S-23

Pemanfaatan Software *Open Source* R dalam pemodelan ARIMA

Dedi Rosadi

Program Studi Statistika, FMIPA UGM
Email: dedirosadi@ugm.ac.id

Abstrak

R (R Development Core Team, 2009) merupakan salah satu *software open source* yang terpopuler dan telah menjadi "lingua franca" atau "bahasa standar" untuk keperluan komputasi statistika saat ini. Dalam tulisan ini, akan dikenalkan dan dibahas penggunaan R untuk komputasi model ARIMA, yang merupakan salah satu model standar yang dikenalkan dalam kuliah analisa runtun waktu. Pengenalan dilakukan dengan menggunakan data empiris dimana komputasi model ARIMA dilakukan dengan menggunakan R versi CLI (command line interface) dan versi GUI (Graphical User Interface) yang merupakan hasil pengembangan terbaru dalam Rosadi, Marhadi dan Rahmatullah (2009). Dalam metodologinya, dikenalkan teknik pemodelan standar dengan menggunakan metode Box-Jenkins, maupun teknik pemilihan model automatik menggunakan ukuran kriteria informasi, seperti yang dibahas di Hyndman dan Khandakar (2008).

Kata-kata kunci: R Commander Plug-in, *Open Source*, automatic ARIMA

1. Pendahuluan

R merupakan salah satu *software open source* yang terpopuler dan telah menjadi bahasa "standar" untuk keperluan komputasi statistika saat ini. Banyak keunggulan yang ditawarkan oleh *software* ini, misalnya bersifat *multiplatforms* (tersedia untuk sistem operasi Windows, Linux, Macintosh dan Unix), reliabilitas dari *software* yang baik, ketersediaan *update* dan *library* yang lengkap, fasilitas *help* untuk *user* yang bersifat *free of charge*, dan lain-lain.

R yang telah dikenal dalam versinya yang sekarang ini (versi terakhir per 25 November 2009 adalah versi 2.10.0) merupakan suatu sistem analisa statistika yang relatif komplet. Versi paling awal R dibuat tahun 1992 di Universitas Auckland, New Zealand oleh Ross Ihaka dan Robert Gentleman (yang menjelaskan asal muasal akronim nama R untuk software ini). Pada awalnya R dikembangkan menggunakan bahasa *LISP* dan di implementasikan berdasarkan sistem semantik bahasa *Scheme* di bawah sistem operasi Macintosh. Saat ini *source code kernel* R dikembangkan oleh R Core Team, yang beranggotakan sejumlah statistisi dari berbagai penjuru dunia (lihat http://www.r-

project.org/contributors.html), dan oleh masyarakat statistisi di seluruh penjuru dunia yang memberikan kontribusi berupa kode, melaporkan *bug* dan membuat dokumentasi untuk R. R bersifat *multiplatforms*, dengan file instalasi binary/file *source tarball* tersedia untuk sistem operasi Windows, Mac OS, Mac OS X, Free BSD, NetBSD, Linux, Irix, Solaris, AIX dan HPUX.

Fungsionalitas dan kemampuan dari R sebagian besar diperoleh dari *Add-on packages*/*library*. Suatu *library* adalah kumpulan perintah/fungsi yang dapat digunakan untuk melakukan analisa tertentu (mirip halnya dengan *Toolbox* dalam MATLAB). Instalasi standar dari R memuat berbagai *library* dasar seperti stats, graphics, utils, datasets dan base. Diluar *library*-*library* dasar ini, terdapat sejumlah besar *library* hasil kontribusi dari pengguna R. Daftar semua library yang tersedia dapat diakses dari *link download* CRAN pada alamat http://cran.rproject.org/.

Untuk keperluan analisa runtun waktu/ekonometri, telah tersedia cukup lengkap paket/library dari R (lihat *taskviews Econometrics, Finance* dan *Time Series* pada CRAN) dengan interaksi berupa R-CLI/ Command Line Interface. Dalam tulisan ini, akan dikenalkan dan dibahas penggunaan R untuk komputasi model ARIMA, yang merupakan salah satu model standar yang dikenalkan dalam kuliah analisa runtun waktu. Pengenalan dilakukan dengan menggunakan data empiris dimana komputasi model ARIMA dilakukan dengan menggunakan R versi CLI (command line interface) dan versi GUI (Graphical User Interface) yang merupakan hasil pengembangan terbaru dalam Rosadi, Marhadi dan Rahmatullah (2009). Dalam metodologinya, dikenalkan teknik pemodelan standar dengan menggunakan metode Box-Jenkins, maupun teknik pemilihan model automatik menggunakan ukuran kriteria informasi, seperti yang dibahas di Hyndman dan Khandakar (2008).

Tulisan ini diorganisasikan sebagai berikut. Pada bagian ini, diberikan latar belakang masalah dari tulisan ini. Pada bagian 2, diberikan review singkat metodologi Box-Jenkins untuk pemodelan ARIMA. Kemudian pada bagian 3 dan 4 diberikan komputasi model ARIMA menggunakan R-CLI. Pada bagian akhir, diberikan review singkat mengenai metodologi komputasi ARIMA dengan R-GUI.

2. Metodologi Box-Jenkins untuk pemodelan ARIMA

Salah satu teknik pemodelan yang digunakan pemodelan ARIMA adalah metodologi Box-Jenkins, yang terdiri atas empat langkah berikut:

2.1. Preprocessing data dan identifikasi model stasioner

Dalam tahap awal, dilakukan identifikasi model runtun waktu yang mungkin digunakan untuk memodelkan sifat-sifat data. Identifikasi secara sederhana dilakukan secara visual dengan melihat plot dari data, untuk melihat adanya trend, komponen musiman, nonstasioneritas dalam variansi, dan lain-lain. Tahapan ini dapat juga digunakan untuk

melihat teknik *preprocessing* data manakah, yang jika diperlukan, dapat digunakan untuk membentuk data yang stasioner. Beberapa teknik preprocessing data yang umum dilakukan adalah seperti membuang *outlier* dari dalam data, *filtering* data menggunakan model/teknik statistika tertentu, transformasi data (seperti transformasi logaritma, atau lebih umum, transformasi Box-Cox), melakukan operasi *difference*, *detrend* (membuang trend), de*seasonal*-isasi (membuang komponen musiman), dan lain-lain. Stasioneritas dari data dapat dilihat dari bentuk fungsi estimator fungsi autokorelasi (sampel ACF/Autocorrelation function) dan estimator fungsi autokorelasi parsial (sampel PACF/Partial ACF), ataupun dengan melakukan uji unit root terhadap data.

Selanjutnya, jika telah dilakukan preprocessing terhadap data sehingga menghasilkan data yang stasioner, dapat di tentukan bentuk model ARMA (autoregressive moving average) yang tepat dalam menggambarkan sifat-sifat data, dengan cara membandingkan plot sampel ACF/PACF dengan sifat-sifat fungsi ACF/PACF teoritis dari model ARMA. Rangkuman bentuk plot sampel ACF/PACF dari model ARMA diberikan pada tabel berikut:

Proses	Sampel ACF	Sampel PACF
White noise (error random)	Tidak ada yang melewati batas interval pada lag >0	Tidak ada yang melewati batas interval pada lag >0
AR(p)	Meluruh menuju nol secara eksponensial	Diatas batas interval maksimum sampai lag ke p dan dibawah batas pada lag >p
MA(q)	Diatas batas interval maksimum sampai lag ke q dan dibawah batas pada lag >q	Meluruh menuju nol secara eksponensial
ARMA(p,q)	Meluruh menuju nol secara eksponensial	Meluruh menuju nol secara eksponensial

2.2. Estimasi model

Setelah ditentukan bentuk model yang kira-kira sesuai untuk data, selanjutnya dilakukan estimasi terhadap parameter dalam model, seperti koefisien dari model ARMA dan nilai variansi dari residual. Estimasi dari model ARMA dapat dilakukan dengan menggunakan metode Maksimum Likelihood Estimator (MLE), Least Square, Hannan Rissanen, metode Whittle dan lain-lain. Kajian statistika detail dari metode-metode tersebut dapat ditemukan pada berbagai literatur runtun waktu. Dalam pemodelan, juga sering dilakukan analisa overfitting, dengan cara mengkaji dan menganalisa model runtun wkatu yang memiliki order yang lebih tinggi daripada model yang telah diidentifikasi pada bagian 1. Untuk pengujian apakah koefisien hasil estimasi signifikan atau tidak (yakni uji hipotesa null koefisien bernilai 0 vs hipotesa alternatif koefisien tidak nol) dapat digunakan

pengujian dengan statistik uji t yang akan berdistribusi student-t dengan derajat bebas n-1, n=banyaknya sampel. Jika terdapat koefisien yang tidak signifikan, maka koefisien/order lag tersebut dapat dibuang dari model dan model diestimasi kembali tanpa mengikutkan order yang tidak signifikan.

2.3. Diagnostic check dan pemilihan model terbaik

Langkah selanjutnya adalah melakukan *diagnostic check* dari model yang telah diestimasi dibagian 2 diatas, yakni melakukan verifikasi kesesuaian model dengan sifat-sifat data. Jika model merupakan model yang tepat, maka data yang dihitung dengan model (fitted value) akan memiliki sifat-sifat yang mirip dengan data asli. Dengan demikian, residual yang dihitung berdasarkan model yang telah diestimasi mengikuti asumsi dari error dari model teoritis, seperti sifat white noise, normalitas dari residual (walaupun asumsi ini dapat diabaikan, tidak sepenting asumsi pertama) dan lain-lain. Untuk melihat apakah residual bersifat white noise, dapat dilakukan dengan dua cara, yakni pertama dengan melihat apakah plot sampel ACF/PACF residual yang terstandardisasi (residual dibagi estimasi standar deviasi residual) telah memenuhi sifat-sifat proses white noise dengan mean 0 dan variansi 1. Cara kedua adalah dengan melakukan uji korelasi serial, yakni menguji hipotesa $H_0 : \rho_1 = \rho_2 = \cdots = \rho_k, k < n$ (tidak terdapat korelasi serial dalam residual sampai lag-k, k <n). Uji ini dapat dilakukan dengan menggunakan statistik uji Box-Pierce $Q = n\sum_{j=1}^{k} \hat{\rho}(j)^2$, atau Ljung Box $Q = n(n+2)\sum_{j=1}^{k} \hat{\rho}(j)^2 / (n-j)$, yang akan berdistribusi $\chi^2(k-(p+q)), k > (p+q)$. Disini $\hat{\rho}(j)$ menunjukkan nilai sampel ACF pada lag-j sedangkan p dan q menunjukkan order dari model ARMA (p,q). Apabila hipotesa diagnostic check ditolak, maka model yang telah diidentifikasikan diatas tidak dapat digunakan, dan selanjutnya dapat diidentifikasikan kembali model yang mungkin sesuai untuk data.

Selanjutnya, dalam praktek akan banyak model yang memenuhi pengujian diagnostik diatas. Untuk memilih model terbaik, dapat dipilih model yang meminimumkan ukuran kriteria informasi seperti Akaike Information Criteria, $AIC = n\ln(\hat{\sigma}_\varepsilon^2) + 2(p+q+1)$, $\hat{\sigma}_\varepsilon^2 = SSE / n$, dengan SSE=Sum of Squared error yang dapat diestimasi dari jumlahan kuadrat semua nilai residual. Akan tetapi, diketahui untuk model autoregressive, kriteria AIC tidak memberikan order p yang konsisten, sehingga untuk pembanding, dapat digunakan kriteria informasi lain, seperti Schwarzt Bayesian Information Criteria, $SBC = n\ln(\hat{\sigma}_\varepsilon^2) + (p+q+1)\ln n$, ataupun bentuk-bentuk kriteria informasi lainnya yang diusulkan didalam literatur.

2.4. Aplikasi model untuk simulasi, peramalan, dan lain-lain

Setelah model terbaik diperoleh dari langkah-langkah pemodelan diatas, maka model tersebut dapat digunakan untuk meramalkan sifat-sifat data dimasa yang akan datang.

Dalam analisa runtun waktu, seringkali data dibagi menjadi dua bagian yang disebut data *in sample*, yakni data-data yang digunakan untuk memilih model terbaik dengan langkah-langkah pemodelan diatas, dan data *out sample*, yakni bagian data yang digunakan untuk memvalidasi keakuratan peramalan dari model terbaik yang diperoleh berdasarkan data *in sample*. Model yang baik tentunya diharapkan merupakan model terbaik untuk data *in-sample* dan sekaligus merupakan model yang baik untuk peramalan, yang dapat diukur dengan data *out sample*. Beberapa ukuran kebaikan peramalan dapat dikenalkan, seperti ukuran mean square error (MSE), root of MSE (RMSE), Median atau Mean Absolut Deviation (MAD), dan lain-lain.

3. Komputasi pemodelan ARIMA dengan R-CLI

Untuk memberikan ilustrasi dari komputasi model ARIMA menggunakan R, akan digunakan data bulanan harga minyak bumi (yakni variabel World Oil Price) selama periode Januari 1996 sampai Desember 2008, tersimpan dalam file bernama latihan3. Keseluruhan langkah komputasi pemodelan dengan R yang diperlukan dapat dilihat pada Gambar 1. Menggunakan plot ACF/PACF, diperoleh beberapa model yang mungkin untuk digunakan bagi data World Oil Price ini, yakni:

- Model 1: log(World_Oil_Prices) adalah model ARIMA(1,1,0)
- Model 2: log(World_Oil_Prices) adalah model ARIMA(0,1,1)

dan model *overfitting* dari dua model diatas

- Model 3: log(World_Oil_Prices) adalah model ARIMA(1,1,1)
- Model 4: log(World_Oil_Prices) adalah model ARIMA(2,1,0)
- Model 5: log(World_Oil_Prices) adalah model ARIMA(0,1,2)

```
>#plot data
>latihan3$World_Oil_Prices <- ts(latihan3$World_Oil_Prices,start=c(1996,1),
  freq=12) # data diubah menjadi bertipe time series
>ts.plot(latihan3$World_Oil_Prices,col="blue",main="Time Series Plot")
> adf.test(latihan3$World_Oil_Prices) # uji stasioneritas
> win.graph()
> par(mfrow=c(2,1))
> acf(latihan3$World_Oil_Prices,na.action=na.pass) #plot ACF/PACF
> pacf(latihan3$World_Oil_Prices,na.action=na.pass) #plot ACF/PACF
> # transformasi difference dan plot
>latihan3$World_Oil_Prices.Diff1 <- diff(latihan3$World_Oil_Prices,
differences=1)
>ts.plot(latihan3$World_Oil_Prices.Diff1,col="blue",main="Time Series Plot")
>#transformasi difference dari log dan plotnya
>latihan3$World_Oil_Prices.Difflog1 <- diff(log(latihan3$World_Oil_Prices),
differences=1)
```

Gambar 2. Script R untuk Komputasi model ARIMA dengan menggunakan R-CLI

Rangkuman harga estimasi dari koefisien, standard error dari koefisien dan harga-harga statistik untuk diagnostic checking (beserta harga p-value nya untuk uji yang bersesuaian didalam kurung) untuk model-model yang kita amati diatas, dapat dirangkum kedalam tabel berikut:

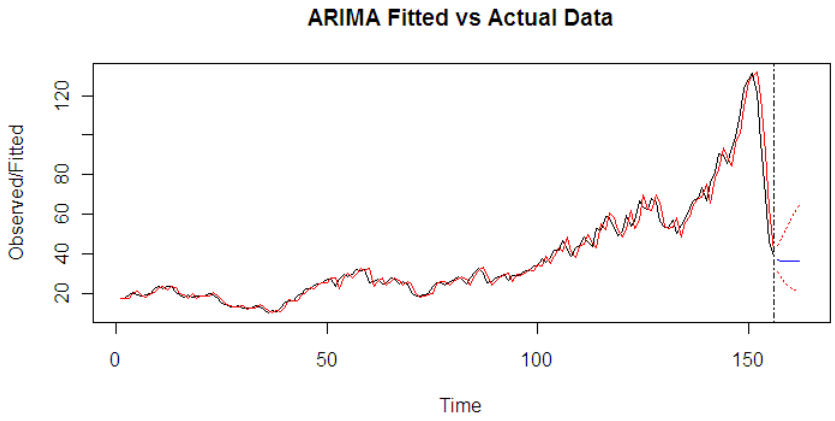
Tabel 1. Rangkuman hasil komputasi

	ARIMA(1,1,0)	ARIMA(0,1,1)	ARIMA(1,1,1)	ARIMA(2,1,0)	ARIMA(0,1,2)
a1	0.2780 SE= 0.0777		0.4419 SE=0.3306	0.2713 SE=0.0801	
a2				0.0287 SE=0.0854	
b1		0.2642	-0.1767		0.2666

		SE=0.0762	SE=0.3592		SE=0.0795
b2					0.0536 SE=0.0851
RMSE	0.091641393	0.091903319	0.091578215	0.091607300	0.091787736
AIC	-295.91	-295.04	-294.12	-294.03	-293.43
SBC/BIC	-289.83	-288.95	-284.99	-284.9	-284.3
Q(12)	11.0819 (0.521)	10.51367665 0.5709904	11.78062946 0.4634546	1.150962e+01 0.4858213	10.99860070 0.5290386
Q(24)	26.918 (0.308)	27.15678994 0.2971979	26.81362388 0.3132154	2.689308e+01 0.3094615	26.88451271 0.3098651
Q(36)	42.929 (0.1985)	42.21472253 0.2201574	43.21439506 0.1903234	4.312413e+01 0.1928927	42.64314171 0.2069993

Berdasarkan rangkuman ini, dengan uji t dan berdasarkan RMSE dan AIC yang paling minimum, dapat disimpulkan model ARIMA(1,1,0) merupakan model terbaik untuk data log(World_Oil_Prices).

Selanjutnya, akan dilakukan peramalan menggunakan model *in-sample terbaik*, yakni model ARIMA(1,1,0) untuk data log(World_Oil_Prices). Komputasi dengan R untuk peramalan diberikan pada gambar 3, sedangkan hasil fitting dan prediksi diberikan pada gambar 2 berikut.



Gambar 2. Plot hasil fitting dan prediksi

```
>pred.data = predict(ArimaModel.1, n.ahead = 6) #prediksi 6 langkah kedepan

>pred.data.low = pred.data$pred - 1.96 * pred.data$se

>pred.data.up = pred.data$pred + 1.96 * pred.data$se

>pred.data=exp(pred.data$pred)

>pred.data.low=exp(pred.data.low)

>pred.data.up=exp(pred.data.up)

>fit.data = fitted(ArimaModel.1)

>fit.data=exp(fit.data) #menghitung nilai fitting untuk World_Oil_Prices

>#plot data

>ts.plot(latihan3$World_Oil_Prices,xlim=c(1,length(latihan3$World_Oil_Prices
)+7))
```

Gambar 3. Script R untuk Prediksi model ARIMA dengan menggunakan R-CLI

4. Pemilihan model terbaik dengan fungsi auto.arima

Untuk melakukan pemilihan model terbaik menggunakan kriteria AIC, dapat juga dilakukan secara otomatis dengan menggunakan fungsi auto.arima (Hyndman dan Khandakar , 2008). Lihat contoh script R berikut

```
>automaticarima <- auto.arima(latihan3$World_Oil_Prices.Log) #pemilihan model

>summary(automaticarima) # menampillkan output
```

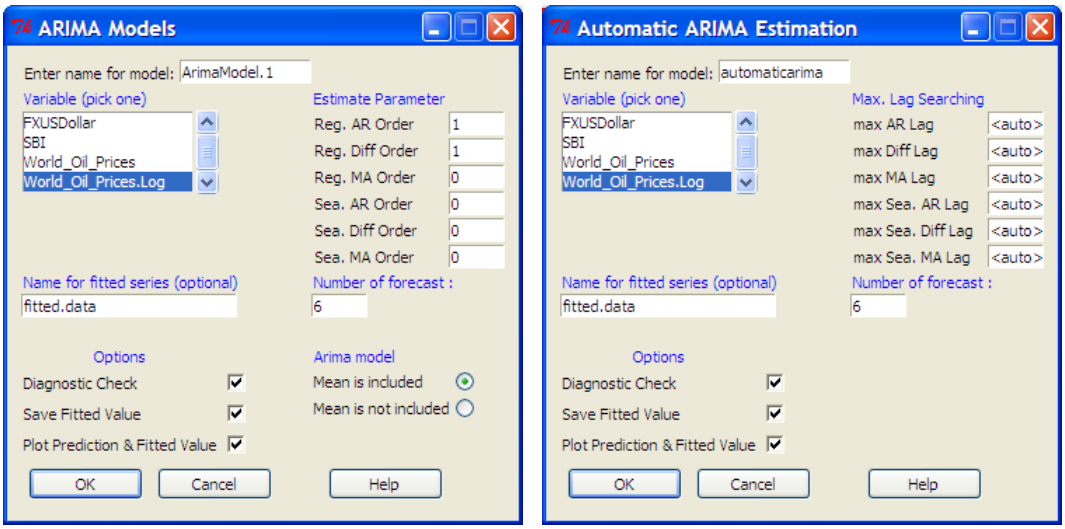
Terlihat model terbaik menurut fungsi ini adalah model ARIMA(0,1,1), sedikit berbeda dengan hasil pemodelan manual diatas. Walaupun demikian, dari tabel 1 terlihat pemodelan dengan ARIMA(0,1,1) dan ARIMA(1,1,0) memiliki nilai RMSE yang tidak berbeda secara signifikan, sehingga kedua model ini secara umum ekuivalen.

5. Arima modeling dengan R-GUI

Komputasi model ARIMA diatas dapat juga dilakukan menggunakan menu pada R-Commander Plugins yang disebut RcmdrPlugin.Econometrics, yang telah dikembangkan dalam Rosadi, Marhadi dan Rahmatullah (2009) menggunakan bahasa Tcl/Tk (Dalgaard, 2001a,b;2002; Welch, Jones dan Hobbs, 2003). Alternatif GUI diberikan pada Hodgess dan Vobach (2008). Menu untuk analisa model ARIMA dapat diberikan sebagai berikut:

- Estimasi dilakukan dengan menu **Econometrics\Univariate TS Analysis\ARIMA\ARIMA Estimation**. Silahkan untuk mengisi order dari model yang akan diestimasi. Pada jendela dialog, dapat pula diisikan/dipilih beberapa opsi lainnya, seperti opsi untuk melakukan diagnostic check, opsi untuk penyimpanan nilai *fitted value*, opsi untuk peramalan, dan lain-lain.
- Pemilihan model automatis dengan menggunakan fungsi auto.arima tersedia pada menu **Econometrics\Univariate TS Analysis\ARIMA\Automatic ARIMA**

Berikut tampilan dialog GUI untuk estimasi model ARIMA dan metode automatic ARIMA



Gambar 4. Jendela GUI untuk pemodelan ARIMA

Terlihat bahwa jendela dialog diatas memuat langkah estimasi, diagnostic check dan prediksi dalam pemodelan ARIMA dan bersifat user-defined, sehingga sangat mudah untuk digunakan dalam keperluan analisa model ARIMA. Dengan demikian diharapkan dimasa yang akan datang, R dan paket RcmdrPlugin.Econometrics dapat menjadi salah satu alternatif pilihan software terbaik untuk keperluan analisa model ARIMA.

Daftar Pustaka

Dalgaard, P., 2001a, *The R-Tcl/Tk interface*. In Kurt Hornik and Fritz Leisch, editors, Proceedings of the 2nd International Workshop on Distributed Statistical Computing, March 15-17, 2001, Technische Universität Wien, Vienna, Austria, 2001. URL http://www.ci.tuwien.ac.at/Conferences/DSC-2001/Proceedings/. ISSN 1609-395X.

Dalgaard, P., 2001b, *A Primer on the R-Tcl/Tk Package,* R News vol. 1/3, September 2001

Dalgaard, P., 2002, *Changes to the R-Tcl/Tk package*, R News vol. 2/3, December 2002

Hodgess, E. dan Vobach, C., 2008. "RcmdrPlugin.epack: A Menu Driven Package for Time Series in R**"** *Paper presented at the annual meeting of the The Mathematical Association of America MathFest, TBA, Madison, Wisconsin*, Jul 28, 2008

Hyndman, R.J. dan Khandakar, Y. 2008, "Automatic time series forecasting: The forecast package for R", ***Journal of Statistical Software***, **26**(3).

R Development Core Team, 2009, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.

Rosadi, D., Marhadi, A. dan Rahmatullah, F., 2009, Pengembangan library *Graphical User Interface* analisa runtun waktu untuk program *Open Source* R menggunakan bahasa *open source* Tcl/Tk dan aplikasinya dalam pemodelan permintaan minyak bumi, Laporan kemajuan penelitian Insentif Riset Terapan Kemenristek, Unpublished

Welch, B., Jones, K. dan Hobbs, J., 2003, *Practical Programming in Tcl and Tk*, 4th eds, Prentice Hall