TESTING FOR OUTLIERS IN RADIONUCLIDE DATA

Barbara Zhao, Wayne A. Woodward, H.L. Gray, and Mark D. Fisk
Southern Methodist University

ABSTRACT

The problem of monitoring atmospheric radionuclides over time is investigated. Such monitoring is desirable for both natural and anthropogenic radionuclides. The statistical problem is one of testing for a time series outlier, and the problem is complicated by the fact that often several observations may be missing. In fact it may be the case that several missing observations may occur immediately prior to a data value that is to be tested as an outlier. Evans (1996) proposes an exponentially weighted moving average (EWMA) approach for detecting these outliers. The EWMA approach is one that is quite popular in practice, but it is restricted to some extent by the fact that it is based on the assumption that the autoregressive integrated moving average model, ARIMA(0,1,1), is a good fit to the data. Evans presents simulation results based on simulated radionuclide data obtained from a model that he fit to Kuwait Be7 data consisting of a sinusoidal component with long period plus an autoregressive component. One problem with Evans' approach is that false alarm rates tend to be high when the data value to be tested as an outlier is preceded by a string of missing observations. In this paper we describe several alternative approaches for outlier detection, and we compare these with the Evans method using a simulation study. In this study, outlier detection capabilities are compared in the case in which no data are missing immediately prior to the data value to be tested as an outlier as well as in the more difficult case in which several data values are missing immediately prior to this value. Our results indicate that an autoregressive-based procedure suggested here has much better control over the false alarm rates than does the Evans procedure, and it has detection capability that is comparable to and sometimes better than that obtained by the Evans approach.

OBJECTIVE

The objective of this research is to investigate the problem of testing for outliers in radionuclide data. Such testing is used for detecting leakage from underground nuclear explosions, and it is also useful for equipment monitoring using natural radionuclides. In particular we develop techniques for testing for outliers in data structures typical of those encountered when monitoring natural radionuclide concentrations. Specifically, we assume that such data are collected at equal time increments but will sometimes involve missing observations. We compare our results with those obtained using the method of Evans (1996).

RESEARCH ACCOMPLISHED

Introduction

We have investigated outlier detection in time series data with the specific application being that of monitoring natural atmospheric radionuclide concentrations. Since these radionuclide concentrations are monitored regularly over time, the problem is that of detecting an outlier in a time series. Consequently, any technique for outlier detection must account for the correlation structure in the data. The radionuclide data sets often have several missing observations, sometimes resulting in contiguous strings of missing observations. Our goal is to develop outlier detection procedures capable of providing acceptable results in the presence of these missing observations. It will also be desirable that those involved in the actual monitoring will not be required to perform the sometimes difficult task of time series modeling as part of the monitoring procedure. In the following we discuss strategies for outlier detection and we compare results using our methods with those obtained using the EWMA technique of Evans (1996) via a simulation study.

Report Documentation Page

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
DEC 2000		**00-00-2000 to 00-00-2000**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Testing For Outliers In Radionuclide Data	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Southern Methodist University,Dallas,TX,75205	

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the 22nd Annual DoD/DOE Seismic Research Symposium: Planning for Verification of and Compliance with the Comprehensive Nuclear-Test-Ban Treaty (CTBT) held in New Orleans, Louisiana on September 13-15, 2000, U.S. Government or Federal Rights.

14. ABSTRACT
See Report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**9**	

Detecting Jumps

Obviously, an outlier should be detected at time k whenever the radionuclide observation at time k is "unusually large". Outlier detection techniques differ on the basis of how they define "unusually large". Letting y_k denote the radionuclide concentration detected at time k, then an outlier should be detected whenever the jump, $y_k - y_{k-1}$, is sufficiently large. If sufficient historical data are available, then a simple outlier detection procedure could be based on the empirical distribution of these "jumps". That is, we can empirically determine the values these jumps typically assume when no outlier is present in historical data. Then for a given level of significance, the $100\alpha\%$ upper critical value of these historical jumps can be estimated and compared with an observed "jump."

A problem occurs whenever it is desired to test whether y_k is an outlier when one or more values directly prior to time k are missing. In this case we can again obtain an appropriate empirical distribution based on historical data. For example, if an observation is to be tested as an outlier and the two preceding values are missing, then the relevant question concerns what type of "jump" typically occurs between the observation at time $k-3$ (i.e. the most recently observed data value in this case) and the observation at time k. Again, based on historical data we can obtain the $100\alpha\%$ upper critical value on the 3-step ahead differences. Obviously, the above procedure can be used in general for k step-ahead differences. In this report, the outlier detection technique based on these historical jump distributions will be called the "jump test".

AR-Based Prediction Method

Another common method of measuring whether an observation, y_k, is unusual is to compare it with the value of the time series at time k that is predicted to occur based on the other data. See Fox (1972) and Ljung (1989, 1993). If the observed value is sufficiently higher than the predicted value, then it is considered to be an outlier. In our implementation, predictions will be based on an autoregressive fit to the data. More generally, a time series is said to be an autoregressive moving average process of orders p and q, denoted ARMA(p, q), if

$$y_t = \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

where a_t is a zero-mean white noise process with constant variance σ_a^2. Suppose that m successive observations $Y_{(T,m)} = (y_{T+1},..., y_{T+m-1})'$ are missing from a time series $Y = (y_1, y_2,..., y_n)'$. Consider the vector $Y_0 = (y_1,..., y_{T-1}, 0,...,0, y_{T+m},..., y_n)'$ that consists of all the actual observations and with zeros placed in the missing positions. The expression

$$Y = Y_0 + XY_{(T,m)}, \tag{1}$$

where X is an $n \times m$ matrix of 1s and 0s, leads to a formula for estimating the missing values, i.e.,

$$\hat{Y}_{(T,m)} = -(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y_0 \tag{2}$$

where Σ is the covariance matrix associated with $(y_1, y_2,..., y_n)'$. Formula (2) also holds in the situation where the missing observations are not successive, say missing at $y_{t_1}, y_{t_2},..., y_{t_m}$. At time k, the difference between observed value y_K and its interpolated value $\hat{y}_k$, which is obtained from (2), can be measured by

$$\frac{(y_k - \hat{y}_k)'\Sigma_{kk}^{-1}(y_k - \hat{y}_k)}{\hat{\sigma}^2}, \tag{3}$$

which has an $F(1, n-1)$ distribution if y_k is not an outlier, where Σ_{kk}^{-1} is the kth diagonal element of Σ^{-1}, and $\hat{\sigma}^2$ is the estimate of the variance of the noise σ^2 in a time series model. Note that Ljung and Box (1979) give a formula for calculating Σ^{-1} based on known parameters in an *ARMA(p,q)* model. In our implementation we follow the suggestion of Ljung (1993), and use the estimates of the auotoregressive parameters to obtain the estimate of Σ. In any autoregressive analysis, model selection is often a key component. In the simulation results shown here, we use an AR(5) to fit the observed series. This allows for flexibility in the autoregressive model chosen and seems to work well. However, further research is needed on the impact of more careful model selection.

Model-Free Prediction Method

In order to avoid any distributional assumptions, we have considered a modification of Ljung's method that is applicable when a "large" number (*N*) of historical observations are available of which *a*% are randomly missing. The (current) data set on which an outlier is to be tested is of length *n*, and we assume that the observation to be tested as an outlier is preceded by *m* missing observations.

The method has three steps:

Step 1. Use an iterative algorithm to estimate the missing observations. At each step of the iteration, the missing observations are estimated using the formula in (2). Note also that at each step we estimate Σ directly from the definition and not based on a fitted model.

Step 2. Suppose we want to predict y_k and that some values proceeding y_k are missing. Using (2) we obtain $\hat{y}_k$ and find the difference $|y_k - \hat{y}_k|$. Using the historical data, we obtain an empirical distribution for $|y_k - \hat{y}_k|$ based upon the missing data structure present in the current realization.

Step 3. If the observed difference $|y_k - \hat{y}_k|$ is larger than the upper αth percentile of the empirical distribution obtained in Step 2, then y_k is found to be an outlier.

Evans' Method

Evans (1996) considered the problem of testing for outliers in the setting considered in this paper. He proposed a technique that is similar to the prediction-based methods presented here, but in this case the prediction is based on an exponentially-weighted moving average (EWMA) model. EWMA prediction is based on an ARIMA(0,1,1) fit to the data. A time series is said to follow an ARIMA(0,1,1) model if $y_t = y_{t-1} + a_t - \theta_1 a_{t-1}$. Evans' method is based on the computation of two prediction-interval filters that measure the average (location or level) and the variability of a process at time *t*. The extent to which the EWMA procedure provides satisfactory predictions depends to some extent on how well an ARIMA(0,1,1) model represents a reasonable fit to the data. However, this model is fairly restrictive and does not well represent many types of time series behavior. Evans reported some simulation evidence concerning the performance of his proposed outlier detector.

Simulation Procedure

We have used a simulation study to compare these four outlier detection methods. Simulations are made to compare the three methods for different time series models and different missing data structures. The details are as follows:

(a) We generate N observations { $y_k, t = 1, 2, \cdots, N$ }. In the simulations reported here, we used $N = 200$. In this study we consider several different models for the data:

(i)   AR(1) with $\varphi_1 = .6$

(II) AR(1) with $\varphi_1 = .6$ plus a sinusoid, f_t, of period six months where

$$f_t = 11\sin(\frac{2\pi t}{180} + U) \text{ and } U \text{ is a random uniform between } -2\pi \text{ and } 2\pi.$$

Note: this is Evans' simulation model based on Kuwait Be7 data.

(iii) AR(1) with $\varphi_1 = .9$

(iv) AR(1) with $\varphi_1 = .9$ + sin term as in (ii)

(v) AR(2) with $\varphi_1 = 1.41$ and $\varphi_2 = -.5$.

(vi) AR(4) with $\varphi_1 = 1.41, \varphi_2 = -1, \varphi_3 = .705$, and $\varphi_4 = -.25$.

In all cases, we use $\sigma_a^2 = 81$. In Figure 1 we show typical realizations of length 200 from each of these models.

(b) The last observation, i.e. the 200th generated observation, is tested as a potential outlier. In our simulations, we consider the case in which the number of missing values immediately preceding the outlier to be tested is $m = 0, 1, 3, 6,$ or 12. In the AR prediction approach the historical values are taken to be all observed data up to but not including the outlier. That is, the number of historical data values is $N - m - 1$. These historical observations are used for model identification and estimation of model parameters. In the case of the model-free and jump procedures, the historical data are taken to be the first $N - m - 11$ data values. In these cases, the historical data are used for obtaining empirical distributions of distributions of interest. In the simulations reported here, the historical data contain no missing values. However, our simulation code allows for missing data in the historical series. Results shown here are similar to those obtained when the historical data contained a moderate amount of missing data, i.e. 10% - 20% missing. If missing data do occur in the historical series, these missing values are "filled in" using the procedure outlined in Step 1 of the discussion of the model-free prediction method. Analysis then proceeds as in the no missing data case.

(c) Critical values for the model-free prediction method and for the jump method are obtained for the significance level of interest. In these simulations we used $\alpha = 0.05$.

(d) We test an augmented version, $x_{200}^{(0)}$, of the last data value as an outlier. Specifically, we consider $x_{200}^{(0)} = x_{200} + h\sigma_x$ where σ_x denotes the estimated standard deviation of the data and where $h = 0, 1, 2, 3,$ and 4.

Simulation Results

In Tables 1-6 we give simulation results. In each case, 1000 realizations of length $N = 200$ are generated and the tabled value is the proportion of times out of the 1000 replications that $x_{200}^{(0)}$ is detected as an outlier. It should be noted that for $h > 0$ this provides information concerning the detection capability of the test, and when $h = 0$ the proportion tabled provides for a check on the false alarm rate of the testing procedure. In the tables it can be seen that the AR-based prediction methods tend to outperform the EWMA approach recommended by Evans (1996). One serious problem with the Evans method is the fact that the false alarm rate rises above the .05 level whenever several values are missing prior to a potential outlier. Even with the higher false alarm rate, the Evans method tended to have lower power than the AR-based method for values of h greater than zero. It should be noted that the model-free method has performance similar to that of the AR-based method, although our more extensive simulation experience has shown that modeling of the data is probably preferable. It should be noted that the jump method maintains the desired false alarm rates but has a tendency to have lower detection probability than the more sophisticated methods.

Conclusions

Results shown here indicate that the AR-based prediction method is superior to the Evans approach for outlier detection in the setting considered here, particularly when missing data values occur immediately prior to the data value to be tested as an outlier. Specifically, methods discussed here are appropriate for monitoring radionuclide data which is collected at regular time intervals and for which an actual reading is obtained at most of the collection times. In the case of anthropogenic radionuclides where it is not uncommon for the atmospheric concentrations to be below a minimum detectable level the majority of the time, other outlier detection techniques must be developed.

Key Words: radionuclide monitoring, outlier testing, time series, autocorrelation

References

Evans, William C. (1996).   Automated Categorization of Airborne Radioactivity Measurements for CTBT Monitoring. *Pacific-Sierra Research Corporation Technical Note 1091*, June 1996.

Fox, A.J. (1972). Outliers in time series. *J. R. Statist. Soc*. **B 34**, 350-363.

Ljung, G.M.(1989). A note on the estimation of missing values on time series. *Commun. Statist.-Simula.***18**, 459-465.

Ljung, G.M.(1993). On outlier detection in time series. *J. R. Statist. Soc*. **B 55**, 559-567

Ljung, G.M. and Box, G.E.P. (1979). The likelihood function of stationary autoregressive moving average models. *Biometrika* **66**, 265-270.

Table 1: Outlier Detection Results using the Six Models Listed in Simulation Section

			(i)					(ii)		
	h	Evans	AR	M-F	Jump	h	Evans	AR	M-F	Jump
	0.0	0.044	0.068	0.070	0.048	0.0	0.045	0.062	0.068	0.061
No	1.0	0.158	0.217	0.237	0.184	1.0	0.243	0.306	0.313	0.260
missing	2.0	0.536	0.679	0.690	0.600	2.0	0.691	0.832	0.837	0.759
value	3.0	0.860	0.954	0.954	0.903	3.0	0.945	0.989	0.991	0.978
	4.0	0.978	0.999	0.999	0.990	4.0	0.996	1.000	1.000	0.998
	0.0	0.046	0.054	0.062	0.046	0.0	0.040	0.055	0.064	0.061
One	1.0	0.173	0.196	0.211	0.152	1.0	0.250	0.259	0.264	0.192
missing	2.0	0.537	0.570	0.586	0.419	2.0	0.706	0.685	0.702	0.563
value	3.0	0.837	0.886	0.889	0.745	3.0	0.956	0.954	0.953	0.886
	4.0	0.972	0.992	0.990	0.932	4.0	0.997	0.997	0.997	0.990
	0.0	0.092	0.076	0.085	0.060	0.0	0.100	0.054	0.068	0.042
Three	1.0	0.197	0.176	0.198	0.110	1.0	0.267	0.197	0.218	0.153
missing	2.0	0.505	0.516	0.536	0.328	2.0	0.650	0.584	0.623	0.417
values	3.0	0.803	0.832	0.839	0.603	3.0	0.923	0.907	0.917	0.761
	4.0	0.945	0.963	0.969	0.844	4.0	0.993	0.995	0.995	0.955
	0.0	0.112	0.052	0.061	0.050	0.0	0.112	0.038	0.057	0.049
Six	1.0	0.240	0.186	0.196	0.115	1.0	0.264	0.157	0.177	0.135
missing	2.0	0.528	0.519	0.531	0.315	2.0	0.655	0.548	0.583	0.406
values	3.0	0.812	0.854	0.851	0.559	3.0	0.918	0.890	0.910	0.745
	4.0	0.949	0.978	0.978	0.800	4.0	0.990	0.993	0.994	0.929
	0.0	0.125	0.057	0.076	0.053	0.0	0.130	0.047	0.069	0.050
Twelve	1.0	0.246	0.169	0.195	0.124	1.0	0.299	0.152	0.204	0.153
missing	2.0	0.525	0.484	0.514	0.298	2.0	0.638	0.520	0.574	0.386
values	3.0	0.805	0.831	0.845	0.560	3.0	0.903	0.853	0.888	0.699
	4.0	0.950	0.976	0.978	0.802	4.0	0.979	0.981	0.988	0.904

Table 1 Continued

	h	Evans	AR (iii)	M-F	Jump	h	Evans	AR (iv)	M-F	Jump
No missing value	0.0	0.042	0.069	0.081	0.061	0.0	0.048	0.052	0.069	0.045
	1.0	0.330	0.545	0.552	0.523	1.0	0.378	0.648	0.641	0.599
	2.0	0.808	0.979	0.979	0.974	2.0	0.865	0.987	0.984	0.985
	3.0	0.969	1.000	1.000	1.000	3.0	0.982	1.000	1.000	1.000
	4.0	0.997	1.000	1.000	1.000	4.0	0.998	1.000	1.000	1.000
One missing value	0.0	0.043	0.056	0.066	0.049	0.0	0.049	0.057	0.076	0.055
	1.0	0.347	0.399	0.422	0.338	1.0	0.386	0.419	0.436	0.361
	2.0	0.835	0.889	0.890	0.851	2.0	0.846	0.887	0.892	0.867
	3.0	0.972	0.995	0.990	0.989	3.0	0.982	0.997	0.995	0.993
	4.0	0.998	1.000	1.000	1.000	4.0	1.000	1.000	1.000	1.000
Three missing values	0.0	0.146	0.089	0.101	0.076	0.0	0.138	0.054	0.074	0.051
	1.0	0.362	0.267	0.281	0.214	1.0	0.382	0.264	0.292	0.222
	2.0	0.752	0.698	0.707	0.594	2.0	0.783	0.720	0.734	0.655
	3.0	0.938	0.933	0.938	0.882	3.0	0.959	0.958	0.958	0.917
	4.0	0.992	0.996	0.996	0.982	4.0	0.999	0.994	0.993	0.988
Six missing values	0.0	0.222	0.074	0.089	0.055	0.0	0.223	0.077	0.085	0.060
	1.0	0.399	0.211	0.234	0.154	1.0	0.424	0.250	0.252	0.187
	2.0	0.732	0.611	0.609	0.466	2.0	0.752	0.619	0.626	0.501
	3.0	0.932	0.893	0.887	0.781	3.0	0.937	0.910	0.904	0.801
	4.0	0.984	0.981	0.979	0.951	4.0	0.992	0.982	0.981	0.947
Twelve missing values	0.0	0.294	0.084	0.114	0.072	0.0	0.314	0.085	0.107	0.068
	1.0	0.428	0.218	0.240	0.140	1.0	0.445	0.218	0.234	0.159
	2.0	0.715	0.534	0.549	0.348	2.0	0.716	0.527	0.539	0.380
	3.0	0.908	0.844	0.847	0.643	3.0	0.911	0.849	0.857	0.673
	4.0	0.975	0.969	0.975	0.865	4.0	0.979	0.967	0.968	0.868

Table 1 Continued

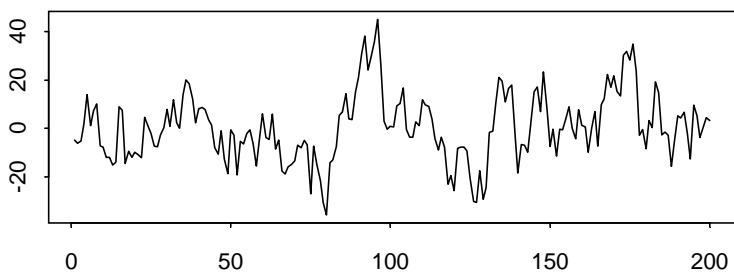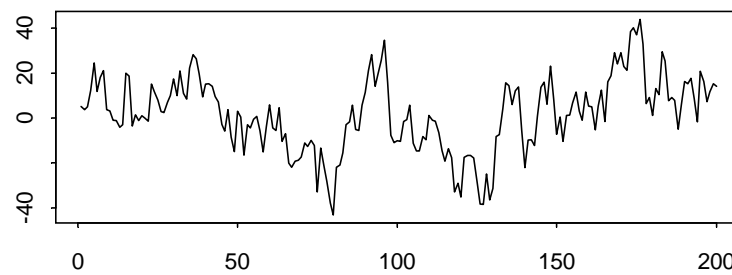|	h	Evans	(v) AR	M-F	Jump		h	Evans	(vi) AR	M-F	Jump
	0.0	0.035	0.066	0.072	0.051		0.0	0.048	0.068	0.069	0.051
No	1.0	0.356	0.875	0.878	0.761		1.0	0.280	0.659	0.665	0.469
missing	2.0	0.784	0.999	0.999	0.997		2.0	0.722	0.989	0.989	0.957
value	3.0	0.958	1.000	1.000	1.000		3.0	0.948	1.000	1.000	1.000
	4.0	0.995	1.000	1.000	1.000		4.0	0.994	1.000	1.000	1.000
	0.0	0.035	0.070	0.080	0.052		0.0	0.048	0.070	0.078	0.055
One	1.0	0.354	0.471	0.487	0.379		1.0	0.279	0.306	0.312	0.241
missing	2.0	0.785	0.949	0.954	0.895		2.0	0.721	0.779	0.788	0.659
value	3.0	0.957	0.997	0.997	0.992		3.0	0.949	0.975	0.976	0.946
	4.0	0.995	1.000	1.000	1.000		4.0	0.994	0.998	0.997	0.995
	0.0	0.175	0.065	0.077	0.053		0.0	0.142	0.065	0.078	0.052
Three	1.0	0.388	0.239	0.253	0.195		1.0	0.335	0.211	0.218	0.158
missing	2.0	0.721	0.694	0.694	0.537		2.0	0.668	0.611	0.627	0.477
values	3.0	0.930	0.950	0.953	0.863		3.0	0.910	0.917	0.925	0.821
	4.0	0.988	0.999	0.998	0.982		4.0	0.991	0.991	0.992	0.969
	0.0	0.267	0.060	0.076	0.071		0.0	0.237	0.069	0.083	0.075
Six	1.0	0.428	0.202	0.208	0.159		1.0	0.371	0.199	0.211	0.145
missing	2.0	0.695	0.551	0.568	0.388		2.0	0.654	0.535	0.553	0.384
values	3.0	0.891	0.867	0.867	0.685		3.0	0.868	0.848	0.855	0.665
	4.0	0.971	0.979	0.980	0.890		4.0	0.973	0.977	0.975	0.874
	0.0	0.287	0.064	0.088	0.062		0.0	0.249	0.072	0.095	0.060
Twelve	1.0	0.428	0.180	0.212	0.129		1.0	0.384	0.178	0.207	0.114
missing	2.0	0.682	0.511	0.541	0.298		2.0	0.651	0.509	0.544	0.308
values	3.0	0.875	0.833	0.847	0.566		3.0	0.856	0.833	0.846	0.561
	4.0	0.958	0.973	0.979	0.799		4.0	0.958	0.971	0.975	0.789

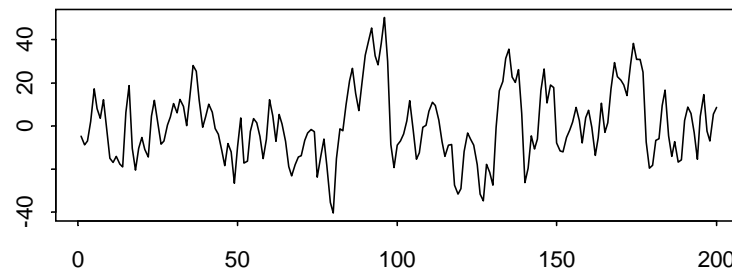**Figure 1. Typical Time Series Realizations for Simulation Models**

(i)

(ii)

(iii)

(iv)

(v)

(vi)