Hong Wang, Eulogio Pardo-Igúzquiza, Peter A. Dowd, Yongguo Yang

Comparison of statistical methods for testing the hypothesis of constant global mean in spatial statistics

Spatial Statistics, 2018; 23:143-159

Final publication at http://dx.doi.org/10.1016/j.spasta.2018.01.001

Comparison of statistical methods for testing the hypothesis of constant global mean in spatial statistics

by Hong Wang[1,2], Eulogio Pardo-Igúzquiza[3], Peter A. Dowd[4]

and YongguoYang[1,2]

[1]Key Laboratory of Coalbed Methane Resources & Reservoir Formation on Process, Ministry of Education, China University of Mining and Technology, Xuzhou, 221008, China, wanghongcumt@hotmail.com; ygyang88@hotmail.com

[2]School of Resources and Geosciences, China University of Mining and Technology, Xuzhou, Jiangsu Province, 221116, China

[3]The Geological Survey of Spain, Ríos Rosas 23, 28003, Madrid, Spain, e.pardo@igme.es

[4]Faculty of Engineering, Computer and Mathematical Sciences, University of Adelaide, Adelaide, SA 5000, Australia, peter.dowd@adelaide.edu.au

Abstract

In spatial statistics in general, and in geostatistics in particular, the choice between a spatial model with drift and a model with constant global mean is often critical, especially when only a small number of samples are available. A statistical test provides an objective means of making this choice. Among the many available statistical tests, a variance-ratio test has been widely used for making this choice because of its good statistical properties but, in addition to a semi-variogram model, it also requires an alternative drift model hypothesis. Another test statistic is the global *D*-statistic, which is a complementary test in the sense that it does not require an alternative hypothesis model. In this paper, we use sparse data from simulated random fields to evaluate and compare the performances of these two methods for testing the hypothesis of constant global mean in spatial statistics. We do so by considering the influence of four factors: the amount of data, the type of random field, the amount of spatial or temporal correlation and parametric drifts. In addition, we evaluate their performances in time series analysis, in which testing the hypothesis of constant global mean is also of significant interest. The two test statistics are compared in terms of their achieved confidence level and achieved power. The better method is the one that achieves the nominal confidence level and has higher power. We discuss departures from the nominal values and the results are used to highlight the importance of this problem in spatial statistics.

Key words: Test statistics; semi-variogram; parametric drift; non-linear drift; Monte Carlo simulation.

1 Introduction

Trend detection is important in many disciplines ranging from astronomy (Park et al., 2011) to medicine (Avent, Charlton, 1990). In the environmental sciences, detecting trends in global warming is becoming an increasingly important application area (Hall and Rajvidi, 2000; Wu and Zhao, 2007; Renard et al., 2008; Chandler and Scott, 2011). The random function model, comprising the sum of a trend and a residual, has significant practical applications in both geostatistics and time series analysis. The trend corresponds to a regional, or broad-scale, variation and the residual corresponds to local anomalies such as those often interpreted in geophysical applications. In time series analysis, the trend is viewed as a low frequency variation while the residual is the high frequency variation component. In geostatistics, if the interest is solely in providing an interpolated map by kriging with a moving window, there may be no need to model an explicit trend (Journel and Rossi, 1989) provided that an appropriate moving window can be defined that will implicitly accommodate a changing local mean. For many other applications, however, the modelling of a trend is important (Lark and Webster, 2006; Ma et al., 2010; Oy and Deutsch, 2004; Visser et al., 2009). In time series analysis, trend modelling is important in applications for detecting systematic temporal changes (Sethi et al., 2015; Tomozeiu et al., 2000; Unal et al., 2012; Yin et al., 2015; Zhang et al., 2009; Zhou et al., 2015) or for spectral analysis (Joshi et al., 2011; Posa and Rossi, 1989; Pardo-Igúzquiza and Rodríguez-Tovar, 2000, 2012). In these applications, the first step is to decide whether there is a significant trend (non-constant global mean). A constant global mean (absence of a trend) implies that the random function is stationary (Myers, 1989). It is good practice to have a statistical test for deciding whether the constant global mean hypothesis is supported by the data.

In spatial statistics, several statistical tests have been proposed for testing the constant global mean hypothesis (e.g., Fuentes, 2005; Jun and Genton, 2012; Kitanidis, 1991,1997; Leung, 2000; Pardo-Igúzquiza and Dowd, 2003). Many of these tests were devised to address specific problems and are thus somewhat application-specific. For example, Leung (2000) proposed a statistical test for spatial non-stationarity based on a geographically weighted regression model. However, this is an extension of ordinary least squares for the case in which the residuals are independent Gaussian variables. In the work presented here we focus on the more general case in which the residuals are correlated. Fuentes (2005) proposed a spectral method for testing a given spatial process for stationarity and isotropy. The spectral method works for gridded data whereas here we address the general case in which the data may be irregularly, and sparsely, distributed. Jun and Genton (2012) focus on the non-stationarity of the covariance structure, assuming the mean of the random field is zero. Here we deal with the identification of non-stationarity in the mean.

We consider two geostatistical tests: a variance-ratio (VR) test (Kitanidis, 1997) and the global *D*-statistic test (Pardo-Igúzquiza and Dowd, 2003). The VR test is based on the sum of squared orthonormal residuals (Kitanidis, 1991). In the application proposed here, it requires a semi-variogram model and a test in which the null hypothesis is a constant global mean and the alternative hypothesis is an hypothesized drift model. The global *D*-statistic requires only a semi-variogram model with no need for specifying the form of the alternative drift hypothesis model.

The purpose of this paper is to evaluate and compare the performances of both methods in testing the hypothesis of constant global mean. For geostatistical applications we use several two-dimensional simulated random fields each with a different spatial covariance model and/or drift model; the data for the tests are obtained by sparsely sampling these fields. For one-dimensional time series applications, we compare the two methods using the widely-acknowledged Mann-Kendall (MK) test. Finally, we calculate and compare the achieved confidence level and the achieved power in a Monte Carlo experiment with varying amounts of data, types of random field, spatial correlation (semi-variogram ranges) and parametric linear and non-linear drifts.

The remainder of this paper is organised as follows. Section 2 introduces the methodology used in the study, the simulation study is presented in section 3 and the conclusions are given in section 4.

2 Methodologies

2.1 The global *D*-statistic

The global *D*-statistic test was proposed by Pardo-Igúzquiza and Dowd (2003) in which a detailed description of the procedure can be found. This test statistic is based on differences of pairs of data values and the test is whether the null hypothesis of a constant global mean is supported by the available data. The alternative hypothesis is that the constant global mean cannot be accepted; this alternative hypothesis does not require the specification of a particular drift model.

The usual way of estimating the drift (or spatial *D*-statistic) from a set of values of a variable Z is:

$$D(d_k^+) = \frac{1}{N(k)} \sum_{l=1}^{N(k)} [Z_1^l - Z_2^l] \qquad \text{for } k = 1, K \tag{1}$$

where K is the number of separation distance (lag) classes for which the drift is calculated; $N(k)$ is the number of pairs of values of Z in the k^{th} distance class defined by the tolerance interval $h_k \pm \varepsilon$, where h_k is the nominal lag and ε is a tolerance; the values of h_k and ε would normally be the same as those used in the associated semi-variogram calculation; d_k^+ is the mean separation distance and the '+' superscript denotes orientation, which must be constant; for example, if the *D*-statistic is calculated for the North-South direction, the first datum in a data pair is always the more northerly (or always the more southerly); Z_1^l and Z_2^l are the first and second values of the orientation-ordered l^{th} data pair used in the computation of $D(d_k^+)$. If the data are regularly spaced, $\varepsilon = 0$ and $d_k^+ = h_k^+$. Note that for directional calculations a direction tolerance would also be required.

The global D-statistic is obtained by setting the lag tolerance ε to a value greater than the maximum distance between all data pairs, so that there is a single class that includes all data and a single (global) value for the D (drift) statistic defined as:

$$D(d_G^+) = \frac{1}{N_T} \sum_{l=1}^{N_T} [Z_1^l - Z_2^l] \ , \tag{2}$$

where N_T is the total number of pairs of values for all separation distances; d_G^+ is the mean distance between all data locations.

The test statistic, $\widetilde{D}(d_G^+)$, is the standardized global D-statistic and is calculated as:

$$\widetilde{D}(d_G^+) = \frac{D(d_G^+)}{\sqrt{\text{var}[D(d_G^+)]}} \ . \tag{3}$$

A major reason for using the global D-statistic rather than the lag-based version is that the distribution of $\widetilde{D}(d_G^+)$ is, for all practical purposes, standard Gaussian, irrespective of the number of data and the underlying distribution of the random function, as demonstrated in Pardo-Igúzquiza and Dowd (2003). Thus, the null hypothesis of a constant global mean can be accepted with a given significance level (α) if $\widetilde{D}(d_G^+)$ falls in the 100(1-α) % confidence interval as determined from the standard Gaussian distribution.

2.2 The variance-ratio test

The variance-ratio (VR) test (Kitanidis, 1991,1997) is based on the sum of squared orthonormal residuals and tests a simpler basic drift model against a more complex alternative drift model using the standard forms of drift models (Matheron, 1971). The sum of squared orthonormal residuals (Kitanidis, 1997) of the two models are given by

$$WSS_0 = \mathbf{z}^T(\mathbf{Q}^{-1} - \mathbf{Q}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{Q}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Q}^{-1})\mathbf{z} \tag{4}$$

$$WSS_1 = \mathbf{z}^T(\mathbf{Q}^{-1} - \mathbf{Q}^{-1}\mathbf{X}_1(\mathbf{X}_1^T\mathbf{Q}^{-1}\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{Q}^{-1})\mathbf{z} \tag{5}$$

where $\mathbf{z}$ is an $n \times 1$ vector of the data; $\mathbf{Q}$ is the $n \times n$ covariance matrix of the residuals of the n data; $\mathbf{X}$ is the $n \times p$ drift matrix of the basic model and $\mathbf{X}_1$ is the $n \times (p+q)$ drift matrix of the alternative model; p is the number of drift coefficients of the basic model; $p+q$ is the number of drift coefficients of the alternative model. A requirement of the model is that the p columns of $\mathbf{X}$ can be obtained from $\mathbf{X}_1$ by eliminating, or linearly combining, q of its columns. Here we take the basic model to be a constant global mean, which satisfies the requirement. The null hypothesis (basic model) is a constant spatial global mean and the alternative hypothesis (alternative model) is a variable global mean that must be specified as a particular parametric model. The normalized relative difference of the two models is

$$\nu = \frac{(WSS_0 - WSS_1)/q}{WSS_1/(n-p-q)} \; . \tag{6}$$

If $\nu > F(q, n-p-q; (1-\alpha))$, where $F(q, n-p-q; (1-\alpha))$ is the corresponding value of the F-distribution (Draper and Smith 1998), the null hypothesis of a constant mean can be rejected with a significance level of α.

2.3 A classical non-parametric test in 1D: the Mann-Kendall test

The Mann-Kendall (MK) test is a widely used standard test in time series analysis and we use it here to compare the two methods in time series applications. The MK test (Mann, 1945; Kendall, 1975) is a rank-based, non-parametric method. In this test, the null hypothesis is the constant global mean and the alternative hypothesis is that there is a trend in the data. The MK statistic S is calculated as

$$S = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \operatorname{sgn}(x_j - x_i) \, , \tag{7}$$

where x_j are the sequential data values, n is the length of the data set, and

$$\operatorname{sgn}(x_j - x_i) = \begin{cases} 1 & \text{if } x_j - x_i > 0 \\ 0 & \text{if } x_j - x_i = 0 \\ -1 & \text{if } x_j - x_i < 0 \end{cases}. \tag{8}$$

For $n \geq 8$, the statistic S is approximately normally distributed with variance:

$$\mathrm{Var}(S) = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{k=1}^{m} t_k(t_k - 1)(2t_k + 5)\right], \tag{9}$$

where m is the number of tied groups and t_k is the number of data in tied group k.

The standardized test statistic Z is computed as

$$Z = \begin{cases} \dfrac{S-1}{\sqrt{\mathrm{Var}(S)}} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \dfrac{S+1}{\sqrt{\mathrm{Var}(S)}} & \text{if } S < 0 \end{cases}. \tag{10}$$

The standardized test statistic Z follows the standard normal distribution $Z \sim N(0,1)$.

If $Z > Z_{1-\frac{\alpha}{2}}$ or $Z < Z_{\frac{\alpha}{2}}$, the null hypothesis of constant global mean is rejected with the specified significance level of α.

The MK test assumes serial independence of the data and the presence of serial correlation in the data increases the rejection rate of the null hypothesis because of the redundancy of correlated data. The application of the MK test to realizations of correlated stochastic processes requires a pre-whitening procedure to remove, or least significantly reduce, the influence of serial correlation (Bayazit and Önoz, 2007; Von Storch, 1995).

2.4 The achieved confidence level and achieved power

The confidence level (equal to 1-α, where α is type I error) is the probability of accepting the null hypothesis when it is true. The power (equal to 1-β, where β is type II error) is the probability of rejecting the null hypothesis when it is false. We estimate the achieved confidence level and achieved power in a Monte Carlo experiment as follows.

Given a stationary random field of N values of a variable $Z = \{Z_1, Z_2, \dots, Z_N\}$, a Monte Carlo sample, denoted by $Z^* = \{Z_1^*, Z_2^*, \dots, Z_M^*\}$, can be obtained by randomly sampling the field M times (M<N) without replacement. The test statistic of interest is applied to each of the Monte Carlo samples. The achieved confidence level (ACL) is then estimated by:

$$ACL = \frac{S_{acp}}{S}, \tag{11}$$

where S is the total number of Monte Carlo samples and S_{acp} is the number of samples (experiments) for which the null hypothesis is accepted.

In the same way, the achieved power (AP) can be estimated using the procedure of estimating the achieved confidence level but using the original random field plus drift. It is calculated as

$$AP = \frac{S_{rej}}{S}, \tag{12}$$

where S is the total number of Monte Carlo samples and S_{rej} is the number of samples (experiments) for which the null hypothesis is rejected.

The simulation settings and performances of each test in terms of achieved confidence level and achieved power are given in the following section.

3 Simulation study

3.1 Simulation setting

To assess the performances of the statistical tests of the hypothesis of constant global mean, we generate multiple simulated random fields with the following specifications (similar to the experiment set up by Russo and Jury, 1987):

(i) Two types of spatially correlated random field: a Gaussian random field and a Chi-square random field. The realizations of the Gaussian random field are generated by using Sequential Gaussian Simulation (Remy et al., 2020) and the realizations of the Chi-square random field are obtained by squaring the corresponding Gaussian random field to produce a highly-skewed distribution of values.

(ii) Two values of the semi-variogram ranges, equal to 10% and 30% of the side-length of the geometric field, respectively. These are representative short and long ranges; it is highly unlikely that any natural processes of interest would be completely spatially uncorrelated (i.e., zero range).

(iii) Three types of parametric drift: two linear and one non-linear drift. The two types of linear drift are polynomial, linear and quadratic, defined as:

$$m(x, y) = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 xy + \beta_4 x^2 + \beta_5 y^2$$

In the one-dimensional case, for a linear drift, we set $\beta_0 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$; for a quadratic drift $\beta_0 = \beta_2 = \beta_3 = \beta_5 = 0$ and in the two-dimensional case, for a linear drift, $\beta_0 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$. For the quadratic drift, we have set $\beta_0 = \beta_3 = 0$ and $\beta_1 = \beta_2 = -\beta_4 = -\beta_5$. The actual values of the parameters are calculated in each case to satisfy a given relationship between the variance of the drift and the variance of the residual.

The non-linear drift is the step function along the *x* direction:

$$m(x, y) = \begin{cases} 0 & x \leq L_x \\ \beta_0 > 0 & x > L_x \end{cases}$$

where L_x is 50% of the side-length of the geometric field along the *X* direction and β_0 is the size of the step, which is defined to satisfy a given relationship between the variance of the drift and the variance of the residual.

(iv) Four values of the ratio σ_L/σ_H , equal to 0.0, 0.5, 1.0 and 2.0, where σ_L and σ_H denote the standard deviation of the drift and the standard deviation of the residual random field, respectively.

(v) Four different sample sizes: 30, 40, 50 and 60 locations, the co-ordinates of which are selected randomly from the random field (the size of the random field is 1×128 for the 1D case and 64×64 for the 2D case). All four sample sizes are considered small.

For each of the 40 (2×2×3×3+4) different random fields, the four different-sized samplings (v) are repeated 200 times. This generates 80,000 different data sets for each of the 1D and 2D cases. For illustrative purposes, four of these are shown in Figs. 1 and 2.

For both the 1D and 2D cases, the achieved confidence level is estimated using the data sets comprising four simulated random fields without drift. These are: Gaussian field with short range (Gaussian1), Gaussian field with long range (Gaussian2), Chi-square field with short range (Chi-square1), and Chi-square field with long range (Chi-square2); where the short and long ranges are, respectively, 10% and 30% of the side-length of the geometric field. The achieved power is evaluated using the data sets of the four cases with added drift. The added drifts are denoted by 1 for linear, 2 for quadratic and 3 for non-linear; A, B and C denote the three ratios σ_L/σ_H, of 0.5, 1.0 and 2.0, respectively.

3.2 Results

In this section, we present and discuss, in four subsections, the experimental results for the two methods. The semi-variogram model for each data set, which is required by the global *D*-statistic test and the VR test, is estimated by maximum likelihood (Pardo-Igúzquiza, 1997) and, for all tests, the confidence intervals of the test statistics are established at $\alpha = 0.05$.

3.2.1 Achieved confidence level for the 1D case

The achieved confidence levels of all the tests are calculated by Eq. (11) and shown in Fig. 3. Note that the VR test requires an alternative hypothesis even for the simulated random fields that do not include drift (although in practical applications, the presence or absence of drift would not be known). Here, the alternative hypothesis for the VR test is specified as the drift model $m(x) = \beta_0 + \beta_1 x$ (denoted Drift_x). In these Monte Carlo simulations, the nominal significance level is set to 0.05 and the corresponding confidence level is 95%. The achieved confidence levels of both the global *D*-statistic (Fig. 3a) and the VR test (Fig. 3b) approximate, or exceed, the nominal value for all cases from which it may be concluded that there is no difference between the two methods with respect to the achieved confidence level. The results are good for the different data sizes that have been tried and for the two types of random function. On the other hand, although the achieved confidence level of the MK test with pre-whitening (Fig. 3d) achieves the nominal confidence level, Fig. 3c shows that its value is much lower without pre-whitening, making this procedure ineffective. This is because the positive serial correlation increases the probability that the MK test will detect trend when none is present (von Storch 1995). Thus, this study indicates that both the global *D*-statistic and the VR tests perform well and are superior to the MK test without pre-whitening and slightly better than the MK test with pre-whitening.

3.2.2 Achieved power for the 1D case

In this case, we specify two alternative hypotheses for the VR test - the real drift model (denoted Drift_r; for example, the alternative hypothesis is linear drift when the real drift is linear), and a drift model that differs from the real drift (denoted Drift_dr; for example, the alternative hypothesis is quadratic when the real model is linear or vice versa). The achieved powers of the two test statistics are calculated using Eq. (12). Figs. 4, 5 and 6 show the achieved powers of the global *D*-statistic and the VR test with the two different alternative hypotheses. The results in these three figures show that the achieved power of both methods is 1.0 for random fields with ratios σ_L/σ_H of 1.0 and 2.0; while for cases with a ratio σ_L/σ_H of 0.5, the achieved power is lower and much more variable. To provide a simpler and more compact comparison, we summarize the results in each of Figs. 4-6 by summing all the achieved powers for each of the four random fields to provide four overall values for each test. The four summed achieved power values for each test are given in Table 1 from which it is evident that the global *D*-statistic and the VR test with two different alternative hypothesis have similar achieved powers for the corresponding cases except the Gaussian 2 field, for which the achieved power of the VR test with real drift model is higher. However, in real situations the real drift model is usually unknown; thus, the results demonstrate that both tests perform similarly. In addition, note that, in these three figures, for the Gaussian field with long range and linear drift with σ_L/σ_H ratio equal to 0.5, for the Gaussian field with long range and non-linear drift with σ_L/σ_H ratio equal to 0.5 and for the Chi-square field with long range and non-linear drift with σ_L/σ_H ratio equal to 0.5, the estimated achieved power (ranging from 0.1 to 0.4) is much lower than the values for the other cases. This is because the variability introduced by the drift is very small with respect to the variability of the noise. The variability introduced by the drift could equally be attributed to the variability introduced by the long range of the underlying random field.

As a reference for the two compared methods, the MK test is also used in time series analysis to estimate the achieved power. Figs. 7a-d and 7e-h show the achieved power of the MK test without and with pre-whitening, respectively, for the four random fields. In Fig. 7, the MK test without pre-whitening has a high achieved power (around 1.0) for all cases; however, if pre-whitening is applied, the achieved power is much lower. The four summed achieved power values for the MK test with and without pre-whitening were also calculated and are shown in Table 1. By comparing the summed achieved power values of the global *D*-statistic and the VR test, we can conclude, in this case, that the global *D*-statistic and the VR tests generally outperform the MK test with or without pre-whitening. This is because, the MK test with pre-whitening has a very low power and the MK test without pre-whitening has a very low confidence level. Practically in all the cases the power increases as the number of data increases except for the Chi-square model with short range and non-linear drift with σ_L/σ_H ratio equal to 0.5, for which the power decreases with the sample size. The latter is an anomalous case because of the small variability introduced by the drift in that particular setting.

3.2.3 Achieved confidence level for the 2D case

Figs. 8a and 8b show the achieved confidence level of the VR test for which the alternative hypotheses are specified as drift models $m(y) = \beta_0 + \beta_1 y$ (denoted by Drift_y) and $m(x, y) = \beta_0 + \beta_1 x + \beta_1 y$ (denoted by Drift_xy). Fig. 9 shows the achieved confidence level of the global *D*-statistic test. Again, in these Monte Carlo experiments, the nominal confidence level is set to 95%. It can be seen in Fig. 9 that the achieved confidence levels obtained from the VR test, for which the alternative hypotheses are different linear drift parameterisations, are similar. Considering all the results in Figs. 8 and 9, the achieved confidence levels of both tests approach, or exceed, the nominal value in all cases, which indicates that, in terms of achieved confidence level, there is no difference between the two methods.

3.2.4 Achieved power for the 2D case

The achieved powers of the VR test with three different alternative hypotheses are shown in Figs 10, 11 and 12. In this case, the three alternative hypotheses are specified as real drift model (see Sect. 3.1), Drift_y (see Sect. 3.2.3) and Drift_xy (see Sect. 3.2.3). The achieved powers of the global *D*-statistic test are shown in Fig. 13. Results in Figs 10, 11 and 12 show that the VR test with the real drift model has the highest achieved power, followed by Drift_xy, and the lowest with Drift_y. This may indicate that the VR test with perfect knowledge of the drift model (alternative hypothesis) gives better results. To provide a simple and more compact further comparison of the achieved powers in Figs. 10, 11, 12 with those in Fig. 13, as was done in the 1D cases, the results in these Figures have been summarized by summing the power values for each of the four random fields of each test. The results are given in Table 2 from which it can be seen that the achieved powers of the global *D*-statistic test are lower than the corresponding values of the VR test when the alternative hypothesis is the real drift model, but higher than the corresponding achieved powers of the VR test for the other two alternative drifts. However, since, in practical applications, the exact form of the drift is usually unknown, these results demonstrate conclusively that the global *D*-statistic outperforms the VR test in both its simplicity and higher performance. In addition, as with the 1D case, the abnormally lower powers in Fig. 10b with 1-A, Fig. 13b with 1-A and 3-A, Fig.13d with 3-A drifts are caused by the fact that the variability of the drift is attributed to the variability of the residual. In all cases the achieved power increases with the sample size.

4 Conclusions

In this work, we present a flexible simulation framework to compare two test statistics for trend detection: the global *D*-statistic test and a variance-ratio test. For time series (1D case) we have used the Mann-Kendal (MK) test with and without pre-whitening as a benchmark for comparison purposes. In this framework, the influences of four factors have been considered: the amount of data, the type of random field, the amount of spatial or temporal correlation and different parametric drifts. The performance of the tests has been evaluated and compared with the achieved confidence level and achieved power. Experimental results from the simulation study have demonstrated that for the 1D case, the VR and the global *D*-statistic tests perform similarly and are preferable to the MK test. This is an important result because the MK test is one of the most used by many practitioners. For the 2D case, the global *D*-statistic and the VR test with the real drift model provided similar results, which implies that the global *D*-statistic is better since it does not require a trend model as an alternative to the global constant mean and, in practice, the true trend would never be known. The results (the estimated confidence levels and powers) are good for the sample sizes used and which, ranging from 30 to 60 data, can be considered as small data sets. Overall, on grounds of simplicity of use and better performance, the global *D*-statistic test is the preferred choice.

Acknowledgements

The first author is a visiting scientist at the Geological Survey of Spain. She and the fourth author are supported by NSFC (41672324, 41430317), the Priority Academic Program Development of Jiangsu Higher Education Institutions, the National Science and Technology Major Projects (2016ZX05044-002) and Surface Well Placement Optimization via the Topology Analysis of Well Spatial Form (41402291). The work of the second author was supported by the project KARTSCLIMA, CGL2015-71510-R, from the Ministerio de Economía y Competitividad of Spain. The work of the third author was supported by Australian Research Council Discovery grant DP110104766. We thank the anonymous reviewers for providing constructive criticism that has assisted the authors to improve the final version of the paper.

References

Avent, R.K., Charlton, J.D., 1990. A critical review of trend-detection methodologies for biomedical monitoring systems. Crit. Rev. Biomed. Eng. 17(6), 621-659.

Bayazit, M., Önoz, B., 2007. To prewhiten or not to prewhiten in trend analysis? Hydrolog. Sci. J. 52, 611-624 .

Chandler, R.E., Scott, E.M., 2011. Statistical methods for trend detection and analysis in the Environmental Sciences. Wiley and Sons, Chichester, UK, p388.

Draper, N.R., Smith, H., 1998. Applied regression analysis. 3rd edn. Wiley-Interscience, New York.

Fuentes, M., 2005. A formal test for nonstationarity of spatial stochastic processes. J. Multivariate Anal. 96, 30-54.

Hall, P., Tajvidi, N., 2000. Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data. Stat. Sci. 15 (2), 153-167.

Joshi, M.K., Pandey, A.C., 2011. Trend and spectral analysis of rainfall over India during 1901-2000. J. Geophys. Res. 116, 1-13.

Journel, A.G., Rossi, M.E., 1989. When do we need a trend model in kriging ? Math. Geosci. 21, 715-739.

Jun, M., Genton, M.G., 2012. A test for stationarity of spatio-temporal random fields on planar and spherical domains. Stat. Sinica. 22, 1737-1764.

Kendall, M.G., 1975. Rank correlation methods. Griffin, London

Kitanidis, P.K., 1991. Orthonormal residuals in geostatistics: model criticism and parameter estimation. Math. Geosci. 23, 741-758.

Kitanidis, P.K., 1997. A variance-ratio test for supporting a variable mean in Kriging. Math. Geosci. 29, 335-348.

Lark, R.M., Webster, R., 2006. Geostatistical mapping of geomorphic variables in the presence of trend. Earth Surf. Proc. Land, 31, 862-874.

Leung, Y., 2000. Statistical tests for spatial nonstationarity based on the geographically weighted regression model. Environ. Plann. A., 32, 9-32.

Ma, F., Wang, Y., Li, Y., Ye, C., Xu, Z., Zhang, F., 2010. The application of geostatistics in grain size trend analysis: A case study of eastern Beibu Gulf. J. Geogr. Sci. 20, 77-90.

Mann, H.B., 1945. Nonparametric tests against trend. Econometrica, 13, 245-259.

Matheron, G., 1971. The theory of regionalized variables, Les cahiers du Centre de Morphologie Mathématique de Fontainebleau. No 5, 211.

Myers, D.E., 1989. To be or not to be... stationary? That is the question. Math. Geosci., 21, 347-362.

Oy, L., Deutsch, C.V., 2004. Transformation of residuals to avoid artifacts in geostatistical modelling with a trend. Math. Geosci. 36, 287-304.

Pardo-Igúzquiza, E., 1997. MLREML: a computer program for the inference of spatial covariance parameters by maximum likelihood and restricted maximum likelihood. Comput. Geosci. 23, 153-162.

Pardo-Igúzquiza, E., Dowd, P., 2003. Testing for constant spatial mean using the global *D*-statistic. Comput. Geosci. 29, 1057-1068.

Pardo-Igúzquiza, E., Rodríguez-Tovar, F.J., 2000. The permutation test as a non-parametric method for testing the statistical significance of power spectrum estimation in cyclostratigraphic research. Earth Planet. Sci. Lett. 181, 175-189.

Pardo-Igúzquiza, E., Rodríguez-Tovar, F.J., 2012. Spectral and cross-spectral analysis of uneven time series with the smoothed Lomb-Scargle periodogram and Monte Carlo evaluation of statistical significance. Comput. Geosci. 49, 207-216.

Park, C., Ahn J., Hendry M. and Jang W., 2011. Analysis of long period variable stars with nonparametric tests for trend detection. J. Am. Stat. Assoc. 106, 832-845.

Posa, D., Rossi, M.E., 1989. Applying Stationary and Non-Stationary Kriging. Metron 47, 295-312.

Remy, N., Boucher, A., Wu, J., 2009. Applied Geostatistics with SGeMS: A User's Guide. Cambridge University Press. Cambridge, UK, p264 .

Russo, D., Jury, W.A., 1987. A theoretical study of the estimation of the correlation scale in spatially variable fields: 2. Nonstationary fields. Water Resour. Res. 23, 1269-1279.

Sethi, R., Pandey, B.K., Krishan, R., Khare, D., Nayak, P.C., 2015. Performance evaluation and hydrological trend detection of a reservoir under climate change condition. Model. Earth Syst. Environ. 1, 33.

Tomozeiu, R., Busuioc, A., Marletto, V., Zinoni, F., Cacciamani, C., 2000. Detection of changes in the summer precipitation time series of the region Emilia-Romagna, Italy. Theor. Appl. Climatol. 67, 193-200.

Unal, Y.S., Ali, D., Toros, H., Incecik, S., 2012. Temporal and spatial patterns of precipitation variability for annual, wet, and dry seasons in Turkey. Int. J. Climatol. 32, 392-405.

Visser, A., Dubus, I., Broers, H.P., Brouyère, S., et al., 2009. Comparison of methods for the detection and extrapolation of trends in groundwater quality. J. Environ. Monitor 11, 2030-2043.

Von Storch, V.H., 1995. Misuses of statistical analysis in climate research. In analysis of climate variability: applications of statistical techniques, von Storch H, Navarra A (eds). Springer-Verlag: Berlin, 11-26.

Wu, W.B. and Zhao, Z., 2007. Inference of trends in time series. J. R. Stat. Soc., Series B, 69, Part 3, 391-410.

Yin ,Y., Liu, H., Yi, X., Liu, W., 2015. Spatiotemporal variation and abrupt change analysis of temperature from 1960 to 2012 in the Huang-Huai-Hai plain, China. Adv. Meteorlo. 1-11.

Zhang, Q., Xu, C.Y., Becker, S., Zhang, Z.X., Chen, Y.D., Coulibaly, M., 2009. Trends and abrupt changes of precipitation maxima in Pearl river basin, China. Atmos. Sci. Let. 10, 132-144.

Zhou, J., Liang, Z., Liu, Y., Guo, H., He, D., Zhao, L., 2015. Six-decade temporal change and seasonal decomposition of climate variables in Lake Dianchi watershed (China): stable trend or abrupt shift? Theor. Appl. Climatol. 119, 181-191.

List of tables

Table 1 The summary of achieved powers for each of four random fields (1D) with different tests. A perfect test with a power of one for all the 36 cases will have a score of 36. This appears to be achieved by the MK test without prewhitening. However this is an illusion as this test is practically useless because it provides a confidence level much lower than the nominal.

Table 2 The summary of achieved powers for each of four random fields (2D) with different tests. A perfect test with a power of one for all 36 cases will have a score of 36.

Table 1

Random Field	Global *D*-statistic	VR Test		MK Test	
		Drift_r	Drift_dr	without Pre-whitening	Pre-whitening
Gaussian1	34.06	33.83	33.63	35.35	20.05
Gaussian2	28.43	32.03	28.42	35.86	20.58
Chi-square1	33.28	33.06	34.22	35.99	27.17
Chi-square2	29.99	30.66	31.28	35.98	21.49

Table 2

Random Field	Global *D*-statistic	VR Test		
		Drift_r	Drift_y	Drift_xy
Gaussian1	30.35	32.97	8.55	28.64
Gaussian2	24.88	28.78	7.39	18.24
Chi-square1	31.05	33.97	8.26	30.29
Chi-square2	29.79	34.11	11.49	22.93

List of Figures

quadratic drift, σ_L/σ_H = 2.0; 3-A: non-linear drift, σ_L/σ_H = 0.5; 3-B: non-linear drift, σ_L/σ_H = 1.0; 3-C: non-linear drift, σ_L/σ_H = 2.0;

Fig. 7 Achieved power of hypothesis test using the MK test without pre-whitening for four random fields. **a**: Gaussian1, **b**: Gaussian2, **c**: Chi-square1, **d**: Chi-square2 and with pre-whitening for four random fields. **e**: Gaussian1, **f**: Gaussian2, **g**: Chi-square1, **h**: Chi-square2

Legend: 1-A: linear drift, σ_L/σ_H = 0.5; 1-B: linear drift, σ_L/σ_H = 1.0; 1-C: linear drift, σ_L/σ_H = 2.0; 2-A: quadratic drift, σ_L/σ_H = 0.5; 2-B: quadratic drift, σ_L/σ_H = 1.0; 2-C: quadratic drift, σ_L/σ_H = 2.0; 3-A: non-linear drift, σ_L/σ_H = 0.5; 3-B: non-linear drift, σ_L/σ_H = 1.0; 3-C: non-linear drift, σ_L/σ_H = 2.0;

Fig. 8 Achieved confidence level for four random fields without drift: Gaussian1, Gaussian2, Chi-square1, Chi-square2 using the VR test at a 95% confidence interval ($\alpha = 0.05$). **a**: Assuming a drift in the Y direction only (Drift_y). **b**: Assuming a drift in both the X and the Y directions (Drift_xy)

Fig. 9 Achieved confidence level for four random fields without drift: Gaussian1, Gaussian2, Chi-square1, Chi-square2 using the global D-statistic at a 95% confidence interval ($\alpha = 0.05$)

Fig. 10 Achieved power of the hypothesis test using the VR test with the alternative hypothesis specified as the real drift model for different random fields. **a**: Gaussian1, **b**: Gaussian2, **c**: Chi-square1, **d**: Chi-square2

Legend: 1-A: linear drift, σ_L/σ_H = 0.5; 1-B: linear drift, σ_L/σ_H = 1.0; 1-C: linear drift, σ_L/σ_H = 2.0; 2-A: quadratic drift, σ_L/σ_H = 0.5; 2-B: quadratic drift, σ_L/σ_H = 1.0; 2-C: quadratic drift, σ_L/σ_H = 2.0; 3-A: non-linear drift, σ_L/σ_H = 0.5; 3-B: non-linear drift, σ_L/σ_H = 1.0; 3-C: non-linear drift, σ_L/σ_H = 2.0;

Fig. 11 Achieved power of hypothesis test using the VR test with the alternative hypothesis specified as Drift_y for four different random fields. **a**: Gaussian1, **b**: Gaussian2, **c**: Chi-square1, **d** :Chi-square2

Legend: 1-A: linear drift, σ_L/σ_H = 0.5; 1-B: linear drift, σ_L/σ_H = 1.0; 1-C: linear drift, σ_L/σ_H = 2.0; 2-A: quadratic drift, σ_L/σ_H = 0.5; 2-B: quadratic drift, σ_L/σ_H = 1.0; 2-C: quadratic drift, σ_L/σ_H = 2.0; 3-A: non-linear drift, σ_L/σ_H = 0.5; 3-B: non-linear drift, σ_L/σ_H = 1.0; 3-C: non-linear drift, σ_L/σ_H = 2.0;

Fig. 12 Achieved power of hypothesis test using the VR test with the alternative hypothesis specified as Drift_xy for four different random fields. **a**: Gaussian1, **b**: Gaussian2, **c**: Chi-square1, **d**: Chi-square2

Legend: 1-A: linear drift, σ_L/σ_H = 0.5; 1-B: linear drift, σ_L/σ_H = 1.0; 1-C: linear drift, σ_L/σ_H = 2.0; 2-A: quadratic drift, σ_L/σ_H = 0.5; 2-B: quadratic drift, σ_L/σ_H = 1.0; 2-C: quadratic drift, σ_L/σ_H = 2.0; 3-A: non-linear drift, σ_L/σ_H = 0.5; 3-B: non-linear drift, σ_L/σ_H = 1.0; 3-C: non-linear drift, σ_L/σ_H = 2.0;

Fig. 13 Achieved power of hypothesis test using the global *D*-statistic for different drifts and random fields. **a**: Gaussian1, **b**: Gaussian2, **c**: Chi-square1, **d**: Chi-square2

Legend: 1-A: linear drift, σ_L/σ_H = 0.5; 1-B: linear drift, σ_L/σ_H = 1.0; 1-C: linear drift, σ_L/σ_H = 2.0; 2-A: quadratic drift, σ_L/σ_H = 0.5; 2-B: quadratic drift, σ_L/σ_H = 1.0; 2-C: quadratic drift, σ_L/σ_H = 2.0; 3-A: non-linear drift, σ_L/σ_H = 0.5; 3-B: non-linear drift, σ_L/σ_H = 1.0; 3-C: non-linear drift, σ_L/σ_H = 2.0;

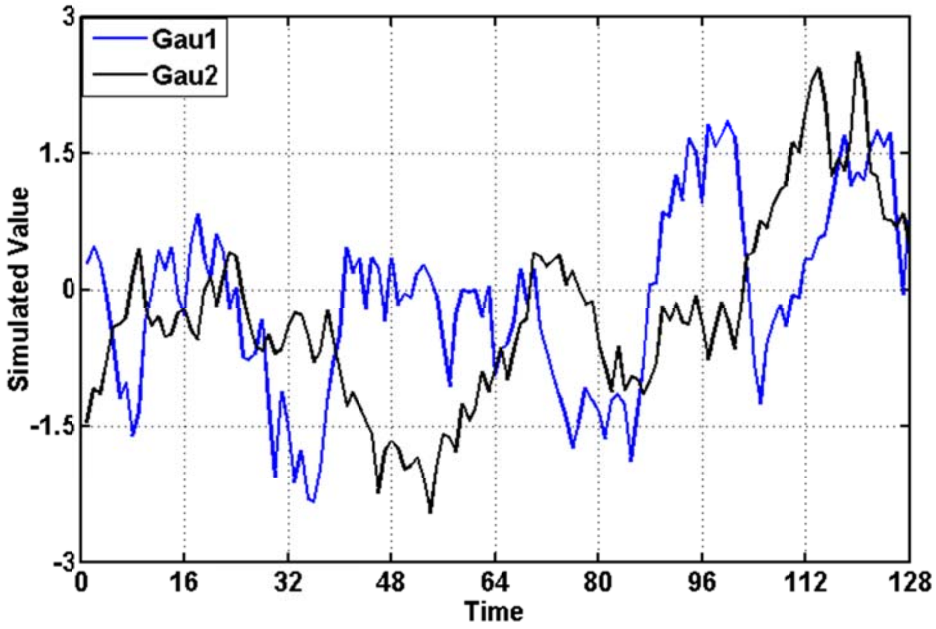

Gau1
Gau2
3
1.5
0
-1.5
-3
Simulated Value
0
16
32
48
64
80
96
112
128
Time

Fig. 1

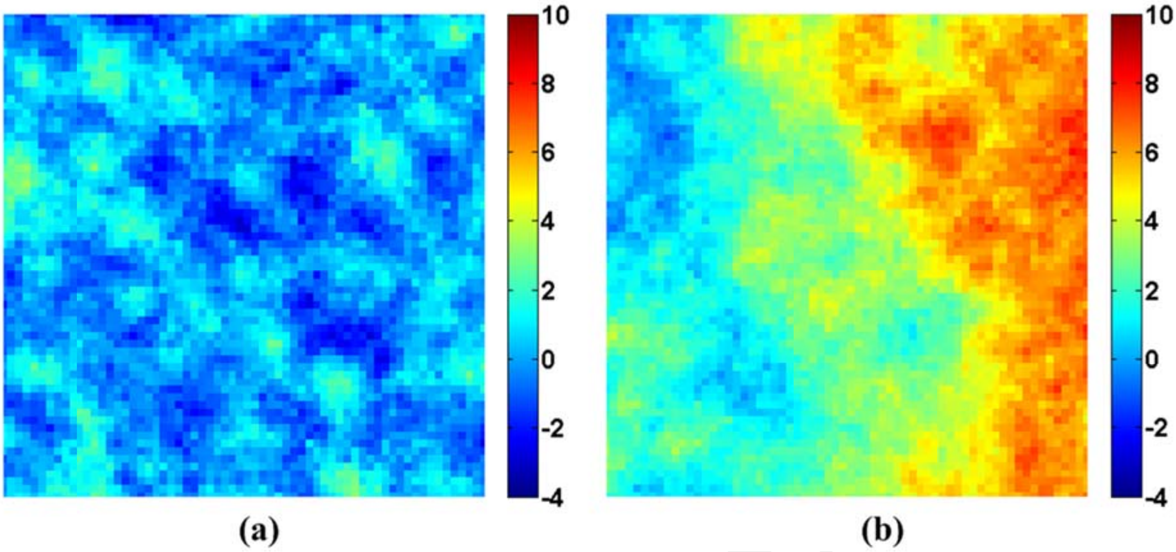
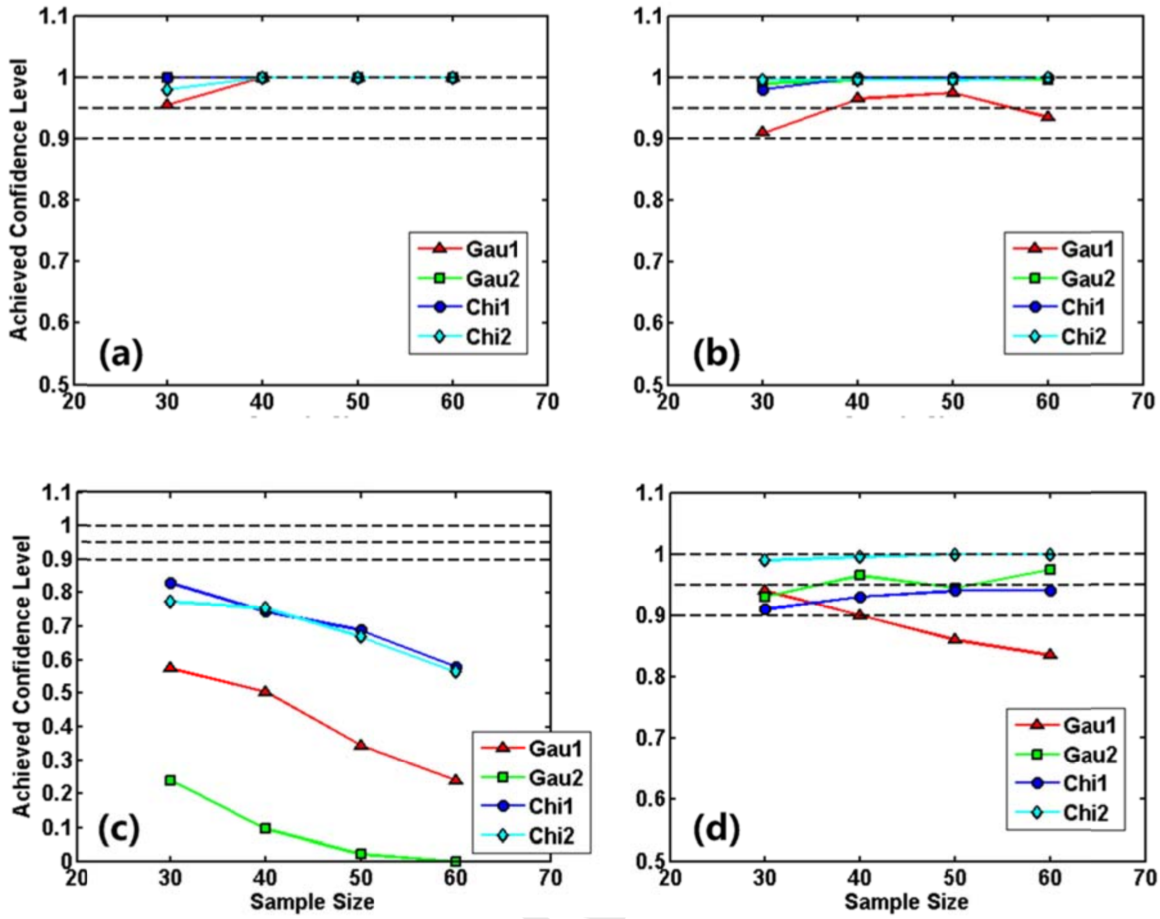


Fig. 2



Achieved Confidence Level
Sample Size
(a)
(b)
(c)
(d)
Gau1
Gau2
Chi1
Chi2

Fig. 3

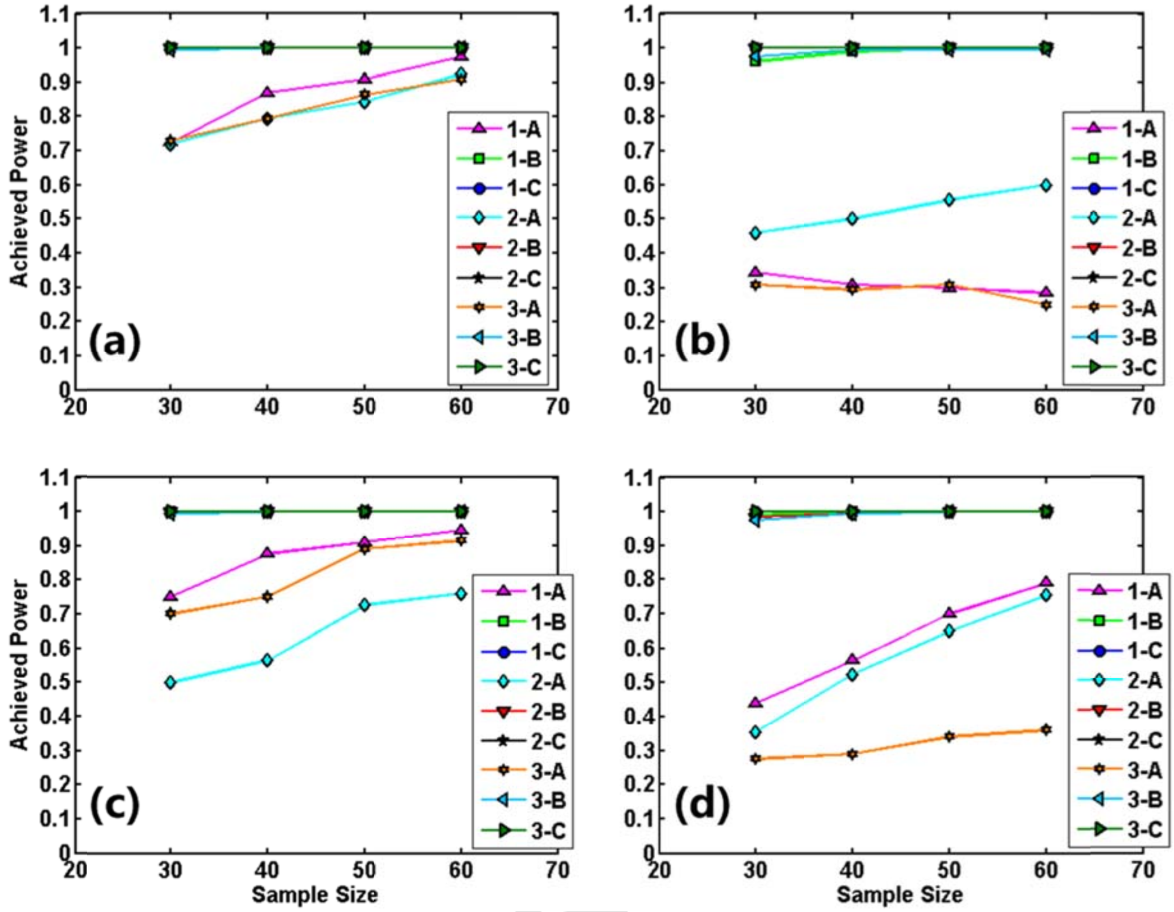


Fig. 4

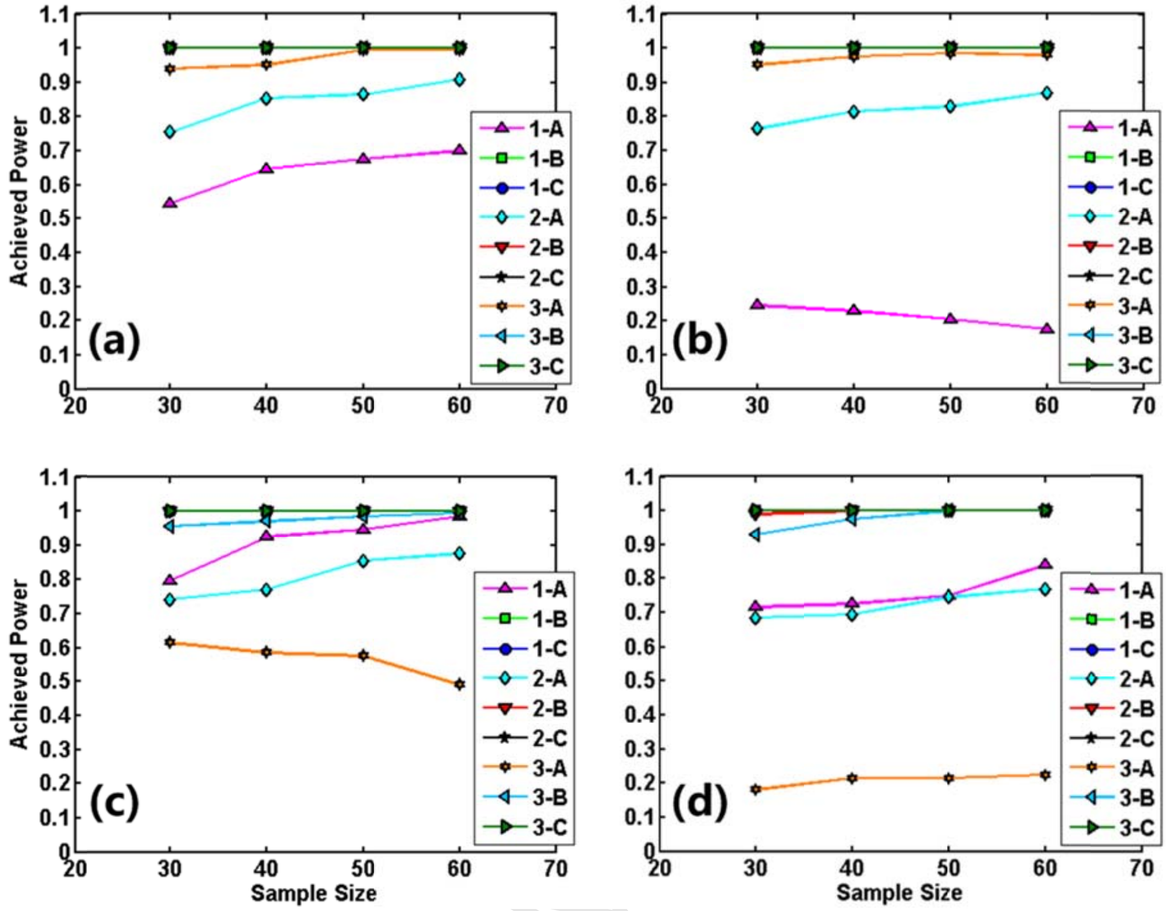

(a)
(b)
(c)
(d)
Achieved Power
Sample Size
1-A
1-B
1-C
2-A
2-B
2-C
3-A
3-B
3-C
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
20
30
40
50
60
70

Fig. 5

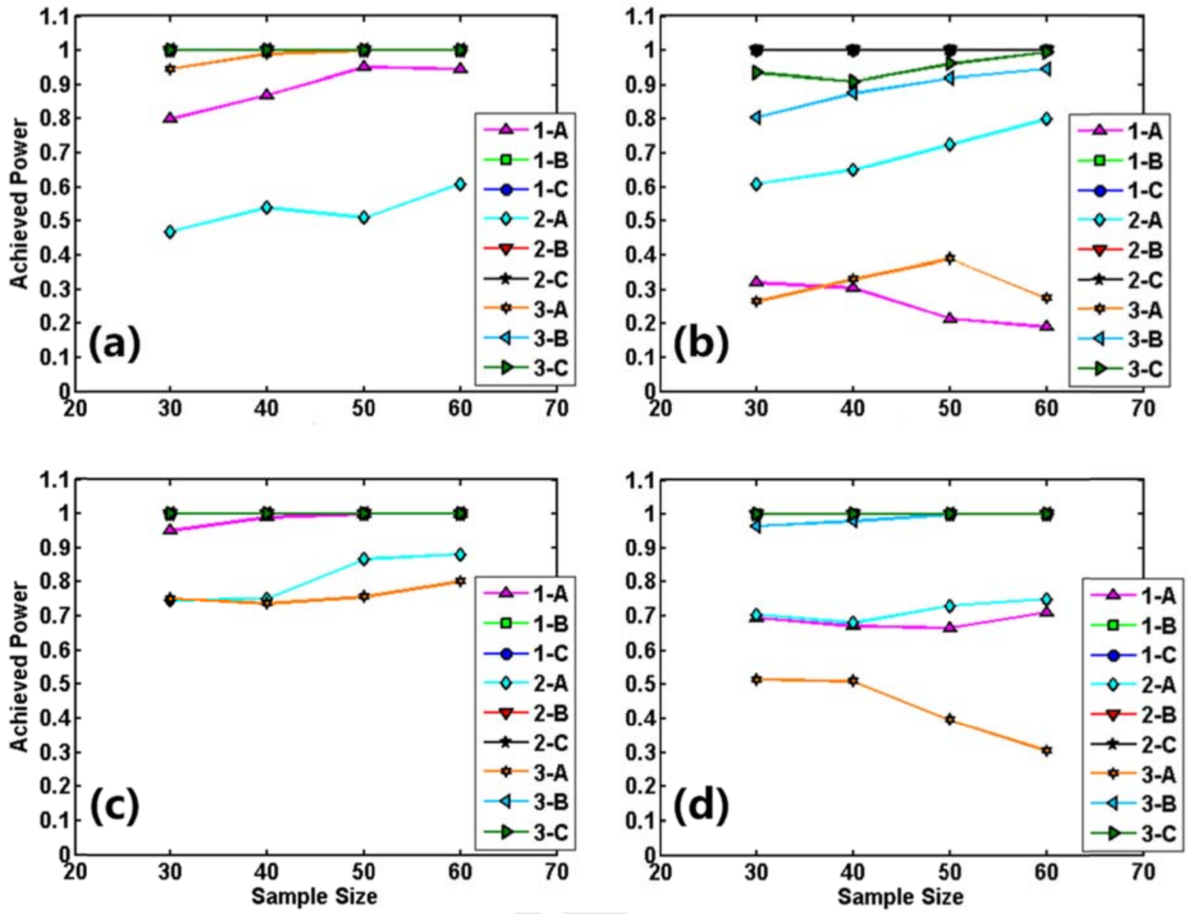


Fig. 6

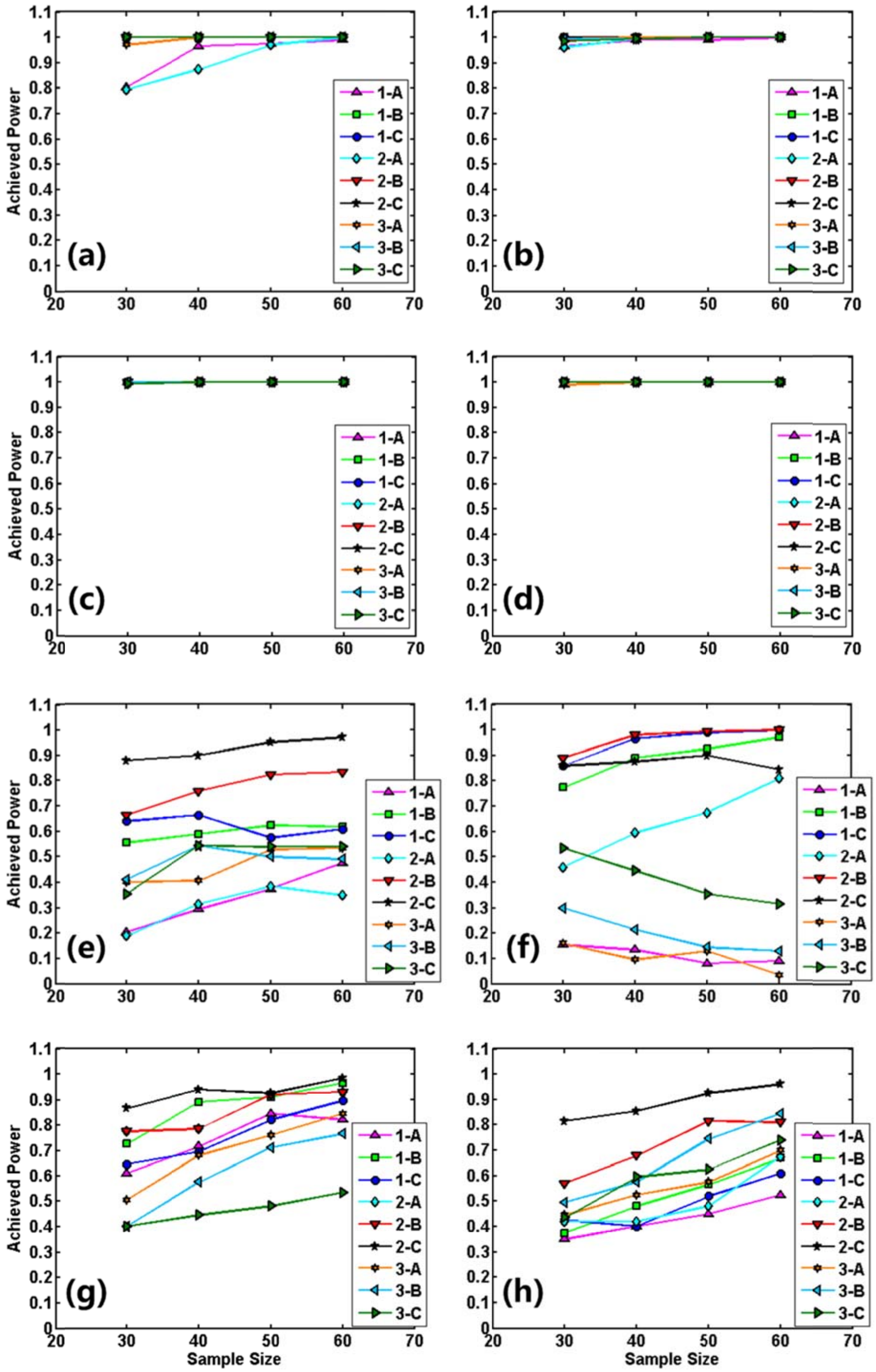
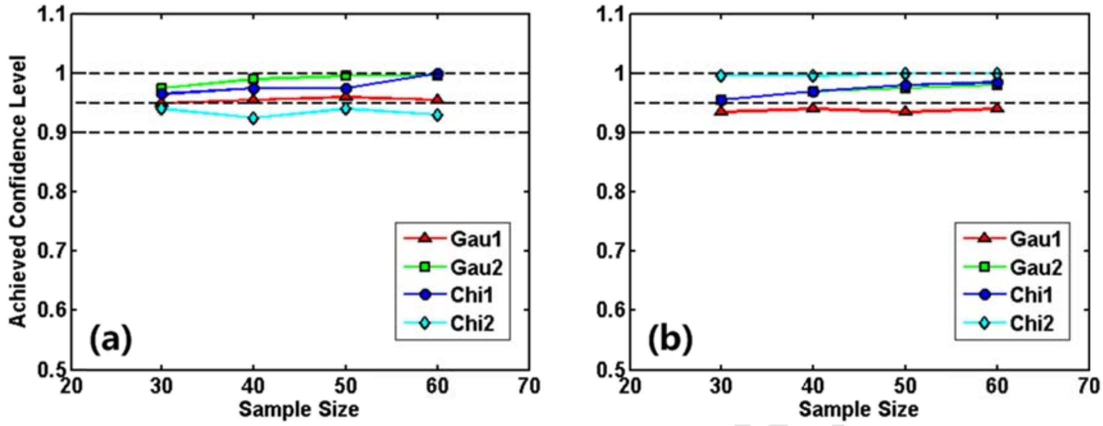

(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
Achieved Power
Sample Size
1-A
1-B
1-C
2-A
2-B
2-C
3-A
3-B
3-C

Fig. 7



Achieved Confidence Level
Sample Size
(a)
(b)
Gau1
Gau2
Chi1
Chi2
1.1
1
0.9
0.8
0.7
0.6
0.5
20
30
40
50
60
70

Fig. 8

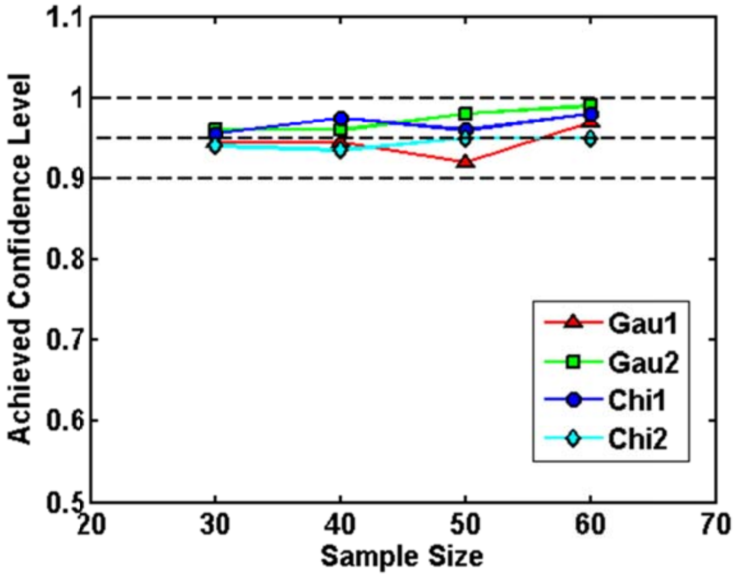

1.1
1
0.9
0.8
0.7
0.6
0.5
20
30
40
50
60
70
Sample Size
Achieved Confidence Level
Gau1
Gau2
Chi1
Chi2

Fig. 9

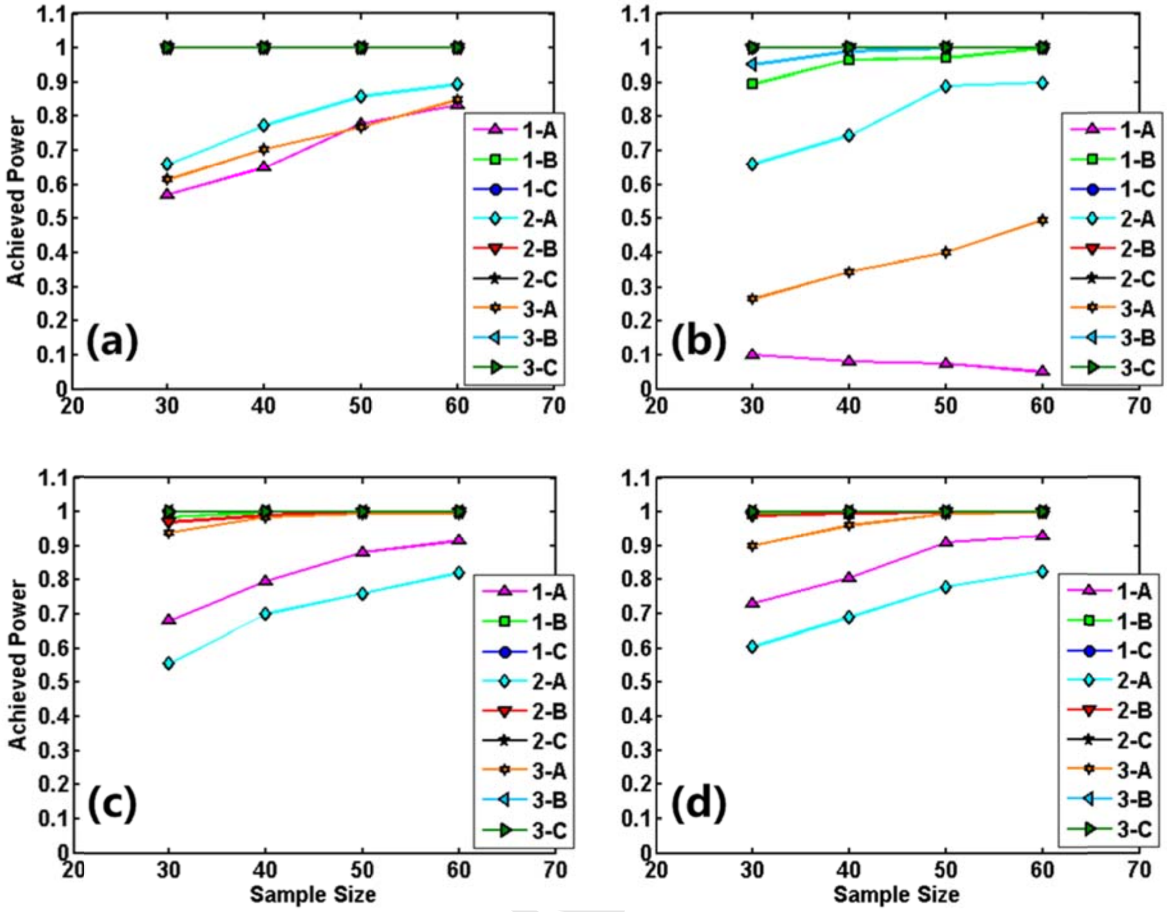


Fig. 10

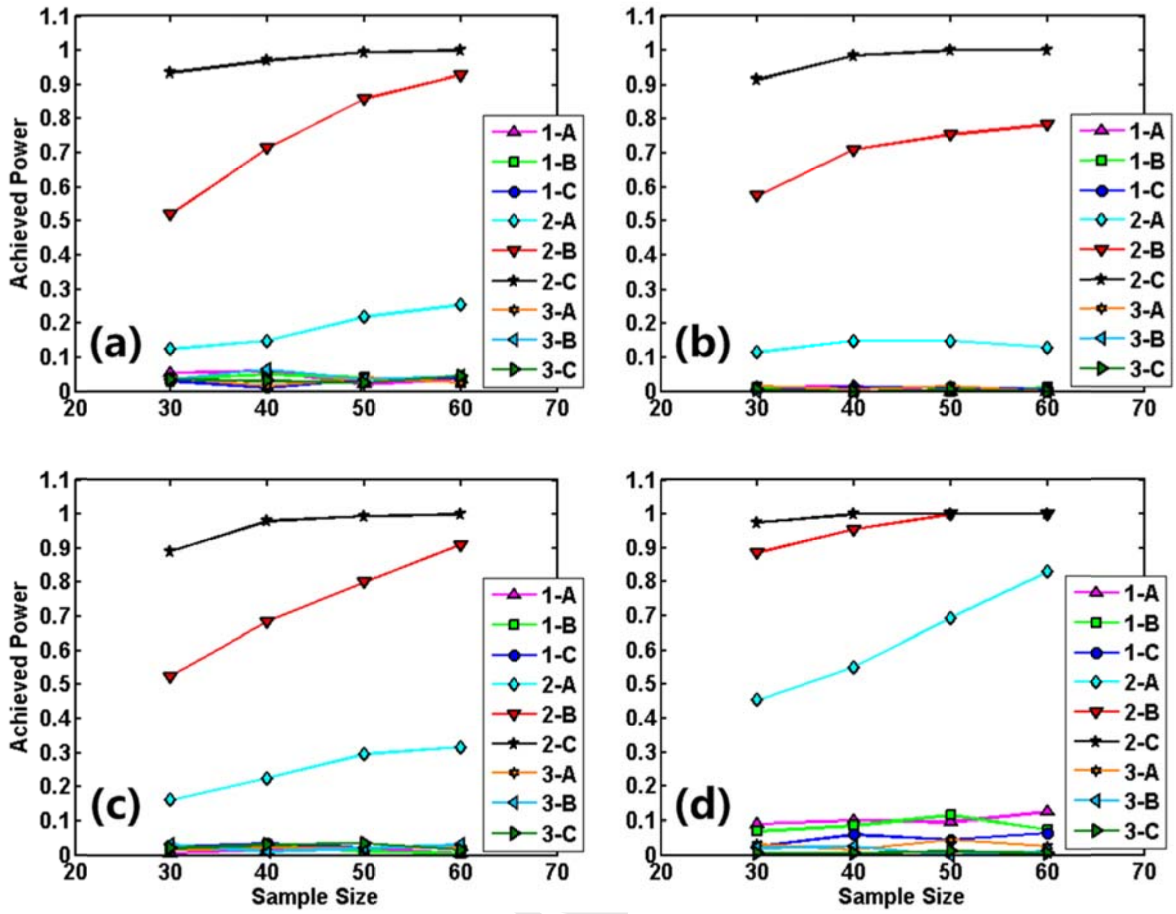


Fig. 11

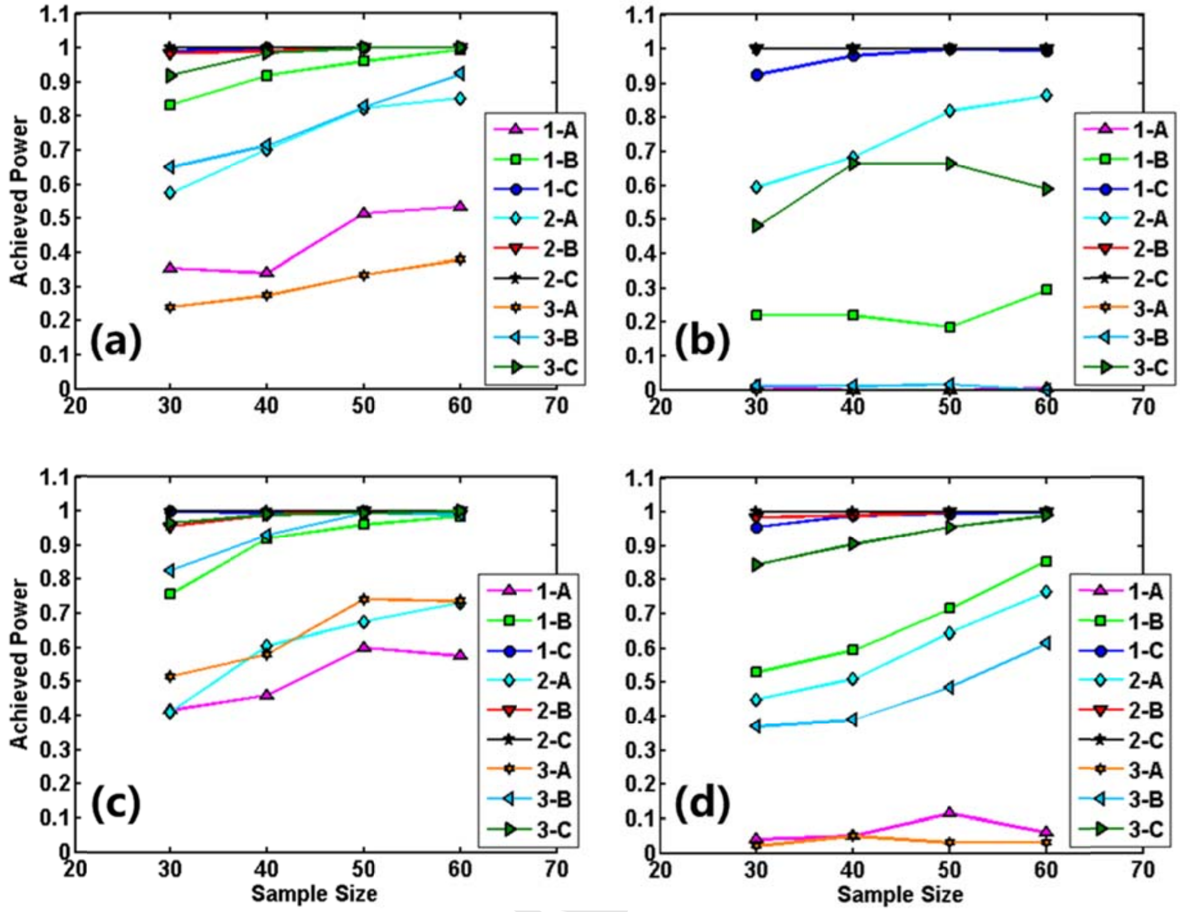

(a)
(b)
(c)
(d)
Achieved Power
Sample Size
1-A
1-B
1-C
2-A
2-B
2-C
3-A
3-B
3-C
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
20
30
40
50
60
70

Fig. 12

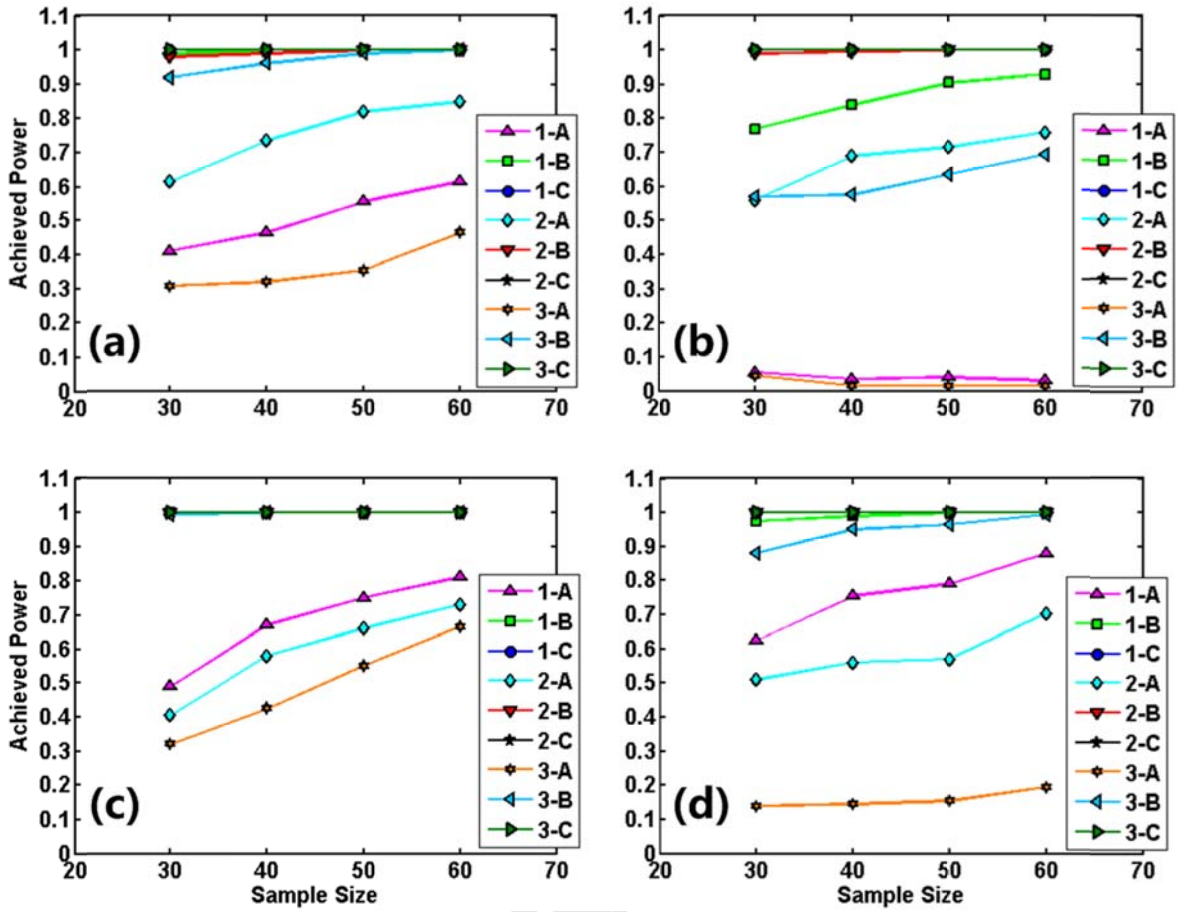


Fig. 13