.2536	-0.2930	0.4019
Fly_Ash	-0.4133	0.1711	-0.4317	-0.2214	-0.5450
Water	0.5641	0.0195	-0.2194	0.2316	-0.3972
Superplasticizer	-0.5245	-0.3613	0.1449	-0.3208	-0.0962
Coarse_Aggregate	0.0256	0.6658	-0.0064	-0.4112	0.4653
Fine_Aggregate	-0.4206	0.0064	-0.0155	0.7304	0.3161

Table 2: Cumulative variance percentage

Comp. 1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
0.3115	0.5147	0.7060	0.8451	0.9718	0.9961	1.0000

3. Model Predictions

3.1. Classification Tree

Aiming at finding the relationship between age and other variables, the approach of classification tree is applied to analyze the data, mapping concrete strength to variables including cement, fly ash and water. In addition, using classification tree can provide a more generous and clear model to make prediction. After searching information about concrete strength, it is found that specific numbers of strength is not as useful as classifying them into different levels.

To decide certain criteria of classification, we firstly performed a regression tree model with all variables and plot the relationship between fitted values and real response. Controlling the complexity of the classification tree, unimportant variables will be ignored by the model automatically, so there is no use to perform on the dataset after principal component analysis. As is shown in the plot (See Figure 1), real response ranges from 0 to 80, while the range of responses in each group are about 30, indicating that using current variables cannot predict response accurately. As a result, to split responses whose conditions are familiar into different levels, we drew the histogram of response to find suitable splits (See Appendix Figure 5).

The number of responses between [15, 20], [28, 32], [48, 54], [62, 64] are of lower level compared to neighborhood region. So, 15, 32 and 54 are chosen to be the criteria.

In addition, to avoid over-fitting, we control the complexity to 0.03, limiting number of variables used and the depth of tree (See Appendix Figure 6).

In the classification tree, there are four variables deciding concrete strength simultaneously. When age is less than 21, strength level can be 1, 2, and 3, and when it is larger than 21, the level can be 2, 3, and 4, which implies the importance of the age to concrete strength. According to the description of the researcher who conducted this concrete-strength experiment (Yeh, 1998), certain ages are used in the experiment to see the specific relationship between age and concrete strength.

This fact may provide implication that there may be strong relationship between these two attributes. While age has strong positive impact on concrete strength, cement and super plasticizer also have important positive effects on strength, but fine aggregate is negative associated with it.

When fitting the categorical tree, 700 out of 1035 data contain training dataset. The remaining 335 data are used as verification. The fitted results are shown in Table 3.

The average error rate is about 39.02%. There are two possible reasons to explain the large error rate. Firstly except age, other variables have complex effects on concrete strength, so it is hard to fit the model preciously using binary trees. In addition, the clean dataset only consists of less than 1100 data with large variation, which may not provide enough information for the real relationship.

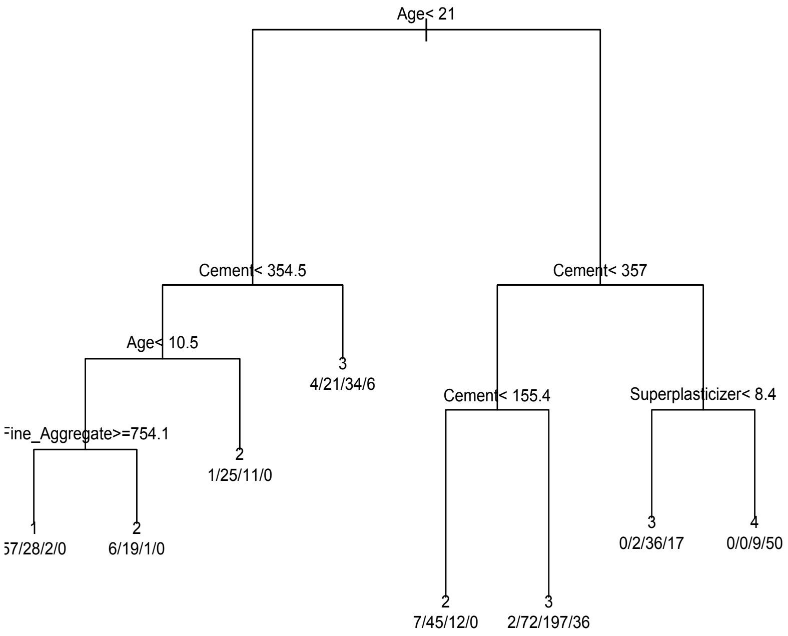

Figure 1: Categorical Tree Plot

Table 3: Fitted Result of Categorical Tree

Test	1	2	3	4	Sum	percent
1	20	18	0	0	38	0.5263
2	14	41	15	0	80	0.5125
3	2	50	106	19	180	0.5889
4	0	0	0	19	19	1.0000

3.2. Artificial Neural Network

Due to the relatively bad performance and inadequate testing accuracy of the previous classification tree model in predicting level of concrete strength, the artificial neural network will be implemented to the concrete data set in order to facilitate its remarkable capacity to derive meaning from complicated data with large noise.

The feed-forward neural network with single hidden layer will be applied to the identical clean data set used in classification tree. The input variables include all 8 continuous variables. Since the variable vary significantly in range, all input variables will be scaled into range [0, 1] to ensure each input has approximately equal importance. The output variables are labeled into 4 levels according to the critical strength 15, 32 and 54 in the same way as the previous classification tree. Observations in cleaned dataset are randomly partitioned into training set and testing set. The training set contains 80% of all observations and remaining parts are used as testing set. The reason of allocating a relatively high weight to the training set is the limitation from the size of the whole data set.

Over-fitting problem is commonly encountered in training the neural network, especially when the model is highly complicated with large number of hidden layer nodes. The feed-forward neural network model with single hidden layer with hidden layer size ranging from 1 to 16 will be fitted repeatedly to the training set. The upper bound is set at twice as large as the number of input variables in the model. The visualization technique to plot layer size against corresponding training and testing error rates will be implemented to explore the relationship between prediction performance and hidden layer size; thus, help the determination of suitable hidden layer size. Under each scenario with hidden layer size, we perform repetitive training processes and select the model with the result. From Figure 7 (See Appendix), it can be found that when starting with small hidden layer size, both the training and testing error rates is relatively high over 30%. As the size increases up to 5 both the training and testing error rate naturally drop in accordance with the more complex architecture of the model. When the hidden layer size exceeds 5, although training error continues to descend, the testing error rate becomes unstable and occasionally even larger, which is a clear sign of over-fitting. Thus the model with hidden layer size at 5 will be chosen and fitted more deeply later.

The maximum number of iteration is determined by visualize the model fitting with the targeting variable. From the Figure 8 and Figure 9 (See Appendix), which illustrate the dynamic of training and testing error rate against the number of iteration during the training process, we can find that when the iteration exceeds around 450 times, the testing error tends to increase, another sign of over-fitting. Hence, in the final model, the maximum iteration number is set at 450.

Finally, we come up with an 8-5-4 neural network model to predict four classes of concrete strength using 8 attributes of the concrete. The total number of weights is 69, around 8.58 % of the size of training data set. To avoid being trapped in the local minimum due to the random assignment of initial weights, we select the model with lowest fitted error from 300 repetitive training processes with maximum iteration number up to 800. The summary of the final model is in the appendix (See Appendix Table 2 and Table 3).

The resulting training error rate is 13.09% and the testing error rate is 19.91% Compared with the 32.9% error rate from classification tree, the neural network provides a noticeable improvement in the predictive performance.

3.3. Dummy Variable Analysis

As mentioned, age is an external factor of the compressive strength compared with other material elements. Intuitively, the solidification of concrete makes it rigid, which implies that we should expect concrete with larger age has large compressive strength. Before we get into classification models, we investigate effects of age on regression models to find out whether age can improve our predictions after some data transforming.

Simply plotting age attributes against target (Figure 2), we can observe similar relationships between age and strength as expected. Although a lot of overlaps, for concrete of age before 100 days, roughly speaking, age positively affects the strength of concrete. To fully use this positive relationship, we try to introduce dummy variables to get rid of the effect of neutral relationship between age and strength for concrete of large age.

After mass of attempts, we finally decided to divided age in 6 groups, namely 1 to 3, 7 to 14, 28, 56, 90 to 120, and over 180. See the box-plot of 6 groups, median of concrete strength in earlier groups follows an increasing trend (Figure 3). We transform age into 5 dummy variables, and test by using simple linear regression model. Comparing two models:

M1: Concrete compressive strength ~ all attributes including age

M2: Concrete compressive strength ~ all attributes without age + 5 dummy variables

We found out that introducing dummy variables improve the R square of regression a lot (from 0.6155 to 0.8085). In M2, coefficients of dummy variables are relatively higher than those of other attributes, but M1 indicates age have coefficient on the same level of other attributes (Table 4).

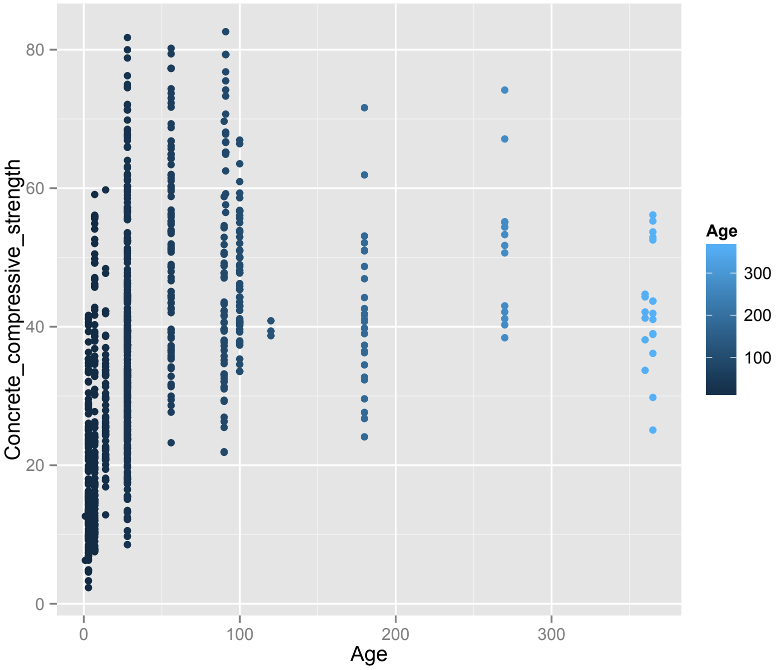

Figure 2: Simple plot – Age *against* Concrete Compressive Strength

One reason maybe the introducing of dummy variable reduce the age attributes from (1, 365) to 0 or 1, thus amplifying the effects of age on compressive strength. But more reasonable reason is that age does have huge effect on strength of concrete. Furthermore, comparing coefficients of dummy variables in M2, we observed that from V1 to V5, coefficients are increasing but V3, V4 and V5 have similar coefficients. It indicates that larger age have more effects on strength of concrete while this positive effect stop to increase when age is large.

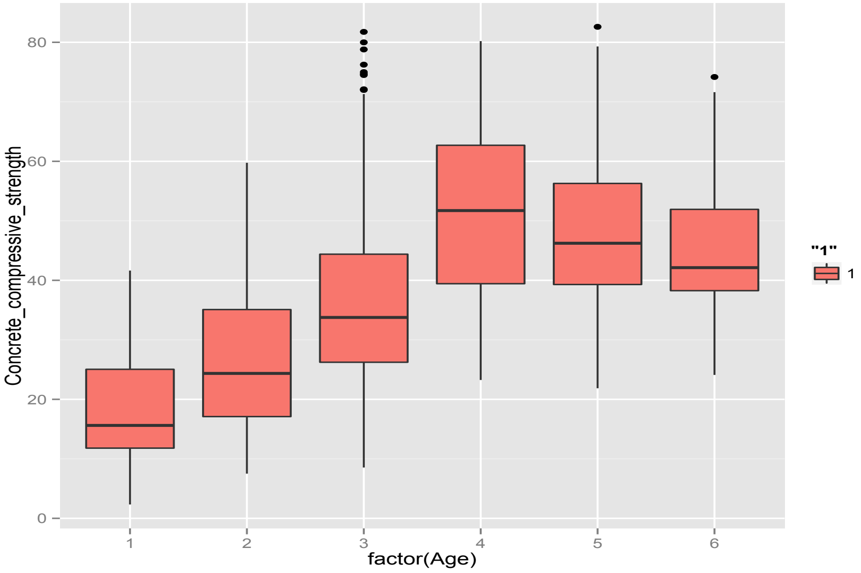

Figure 3: Box-plot of concrete compressive strength on different age groups

Coefficient	Int.*	1	2	3	4	5	6	7	Age	V1	V2	V3	V4	V5
M1	-23.33	0.12	0.10	0.09	-0.15	0.29	0.02	0.02	0.11	-	-	-	-	-
M2	-60.51	0.13	0.11	0.09	-0.12	0.12	0.03	0.03	-	8.62	19.54	28.57	31.26	31.94

Table 4: Coefficients of attributes in two simple linear regressions

* Int.: Intercept 1 to 7: Material factors V1 to V5: 5 dummy variables

In conclusion, age, though regarded as external effects, should be considered as a good indicator for concrete compressive strength. Because the whole data set is retrieved from experiment and the amount of items are not that sufficient, we should include the age while doing classification in case we lose some important information so as to build a more satisfying model.

3.4. Multinomial Logistic Regression

After some analysis on external factor of age, we turn to classification part. As disclosed in the introduction, concrete applied in different fields can be classified into different types according to compressive strength, in the sense that we would not prefer concrete with high compressive strength as long as the compressive strength is enough for specific use of concrete. More intuitively, considering the cost, the higher the compressive strength, the more expensive the concrete is. If concrete of compressive strength 35Mpa/area is enough for someone to build their residence, he will not bother to change to a 60Mpa one as it cost a lot. Under this important assumption, different level of compressive strength is not a sequential quality, thus we can transform the continuous data of compressive strength attribute into multivariate one to continue the classification work. The objective of this module is to build up some prediction rules to classify different type of concrete with related level of compressive strength given attributes of concrete.

Firstly we split the compressive strength by 15, 32 and 54, which means 4 class of concrete: minimum purpose (compressive strength of concrete lower than 15 Mpa/area), general purpose (between 15 and 32), moderate purpose (32 to 54) and advanced purpose (higher than 54).

We use 700 of the 1005 cleaned dataset as training dataset and others as testing dataset to do multinomial logistic regression(R-output 1). But the result is not satisfactory, with large training error of 26% and testing error of 29%. This result might be caused by the ordinal nature of the variable.

4. Conclusion

In the analysis, Categorical Tree provides an unsatisfactory 60.98% accuracy on prediction due to the complex effect of age on other attributes and the lack of enough data evidence, which might lead to a slight bias in analysis. Moreover, the multinomial logistic regression with age as dummy variable is not appropriate to fit data, because there is an ordinal relation between each level of the response. So, the error rate is 26% for training and 29% for testing. Comparatively, Artificial Neural Network can provide a nice prediction with training error of 13.09% and 19.91% testing error. However, the absence of practical interpretation and implication from ANN model is also critical. It is also difficult to formulate a stable model in the analysis, which might largely owe to the noisy structure in the original dataset. More experiments should be conducted to provide more solid evidence and support on explaining relationships between different attributes and compressive strength of concrete.

References

Alilou, V. K. & Teshnehab, M. (2010). Prediction of 28-day compressive strength of concrete on the third day using artificial neural networks. *International Journal of Engineering, 3*(6), 565-576.

Cement Organization (n.d) ? Concrete Technology [PDF document]. Retrieved from http://www.cement.org/tech/faq_strength.asp

Erdal, H. I. (2013). Two-level and hybrid ensembles of decision trees for high performance concrete compressive strength prediction. *Engineering Applications of Artificial Intelligence, 26*(7), 1689-1697. doi:http://dx.doi.org/10.1016/j.engappai.2013.03.014

Topcu, I. B. & Sarıdemir, M. (2008). Prediction of compressive strength of concrete containing fly ash using artificial neural networks and fuzzy logic. *Computational Materials Science, 41*(3), 305-311. doi:10.1016/j.commatsci.2007.04.009

Yeh, I. C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research, 28*(12), 1797-1808.

Appendices

Figures

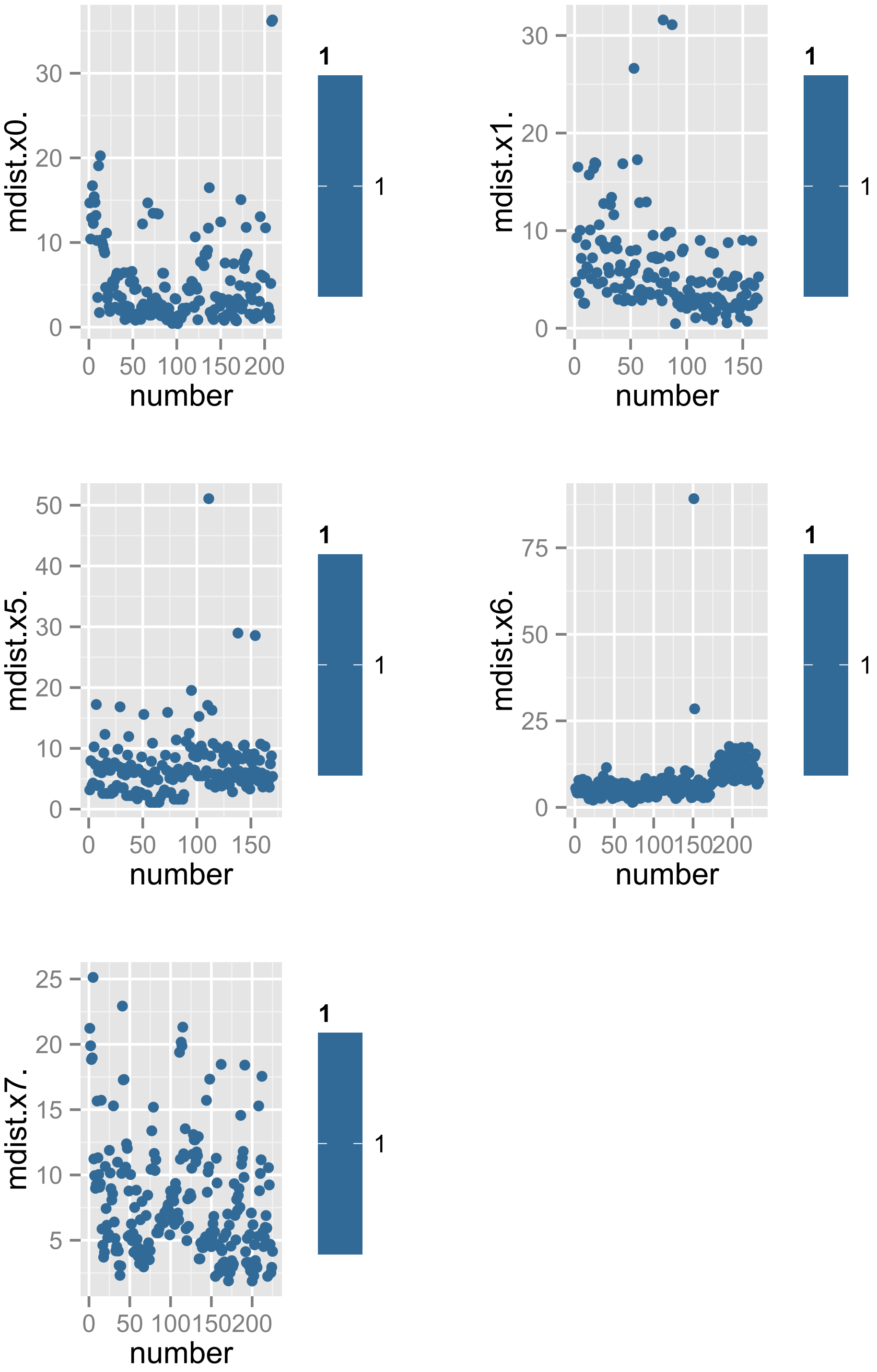

Figure 1: Outlier detection plot using Mahalanobis distance.

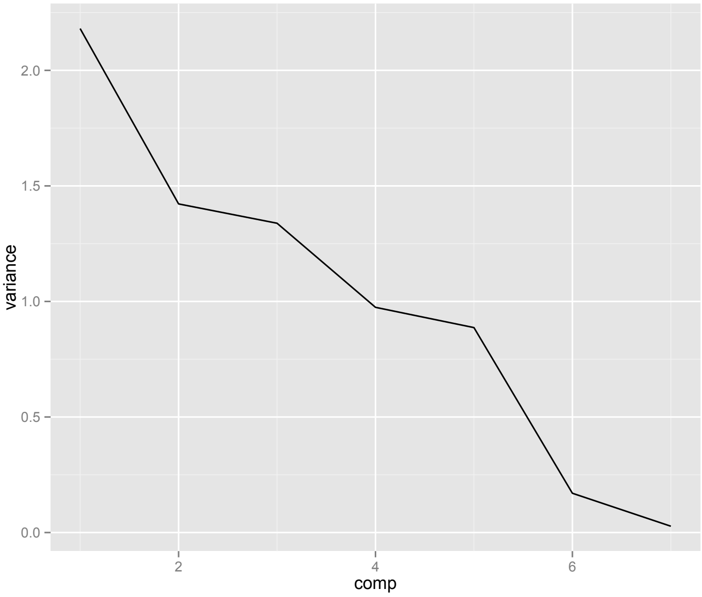

Figure 2: Scree plot of PCA analysis of concrete strength level data.

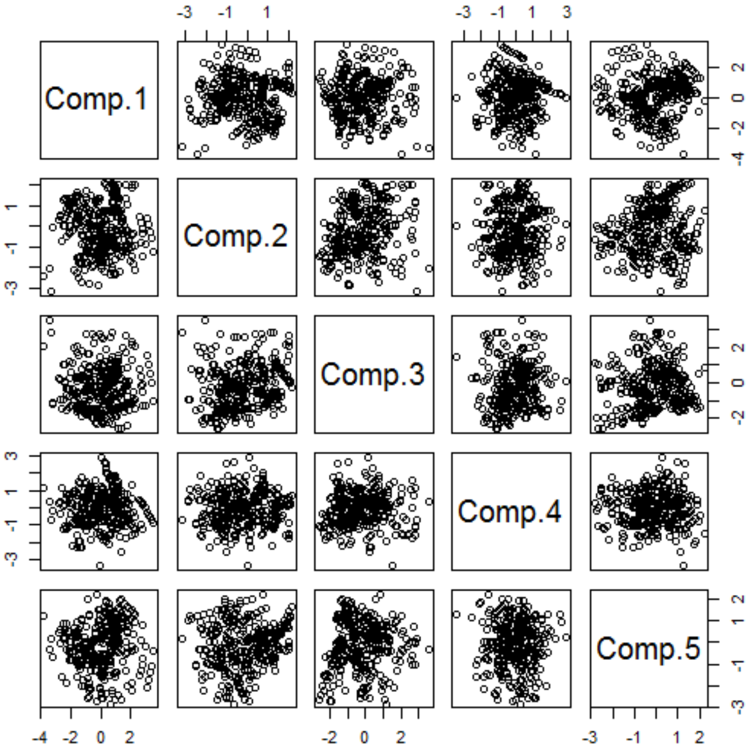

Figure 3: Scatterplot for PCA scores (first 5 components).

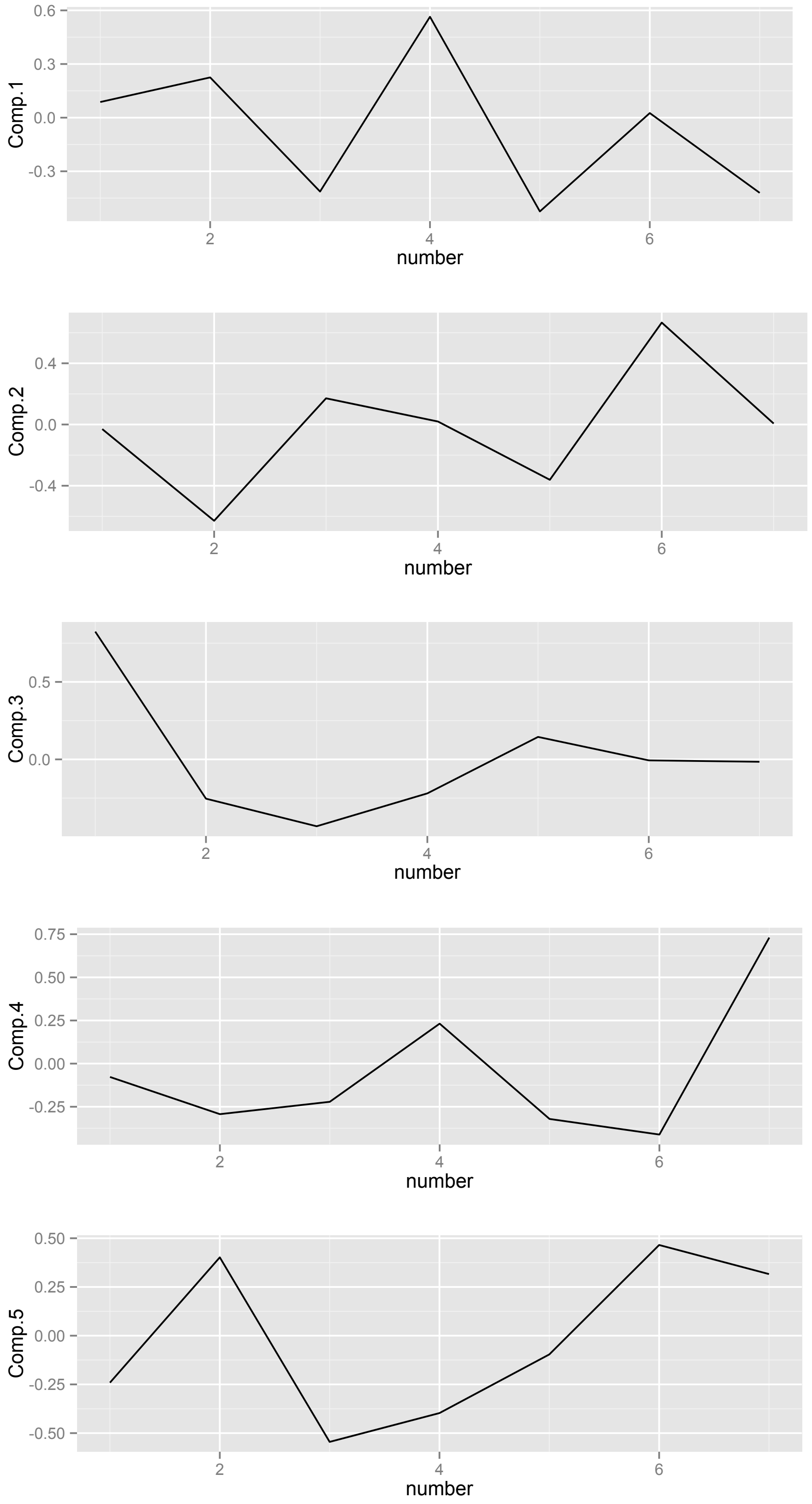

Figure 4: PCA loadings plot for first 5 components.

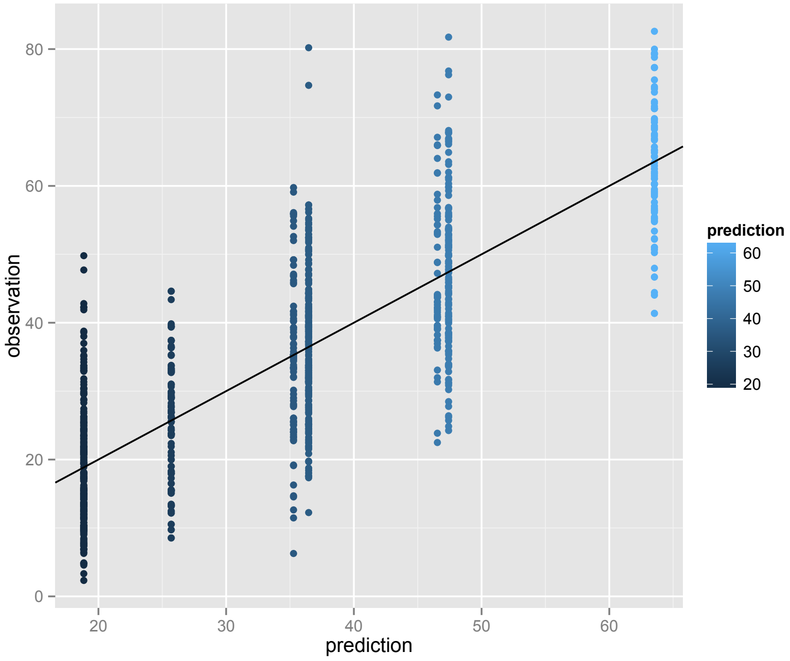

Figure 5: Categorical Tree-prediction and observation plot.

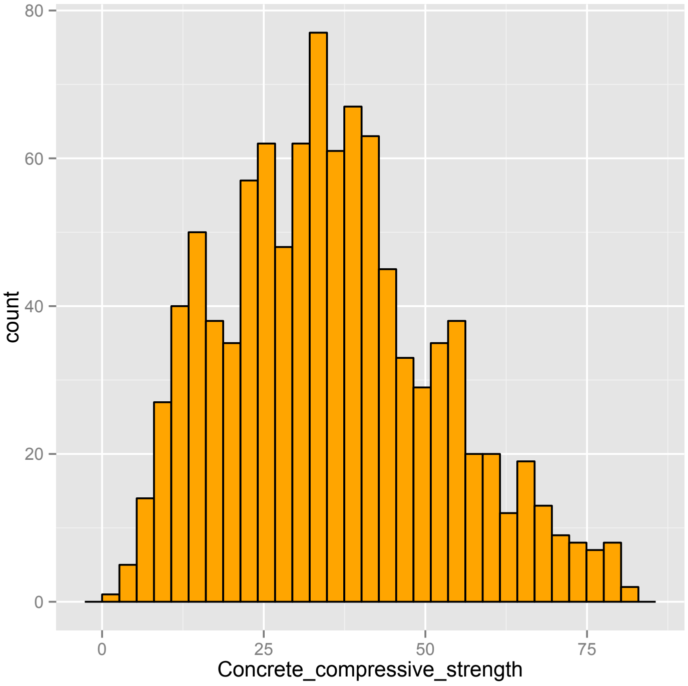

Figure 6: Histogram of Concrete compressive strength.

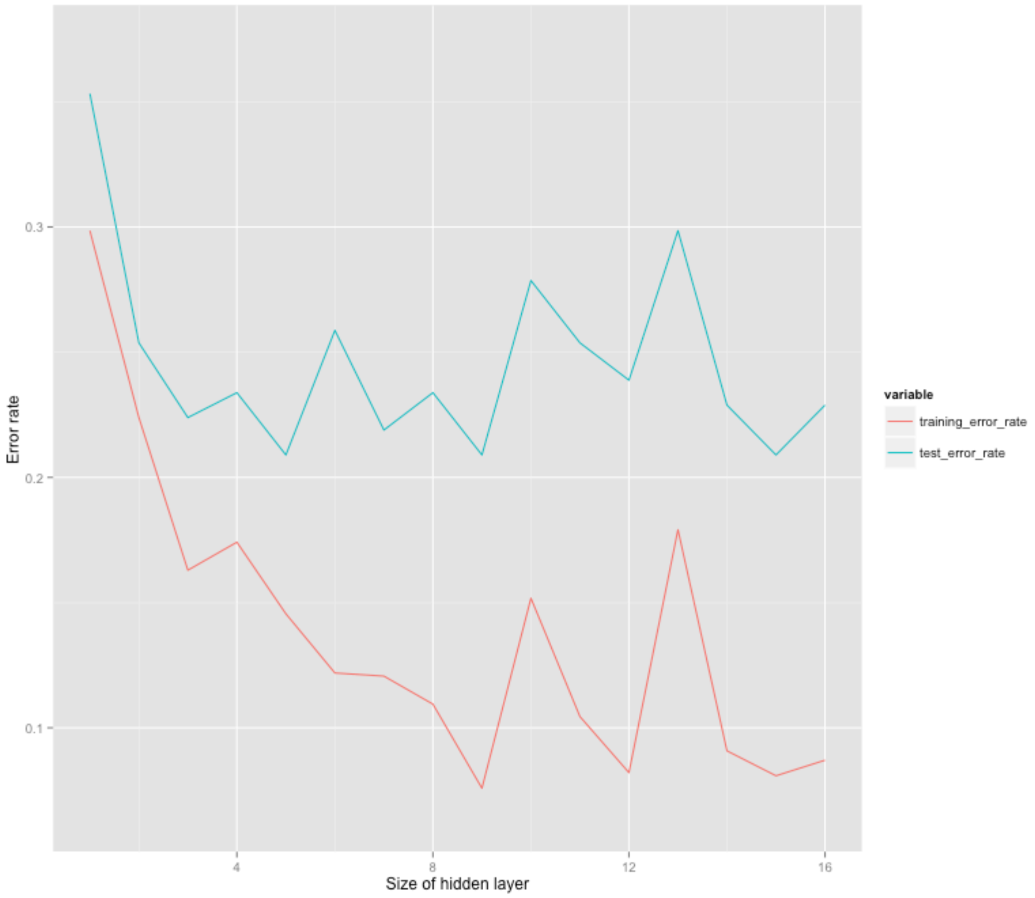

Figure 7: Training and testing error rate under various hidden layer size.

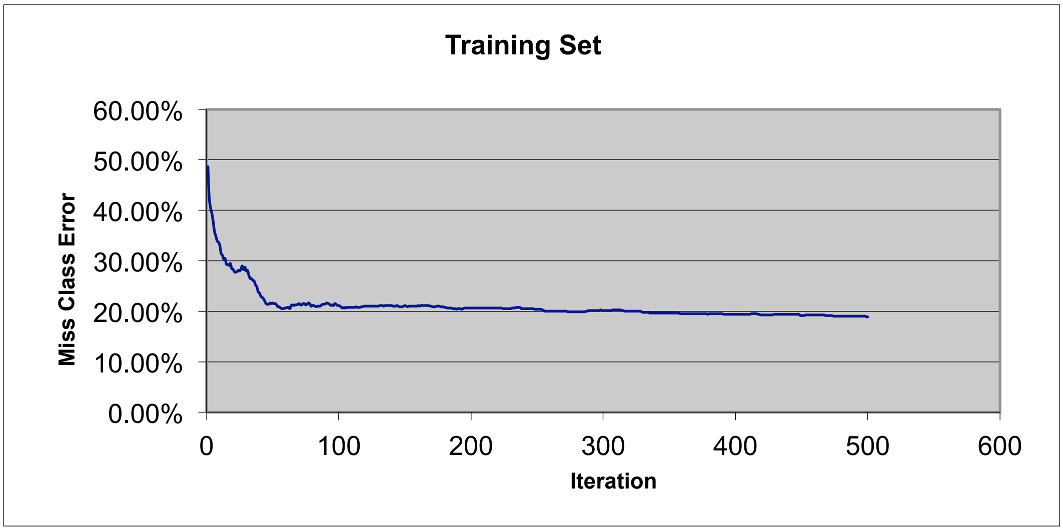

Figure 8: Training error rate against iteration rounds.

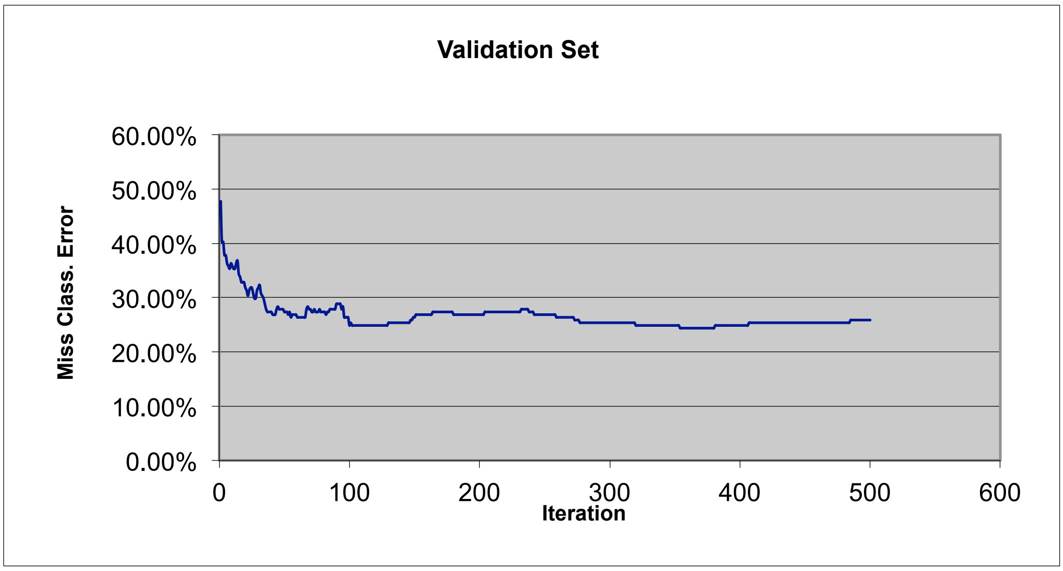

Figure 9: Testing error rate against iteration rounds.

Tables

Table 1: Outlier detection result summary

Type of combination	Total Num. of observation	Num. of Outliers Deleted
(1), (2), (3)	209	7
1, (2), (3)	164	7
(1), 2, (3)	6	
(1), (2), 3	23	
1, 2, (3)	0	
1, (2), 3	170	4
(1), 2, 3	233	2
1, 2, 3	225	5

Note:

1. “1”= Blast furnace slag, “2”=Fly ash, “3”=super plasticizer.

2. Numbers with “()” means categories without corresponding ingredients.

3. For categories with small number of observations, detection is skipped.

Table 2: Classification table for training data

Training	1	2	3	4	Sum	Percent
1	76	14	1	0	91	0.8352
2	7	208	23	0	238	0.8739
3	3	32	308	7	350	0.8800
4	2	0	16	107	125	0.8560

Table 3: Classification table for testing data

Testing	1	2	3	4	Sum	Percent
1	21	4	0	0	25	0.8400
2	3	44	4	0	51	0.8627
3	0	18	68	5	91	0.7473
4	1	1	4	28	34	0.8235

Summary of final 8-5-4 neural network model

```
a 8-5-4 network with 69 weights
options were -
   b->h1   i1->h1   i2->h1   i3->h1   i4->h1   i5->h1   i6->h1   i7->h1   i8->h1 
   10.45    -2.47    -4.00    -0.58    -6.30     6.56    -6.01    -6.88     1.89 
   b->h2   i1->h2   i2->h2   i3->h2   i4->h2   i5->h2   i6->h2   i7->h2   i8->h2 
  -20.13    27.78    23.23    10.97    -7.20     6.59     2.28     4.48     0.04 
   b->h3   i1->h3   i2->h3   i3->h3   i4->h3   i5->h3   i6->h3   i7->h3   i8->h3 
  -48.80    50.27    40.70    19.20    -1.94     5.99    11.54    16.97    41.82 
   b->h4   i1->h4   i2->h4   i3->h4   i4->h4   i5->h4   i6->h4   i7->h4   i8->h4 
  -42.19    23.39    20.13     9.25     9.78    -6.38    15.85    19.65    68.97 
   b->h5   i1->h5   i2->h5   i3->h5   i4->h5   i5->h5   i6->h5   i7->h5   i8->h5 
  -10.50    11.76     7.58     4.40    -1.34     3.41     2.73     3.96    49.52 
   b->o1   h1->o1   h2->o1   h3->o1   h4->o1   h5->o1 
   67.18   292.94    73.59 -1561.18   281.06  -407.74 
   b->o2   h1->o2   h2->o2   h3->o2   h4->o2   h5->o2 
 -203.14  -954.56  1793.03 -1898.67  -795.85  1264.45 
   b->o3   h1->o3   h2->o3   h3->o3   h4->o3   h5->o3 
 -794.01  -212.61  -389.64   441.83  -169.43   984.70 
   b->o4   h1->o4   h2->o4   h3->o4   h4->o4   h5->o4 
-1019.65   697.36  1405.25  -492.73   593.67  -688.14 
```