Estimation of Post Milling Particle Size Distribution Using Linear Transformation in Parameter Space

Jesang Yim
Kimball Union Academy; NH, United States of America
Email: yimejsang@gmail.com

Abstract - The research attempts to predict the change in the size distribution of the particles after the milling process. The data for the research was sourced from https://www.innocentive.com/. It consisted of pre and post-milling data of the size distributions of the particles. This research made two assumptions: the distribution of particle sizes can be described by combining several individual normal distribution graphs with different parameters. Curve fitting was used to find out the best fitting parameters of the component normal distribution functions. We also assumed that the formula for predicting post-milling distribution from pre-milling distribution is linear, i.e., a matrix operation on the pre-milling parameters will predict the post-milling one. A single matrix should be able to convert all 26 *datasets*. However, only 25 of the datasets were used in order to test the remaining one for cross-validation. Four normal distributions were determined to be optimal for describing the particle size distribution in the data. The prediction had a sum of squared errors ratio ranging from 4.95% to 31.2%, with the average value of 18.7%, which shows the potential of this approach. With more data, or with an ability to conduct the experiment, higher and more consistent particle size predictions should be within reach.

Key Words - Milling, Linear Transformation, Parameter Space, Particle Size Distribution

INTRODUCTION

The research is about predicting the distribution of sizes of the particles after the milling process using normal distribution functions and linear transformation.

I. Hypothesis

The hypothesis of this research is that the particle size distribution in post-milling data can be predicted using normal distribution functions of particle sizes of pre-milling data. And that there exists a single matrix which can be used to convert all 26 parametrized datasets from pre-milling data to post-milling data.

II. Background Research

a) Milling Process and Different Types of Milling

Milling is a process of making particles into smaller pieces. There are various types of milling, such as ball mill, rod mill, pebble mill, burrstone mill, or tower mill. Slurry milling is a type of milling where particles are mixed with liquid. Typically, metal particles such as zinc, lead, silver, aluminum, and nickel are milled for industrial processes, and grains are milled in milling machines through various types of mills. [1]

b) Normal Distribution and Different Types of Distribution functions

The normal (Gaussian) distribution is a function that shows how the probability of variables is distributed [2]. The normal distribution is also called a bell curve, and it can be applied to various random processes. The bell curve is one of the most commonly used forms of distributions. The density is higher in the middle and lower at both ends of the graph. The middle part of the graph has values closer to the mean of the values. The bell curve is symmetric, and the center of the curve indicates the mean. [3]

Even non-normal distribution functions can be shown as a sum of multiple normal distribution functions with different parameters. In normal distribution, the mode, mean, and the median are identical and at the center, and the graph is symmetric with respect to the center. [4] Besides normal distribution, there are several different types of distributions such as Cauchy distribution, power distribution, and Poisson distribution.

Perhaps it is possible that this research can be repeated with other probability distributions. Normal distribution was chosen because of its familiarity, and its ubiquitous implementations in all computer languages.

III. Existing Literature Related to this Research

Existing literature consists of research on the effect of ball and feed particle size distribution and the effect of slurry density on ball milling. [5] This research is different from those previous studies, as the particle size distributions after the milling process will be predicted by a mathematical analysis.

This idea of modeling an arbitrary distribution as a combination of multiple normal distributions is an idea of this author's mentor, and was not published anywhere to the best of our knowledge. Thus it is uncited.

IV. Materials for Research

a) Mathematica was used in this research as a programming language [6].
b) There were 26 datasets provided by Innocentive that show the size and number of particles before and after the milling process. The policy of Innocentive does not disclose who the SEEKERS are. All that the company provided was the data without any mention of the materials or machines used. Therefore, this research was all based on pure mathematical work.

EXPERIMENTAL PROCEDURE

This research consists of four parts:

a) Deciding the number of normal distributions based on the sum of squared estimates of errors (SSE) to describe the particle distribution.
b) Finding the parameters, i.e., the heights, means, and the standard deviations of the component normal distributions.
c) Determining the matrix to transform pre-milling distribution data to post-milling distribution data while minimizing error.
d) Measurement of performance by measuring errors in the predictions.

The graphs of 26 datasets do *not* look like normal distributions. However, we assume that this non-normal distribution looking graph is composed of several normal distributions. In order to increase the prediction accuracy, we find out the number of normal distributions consisting of the data graph.

I. Importing the Dataset

In the first part of the research, we import pre-milling and post-milling data of particles' sizes in a .CSV file. The file contains 26 datasets, and each of the dataset contains particle size and the number of particles underneath each size. The particle numbers are normalized to add up to 100. Since the exponential scale was used to measure the particle sizes, we took the log of each particle number to make it linear.

II. Fitting the Parameters and Solving the Equation by Using FindFit Function

The default normal distribution function is defined as (a $exp(-((x-b)/c)^2)$), where {a, b, c} are the parameters: a is the amplitude, b is the mean, and c is the standard deviation of the normal distribution. The equations change according to the number of normal distributions. In the normal distribution graph, a is the height of the graph, b is the center of the graph, and c indicates the spread of the graph. Equation 1 and 2 below show how the function looks like in different numbers of normal distributions.

#(2)

We use the FindFit function, which calculates and finds out the predicted parameters according to the given equation and variables, to find the values of parameters of the distribution. FindFit function takes the data, {equation, constraint}, {parameter, starting value} and the variable used in the equation, and returns the parameters. Constraints limit the range of parameters in the equation, and the starting value sets the starting point of the parameters. Without the constraints, the fit value can be wildly off the mark. Some constraints were trivial such as the heights, means and standard deviations being positive numbers. Others were determined by trial and error. A single set of constraints was used for all data.

```
FindFit[sdPreMillingDataPaired[[i]], {sfexponential,
 sfConstraint}, sfparameter, x, Method → "NMinimize"]
```

Fig. 1. FindFit Function

```
sfConstraint4 = {6 ≤ a1 ≤ 8, 0 < a2 < 0.5, 0.5 < a3 < 1.7,
   0.1 < a4 < 2.5, 4 < b1 < 5, -1.5 < b2 < 1, 5 < b3 < 6.5, 3 < b4 < 3.9,
   0.8 < c1 < 1, 0.6 < c2 < 1.65, 0.5 < c3 < 0.65, 0.7 < c4 < 2.5};
```

Fig. 2. Example of Constraints

```
sfparameter4 = {{a1, 6}, {b1, 4.5}, c1, a2, {b2, 0.5}, c2, a3,
   {b3, 6}, c3, a4, {b4, 3}, c4};
```

Fig. 3. Parameters and their starting points

Fig. 1, Fig. 2, and Fig. 3 above show how FindFit function, constraints, and starting values work. Fig. 4 below is the result of using FindFit function.

$$\{a1 \to 0.623848,\ b1 \to 3.49757,\ c1 \to 2.49512,\ a2 \to 2.27611 \times 10^{-19},\ b2 \to 1.27016,\ c2 \to 0.143678,\ a3 \to 6.59942,\ b3 \to 4.47633,\ c3 \to 0.943641,\ a4 \to 0.14814,\ b4 \to -0.0905873,\ c4 \to 0.355804\}$$

Fig. 4. Resulting Parameters After Using FindFit Function

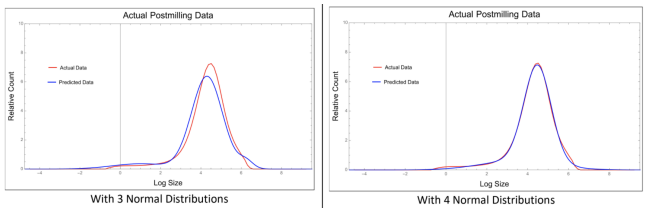


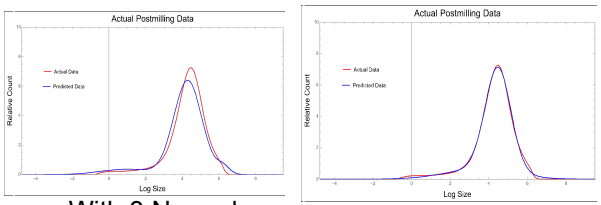
Fig. 5. Addition of Components

Fig. 5 above shows the addition of 4 different components. The 4 different normal distribution graphs add up to each other and form one non-distribution function, which looks like a distribution function.

Fig. 6. Component of Normal Distributions

Fig. 6 above shows the components of the normal distributions inside the particle distribution graphs. It clearly indicates that the particle distribution graph actually consists of several components. As the graph shows, three components are not enough to cover the full distribution graph.

Fig. 7 below shows how the actual and predicted post-milling data look like when using 3 and 4 normal distributions. This was typical.



With 3 Normal Distributions

With 4 Normal Distributions

Fig. 7. Actual Data with Predicted Data using 3 and 4 Normal Distributions

III. Calculating Errors and Deciding the Number of Distributions

The number of normal distributions we tried ranges from one to five. The equations for each normal distribution function are used to plot a graph. The graph includes both predicted and real values of particles' sizes and their numbers. Errors are defined by calculating the sum of error squared between the predicted value and the actual value of each dataset. was used as the equation for calculating errors. The number of graphs with minimum error is chosen as the number of the normal distribution function. Fig. 8 and Fig. 9 below show the predicted and actual data with errors, respectively using 3 and 4 normal distributions.

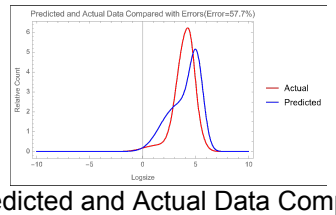


Fig. 8. Predicted and Actual Data Compared with Errors using 3 Normal Distributions

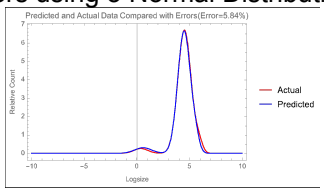


Fig. 9. Predicted and Actual Data Compared using 4 Normal Distributions

IV. Creating Matrix with Optimized Parameters from FindFit

In the second part, we use a matrix to predict the formula that shows the relationship between the pre-milling data and the post-milling data. Parameters of the pre-milling and post-milling normal distributions are expressed in a matrix. There are 3 parameters per normal distribution and 4 normal distributions. Therefore, the matrix of parameters will have matrix.

A key to deriving the matrix for the post-milling data is by multiplying the key with the pre-milling data. Matrix of size is composed of parameters that are fitted in the previous step. pre-milling data parameters will be multiplied to matrix, and the result should be the matrix composed of parameters from post-milling data. Fig. 10 below shows the relationship of the matrix and parameters of pre-milling and post-milling distributions.

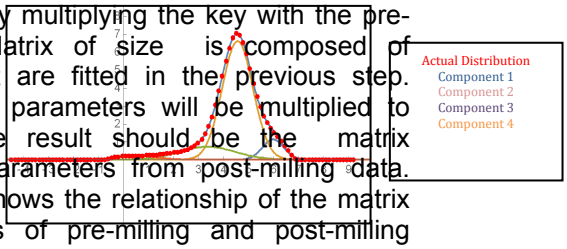


Fig. 10. Matrix and Parameters

V. Calculating Errors and testing the matrix

In the third part, we review the errors from the previous parts. For the cross-validation, since we only have 26 datasets, we use 25 of the datasets to test our predicted post-milling dataset against the remaining single data set. Cross validation is needed to prevent the overfitting and to test if the predicted matrix fits well to the actual data. We repeat this for all 26 datasets by excluding 1 different dataset on each iteration.

```
eq01 = matrix01.sfparameter;
data02 = Table[Flatten[{preParamsTable[[i]], postParamsTable[[i, rowNumber]]}],
  {i, 1, useTheseData}];
matrix = Values@FindFit[data02, eq01, matrix01, sfparameter]
```

Fig. 11. How Matrix Can Be Found using FindFit Function

DATA ANALYSIS AND DISCUSSION

The data chart shows the name of particle samples and their numbers according to each size. Using those 26 datasets about the particles' sizes and their counts, we created the normal distributions.

I. Compare the Graphs with Different Numbers of Normal Distribution Functions

The distribution graphs are composed of several individual normal distribution graphs. These individual graphs cumulate to create the predicted distribution graph. Referring to the figure below, four individual normal distributions in the color [blue, green, orange, and brown] accumulate to the distribution graph in red dots.

In this procedure, as shown above in Fig. 7, we combined different numbers of individual normal distributions ranging from 1 to 4. The cumulative distribution graph with 4 individual graphs had the smallest error compared to the others.

II. Set the Constraints and Starting Values for Each Parameter

Fig. 13 below shows how parameters and the starting values for the parameters are set.

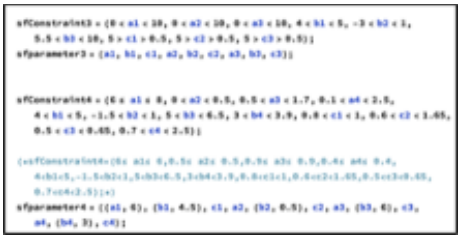


Fig. 13. Constraints and Starting Values for Each Parameters

III. Calculate the Errors from Each Graph and Decide What Number of Distributions to Use

We found that a cumulative distribution graph consisting of 4 individual normal distribution graphs generates the smallest error. As mentioned in the previous section, we use to compute the error. The error generated from a graph consisting of 4 individual normal distribution graphs is 1.49% (3 significant figures). Fig. 14 below shows the comparison of predicted and actual data of 4 different normal distribution graphs.

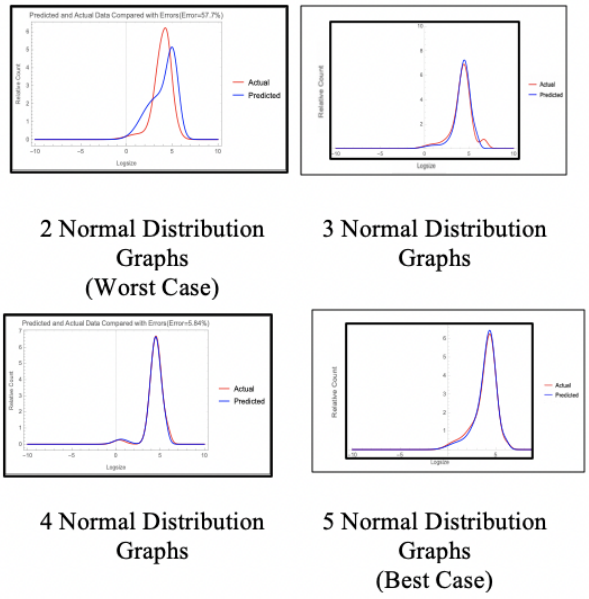


Fig. 14. Predicted and Actual Data Compared with Errors

IV. Find the Matrix that Can Be Used to Predict the Post Milling Data from Pre-milling Data

As the error from 4 normal distribution graphs is minimal with the error of 18.7 percent, we can deduce that we can predict an accurate enough post-milling distribution from pre-milling distribution. We can do this by multiplying the key

matrix with the pre-milling data parameters and predicting the parameters for post-milling data. Therefore, the hypothesis that we can predict the distribution of post-milling data from pre-milling data is true.

```
eq01 = matrix01.sfparameter;
data02 = Table[Flatten[{preParamsTable[[i]], postParamsTable[[i, rowNumber]]}],
  {i, 1, useTheseData}];
matrix = Values@FindFit[data02, eq01, matrix01, sfparameter]
```

Fig. 15. How Matrix Can Be Found Using FindFit Function

-3.29084	12.3884	-16.1371	11.8465	6.76849	-12.7256	-5.00342	2.84763	-12.8439	-5.58792	1.15201	-6.88535
0.221241	0.546776	0.657685	1.60003	0.404921	-0.545993	0.0697963	-0.114201	-0.117588	0.263331	0.0662485	0.148325
-0.334875	0.868442	-0.894564	0.0665116	0.313898	-0.585628	-0.330966	0.328193	-0.647782	-0.452887	0.0424126	-0.417751
0.0640382	-0.0711243	1.53481	-1.48019	-0.458339	0.678978	0.162593	-0.347125	-0.0884997	0.17324	0.00463856	0.254726
0.126539	-1.26127	4.15178	-8.1663	-3.03514	2.77444	0.466588	0.0583324	-0.516545	0.366715	0.619857	-0.543542
0.112207	-0.818459	2.22368	-5.4179	-2.22371	2.04899	0.29586	0.170248	-0.591621	0.234708	0.484879	-0.38954
-0.928337	-8.27172	-42.496	-33.9347	-14.56	29.1393	-0.934997	12.6569	7.16157	-2.55404	-6.069	5.74923
-5577.62	-197151.	-1.11455×10^{6}	-181611.	-232781.	368529.	-80799.4	332774.	117544.	-68220.3	-54276.	-22224.1
-0.393938	-6.63298	-17.3934	-7.55592	-7.42545	13.4956	-1.49002	7.97738	4.86888	-1.01312	-3.43109	2.27855
2.56906	-8.85181	14.4596	-8.85669	-4.47283	7.91941	3.65985	-2.11631	7.53825	4.32915	-0.447386	4.12479
0.70263	-3.83566	5.53184	-8.34919	-4.56479	4.45646	0.493636	1.0803	2.15612	1.09686	0.650686	-0.237217
0.175282	-0.310897	-2.34148	2.83078	0.821019	0.0371577	-0.429454	0.410772	2.497	0.0438293	-0.59987	0.926121

Fig. 16. 12 X 12 Matrix Answer

CONCLUSION

I. Results

From this research, we can conclude that the hypothesis is true: it is possible to find a matrix that can transform the pre-milling particle size distribution to the post-milling particle size milling distribution. We found that 4 normal distributions fit the original normal distribution with minimum error, and provided the most accurate matrix formula that can be used for a precise prediction of sizes of post-milling data. Multiplying the matrix with the parameters of pre-milling data from 4 different normal distributions gave the parameters of predicted post-milling data with 18.7% accuracy.

The significance of this 18.7% accuracy is unusually difficult to evaluate due to the circumstances in which the data was obtained. The data was posted on the innocentive.com site by an anonymous company that was referred only as SEEKER. The SEEKER didn't specify what the error rate they were currently getting, nor what accuracy they were looking for.

Thus, this research has to settle with the success in demonstrating the feasibility of using multiple normal distributions to model and compute transformations in distribution. It is not possible to claim any benefit without knowing the current base rate and the cost of error.

II. Challenges in the Research

The challenges in this research were mainly from the fitting part. The values from the function were different although the codes were exactly the same. The hardest part of the research was debugging the code. Since the codes were repeated several times to compare different values, the entire result came out to be wrong if the initial code was not correct. One of the errors was in the code generating the code. Each parameter needed different constraints each time, but the parameters were only given once. Once the debugging was complete, the code worked all together.

The other challenge in the research was using the FindFit function. Without the specified initial conditions, it would find the fit in the wrong places. The constraints were needed in a known situation of the graph so that the most numerous particle sizes would go to each of the parameters and form different normal distribution graphs that fit into the total distribution. Without the constraint, the parameters found using the function changed every time the code was run. Taming the constraints to find out the intuitive result took many trials and errors.

IDEAS FOR FUTURE RESEARCH

I. Improving Accuracy by Using a Larger Set of Data on the Specific Research

There could be two types of future research. One could be building further on this specific research with a larger set of data to improve the accuracy. With a larger dataset, we could increase the number of normal distribution graphs that form the cumulative distribution graph. We could check if increasing the number of individuals beyond 4 would reduce errors. A larger dataset will also enable us to increase the size of the matrix, which can be tested to see if it has a higher accuracy and precision compared to 4 individual normal distribution functions.

Another research could be done using multiple normal distribution fitting techniques to represent non-normal distribution looking patterns and predict the post-event dataset from the given pre-event dataset. Any strange looking non-normal distributed dataset can be turned and analyzed into multiple normal distributions, resulting in the prediction of post-event distributions.

II. Application of the Method to Predict Post-event Patterns in the Real World

This method can be also used to predict post-event patterns in the real world. For example, an online shopping mall can analyze how well their new advertisement worked by comparing the clicking rates of customers before and after the advertisement. Real world events that have changes in-pre-event data and post-event data can use this method to predict the pattern.

The methodology in this research can be used to find ways to predict the change in the distribution of other pre-events and post-events as well. This research shows that events that display complex distribution functions that do not fit any known distribution functions can still be represented using different weighted sums of known distributions and their parameters, extending this method to areas that deal with unusual distributions. For example, the effect of a certain farming method to the crops' sizes can be tested before and after the use of the method by examining the distribution of crop sizes. This future research is expected to be less challenging after this research dealing with a number of size categories. This future research would have its own challenges as it takes a long time to produce the crops after using the farming method. After predicting the size distributions, farmers can produce bigger crops and leverage the economic profit using the results of this research. Although confronted with unknown distribution models, this research enables to solve it by making weighted sums of known distributions.

References

[1] M.Baron, A.Chamayou, L.Marchioro, J.Raffib, "Radicalar probes to measure the action of energy on granular materials," *Advanced Powder Technology,* vol. 16, no. 3, pp. 199-211, 2005.

[2] G. Casella and R. L. Berger, Statistical Inference, Duxbury: Thomson Learning, 2002.

[3] A. Lyon, "Why are Normal Distributions Normal?," vol. 65, pp. 621-649, 2014.

[4] Casella, George; Berger, Roger.L, Statistical Inference, Duxbury, 2001.

[5] D. S. M. ,. S. S. N. ,. E. A. Augustine B. Makokha, "Effect of Slurry Solids Concentration and Ball Loading on Mill Residence Time Distribution," *International Journal of Mining Engineering and Mineral Processing,* vol. 3, no. 2, pp. 21-27, 2014.

[6] "Wolfram Mathematica," [Online]. Available: http://www.wolfram.com/mathematica/.