Estimating Missing Values from Noisy Price Data with Interpolation

Case Study: Bandung Groceries Price

Sandy Socrates (23516055)
Informatics Graduate Program
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung
23516055@std.stei.itb.ac.id

***Abstract*— Estimating price trends can be done using previously obtained data. But often the data is incorrect because of noise. Noise found in data can be found as missing values or wrong format. To solve this problem, we can use interpolation to replace the noise with estimation based on the neighboring data. In this paper, we are exploring several methods of interpolation used for estimating missing values from groceries price data in Bandung.**

Keywords—groceries price, interpolation

I. INTRODUCTION

Groceries price changes frequently. [1] explains the factors causing the change of prices of groceries in Indonesia. They are production and consumption rate, and pricing policy. The price prediction and control done by government needs previous price data observed in the market. But sometimes there are noises like missing data or invalid data detected. We can see on Fig. 1. below from [2].

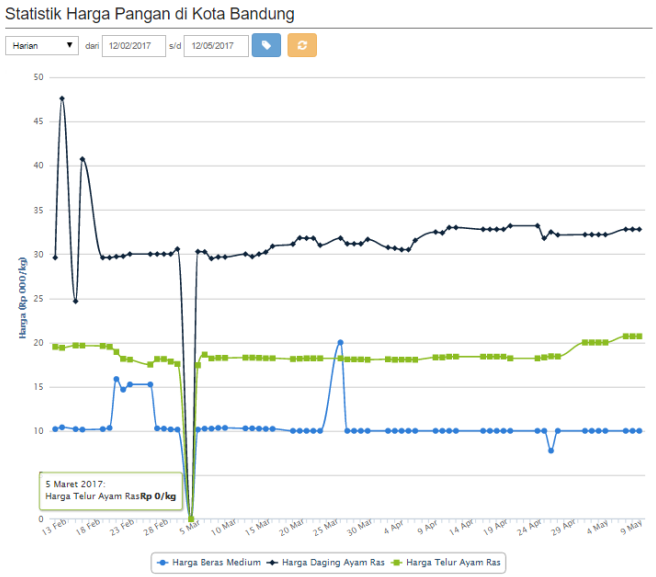

Fig. 1. Some groceries data from [2]

There is this anomaly can be noticed on 5 May 2017 where all three lines go down to 0. This is an impossible value and can be considered noise, invalid data. We can also notice that the data are collected mainly on weekdays, but some weekdays also missing data. This kind of condition is what we are trying to fill with interpolation. With complete data, better prediction and policy can be made to control the price.

II. INTERPOLATION

[3] explains that interpolation is the act of estimating values between precise data points. This is exactly what we are trying to do with the missing and noise data. With interpolation, we can make an estimation based on the surrounding points. This will smooth over the graph built from the data around the missing value. While the returned value itself is only an estimation, this is often close to the nature of the graph itself. There are several ways to do interpolation.

A. Linear Interpolation

Linear interpolation takes two points as reference to generate values between them. You can imagine it as creating a straight line between two points. The function is estimated using (1) below and its usage to estimate f(x) on Fig. 2. .

$$p_1(x) = y_0 + \frac{(y_1 - y_0)}{(x_1 - x_0)}(x - x_0) \qquad (1)$$

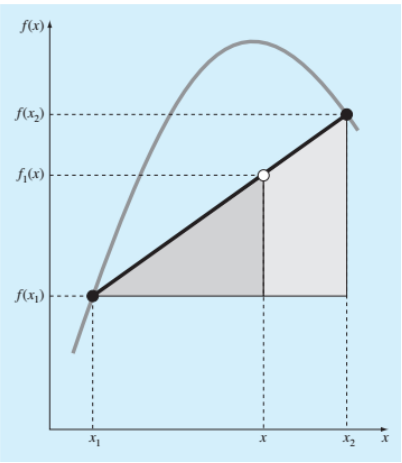

Fig. 2. Linear interpolation from [3]

B. *Quadratic Interpolation*

As we can see previously, linear interpolation cannot fit well for non-linear function. Quadratic interpolation takes three points to gives quadratic function estimation to give better estimation. The etimation is given using (2)(3)(4).

$$a_0 + a_1 x_0 + a_2 x_0^2 = y_0 \quad (2)$$
$$a_0 + a_1 x_1 + a_2 x_1^2 = y_1 \quad (3)$$
$$a_0 + a_1 x_2 + a_2 x_2^2 = y_2 \quad (4)$$

Solve *a*, *b*, and *c* given (x_0, y_0), (x_1, y_1), and (x_2, y_2). Then use the resulted function to estimate the values between the three points. Illustration can be seen on Fig. 3.

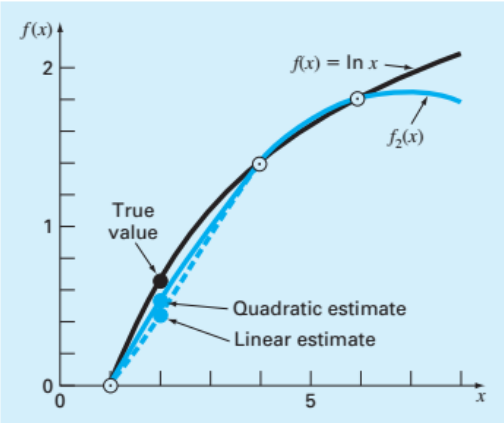

Fig. 3. Quadratic interpolation from [3]

C. *Cubic Interpolation*

We can grasp the concept from the two interpolation before to conclude that quadratic interpolation needs four points. First we will need to represent the estimation function as cubic function and input the known points. And similar step will also be done with (n+1) known points. Illustraion can be foind on Fig. 4.

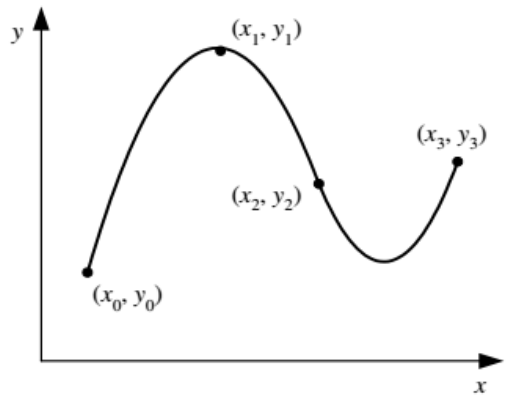

Fig. 4. Cubic interpolation from [4]

D. *Lagrange Interpolation*

To understand Lagrange interpolation, we need to look back into linear interpolation (1). We can rearrange this function into (5).

$$p_1(x) = \frac{(x-x_1)}{(x_0-x_1)}(x - x_0) + \frac{(x-x_0)}{(x_1-x_0)} \quad (5)$$

We can then change it into (6), (7), (8), (9), and (10).

$$p_1(x) = a_0 L_0(x) + a_1 L_1(x) \quad (6)$$
$$a_0 = y_0 \quad (7)$$
$$L_0(x) = \frac{(x-x_1)}{(x_0-x_1)} \quad (8)$$
$$a_1 = y_1 \quad (9)$$
$$L_1(x) = \frac{(x-x_0)}{(x_1-x_0)} \quad (10)$$

From these equations we can then create general form of Lagrange Interpolation (11).

$$p_n(x) = \sum_{i=0}^{n} a_i L_i(x) \quad (11)$$

With $a_i = y_i$ and (12).

$$L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{(x-x_j)}{(x_i-x_j)} \quad (12)$$

Illustration for the first order Lagrange interpolation can be found on Fig.5.

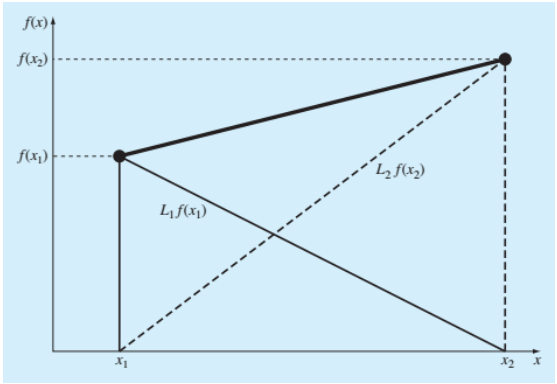

Fig. 5. First order Lagrange interpolation from [3]

E. *Newton Interpolation*

Lagrange interpolation stated before is not liked for these reasons:

1. It needs big computation for interpolatimg a point in the graph. Each interpolation needs to be computed independently as there is no reusable component for the computation.
2. If the number of reference point is changed then the estimations done from previous interpolations also need to be changed.

Newton interpolation is an alternative to answer those problems. The function is a recursive function defined on (13) and (14)

$$p_n(x) = p_{n-1}(x) + a_n \prod_{i=0}^{n-1}(x - x_i) \quad (12)$$
$$p_0(x) = a_0 \quad (13)$$
$$a_0 = f(x_0) \quad (14)$$

a is defined on (15), (16), and (17)

$$a_1 = f[x_1, x_0] \quad (15)$$
$$a_2 = f[x_2, x_1, x_0] \quad (16)$$
$$a_n = f[x_n, x_{n-1}, \dots, x_1, x_0] \quad (17)$$

f[] is defined on (18),(19), and (20)

$$f[x_i, x_j] = \frac{f(x_i) - f(x_j)}{x_i - x_j} \quad (18)$$
$$f[x_i, x_j, x_k] = \frac{f[x_i, x_j] - f[x_j, x_k]}{x_i - x_k} \quad (19)$$
$$f[x_n, x_{n-1} \dots, x_0] = \frac{f[x_n, x_{n-1}, \dots, x_1] - f[x_n, x_{n-1}, \dots, x_1]}{x_n - x_0} \quad (20)$$

The functions are best described using a divided difference table depicted on Table 1.

TABLE I. 3RD ORDER NEWTON INTERPOLATION DIVIDED DIFFERENCE TABLE

x_i	$f(x_i)$	First	Second	Third
x_1	$f(x_1)$	$f[x_2, x_1]$	$f[x_3, x_2, x_1]$	$f[x_4, x_3, x_2, x_1]$
x_2	$f(x_2)$	$f[x_3, x_2]$	$f[x_4, x_3, x_2]$	
x_3	$f(x_3)$	$f[x_4, x_3]$		
x_4	$f(x_4)$			

The differences then can be used to be substituted to (12) to estimates *x*. Illustration of the interpolation can be seen on Fig.6.

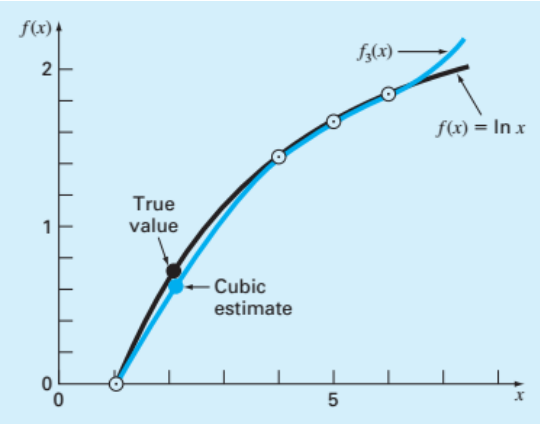

Fig. 6. Third order Newton interpolation of ln(x) from [3]

F. *Newton-Gregory Interpolation*

Newton-Gregory interpolation is a special case of Newton interpolation. It's Newton interpolation with consistent distance between each reference points. Because of the consistent distance, the table is easier to buid. We call the table as difference table. As there are two kind of difference, forward and backward, the interpolation also divided into two kind: forward Newton-Gregory and backward.

III. EXPERIMENTS

As stated before on Introduction, we will work on some groceries price data in Kota Bandung to estimate the value to replace the noise.

A. *Data Source*

Data is available from [2]. [2] provides reports for selection of groceries price in West Java, Indonesia. This service is run by Forum Koordinasi Pengendalian Inflasi and sponsored by BKPD Jawa Barat, Bulog Jawa Barat, Dinas Pertanian Provinsi Jawa Barat, Disperindag, and FKPI Provinsi Jabar. Data provided here can be downloaded as CSV files.

The data we worked on are chicken, rice, and eggs prices from 12 February 2017 to 12 May 2017 in Kota Bandung[1].

B. *Experiment Tasks*

The tasks:

1. Run each interpolation method for known value

   Method used will be Lagrange, Newton, forward Newton-Gregory, and backward Newton-Gregory

2. Calculate the errors

   Mean of the errors for each grocery type is then taken

3. Select best method based on the error mean

   The best method is then used to estimate the real noise

C. *Hypothesis*

Based on the explanation on II, we think the best method is Newton-Gregory interpolation (given the same order used on testing) for estimating value of noises. Because the distance is constant hence the details in the middle of the graph is higher.

D. *Data*

To get the errors of each method we will use data from Table II.

TABLE II. TESTING DATA

Date	t	Rice	Chicken Meat	Eggs
3 March 2017	-2	10.143	30.571	17.571
6 March 2017	1	10.143	30.286	17.429
7 March 2017	2	10.250	30.250	18.625
8 March 2017	3	10.250	29.500	18.188

[1] Obtained from http://priangan.org/publik/datapercity/1

Date	t	Rice	Chicken Meat	Eggs
9 March 2017	4	10.333	29.667	18.250
10 March 2017	5	10.333	29.667	18.250

8 March 2017 will be the data used to calculate the error. Error is then defined as the mean of the three errors per method.

E. Results

These are the experiment results

1) *Lagrange Interpolation*

Order number used is 3 with data on Table III.

TABLE III. TESTING DATA FOR LAGRANGE INTERPOLATION

Date	t	Rice	Chicken Meat	Eggs
3 March 201	-2	10.143	30.571	17.571
6 March 2017	1	10.143	30.286	17.429
7 March 2017	2	10.250	30.250	18.625
9 March 2017	4	10.333	29.667	18.250

Interpolation results provided on Table IV.

TABLE IV. LAGRANGE INTERPOLATION RESULTS

	Rice	Chicken Meat	Eggs
Estimation	10.033	30.077	19.156
Real Value	10.250	29.500	18.188
Difference (abs)	217	577	968
Error (/Estimation)	0.0212	0.0196	0.0532

Errors is calculated as mean below.

$$Mean = \frac{0.0212 + 0.0196 + 0.0532}{3} = 3.13\%$$

2) *Newton Interpolation*

Order number used is 3 with data on Table V.

TABLE V. TESTING DATA FOR NEWTON INTERPOLATION

Date	t	Rice	Chicken Meat	Eggs
3 March 201	-2	10.143	30.571	17.571
6 March 2017	1	10.143	30.286	17.429
7 March 2017	2	10.250	30.250	18.625
9 March 2017	4	10.333	29.667	18.250

Interpolation results provided on Table VI.

TABLE VI. NEWTON INTERPOLATION RESULTS

	Rice	Chicken Meat	Eggs
Estimation	10.033	30.077	19.156
Real Value	10.250	29.500	18.188
Difference (abs)	217	577	968
Error (/Estimation)	0.0212	0.0196	0.0532

Errors is calculated as mean below.

$$Mean = \frac{0.0212 + 0.0196 + 0.0532}{3} = 3.13\%$$

3) *Forward Newton-Gregory Interpolation*

Order number used is 3 with data on Table VII.

TABLE VII. TESTING DATA FOR FORWARD NEWTON-GREGORY INTERPOLATION

Date	t	Rice	Chicken Meat	Eggs
6 March 2017	1	10.143	30.286	17.429
7 March 2017	2	10.250	30.250	18.625
9 March 2017	4	10.333	29.667	18.250
10 March 2017	5	10.333	29.667	18.250

Interpolation results provided on Table VIII.

TABLE VIII. FORWARD NEWTON-GREGORY INTERPOLATION RESULTS

	Rice	Chicken Meat	Eggs
Estimation	10.333	29.667	18.250
Real Value	10.250	29.500	18.188
Difference (abs)	83	167	62
Error (/Estimation)	0.0081	0.0057	0.0034

Errors is calculated as mean below.

$$Mean = \frac{0.0081 + 0.0057 + 0.0034}{3} = 0.57\%$$

4) *Backward Newton-Gregory Interpolation*

Order number used is 3 with data on Table IX.

TABLE IX. TESTING DATA FOR BACKWARD NEWTON-GREGORY INTERPOLATION

Date	t	Rice	Chicken Meat	Eggs
6 March 2017	1	10.143	30.286	17.429
7 March 2017	2	10.250	30.250	18.625
9 March 2017	4	10.333	29.667	18.250
10 March 2017	5	10.333	29.667	18.250

Interpolation results provided on Table X.

TABLE X. BACKWARD NEWTON-GREGORY INTERPOLATION RESULTS

	Rice	Chicken Meat	Eggs
Estimation	10.250	30.250	18.625
Real Value	10.250	29.500	18.188
Difference (abs)	0	750	437
Error (/Estimation)	0	0.0254	0.0240

Errors is calculated as mean below.

$$Mean = \frac{0 + 0.0254 + 0.0240}{3} = 1.64\%$$

F. *Best Method*

The errors from the experiments are provided on Table XI.

TABLE XI. EXPERIMENT RESULTS

Interpolation Method	Error
Lagrange	3.13%
Newton	3.13%
Forward Newton Gregory	0.57%
Backward Newton Gregory	1.54%

The method yields best is Forward Newton-Gregory. This method outperforms the other methods for our data test.

But this method cannot be used to estimate the prices on 5 May 2017. It is caused by no data is available on 4 May 2017 (Refer to Table II.). And the method needs the estimated price to be in the middle of the graph to have best estimation.

G. *Noise Value*

The best method based on the errors on our experiments is the Forward Newton-Gregory. But there is no available data on 4 May 2017 hence we cannot calculate the estimation using both Forward and Backward Newton Gregory Interpolation. Hence the estimation is done using Lagrange and Newton Interpolation

Order number used is 3 with data on Table XII.

TABLE XII. REFERENCE DATA FOR FIXING NOISE VALUE

Date	t	Rice	Chicken Meat	Eggs
3 March 2017	-2	10.143	30.571	17.571
6 March 2017	1	10.143	30.286	17.429
7 March 2017	2	10.250	30.250	18.625
8 March 2017	3	10.250	29.500	18.188

Lagrange and Newton Interpolation results provided on Table XIII.

TABLE XIII. INTERPOLATION RESULTS

	Rice	Chicken Meat	Eggs
Lagrange	10.025	30.054	15.953
Newton	10.025	30.054	15.953

Plotted data can be seen on Fig.7.

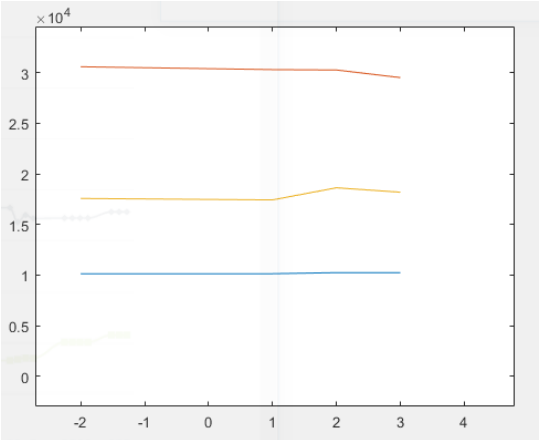

Fig. 7. Plotted data

IV. CONCLUSION AND SUGGESTION

A. *Conclusion*

These are some conclusions we take from the experiments:

1) The best method tested from our experiments is Forward Newton-Gregory interpolation. This method's error scorre id 0.57% and this outperforms the other method we used.

2) Newton-Gregory method depends on the availability of data surrounding the noise value. It will not be able to give results if the neighboring data are missing. Also if the neighboring data ae also noises, then there is no meaning to our interpolation.

3) For data which is not updated regulary, Newton Gregory will not be able to give estimation Hence it will also need another interpolation method for special cases like this. For noises with complete neighbours we can use Newton-Gregory. Otherwise, use another methods. In our case, we use both Lagrange and Newton methods.

B. *Suggestions*

These are some suggestion from the experiments:

1) A price monitoring service like [2] should update prices regularly. Missing data will need to be eliminate, as it will be pointless if we don't have regular data.

2) Next experiments should add more interpolation methods for better comparison.

3) Noises also should be able to be spotted automatically. Instead of manually reading the data and slicing them for interpolation methods.

ACKNOWLEDGMENT

Special thanks for Stan and Tisha for the support during the writing of the paper.

REFERENCES

[1] AH. Alian, S. Mardianto, and M Ariani, "Factors influencing production, consumption, and price of rice and groceries inflation", "Faktor-faktor yang mempengaruhi produksi, konsumsi, dan harga beras serta inflasi bahan makanan", Jurnal Agro Ekonomi, Vl 22, No 2, 2004.

[2] Forum Koordinasi Pengendalian Inflasi, "Priangan – Groceries price information portal", "Priangan – Portal informasi harga pangan", service available at http://priangan.org

[3] SC. Chapra, "Applied numerical methods with MATLAB® for engineers and scientist", Third Edition, McGraw Hill

[4] R Munir, Lecture slide:"07 Polynomial Interpolation", Slide Kuliah:"07 Interpolasi polinom"

NON PLAGIARISM STATEMENT

I hereby declare that this paper written by me in my own words, not an adaptation or translation of another paper, and certainly not a result of plagiarism.

Bandung, 12 May 2017

Sandy Socrates
23516055