ANALYSIS AND DESIGN SIDOARJO ON HANDS (SOH) SYSTEM FOR SUPPORTING SIDOARJO POTENTIAL PROMOTION

[1]Rani Purbaningtyas, [2]Arif Arizal, [3]Tri Wardoyo
[1,2]Lecturer, Department of Informatics Engineering, Bhayangkara University of Surabaya
[3]Lecturer, Department of Civil Engineering, Bhayangkara University of Surabaya
Jl. Ahmad Yani no. 114, Surabaya

e-mail: raniubhara@gmail.com

ABSTRACT

SoH is mobile application android based which has importance in a discipline oriented to help people get information about local potential of Sidoarjo. The analysis and design SoH presented here carry out objective to support promotion of potential of Sidoarjo.

Specific analysis and design tools will be essential to assure the development of SoH itself. This paper addresses the results of analysis and design SoH. The result includes 5 steps : digitized map of Sidoarjo, design of system architecture, flowchart for describing SoH system flow, design of SoH database using entity relationship diagram (ERD), data flow diagram (DFD) for describing relation between system flow and SoH database, and design of SoH user interface

We evaluate the result of analysis and design of SoH. The final result is used as blueprint to develop an effective SoH system. In this paper, the final result has satisfied user requirements. The result will be used as preliminary study on the future research.

Keywords : analysis and design system, Sidoarjo on Hands, potential promotion

Introduction

Sidoarjo is one of district in Jawa Timur province. Its coverage area is 714,243 km^2 (Bagian Telekomunikasi dan Informatika Kabupaten Sidoarjo, 2015). Sidoarjo provides many kind of local potential which can attract tourists for visiting Sidoarjo. The local potentials are including potential of industry, agriculture, fishery, crafts, tourism and culinary.

The local government has implemented various way to promote potentials of Sidoarjo using conventional media promotion such as newspaper, billboard, and local television. But it seems the promotion method is less optimum. The official website is owned by local government and tourism department has less known even by local population itself. Besides that, the information which is shown inside less complete and uninformative.

With the increase of mobile application technology, it was possible to implement this technology for promoting potentials of Sidoarjo. We can take advantages of Sidoarjo on Hands (SoH) mobile application as a promotion medium. Thus the need arose to study the analysis and design SoH that could help in introducing potentials of Sidoarjo. Next, we can get all information about local potentials of Sidoarjo easily. So, people who wants to know deeply about Sidoarjo can retrieve the information using SoH.

Research Methodology

The following are the procedures for finishing the analysis and design phase of SoH :

1. Analyze and map all of potentials of Sidoarjo. This first phase aimed to classify each potential into its main category.
2. Collect primary data of local potentials at each districts in Sidoarjo.
3. Collect and validate secondary data from related departments of local government at each district in Sidoarjo.
4. Combine analysis between primary and secondary data.
5. Digitize map of Sidoarjo areas.
6. Set marking point at the Sidoarjo digital maps.
7. Create design of SoH system architecture
8. Describe of SoH system flow using flowchart.
9. Design of SoH database using entity relationship diagram (ERD)
10. Design of SoH of user interface.
11. Evaluate the result of analysis and design phases above.

Discussion

The first phase is analyze and map all of potentials of Sidoarjo. At the first time, we had six main categories for local potentials of Sidoarjo. They are potential of industry, agriculture, fishery, crafts, tourism and culinary. We used this pattern when collected primary and secondary data of local potentials at each districts in Sidoarjo. Then we took the combination of those data for analysis. We found that some categories need to be decomposed into some sub categories. They are :

1. Main category of industry potential has three sub categories which are home appliances, exhaust, and iron tools.
2. Main category of craft potential has six categories which are bag and suitcase, embroidery, batik, convection, hat, and slippers.
3. Main category of tourism potential has two sub categories which are natural tourism and artificial tourism.
4. Main category of culinary potential has seven sub categories which are crackers, tofu, tempe, salted egg, petis, nugget, and sausage.

The fourth phase is digitize map of Sidoarjo areas. We implemented GoogleMaps API library. Then, we marked the Sidoarjo map using pointer to specific area according to the result of second phase. As general, the architecture of SoH system as shown in the figure 1 below :

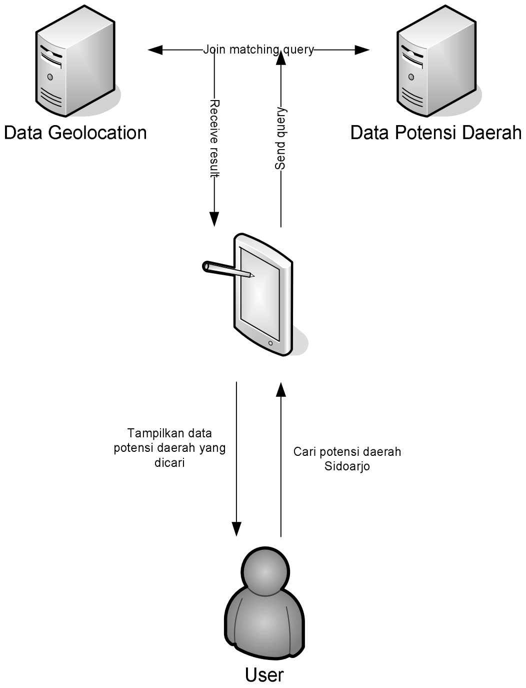


Figure 1. The architecture of SoH system

Figure 1 showed that user of SoH will retrieve the information about potentials of Sidoarjo easily. User only need to entry data about any kind of potential that he wanted. System will use the input data as predefined data searching inside server. System implemented join matching query as searching method. This searching method is combine between text data searching at Sidoarjo potential server and maps data searching at geo spatial server. The searching result will be shown at client side of SoH application itself.

As stated previously, in order to derive an effective SoH system, we provide an optimum database design of SoH. Following is entity relationship diagram of SoH :

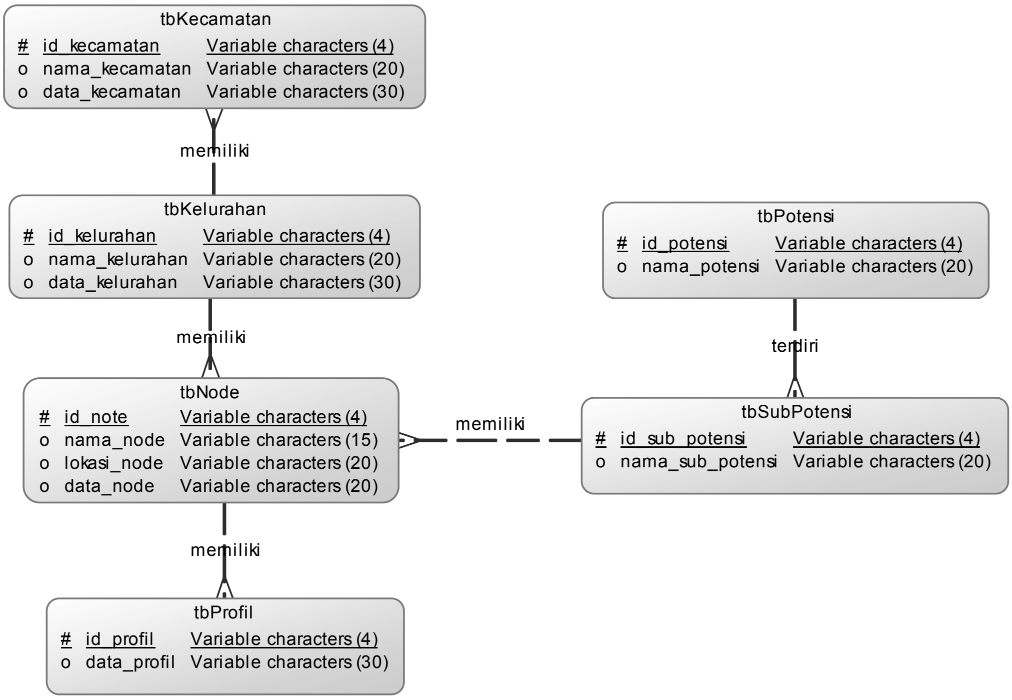
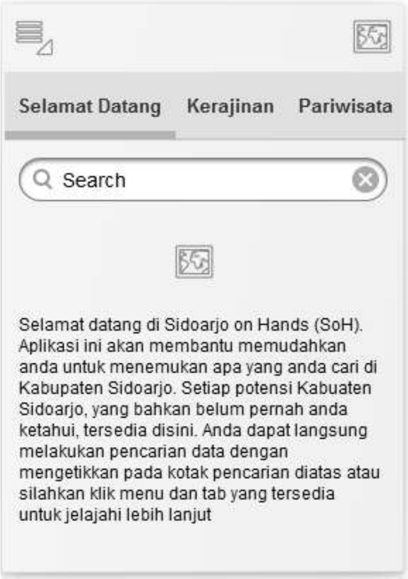

tbKecamatan
# id_kecamatan Variable characters (4)
o nama_kecamatan Variable characters (20)
o data_kecamatan Variable characters (30)
memiliki
tbKelurahan
# id_kelurahan Variable characters (4)
o nama_kelurahan Variable characters (20)
o data_kelurahan Variable characters (30)
tbPotensi
# id_potensi Variable characters (4)
o nama_potensi Variable characters (20)
memiliki
terdiri
tbNode
# id_note Variable characters (4)
o nama_node Variable characters (15)
o lokasi_node Variable characters (20)
o data_node Variable characters (20)
memiliki
tbSubPotensi
# id_sub_potensi Variable characters (4)
o nama_sub_potensi Variable characters (20)
memiliki
tbProfil
# id_profil Variable characters (4)
o data_profil Variable characters (30)

Figure 2. Design of SoH entity relationship diagra

Each table has own function as shown as below :

1. Table of Kecamatan was used to save data of 18 districts in Sidoarjo. They are Balongbendo, Buduran, Candi, Gedangan, Jabon, Krembung, Krian, Porong, Prambon, Sedati, Sidoarjo, Sukodono, Taman, Tanggulangin, Tarik, Tulangan, Waru dan Wonoayu (Bagian Telekomunikasi dan Informatika Kabupaten Sidoarjo, 2015).
2. Table of Kelurahan has function to keep data of village in each districts of Sidoarjo. Sidoarjo has amount of 322 villages.
3. Table of Potensi was used to save data of main category of potentials of Sidoarjo. We have six main category of potentials data which are potential of industry, agriculture, fishery, crafts, tourism and culinary.
4. Table of SubPotensi was used to save subcategory data of each main category.
5. Table of Node was used to save all of location points which contain specific potential.
6. Table of Profil was used to save supporting data for each location which has own potential.

Following are design of SoH user interface which are including startup form and main forms of SoH :



Selamat datang di Sidoarjo on Hands (SoH). Aplikasi ini akan membantu memudahkan anda untuk menemukan apa yang anda cari di Kabupaten Sidoarjo. Setiap potensi Kabuaten Sidoarjo, yang bahkan belum pernah anda ketahui, tersedia disini. Anda dapat langsung melakukan pencarian data dengan mengetikkan pada kotak pencarian diatas atau silahkan klik menu dan tab yang tersedia untuk jelajahi lebih lanjut

Figure 3. Welcome screen

While SoH was running, startup form of SoH was shown by figure 3. This form provided a brief description of SoH and simple guidance for operating SoH. User can use the searching box for finding general information which is provided by SoH.

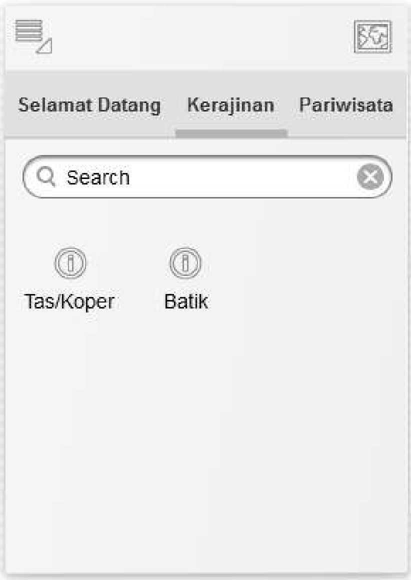


Figure 4. Design form for each main category

Figure 4 showed design of main form of SoH. Each SoH main category will be displayed using tab dialog. Each tab dialog contains sub category of each main category. The sub category will be displayed using mini icon. If user chooses one of mini icon, then user will pursue to the next form as shown in figure 5. User can use searching box in this form for finding the information in accordance with the chosen main category.

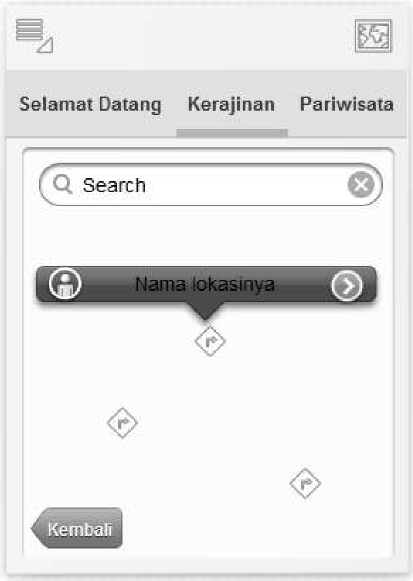


Figure 5. Design form for each sub category

Figure 5 showed form within Sidoarjo map inside. The digitize map had completed with points. The points which is shown on the map in accordance with sub category had been chosen by user from the previous form. Each point displayed the name of each potential in Sidoarjo. If user chose one of these points, then SoH will display the profile of its potential on the next form. See figure 6.

Figure 6. Design form of potential profile

The result of analysis and design phases above will be evaluated. Below are the result of evaluation phase :

1. There are two searching box with different purpose. Searching box in the welcome screen was used to find general information inside SoH. And the other one, which is in the main form, can be used to find information which is suitable with user chosen category.
2. System will search local potential data using join matching query as searching method. This searching method is combine between text data searching at Sidoarjo potential server and maps data searching at geo spatial server.
3. System will implement Apriori TID algorithm on the future research. It is aimed to optimize result of SoH searching (Astuti, 2016)
4. SoH doesn’t provide login form. System will create user log activity by Universal Unique Identifier (UUID). System will use this identifier as the searching key on the next research.

Conclusion

This research discuss analysis and design of SoH system using Structured Analysis and Design (SAD) approach. It presents design of SoH system in flowchart, entitiy relationship diagram, and data flow diagram. On the next discussion, SoH will implement Apriori TID algorithm for searching engine inside. Also, SoH will use UUID as key of data searching. The output of this research will be used as a blueprint for developing an effective SoH. Based on this research, we conclude the final result has satisfied user requirements.

Acknowledgement

DRPM Research and Higher Education Ministries of Indonesia for providing research funds

References

[1] Anonimus, 2015, Sidoarjo Dalam Angka 2015, BPS Kab Sidoarjo.

[2] Astuti, Femi Dwi, Widyastuti Andriyani, 2016, *Optimasi Pemrograman Query Untuk Algoritma Apriori Berbasis Asosiasi Data Mining*, Jurnal Riset Sistem Informasi & Teknik Informatika (JURASIK) Vol. 1 No. 1 Juli 2016, STIKOM Tunas Bangsa, Pematang Siantar.

[3] Bagian Telekomunikasi dan Informatika Kabupaten Sidoarjo, 2015, Website Resmi Pemkab Sidoarjo, www.sidoarjokab.go.id

[4] Jogiyanto, H.M., 2005, *Analisis & Desain Sistem Informasi: Pendekatan Terstruktur, Teori, dan Aplikasi Bisnis*, Edisi Ketiga. Andi Offset, Yogyakarta.

POWER FACTOR CORRECTION OF THREE PHASE AC-DC CONVERTER VIA CURRENT CONTROLLER AND PWM TECHNIQUE

SAIDAH,

Electrical Engineering Department, Bhayangkara University

Ahmad Yani 114, Surabaya, East Java, Indonesia

saidah@ubhara.ac.id

ABSTRACT

Use of three phase AC-DC converters for DC loads can cause high harmonics on the AC side, reducing power factor. To improve the power factor can not be done by adding a passive filter on the input side of the conveter. This paper proposes the improvement of power factor through the regulation of current on the AC side and the PWM technique . The current control method uses PI and the exact PWM technique is SVPWM. The System behavior and power factor observations are studied through simulations.

Keywords: *AC-DC Converter, SVPWM, Current Controller*

1. INTRODUCTION

The AC/DC power converters are extensively used in various applications like household electric appliances, power conversion, dc motor drives, adjustable-speed ac drives, HVDC transmission and UPS. The main problem faced by using switchs on an AC-DC converter is to generate high harmonics on the AC side, reducing power factor in low or medium power applications. Normally the input voltage to an AC-to-DC converter is sinusoidal but the input current is non-sinusoidal i.e. harmonic currents are present in the ac sides. Harmonics have a negative effect on the power factor as well. The addition of harmonic currents to the fundamental component increases the total rms current hence harmonics will affect the power factor of the circuit. Unity power factor, lower harmonic current or low input current THD and fixed DC output voltage with minimum ripple are the important parameters in rectifier.

The switch used in this research already uses the transistor family and has left diode and thiristor [1-2]. The use of switch components is IGBT, because IGBT has a higher switching speed / working frequency than other transistors. That's why IGBT is often used in drivers (motor drives) that require large currents and operate in high voltage, because it has better efficiency than other transistor types [3-4]. In figure 1. A three phase phase rectifier with an igbt switch component has been shown. To flatten the ripple on the DC side added capacitor components.

Several ways have been done by researchers for power factor improvements such as by adding passive filters on the input side [5-12]. But the addition of this passive filter does not solve the problem of power factor decline. The method used in this paper by SVPWM technique is a pulse width modulation technique on the IGBT switch component and adjusts the current amplitude on the AC side.

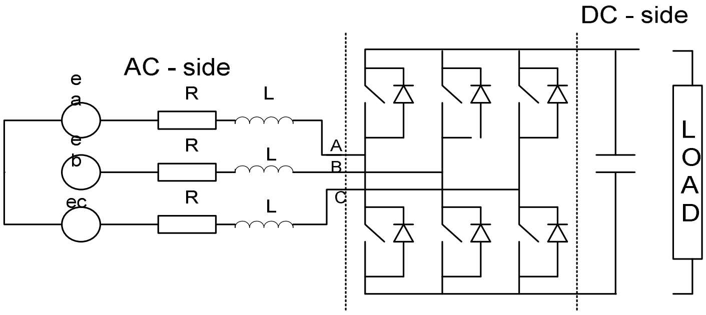


Fig. 1 System of Three-Phase AC-DC converter

2. DESCRIPTION OF SYSTEM

System of the three phase AC-DC converter is shown in Fig. 1. The System is consist of the AC voltage is a balanced three phase supply, IGBT is ideal switch and lossless. Where e_a , e_b and e_c are the phase voltage of three phase balanced voltage source , R and L are mean resistance and inductance respectively, C is smoothing capacitor across the DC bus, R_L is DC side Load.

The following equation describe the dynamic behaviour of the AC-DC converter using space vector :[13]

$$v(t) = e(t) - Ri(t) – L\, di(t)/dt \tag{1}$$
$$v(t) = ½\, s(t)v_o(t) \tag{2}$$
$$i_o(t) = (3/4)\, \mathrm{Re}\{s(t)i^*(t)\} \tag{3}$$
$$i_c(t) = C\, dv_o/dt = i_o(t) - i_L(t) \tag{4}$$

Where
$v(t)$ = rectifier input voltages
$i(t)$ = rectifier input currents
$i^+(t)$ = complex conjugate of i(t)
$e(t)$ = the input line voltages
$i_o(t)$ = rectifier output current
$i_c(t)$ = capacitor current
$i_L(t)$ = load current
$s(t)$ = switching function
$v_o(t)$ = output voltage

3. METHODE OF THE SYSTEMS

3.1 SPACE VECTOR PWM (SVPWM)

SVPWM is to form or encode analog signals into on-off signals for switching using space vector. A Three Phase AC-DC converter has eight active states comprising six active states and two zero states. If depicted in the space vector field, the space vector voltage of the six active states divides the space vector into the same six sectors while the space vector voltage of the zero state lies at the center of the reference plane.

Figure 2 shows the voltage space vector for the AC-DC converter. Vector Vs1, Vs2 .... Vs8 is a vector stationer, meaning it does not rotate by time so its magnitude is determined by dc voltage and its argument is unchanged. Using the PWM modulation technique, it is possible to generate dc voltage with reference voltage vector Vs through vector state synthesis

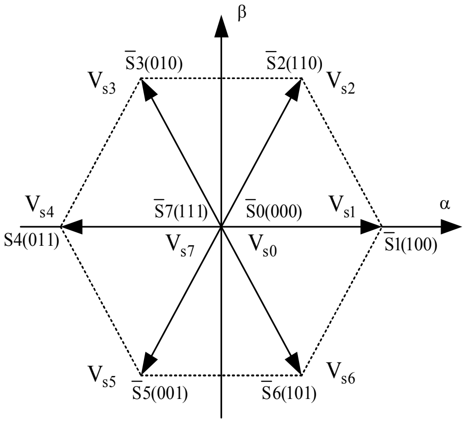


Fig. 2. Space Vector PWM

The six sectors are

$$0^0 < \text{sektor I} < 60^0 \qquad 180^0 < \text{sektor IV} < 240^0$$
$$60^0 < \text{sektor II} < 120^0 \qquad 240^0 < \text{sektor V} < 300^0$$
$$120^0 < \text{sektor III} < 180^0 \qquad 300^0 < \text{sektor VI} < 360^0$$

.

When state 1 and state 2 voltages are generated alternately and at high speeds with respective durations are T1 and T2 in a given time interval, the space vector voltage generated by state1 and state 2 forms the resultant space vector voltage Vs as:

$$V_s = V_\alpha + V_\beta \qquad (5)$$
$$V_\alpha = T_1\ V_{s1} \quad \text{and} \quad V_\beta = T_2\ V_{s2} \qquad (6)$$

Generally the space vector voltage Vs is formed by the vector of adjacent states in the sector where the desired vector voltage space is located. Thus the space vector voltage Vs in sector 2 is formed by vector state 2 and state 3 and so on in other sectors.

The process of calculating the time duration of the active states and the zero state is performed, when both components of the reference voltage vector and sector position are known. Time state zero T0, active state time T1 and T2 can be calculated by considering triangle on sector 1

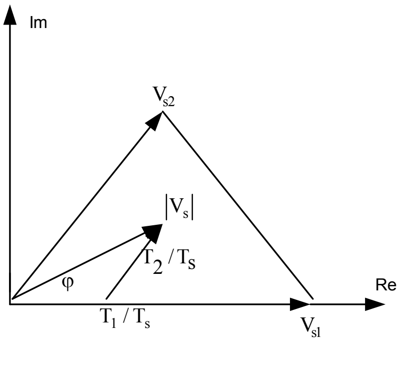


Fig. 3. Determine the switching time of each Sektor I

T 1 dan T 2, can be obtained by considering the following relationship.:

$$\frac{T_1/T_s}{\sin(\pi/3-\varphi)}=\frac{T_2/T_s}{\sin\varphi}=\frac{|V_s|}{\sin 2\pi/3} \tag{7}$$

$$\frac{T_1/T_s}{\sin(\pi/3-\varphi)}=\frac{|V_s|}{\sin 2\pi/3}\quad,\quad \text{then}\quad \frac{T_1}{T_s|V_s|}=\frac{\sqrt{3}}{2}\sin(\pi/3-\varphi_n)=\frac{\sqrt{3}}{2}\cos(\varphi_n+\pi/6) \tag{8}$$

$$\frac{T_2/T_s}{\sin\varphi}=\frac{|V_S|}{\sin 2\pi/3}\quad,\text{ then}\quad \frac{T_2}{T_s|V_s|}=\frac{\sqrt{3}}{2}\sin\varphi_n \tag{9}$$

Determine T0 = Ts - Tm - Tm + 1 used for OFF time for AC-DC converter. So in each sampling period there are two active states m and m + 1. T1, T2 and T0 times, it can be determined which combination of switches should be activated according to state and the length of time the active states and zero state.

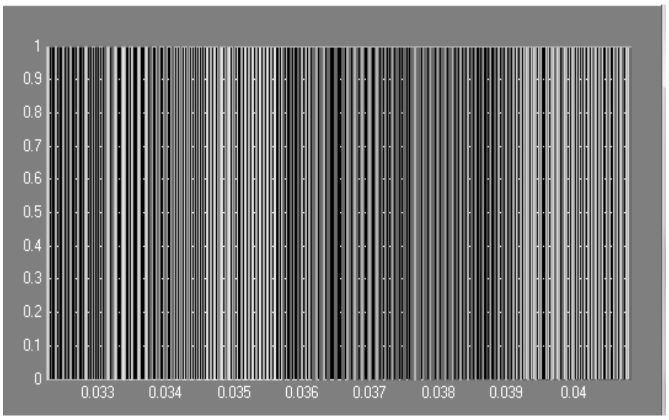


Fig 4. Combinationi T0, T1 and T2

2.2 Current Controller

The control current to get variation of the DC voltage based on SVPWM for three phase of Converter and uses switching variable in d–q frame (s_d , s_q) as input instead of voltage (v_d , v_q) or current (i_d , i_q) that are often used. The system also used a PI controller to control the amplitude of the line current based on the error of the DC voltage. Therefore, a regulated variation of line current amplitude will obtained a low Total harmonic distortion of the line current, a DC voltage variation stable, unity power factor, low ripple. The PI Current Control Proposed was showed in Fig 5.

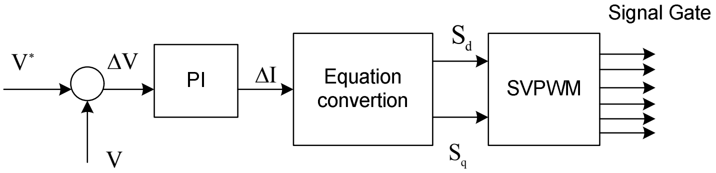


Fig. 5. Block Diagram of the current controller

Convertion Equation

$$s_d=\frac{2}{v_o}\left\{E_m-\frac{L}{T_s}\Delta I_m\right\} \tag{10}$$

$$s_q=\frac{2}{v_o}[-\omega L I_m] \tag{11}$$

4. RESULTS

The simulation has been done using MATLAB/SIMULINK software which it is easy to implement. Various Parameters Used for Simulation Study: AC input voltage = 220V with Supply frequency = 50Hz, Load resistance = 60Ω, DC reference voltage = 600V, Switching Frequency = 9KHz. Figure 6 shows the transient response of the DC link voltage for a switching frequency of 9KHz . The Line current shows on the AC side is stable at 0.04 sec with THD 2.457 on Fig. 7.

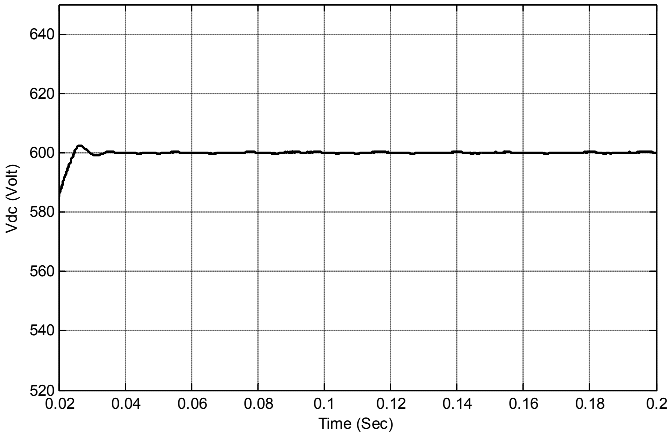


Fig. 6. Simulation result for DC-link voltage dynamics

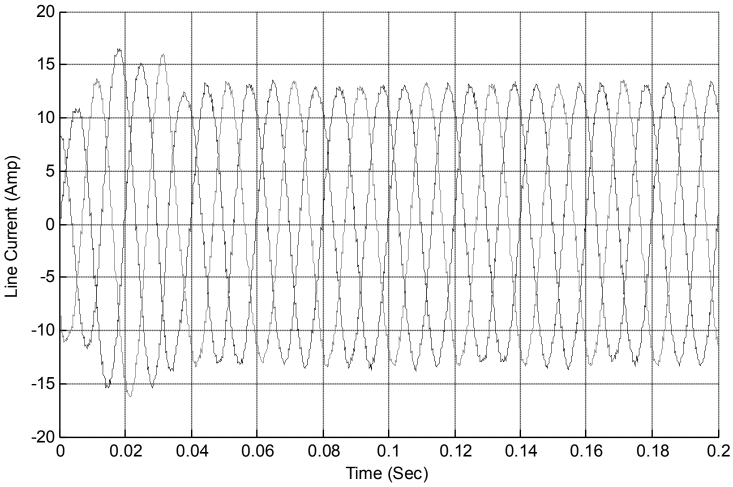


Fig. 7. Simulation result for line current

Fig. 8 shows the voltage and current on line side. We can see the current of sinusoidal wave is the same phase with the voltage. The power factor calculation is achieved using with difference switching frequency. Fig 9. Shows power factor at switching frequency 9 KHz and Fig. 9 shows Power Factor (PF) at switching frequency 5 KHz

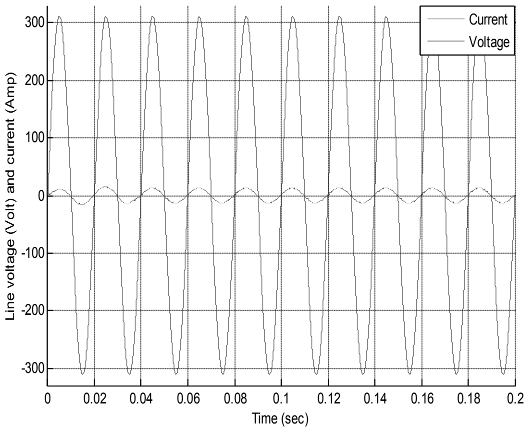
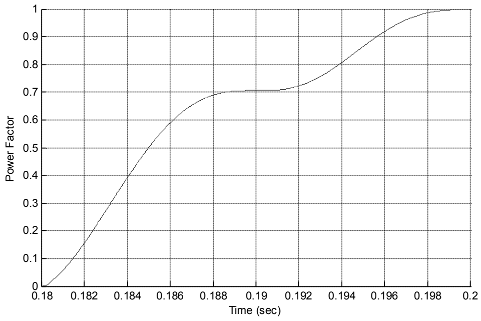


Fig. 8. Simulation result for line current and line voltage



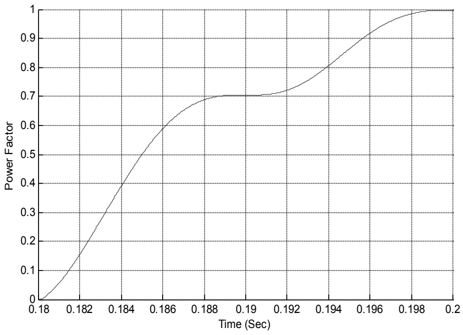


Fig.9 Power Factor (PF) with switching frequency 9 KHz

Power Factor (PF) with switching frequency 5 KHz

5. CONCLUSIONS

The System behavior and power factor observations are studied through simulations using matlab/simulink. From the simulation, power factor obtained for frequency switching 9 khz is 0.998, Total Harmonic Distortion 2.457 and for frequency switching 5 khz is 0.9968, Total Harmonic Distortion 4.136. Thus, using SVPWM method and PI current controller on three phase ac-dc converter are obtained unity power factor and minimum Total Harmonic Distortion, stable DC voltage

REFERENCES

[1] Suma Umesh, L.Venkatesha, Usha A, (2014), " Active Power Factor Correction Technique for Single Phase Full Bridge Rectifier", IEEE

[2] Mukesh kumar, Prof. Gautam Kumar Panda, Prof. Pradip Kumar Saha, (2015), "Power Factor Correction of Three Phase Rectifier with Interleaved Boost Converter", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering", Vol. 4 , Issue 5 , May 2015

[3] K. Georgakas, A. Safacas, (2007), "Power Factor Improvement of an AC-DC Converter via Appropriate sPWM Technique, Proceedings of the 15th Mediterranean Conference on Control & Automation, July 27-29, Athens Greece.

[4] Saidah, M. Hery Purnomo, M. Ashari, (2013), "High Performance of Nonlinear Active Rectifier Voltage and Power Factor Control Using Feedback Linearization", International Review of Electrical Engineering (I.R.E.E.).

[5] Kataoka, T., Mizumachi, K., and Miyairi, S., (1979),“A Pulse-width Controlled AC-to-DC Converter to Improve Power Factor and Waveform of AC Line Current”, IEEE Transactions on Industry Applications, Vol. IA-15, Issue No. 6, pp.670-675,.

[6] Dixon, J. W., and Ooi, B. T., “Indirect Current Control of a Unity Power Factor Sinusoidal Current Boost type Three-phase Rectifier”, IEEE Transactions on Industrial Electronics, Vol. 35, Issue No. 4, pp.508-515, 1988.

[7] Xue, M., and He, M., “Control of Unity Power Factor PWM Rectifier”, Scientific research, Energy and Power Engineering (EPE), Vol. 5, Issue No. 4B, pp.121-124, 2013.

[8] Srivastava, S., and Kumar, S., “Comparative Analysis of Improved Quality Three Phase AC/DC Boost Converters, using SIMULINK”, International Journal of Emerging Technology and Advanced Engineering, Vol. 2, Issue No. 9, pp.427-432, 2012.

[9] Bhat, A. H., and Agarwal, P., “A Comparative Evaluation of Three-Phase High Power Factor Boost Converters for Power Quality Improvement”, IEEE, International Conference on Industrial Technology (ICIT), pp.546-551, 2006.

[10] Zhongjiu, Z., Guofeng, L., and Ninghui, W., “Research on Control Strategy of Three-phase High Power Factor PWM Rectifier”, International Journal of Digital Content Technology and its Applications, vol. 5, Issue No. 8, pp.365-373, 2011.

[11] Balamurugan, R., and Dr. Gurusamy, G., “Harmonic Optimization by Single Phase Improved Power Quality AC-DC Power Factor Corrected Converters”, International Journal of Computer Applications (0975 – 8887), Vol. 1, Issue No. 5, pp.33-40, 2010.

[12] Rajnikanth and Dr. Nagapadma, R., “Simulation of AC/DC/AC Converter for Closed Loop Operation of Three-phase Induction Motor”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue No. 6, pp.7074-7079, 2014.

[13] Saidah, Bambang Purwahyudi, Kuspijani, (2017), “Control Strategy for PWM Voltage Source Converter Using Fuzzy Logic for Adjustable Speed DC Motor”, International Journal of Power Electronics and Drive System (IJPEDS)

AUTHOR PROFILES:

Saidah received her bachelor, master and Ph.D degree from Institut Teknologi Sepuluh Nopember (ITS) Surabaya in 1985, 2005 and 2013 respectively. She has joined Bhayangkara University in Surabaya since 2006. Her research interest on use of artificial intelligent for power electronics, control and electric drives applications.

FORECASTING GOLD PRICES USING SINGLE EXPONENTIAL SMOOTHING METHOD (SES) AND DOUBLE EXPONENTIAL SMOOTHING METHOD (DES)

[1] HIKMAH MAHARANI IQOMATUL HAQ,
[2]M. MAHAPUTRA HIDAYAT, S.KOM., M.KOM, [3]RIFKI FAHRIAL ZAINAL, ST., M.KOM,

Department of Informatics Engineering, Bhayangkara University,
Ahmad Yani street 114, Surabaya

e-mail : [1]hikmahmaharani24@gmail.com,[2]mahaputra@ubhara.ac.id,[3]rifki24@ubhara.ac.id

ABSTRACT

The Condition of gold prices that experienced an unstable increase and decrease, causing people who invest with gold suffered losses. To solve these problem, forecasting gold prices is required. Forecasting is an activity of predicting the future using the condition or data in the past. There are so many method that have been used, but in this study using Single Exponential Smoothing and Double Exponential Smoothing Method. Based on the test from forecasting gold prices data 2013-2015 from Antam, the results of the value of Mean Square Error obtained using single exponential smoothing method is 162,5101 at α = 0,5 and at α = 0,9 is 143,9416. While when using double exponential smoothing method at α = 0,5 is 174,67 and at α= 0,9 is 266,33. The lowest error value -62,84 at α =0,1 is obtained by using single exponential smoothing method. So, the single exponential smoothing method is better to forecast gold prices problem.

Keyword : *Forecasting, Gold, Gold Prices, Single Exponential Smoothing, Double Exponential Smoothing.*

1. INTRODUCTION

Gold is one of the most valuable precious metals to gain optimum benefits. For people who invest in gold, expect to get the lowest price at the time of purchase and high selling price. Due the increase and decrease in gold prices are unstable, sometimes causing people who make investments have a loses. To solve these problems, required a system to predict the price of gold. In this journal, forecasting gold prices using the Single Exponential Smoothing method and Double Exponential Smoothing method. The Single Exponential Exponential Method is the development of a simple moving average method. While the Double Exponential Smoothing method is a method used to solve the difference between actual data and forecasting values if there is a trend in plot data. Forecasting Gold prices using gold price data 2013-2015 from Antam in Dollar unit.

2. METODOLOGY

2.1 System Analysis

This step is aims to explain the problem of the system, analyzing the system requeriment and elements method in process of forescasting gold prices.

2.1.1 Problem Analysis

For people who invest in gold, expect to get the lowest price at the time of purchase and high selling price. Due the increase and decrease in gold prices are unstable, sometimes causing people who make investments have a loses.

2.1.2 Data Analysis

Data used in this study are gold price data 2013-2015 from Antam in Dollar unit.

2.2 System Design

Design system consist of two main parts. They are master data process and Single Exponential Smoothing and Double Exponential Smoothing method calculation process.

2.2.1 Master Data Process

Master data process containing data whitch is going to be used as training data and testing data. In master data process there are function to input data, update data, and delete the data. The input data needed by the system are participant data for training and participant data for testing.

2.2.2 Single Exponential Smoothing and Double Exponential Smoothing Calculation

Development of Single Exponential Smoothing and Double Exponential Smoothing method has several steps to do. Those steps explained in follwing figure.

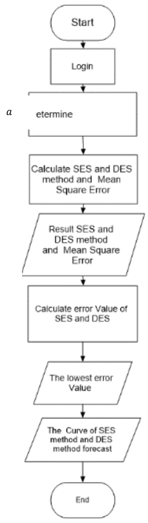

Start
Login
a
etermine
Calculate SES and DES method and Mean Square Error
Result SES and DES method and Mean Square Error
Calculate error Value of SES and DES
The lowest error Value
The Curve of SES method and DES method forecast
End

Figure 1 Flowchart of implementation process

Based on figure 1, Input data for the first process, then determine the value of α. The next step is calculate forecasting value using Single Exponential Smoothing and Double Exponential Smoothing method. Error value is calculated using Mean Square Error (MSE) method and then determine the lowest error value. After the calculation have been done, the results of forecasting among the forecasting value of gold prices, the results of MSE calculations, the lowest error value and curves will shown.

2.2.2.1 Calculation Example Forecasting Gold Prices

To do a calculation process of Single Exponential Smoothing and Double Exponential Smoothing method. These are several data which is going to be used for training.

Table 1 Sample data of Gold Prices and The Result of forecasting gold price using SES and DES for training data

Date	X_t	SES	e_t (SES)	e_t^2 (SES)		S'_t	S''_t	a_t	b_t	F_{t+m}	e_t (DES)	e_t^2 (DES)
01/01/13	1660	1660.00	0.00	0.00		1660.00	1660.00	1660.00	0.000	1660.00	0.00	0.00
02/01/13	1662	1660.00	2.00	4.00		1661.00	1660.50	1661.50	0.500	1660.00	2.00	4.00
03/01/13	1665	1661.80	3.20	10.24		1663.00	1661.75	1664.25	1.250	1662.00	3.00	9.00
04/01/13	1667	1664.68	2.32	5.38		1665.00	1663.38	1666.63	1.625	1665.50	1.50	2.25
05/01/13	1669	1666.77	2.23	4.98		1667.00	1665.19	1668.81	1.813	1668.25	0.75	0.56

The calculation example for 2013-2014 is shown below.

1) Calculation example using single exponential smoothing using α = 0,9:

X_t = 1660 (01-01-2013) / First data

S_t = 1660

$$S_{t+1} = \alpha X_t + (1 - \alpha) S_t$$
$$= 0{,}9\ (1660) + (1 – 0{,}9)\ 1660$$
$$= 1494 + (0{,}1)\ 1660$$
$$= 1494 + 166$$

= 1660 (02-01-2013) / Second data

S_t = 1660

S_{t+1} = $\alpha X_t + (1 - \alpha)S_t$
= 0,9 (1662) + (1 – 0,9) 1660
= 1495,8 + (0,1) 1660
= 1495,8+ 166
= 1661,8 (03-01-2013) / Third data

2) Calculation example using double exponential smoothing using α = 0,5 :

a. Determine value of *Eksponensial Smoothing* (S'_t)

X_t = 1660 (01-01-2013)
X_t = 1662 (02-01-2013)
S'_{t-1} = 1660
S'_t = $\alpha X_t + (1-\alpha)S'_{t-1}$
= 0,5(1662)+(1-0,5) 1660
= 831+ (0,5) 1660
= 831+ 830
= 1661 (02-01-2013)

b. Determine value of *Double Eksponential Smoothing* (S''_t)

S'_t = 1661
S''_{t-1} = 1660
S''_t = $\alpha S'_t + (1-\alpha)S''_{t-1}$
= 0,5 (1661) +(1-0,5) 1660
= 830,5 + (0,5) 1660
= 830,5 + 830
= 1660,5 (02-01-2013)

c. Determine value of *constanta* (α_t)

S'_t = 1661
S''_t = 1660,5
α_t = $2. S'_t - S''_t$
= 2 (1661) – 1660,5
= 3322 – 1660,5
= 1661,5 (02-01-2013)

d. Determine value of *slope* (b_t)

S'_t = 1661
S''_t = 1660,5
b_t = $\frac{\alpha}{1-\alpha}(S'_t - S''_t)$
= $\frac{0,5}{1-0,5}(1661 - 1660,5)$
= 0,5 (02-01-2013)

e. Determine value of *forecast*

α_t = 1661,5
b_t = 0,5
m = 1
F_{t+m} = $\alpha_t + b_t\, m$
= 1661,5 + 0,5 (1)
= 1662 (03-01-2013)

For the calculation forecasting 02-01-2013 using the gold price of 2013-2014 for Double Exponential Smoothing method use the last formula of forecasting $F_{t+m} = \alpha_t + b_t\, m$. Calculation for 01-01-2013 is shown below:

α_t = 1660 (Result α_t from 1-01-2013)
b_t = 0 (Result b_t from 1-01-2013)

m = 1

F_{t+m} = $\alpha_t + b_t\, m$

= 1660 + (0) (1)

= 1660 (Result 02-01-2013)

2.2.2.2 Calculation Example For Error Value Process

1) *MSE* of *Single Eksponential Smoothing* α=0,9

$$\text{MSE} = \frac{\sum_{t=1}^{t=n} = e_t{}^2}{n}$$

$$\text{MSE} = \frac{135006{,}92}{730}$$

= **184,94**

2) *MSE* of *Double Eksponential Smoothing* α=0,9

$$\text{MSE} = \frac{226208{,}90}{730}$$

= **309,88**

3) *MSE* of *Single Eksponential Smoothing* α=0,5

MSE = **239,28**

4) *MSE* of *Double Eksponential Smoothing* α=0,5

MSE = **218,04**

Table 2 The Value of Mean Square Error

α	***MSE***	
	SES	***DES***
0,5	*239,28*	***218,04***
0,9	***184,94***	*309,88*

Based on the table 2 the training result that obtained from the lowest error value of MSE is using Single Expinential Smoothing method at *α* = 0,9 and using Double Expinential Smoothing method at *α* = 0,5. So, the value of a is going to be used for testing is 0,5 and 0,9 .

Table 3 The lowest error value

α	*The lowest error value*	
	SES	***DES***
0,1	***-103,40***	*-91,81*
0,9	*-83,95*	***-123,65***

Based on the table 3 the training result that obtained from the lowest error value is using Single Expinential Smoothing method at *α* = 0,1 and using Double Expinential Smoothing method at *α* = 0,9. So, the value of a is going to be used for testing is 0,1 and 0,9 .

2.3 Data Flow Diagram

Data flow diagram is developed methods of structured data system. DFD describe whole activities inside system clearly.

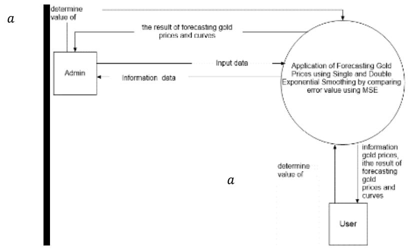

a
determine value of
the result of forecasting gold prices and curves
Admin
Input data
Information data
Application of Forecasting Gold Prices using Single and Double Exponential Smoothing by comparing error value using MSE
information gold prices, ithe result of forecasting gold prices and curves
determine value of
a
User

Figure 2 DFD Level 0

Based on figure 2, Admin has full access of system. It can manage the data for forecasting, it also can access the forecasting process and get the result of forecasting. But user only has access the forecasting process and get the result of forecasting.

3. RESULT

The program is Java Programming language based,interface of the program is shown in the figure 3 and figure 4.

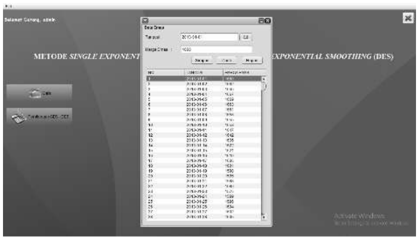


Figure 3 Master data interface of the program

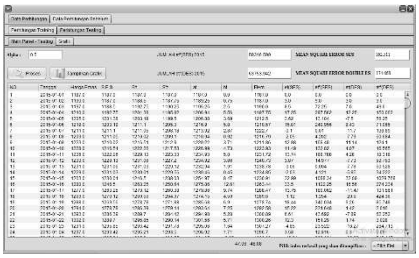


Figure 4 Implementasi Interface

In the Master data process, admin can input data, edit data and delete data.

Figure 4 show the result of forecasting process after input data. The value of α, forecasting value using Single Exponential Smoothing and Double Exponential Smoothing method, the value of MSE, the lowest error value, and curves is shown.

3.1 System Testing

The experiment have been done 2 times using α = 0,9 and α = 0,5 and got the following result:

Table 4 Sample data of Gold Prices and Testing Result using SES and DES Method at α = 0,9

Date	*Xt*	*SES*	*e_t (SES)*	*e_t^2 (SES)*	S'_t	S''_t	a_t	b_t	F_{t+m}	e_t (DES)	e_t^2 (DES)
01/01/15	*1187*	*1187.00*	*0.00*	*0.00*	1187.00	1187.00	1187.00	0.000	1187.00	0.00	0.00
02/01/15	*1190*	*1187.00*	*3.00*	*9.00*	1189.70	1189.43	1189.97	2.430	1187.00	3.00	9.00
03/01/15	*1197*	*1189.70*	*7.30*	*53.29*	1196.27	1195.59	1196.95	6.156	1192.40	4.60	21.16
04/01/15	*1210*	*1196.27*	*13.73*	*188.51*	1208.63	1207.32	1209.93	11.737	1203.11	6.89	47.47
05/01/15	*1205*	*1208.63*	*-3.63*	*13.16*	1205.36	1205.56	1205.17	-1.764	1221.67	-16.67	277.82

Table 5 Sample data of Gold Prices Testing Result using SES and DES Method at α = 0,5

Date	*Xt*	*SES*	*e_t (SES)*	*e_t^2 (SES)*	S'_t	S''_t	a_t	b_t	F_{t+m}	*e_t (DES)*	*e_t^2 (DES)*
01/01/15	*1187*	*1187.00*	*0.00*	*0.00*	*1187.00*	*1187.00*	*1187.00*	*0.000*	*1187.00*	*0.00*	*0.00*
02/01/15	*1190*	*1187.00*	*3.00*	*9.00*	*1188.50*	*1187.75*	*1189.25*	*0.750*	*1187.00*	*3.00*	*9.00*
03/01/15	*1197*	*1188.50*	*8.50*	*72.25*	*1192.75*	*1190.25*	*1195.25*	*2.500*	*1190.00*	*7.00*	*49.00*
04/01/15	*1210*	*1192.75*	*17.25*	*297.56*	*1201.38*	*1195.81*	*1206.94*	*5.563*	*1197.75*	*12.25*	*150.06*
05/01/15	*1205*	*1201.38*	*3.63*	*13.14*	*1203.19*	*1199.50*	*1206.88*	*3.688*	*1212.50*	*-7.50*	*56.25*

Table 6 The Value of Mean Square Error

A	***MSE***	
	SES	***DES***
0,5	*162.5101*	*174.67*
0,9	*143.9416*	*266.33*

Based on the table 6 the results of the Single Exponential Smoothing method is considered better to apply than the Double Exponential Smoothing method because when using the Single Exponential Smoothing method the value of MSE at α = 0,5 ie 162, 5101 and at α = 0 , 9 that is 143,9416, whereas in Double Exponential Smoothing method the value of MSE at α = 0,5 that is 174,67 and at α = 0,9 that is 266,33.

Table 7 The lowest error value

A	The lowest error value		**Date**
	SES	**DES**	
0,1	-62,84	-49,932	7-11-2015 and 6-2-2015
0,9	-44,03	-37,09	27-12-2015 and 31-8-2015

Based on the table 7 the lowest error value is -62,84 using Single Expinential Smoothing method at $\alpha = 0,1$.

4 CONCLUTION

This study applied single exponential smoothing method and double exponential method to the forecasting gold prices using Java and database MySQL. Based on the calculation value forecasting error using MSE is known the best value of α is α =0,9 and α =0,5. The value of MSE using Single expotential smoothing method at α =0,5 that is 162,5101 and at α =0,9 that is 143,9416. While when using double exponential smoothing method, the value of MSE at α =0,5 that is 174,67 and at α =0,9 that is 266,33. The lowest error value is -62,84 using Single Expinential Smoothing method at $\alpha = 0,1$. Single exponential smoothing method and double exponential smoothing method is not good for forecasting gold prices because these method obtained big value of MSE.

REFERENCES

[1] Imbar, Radiant Victor, Yon Andreas. 2012, *Aplikasi Peramalan Stok Barang Menggunakan Metode Double Exponential Smoothing*, Vol 7, No 2, September 2012: 123 – 141 Jurusan Sistem Informasi, Fakultas Teknologi Informasi Universitas Kristen Maranatha

[2] Raharja, Alda, Wiwik Angraeni, S.Si, M.Kom, Retno Aulia Vinarti, S.Kom. 2010, *Penerapan Metode Exponential Smoothing Untuk Peramalan Penggunaan Waktu Telepon Di Pt.Telkomsel Divre3 Surabaya,* Sistem Informasi, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh November

[3] Mahena, Yuliga, Muhammad Rusli, Edy Winarso. 2015, *Prediksi Harga Emas Dunia Sebagai Pendukung Keputusan Investasi Saham Emas Menggunakan Teknik Data Mining,* Volume 2 No.1 Februari 2015 Sistem Informasi, Institut Teknologi dan Bisnis Kalbis, Jakarta

[4] Pratiwi, Yustian Winda. 2015, *Aplikasi Peramalan Penjualan Mobil Hyundai Menggunakan Metode Single dan Double Exponensial Smoothing Dengan Perbandingan Nilai Error Terkecil Menggunakan Mean Square Error*, Fakultas Teknik Informatika Universitas Bhayangkara Surabaya

[5] Firdaus, Affan, Suyanto, Mahmud Dwi Suliiyo. 2012, *Analisis Dan Implementasi Grey Model Untuk Memprediksi Harga Emas,* Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

[6] Biri, Romy, Yohanes A.R. Langi, Marline S Paendong. 2013, *Penggunaan Metode Smoothing Eksponensial Dalam Meramal Pergerakan Inflasi Kota Palu* Vol. 13 No. 1 April 2013 Universitas Sam Ratulangi

[7] Santoso, Denny Achmad. 2013, *Pemodelan Arima Untuk Peramalan Harga Emas,* Sekolah Tinggi Manajemen Informatika & Teknik Komputer Surabaya

[8] Kusumadewi, Felasufah. 2014, *Peramalan Harga Emas Menggunakan Feedforward Neural Network Dengan Algoritma Backpropagation,* Universitas Negeri Yogyakarta

[9] Andriyanto, Teguh. 2017*, Sistem Peramalan Harga Emas Antam Menggunakan Double Exponential Smoothing* Vol.1, No.1, Februari 2017 ISSN: 2549-6824 Sistem Informasi Universitas Nusantara PGRI Kediri

[10] Marthasari, Gita Indah, Arif Djunaidy. 2014, *Optimasi Data Latih Menggunakan Algoritma Genetika Untyuk Peramalan Harga Emas Berbasis Generalized Regression Neural Network,* Jurusan Teknik Informatika Universitas Muhammadiyah Malang

NOISE MINING USING MODIFIED SHARED NEAREST NEIGHBORS ALGORITHM

[1]RIFKI FAHRIAL ZAINAL

[1]Lecturer, Department of Informatics Engineering, University of Bhayangkara Surabaya

Jl. Ahmad Yani 114 Surabaya

ABSTRACT

Removing objects that are noise is an important goal of data cleaning as noise hinders most types of data analysis. Most existing data cleaning methods focus on removing noise that is the result of low-level data errors that result from an imperfect data collection process, but data objects that are irrelevant or only weakly relevant can also significantly hinder data analysis. One of the way to enhance the data analysis as much as possible, is finding and removing the right noise data. Consequently, if the attributes for the noise can be found, a new and better way to remove the noise in a large data set can be applicated.

1. INTRODUCTION

Noise as described in Oxford Dictionary is "Random Fluctuations that obscure or do not contain meaningful data or other information". The term has often been used as a synonym for corrupt data. For most existing data cleaning methods, the focus is on the detection and removal of noise (low-level data errors) that is the result of an imperfect data collection process. However, its meaning has expanded to include any data that cannot be understood and interpreted correctly by machines, such as unstructured text. Any data that has been received, stored, or changed in such a manner that it cannot be read or used by the program that originally created it can be described as noisy.

The need to address this type of noise is clear as it is detrimental to almost any kind of data analysis. However, ordinary data object that are irrelevant or only weakly relevant to a particular data analysis can also significantly hinder the data analysis, and thus these objects should also be considered as noise, at least in the context of a specific analysis. For instance, in document data sets that consist of news stories, there are many stories that are only weakly related to the other news stories. If the goal is to use clustering to find the strong topics in a set of documents, then the analysis will suffer unless irrelevant and weakly relevant documents can be eliminated. Consequently, there is a need for data cleaning techniques that remove both types of noise.

In some cases, the amount of noise in a data set is relatively small. For example, it has been claimed that field error rates for business are typically around 5% or less if an organization specifically takes measures to avoid data errors [24]. However, in other cases, the amount of noise can be large. For example, a significant number of false-positive protein interactions are present in current experimental data for protein complexes. Gavin *et al.* [9] estimates that more than 30% of the protein interactions they detect may be spurious, as inferred from duplicate analyses of 13 purified protein complexes. Although this is an example of a data set that has a large amount of noise due to data collection errors, the amount of noise due to irrelevant data objects can also be large. Examples include the document sets mentioned earlier [7] and Web data [25], [26]. Therefore, data cleaning techniques for the enhancement of data analysis also need to be able to discard a potentially large fraction of the data.

Noisy data can give a lot of problems, but the main problems with noisy data are:

1. Noisy data unnecessarily increases the amount of storage space required and can also adversely affect the results of any data mining analysis. Statistical analysis can use information gleaned from historical data to weed out noisy data and facilitate data mining.
2. Noisy data can be caused by hardware failures, programming errors and gibberish input from speech or optical character recognition programs.
3. Spelling errors, industry abbreviations and slang can also delay in machine reading.

Schema-related data quality problems occur because of the lack of appropriate model-specific or application specific integrity constraints. For examples, due to data model limitations of poor schema design, or because only a few integrity constraints were defined to limit the overheard for integrity control. Instance-specific or application specific problems relate to errors and inconsistencies that cannot be prevented at the schema level. Table 1 and Table 2 shows the single source problem at schema level and single source problems at instance level.

Table 1. Examples for Single-Source Problems at Schema Level

Scope/Problems		**Noise**	**Reasons/Remarks**
Attribute	Illegal Values	Bdate = 01.09.78	Values outside the domain range
Record	Violated Attribute Dependencies	Age=22, bdate=01.09.78	Age = current year-birth year should hold

Record Type	Uniqueness Violation	Emp1=(name="Rifki F", ID="123456")
		Emp2=(name="Hartatik S", ID="123456"
			Uniqueness for ID violated

Source	Referential Integrity Violation	Emp=(name="Rifki F", ID_Dept=127)	Referenced Departement (127) not defined

Table 2. Examples for Single-Source Problems at Instance Level

Scope/Problems		**Noise**	**Reasons/Remarks**
Atribute	Missing Values	Phone=9999-9999	Unavailable values during data entry(dummy values or null)
	Misspelling	City="Sursbaya"	Usually typos, phonetic errors

	Cryptic values, Abbrevations	Experiences="B"
		Occupation="DB Prog"	

	Embedded Values	Name="Rifki F 01.09.78 Surabaya"	Multiple values entered in one attribute
	Misfielded Values	City="Surabaya"	
Record	Violated Attribute Dependencies	City="Surabaya", zip=101200	City and zip code should correspond
Record Type	Word Transpositions	Name1="Rifki F", name2="Hartatik S"	Usually in free form field

	Duplicated	Emp1=(name="Rifki F")
		Emp2=(name="Rifki F")
			Same employee represented twice due to some data entry errors

When multiple sources are integrated, the problems present in single source are aggregated. Each source may contain dirty and inconsistent data and the data in the sources may be represented differently, overlap or contradict. This is because the sources are typically developed, deployed and maintained independently to serve specific needs. This results in a large degree of heterogeneity with data management systems, data models, schema designs and the actual data. Tabel 3 shows the multisource problems.

Table 3. Examples of Multisource Problems at Schema & Instance Level
Data Source 1

CID	Name	Street	City	Sex
11	Rifki Fahrial	Jemursari VI/26	Surabaya	1
24	Hartatik Sri	Kutisari Selatan 2/14	Surabaya	0

Data Source 2

CNO	LastName	FirstName	Gender	Address	Phone/Fax
24	Rahayu	Rejeki	F	Rungkut Sier 16 Surabaya	333-222-6542
493	Fahrial	Rifki	M	Jemursari VI/26 Surabaya	444-555-6666

Integrated Target with Cleaned Data

No	LName	FName	Gender	Street	City	Phone	CID	CNO
1	Fahrial	Rifki	M	Jemursari VI/26	Surabaya	444-555-666	11	493
2	Sri	Hartatik	F	Kutisari Selatan 2/14	Surabaya		24	
3	Rahayu	Rejeki	F	Rungkut Sier 16	Surabaya	333-222-6542		24

Data cleaning routines is to clean the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. Dirty data can cause confusion for the mining procedure, resulting in unreliable and poor output. There is necessity for useful preprocessing step to be used some data-cleaning routines, specially the missing values. The missing values are to be corrected by following measures

a) Ignore the missing values
b) Fill in the missing values manually
c) Use a global constant to fill in the missing values
d) Use the attribute mean to fill in the missing values
e) Use the most probable value to fill in the missing values.

The missing values can be ignored if the class label is missing. This method is not very effective, unless the missing values contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.

The second approach is time consuming and may not be feasible to the large data set with many missing values. To shorten the time of the process, the missing value in the data sets are filled by using a global constant in the missing value and replace all missing attribute values by the same constant. The other way is to fill the missing values with the attribute mean. The last approach, which is using the most probable value to fill in the missing value may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction.

For example, using the other attributes in the soil data set, we may construct a decision tree to predict the missing values for income. This is a popular strategy. In comparison to the other methods, it uses the most information from the present data to predict missing values. This procedure is bias the data. The filled-in value may not be correct. It is important to have a good design of databases or data entry procedures which would minimize the number of missing values or errors.

On the other hand, noisy data or data that contains errors can be corrected with Binning, Regression or Clustering. Binning methods have the tendency to smooth a sorted data by consulting its “neighborhood”, that is, the data around it. The sorted data are distributed into a number of “buckets” or bins. Because binning methods consult the neighborhood of values, they perform local smoothing. In smoothing by bin means each value in a bin is replace by the mean value of the bin.

Smoothing by bin medians can be employed, in which each bin value is replaced by the bin median. In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value. In general, the larger the width, the greater the effect of the smoothing. Alternatively, bins may be equal-width, where the interval range of values in each bin is constant. Binning is also used as a discretization technique.

The Regression methods smoothed the data by fitting the data to a function. Linear regression involves finding the “best” line to fit two attributes (or variables), so that one attribute can be used to predict the other. Multiple linear regression is an extension of linear regression where more than two attributes are involved and the data are fit to a multidimensional surface.

The last methods is by using clustering methods. Outliers or noise may be detected by clustering where similar values are organized into groups or clusters. Intuitively, values that fail outside of the set of clusters may be considered outliers or noise. This method can be used not only to find class for each of the data in the data sets, but also able to clean the noise in its process. But the process of finding the noise among one big data set, is a long and time-consuming process.

In this paper, we propose a way to use the clustering methods to find the noise and examining it, thus finding the attributes for the noise. Then by using the noise attribute, we can do several things such as avoiding filling in the missing values with the noise values, or cleaning the data set with the right attribute for the noise.

2. METHODOLOGY

There are several clustering methods that can be used to find the noise. This paper use the Shared Nearest Neighbors (SNN). SNN is a density based clustering algorithm which is capable of finding clusters of arbitrary shapes, sizes and densities and we need not mention the number of clusters as parameter. This algorithm makes use of a similarity measure which is obtained from the number of neighbors two points share. This can be computed from the k-nearest neighbors of each point. In order to identify the k-nearest neighbors, we need a distance function like Euclidean distance.

2.1 SNN Algorithm

This algorithm requires the following input:

a. K: number of nearest neighbors to be identified for each point.
b. Eps: Density threshold-minimum number of points shared by two points in order to be considered close to each other.
c. MinPts: minimum density a point should have to be considered a core point.

Algorithm includes the following steps:

1) Create the distance matrix using a given distance function and identify for each point, the k nearest neighbors.
2) For each two points, calculate the similarity, which is given by the number of shared neighbors.
3) Establish the SNN density of each point. The SNN density is given by the number of nearest neighbors that share Eps or more neighbors.
4) Identify the core points of the data set. Each point that has a SNN density greater or equal to MinPts is considered a core point.
5) Build clusters from core points. Two core points are allocated to the same cluster if the share Eps or more neighbors with each other.
6) Handle noise points. Points not classified as core points and that are not within Eps of a core point are considered noise.
7) Assign the remaining point to cluster. All non-core and non-noise points are assigned to the nearest cluster.

The inmportant step in the SNN algorithm is identification of k-nearest neighbors. In order to identify the nearest neighbors, we make use of a distance function. Various distance functions that could be applicable to numerical data are:

1) Euclidean Distance:
It is the straight-line distance between two points. It computes the root of squares difference between coordinates of pair of objects.

$$Dist_{xy} = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \qquad (1)$$

2) Manhattan Distance:
Manhattan Distance computes the absolute differences between coordinates of pair of objects.

$$Dist_{xy} = \sum_{i=1}^{n}((|X_i - Y_i|)) \qquad (2)$$

3) Minkowski Distance:
Minkowski Distance can be defined as the generalized metric distance. It is formulated as

$$Dist_{xy} = \sum_{i=1}^{n}((|X_i - Y_i|)^p)^{\frac{1}{P}} \qquad (3)$$

When p=2, it becomes Euclidean Distance.

The algorithm has been implemented by using the above measures in order to obtain the efficient distance measure. It is found that using Manhattan as distance measures, the algorithm doesn't perform well. Using Euclidean it has produced the effective results. Since Minkowski also behaves like Euclidean, it also has produced the same results. Therefore, we have implemented the algorithm by using Euclidean distance in this paper.

In the first step, we create a distance matrix and identify the k-nearest neighbors of each point. Then for each two points, we calculate similarity i.e. number of nearest neighbors two points share. Then SNN density is computed. The points whose SNN density is greater than MinPts are considered as core points. All core points that share Eps or more neighbors are allocated to the same cluster and continue to be core points. Then we start clustering remaining points with the help of core points.

2.2 Modified SNN Algorithm

For the noise mining, we need data considered as the noise in the data set. In this paper, we propose a modified SNN algorithm by using only 1 to 4 steps in the original SNN algorithm. After we form and compute the SNN density, we can find the noise data by using the MinPts. For each points that have SNN density greater than MinPts are considered core points. While the points not classified as core points and that are not within Eps of a core point are considered noise.

This noise data than can be used as the core points for our noise mining or to avoid confusion now called the noise core points. We modified the next step whereas the original is build cluster from core points, we use the noise core points as the core points for the clusters. Two noise core points are allocated to the same cluster if they share Eps or more neighbors with each other. This modified SNN algorithm gives us a fast solution to perform analysis for the noise data.

The modified Algorithm includes the following steps:

1) Create the distance matrix using a given distance function and identify for each point, the k nearest neighbors.
2) For each two points, calculate the similarity, which is given by the number of shared neighbors.
3) Establish the SNN density of each point. The SNN density is given by the number of nearest neighbors that share Eps or more neighbors.
4) Identify the noise core points of the data set. Each point that has a SNN density lower to MinPts is considered a noise core point.
5) Build noise clusters from noise core points. Two noise core points are allocated to the same cluster if the share Eps or more neighbors with each other.

For each noise cluster can gives us a new point of view to see the bigger pictures. It can reveal a new perspective. Thus, opening the way of noise mining.

3. CONCLUSION

The modified SNN algorithm can gives a new way to perform noise mining. By using the result of noise mining, we can find the noise attributes to help us cleaning the data. Another way of using noise mining is to help providing the values of the missing values. In this case, is the value that should not be given to the missing values. Still need a further research to find the full ability of the noise mining.

REFERENCES

[1] Aggrawal, R, Imielinski, T and Swami, A. (1993), *Mining Association Rules between Sets of Items in Large Databases*, In Proc. Of the ACM SIGMOD.

[2] Angiulli, F and Pizzuti, C, (2002), *Fast Outlier Detection in High Dimensional Spaces*, In Proceedings of the Sixth European Conference on the Principles of Data Mining and Knowledge Discovery.

[3] Bay, Stephen D, and Schwabacher, M (2003), *Mining Distance-based Outliers in Near Linear Time with Randomization and A Simple Pruning Rule*, In KDD '03: Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 29-38, New York, NY, USA, ACM Press

[4] Breunig, Markus M, Kriegel, Hans-Peter, Ng, Raymond T and Sander, Jorg, (2002) *Lof: Identifying Density Based Local Outliers*, In Proc. of the 2000 ACM SIGMOD International Conference on Management of Data.

[5] Brodley, Carla E and Friedl, Mark A., (1999), *Identifying Mislabeled Training Data*, Journal of Artificial Intelligence Research, 11:131-167.

[6] Eisen, Michael B, Spellman, Paul T, Browndagger, Patrick O, and Botstein, David, (1998), *Cluster Analysis and Display of Genome-wide Expression Patterns*, Proceeding of the National Academy of Sciences of the United States of America (PNAS).

[7] Ertoz, L, Steinbach, M and Kumar, V, (2003), *Finding Clusters of Different Sizes, Shapes and Densities in Noisy, High Dimensional Data*, In Proceedings of Third SIAM International Conference on Data Mining, San Francisco, CA, USA.

[8] Ester, M, Kriegel, H. P, Sander, J and Xu, X, (1996), *A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, In Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining.

[9] Gavin, A et al, (2002), *Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes*, Nature, 415:141-147.

[10] Gaede, V and Gunther, O, (1998), *Multidimensional Access Methods*, ACM Computing Surveys, 30(2): 170-231.

[11] Galhardas, H, Florescu, D, S, and Simon, E, (2000), *Ajax: An Extensible Data Cleaning Tool*, In Proceedings of the ACM SIGMOD International Conference on Management of Data.

[12] Galhardas, H, Florescu, D, S, Simon, E and Saita, C (2001), *Declarative Data Cleaning: Language, Model and Algorithms*, In Proceedings of the 2001 Very Large Data Bases (VLDB) Conference.

[13] Guha, S, Rastogi, R and Shim, K, (1998), *Cure: An Efficient Clustering Algorithm for Large Databases*, In Laurea M. Haas and Ashutosh Tiwary, editors, SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, June 2-4, 1998, Seattle, Washington, USA, pages 73-84, ACM Press.

[14] Han, E, Boley, D, Gini, M, Gross, R, Hastings, K, Karypis, G, Kumar, V, Mobasher, B, and Moore, J, (1998), *Webace: A Web Agent for Document Categorization and Exploration*, In Proc. Of the 2nd International Conference on Autonomous Agents.

[15] Hernandez, M, and Stolfo, S, (1995), *The Merge/Purge Problem for Large Databases*, In Proceedings of the ACM SIGMOD International Conference on Management of Data, pages 127-138.

[16] Hernandez, M and Stolfo, S (1998), *Real Word Data is Dirty: Data Cleansing and the Merge/Purge Problem*, Data Mining and Knowledge Discovery, 2:9-37.

[17] Hodge, V.J, and Austin, J, (2004), *A Survey of Outlier Detection Methodologies*, Artificial Intelligent Review, 22:85-126.

[18] Jain, A.K, and Dubes, R.C, (1988), *Algorithms for Clustering Data*, Prentice Hall Advanced Reference Series, Prentice Hall, Englewood Cliffs, New Jersey.

[19] Karypis, G, (2006), *Cluto: Software for Clustering High Dimensional Dataset*, Data Mining and Knowledge Discovery, 3:12-16.

[20] Knorr, E.M, Ng, R.T, and Tucakov, V, (2000), *Distance-based Outliers: Algorithms and Applications*, VLDB Journal: Very Large Databases, 8:237-253.

[21] Kohavi, R, and John, G.H, (1997), *Wrappers for Feature Subset Selection*, Artificial Intelligence, 97(1-2):273-324.

[22] Larsen, B, and Aone, C., (1999), *Fast and Effective Text Mining Using Linear-Time Document Clustering*, In Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[23] Lee, M.L, Ling, T.W, and Low, W.L, (2000), *Intellclean: A Knowledge-based Intelligent Data Cleaner*, In Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[24] Monge, A.E, and Elkan, C.P, (1997), *An Effiecient Domain-Independent Algorithm for Detecting Approximately Duplicate Database Record*, In Proc. of the ACM-SIGMOD Workshop on Research Issues on Knowledge Discovery and Data Mining

[25] Yang, Y, (1995), *Noise Reduction in a Statistical Approach to Text Categorization*, In Edward A, Fox, Peter Ingwersen, and Raya Fidel, editors, SIGIR, pages 256-263, ACM Press.

[26] Yi, L, Liu, B, and Li, X, (2003), *Eliminating Noisy Information in Web Pages for Data Mining*, in Lise Getoor, Ted E, Senator, Pedro Domingos, and Christos Faloutsos, editors, KDD, pages 296-305.