Journal of Global Positioning Systems (2012)
Vol.11, No.1 : 33-45
DOI: 10.5081/jgps.11.1.33

Journal of Global Positioning Systems
2012

Regression and Hypothesis Tests for Multivariate GNSS State Time Series

Yanming Feng
Science and Engineering Faculty, Queensland University of Technology, GPO Box 2434, Q4001, Australia

Abstract

A satellite based observation system can continuously or repeatedly generate a user state vector time series that may contain useful information. One typical example is the collection of International GNSS Services (IGS) station daily and weekly combined solutions. Another example is the epoch-by-epoch kinematic position time series of a receiver derived by a GPS real time kinematic (RTK) technique. Although some multivariate analysis techniques have been adopted to assess the noise characteristics of multivariate state time series, statistic testings are limited to univariate time series. After review of frequently used hypotheses test statistics in univariate analysis of GNSS state time series, the paper presents a number of T-squared multivariate analysis statistics for use in the analysis of multivariate GNSS state time series. These T-squared test statistics take the correlation between coordinate components into account, which is neglected in univariate analysis. Numerical analysis was conducted with the multi-year time series of an IGS station to schematically demonstrate the results from the multivariate hypothesis testing in comparison with the univariate hypothesis testing results. The results have demonstrated that, in general, the testing for multivariate mean shifts and outliers tends to reject less data samples than the testing for univariate mean shifts and outliers under the same confidence level**.** It is noted that neither univariate nor multivariate data analysis methods are intended to replace physical analysis. Instead, these should be treated as complementary statistical methods for a prior or posteriori investigations. Physical analysis is necessary subsequently to refine and interpret the results.

Key words: GNSS state time series, univariate analysis, multivariate analysis, T-squared statistics.

---

1. Introduction

With a GNSS-based observation system, users can repeatedly generate state vectors from time to time, despite possibly at different accuracy levels. Typically, the state vectors can include any combination of satellite orbit and clock parameters, ground station coordinates and clock biases, atmospheric delays etc. A local network of GPS stations may be continuously observed to produce high-rate station displacements for monitoring the earthquake in its coverage area Borghi et al. (2009). A regional or global Continuous Operating Reference Stations (CORS) network is more often used to generate daily or weekly station solutions for geodynamics studies, such as crustal deformation monitoring. These solutions are usually given in form of coordinates biases with respect to a certain reference frame such as International Terrestrial Reference Frame 2008-ITRF2008. For instance, the International GNSS Services (IGS) routinely generate a number of weekly, daily and sub-daily products. Station coordinates and velocities, earth rotation parameters (ERPs) and apparent geocentre are among these products generated (Ferland and Piraszewski (2009), Ferland (2006), Altamimi and Collilieux (2008). Form a global CORS stations, the IGS community also generate various GNSS orbital and clock products for the satellites over the same periods, including daily available IGS rapid orbits, final orbits for various applications and services. In recent years, a number of IGS data analysis centres start to generate real time GPS/Glonass orbital and clocks corrections which are precise orbits and clocks given with respect to broadcast orbits and clocks (Caissy et al, 2012).

Generally GNSS permanent station time series show various types of signals, some of which are real whilst the others may not have apparent causes: miss-modelled errors, effects of observational environments, random noise or any other effects produced by GNSS analysis software or operator choices of software parameters and settings of a prior stochastic models for different types of measurements. However, IGS station solutions are basically given in two different ways (i) 3D coordinate time series which reflect the sum of all noises and signal; (ii) the covariance matrices of the stations derived from the estimation systems. It is challenging to extract detailed signals from the limited information. Significant efforts have been made to analyse GNSS time series, including the earlier studies by Mao et al (1999) about

noise in the GPS time-series and Blewitt & Lavallée (2002) on the effect of annual signals on geodetic velocity time series. Williams (2003) described the effect of coloured noise on the uncertainties of the rates estimated from the geodetic time series. Williams et al. (2004) reported significant spatial correlation between GPS time-series. Biagi et al (2006) studied the effects of tidal errors and deformations in regional GPS networks. In geophysical studies, in addition to global models of plate motions, it is widely accepted that the site velocities of permanent GPS stations are determined by a linear regression of individual GPS coordinate time-series. In the work by Amiri-Simkooei et al (2007), a method was used to assess the noise characteristics of univariate GPS coordinate time series. All these analyses were based on the univariate noise assessment for which the time series were estimated individually. In the recent years, multivariate analysis methods have been introduced to the analysis of noise in GNSS time series. Amiri-Simkooei et al (2009) adopted some stochastic models to assess the noise characteristics of multivariate time series. The least-squares variance component estimation (LS-VCE) was then applied to estimate full covariance matrices among different series. The analysis for five IGS station timer series confirmed that the spatial correlation between different stations for individual components is significant both for white and for colored noise components.

However, multivariate analysis of GNSS time series does not limit to the multivariate linear modelling, parameters and variance-covariance component estimation. Hypothesis testing is another important aspect of both univariate and multivariate analyses. Hypothesis testings answer the questions such as whether the signals or biases are statistically significant with respect to the level of noise in the background, whether the parameters selected are statistically significant enough to be included in the model. Statistic tools for such types of analysis are less studied, albeit many traditional statistics, such as univariate mean, standard deviations, spectrum analysis, have been used in the analysis of 3D position time series, such as the t-test procedures as outlined by Kouba (2009).

The rest of the paper is structured as follows. Section 2 gives a review of univariate linear models, regression estimation of state parameters and various testing statistics. Section 3 presents T-square testing statistics for use in multivariate time series. Section 4 provides experimental analysis for the results showing by example how the introduced test can be used to detect the significance of coordinate variations in the solutions. Section 5 summarised the research findings of this work and other potential applications of the testing statistics as concluding remarks.

2. Univariate Analysis of GPS State Time Series: Models, Estimation and Testing Statistics

Although univariate analysis of GPS time series has been reasonably discussed, a systematic review of the regression models, estimation and hypothesis testing problems are provided herein for a number of reasons: (1) univariate modelling and analysis should be used as a preliminary step to analyse the multivariate data; (2) comparison with the multivariate analysis can be made both theoretically in next section and numerically in the section 4; and (3) the procedures available for the studies of residuals, detection of outliers in univariate analysis of time series may be extendable to multivariate analysis.

2.1 Linear models and least square estimation

We consider an individual GNSS coordinate time series, for instance, the daily solution of one coordinate component of a station, having a function model generally expressed as a linear regression model

$$y_i = X_0 + a_{i,1}X_1 + a_{i,2}X_2 + ... + a_{i,p-1}X_{p-1} + e_i, \qquad (1)$$

where$\{y_i, i=1,..,n\}$ is the observable of the coordinate component at the data point i; $\{X_j, j=0,1,..., p-1\}$ are the p-by-1 vector to be estimated as regression coefficients; $\{a_{i,j,.}\ i=1,\ldots,\ n;\ j=0,1,..,\ p-1\}$ are the independent variables in the observation equation (1), which could be given the function of time, depending on the actual physical problems; and $\{e_i,i=1,..,n\}$ is the noise of the observable $\{y_i\}$. For instance, Willims (2003) gives the linear model to describe a component of coordinate time series as function of time:

$$\begin{aligned} &y(t) = X_0 + X_1 t + \\ &\sum_{k=2}^{s}[X_{(2k-1)}\cos(\omega_k t) + X_{(2k)}\sin(\omega_k t)] + e(t) \end{aligned} \qquad (2)$$

where 2s=p, ω_k represents different frequencies of the signals; t is the time variable given with respect to certain time epoch t_0.

Using vector-matrix symbols,

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_1 \end{bmatrix}, \mathbf{X} = \begin{bmatrix} X_0 \\ X_1 \\ \vdots \\ X_{p-1} \end{bmatrix}, \mathbf{A} = \begin{bmatrix} 1 & a_{1,1} & \cdots & a_{1,p-1} \\ 1 & a_{2,1} & \cdots & a_{2,p-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & a_{n,1} & \cdots & a_{n,p-1} \end{bmatrix},$$

Equation (1) is then rewritten as

$$\mathbf{Y = AX + e} \qquad (3a)$$

For GNSS daily state solutions obtained from different sets of measurements, it is straightforward we assume **e** as a white noise vector, the following statistical models

$$\mathrm{E}(\mathbf{e}) = \mathbf{0}, \qquad \mathrm{Cov}(\mathbf{e}) = \sigma^2 \mathbf{I_n} \tag{3b}$$

where I_n is the n-by-n unit matrix . For the statistic analysis, we assume that the noise vector has the multivariate normal distribution:

$$\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n) \tag{4a}$$

Or defining $\mathrm{E}(\mathbf{Y})=\boldsymbol{\mu}_\mathbf{y}$, the observation vector **Y** is distributed according to

$$\mathbf{Y} \sim \mathrm{N}(\mu_\mathbf{y}, \sigma^2 \mathbf{I_n}) \tag{4b}$$

The least squares solutions of the problems (3) are given as follows

$$\hat{\mathbf{X}} = (\mathbf{A}^\mathbf{T}\mathbf{A})^{-1}\mathbf{A}^\mathbf{T}\mathrm{Y} \tag{5}$$

$$\hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{A}\hat{\mathbf{X}})^\mathbf{T}(\mathbf{Y} - \mathbf{A}\hat{\mathbf{X}})}{\mathrm{n} - \mathrm{p}} \tag{6}$$

As a special case, when p=1, the equations (5) and (6) are reduced as

$$\hat{X}_0 = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{7}$$

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2 \tag{8}$$

2.2 Univariate hypotheses testing

After a linear regression model is obtained from the least-square estimation given above, the first question is whether the regression model has properly described the dependences of $\{y_i\}$ on the independent variables $\{a_{i,j}\}$. On one hand, we shall test the experimental models with the extensive real world data sets and give physical interpretation. On the other hand, we can perform statistical hypotheses tests, which may show how **A** is statistically significantly related to **Y**, and whether the dependence of **Y** on some specific variables, such as harmonic functions in (2), is statistically significant. In addition, one may question whether the measurements have any outliers. We outline the required equations and statists for univariate analysis, referring to Jabson (1992) and Yang (2006) in particular.

2.2.1 General formation of regression testing

In general, various hypothesis tests can be expressed as the test of the null hypothesis

$$\mathrm{H}_0\mathbf{:} \qquad \mathbf{HX}=\mathbf{d} \tag{9}$$

where **H** is a full-rank m-by-p matrix to generally represent various testing problems as specified in the later discussion where m≤p; **d** is the m-by-1 constant vector. Under the constrained equation (9), the least square estimate of the **X** is given as follows:

$$\hat{\mathbf{X}}_\mathbf{H} = \hat{\mathbf{X}} - (\mathbf{A}^\mathbf{T}\mathbf{A})^{-1}\mathbf{H}^\mathbf{T}[\mathbf{H}(\mathbf{A}^\mathbf{T}\mathbf{A})^{-1}\mathbf{H}^\mathbf{T}]^{-1}(\mathbf{H}\hat{\mathbf{X}} - \mathbf{d}) \tag{10}$$

Letting

$$\mathrm{SS}_\mathrm{e} = (\mathbf{Y} - \mathbf{A}\hat{\mathbf{X}})^\mathbf{T}(\mathbf{Y} - \mathbf{A}\hat{\mathbf{X}}) \tag{11}$$

$$\mathrm{SS}_\mathrm{He} = (\mathbf{Y} - \mathbf{A}\hat{\mathbf{X}}_\mathbf{H})^\mathbf{T}(\mathbf{Y} - \mathbf{A}\hat{\mathbf{X}}_\mathbf{H}) \tag{12}$$

The F statistic is obtained as

$$F = \frac{(SS_{He} - SS_e)/m}{SS_e/(n-p)} \sim F_{(m,n-p)} \tag{13}$$

For a given confidence level of α, the critical region for testing the null hypothesis (9) is

$$F \geq F_{(m,n-p)}(\alpha) \tag{14}$$

This is the interference for the linear function (9), which can be reduced to several special cases for different testing purposes.

2.2.2 One-sample testing for univariate mean shifts

The first special case is about testing for the hypothesis that some coefficients of the regression model are zeros. For instance,

$$\mathbf{H_0}\mathbf{:}\ X_1=X_2=\ldots=X_{p-1}=0 \tag{15}$$

This means in the null hypothesis (9), the matrix **H** is the following:

$$\mathbf{H}=[\mathbf{0}\ \mathbf{I}_{\mathbf{p}-1}] \text{ and } \mathbf{d}=\mathbf{0} \tag{16}$$

Directly substituting (16) into (10) and (12), we obtain the F-statistic (13) where m=p-1. If this null hypothesis test is accepted, the effects of the independent variable factors in the regression model of the time series are insignificant.

As a special case of (15), we can test the significance of the individual coefficients of the regression model.

$$\mathbf{H_0}\mathbf{:}\ \mathbf{X_i}=\mathbf{0} \qquad i=1, 2,\ldots,p-1 \tag{17}$$

There are p-1 individual matrices:

$$\mathbf{H}=[\,0\ 1\ 0,\ldots,\ 0],\ldots,\ \mathbf{H}=[0\ 0,\ldots,1] \tag{18}$$

Substituting the **H** matrix (18) one by one into (10), we obtain the F-statistic (13), which is however, equivalent to the t-statistic

$$t_i = \frac{\hat{X}_i}{\hat{\sigma}\sqrt{c_{ii}}} \sim t_{n-p} \tag{19}$$

where c_{ii} is the diagonal element of the matrix $\mathbf{(A^T A)^{-1}}$

Many types of GNSS state time series are repeated measurements. In this case, the regression model derived from pervious time epochs may be used as the known model for the current time epoch, if the effects of the residuals are insignificant. This is equivalent to test the hypothesis that the mean shit or bias of the new residuals is equal to zero or to a specific value as the population mean. It is noticed that p=1, and the matrix **A** is

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}^{\mathrm{T}}$$

In this case, p=1, m=1. The null hypothesis is

$$H_0: X_0 = \mu_y \tag{20}$$

where μ_y is the population mean. In this case, the matrix **H**=[1]. The F-statistic (13) is reduced

$$F = \frac{(\bar{y} - \mu_y)^2}{\hat{\sigma}^2 / n} \sim F_{1,n-1} \tag{21a}$$

(21a) can be replaced by the well-known t-statistic, which is the t-distribution with (n-1) degrees of freedom and sample standard deviation:

$$t = F^{\frac{1}{2}} = \frac{(\bar{y} - \mu_y)}{\hat{\sigma} / \sqrt{n}} \sim t(n-1) \tag{21b}$$

where $\bar{y}$ and $\hat{\sigma}$ are computed with (7) and (8), respectively.

2.2.3 Two-sample testing for univariate mean shifts

The next case is to test a two-sample problem. There are two independently observed/sampled data sets, for instance, daily GNSS station solutions over two different months or over two different seasons. The question is whether the regression models derived from monthly/quarterly data sets are identical. For the first set of data, we have the model:

$$\mathbf{Y_1 = A_1 X_1 + e_1}, \qquad \mathbf{e_1} \sim N(\mathbf{0}, \sigma^2 \mathbf{I_{n1}})$$

For the second the set we also have

$$\mathbf{Y_2 = A_2 X_2 + e_2}, \qquad \mathbf{e_2} \sim N(\mathbf{0}, \sigma^2 \mathbf{I_{n2}})$$

We combine the two equations and obtain

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y_1} \\ \mathbf{Y_2} \end{pmatrix}, \mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & 0 \\ 0 & \mathbf{A}_2 \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \mathbf{e} = \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{pmatrix} \tag{22a}$$

$$\begin{pmatrix} \mathbf{e_1} \\ \mathbf{e_2} \end{pmatrix} \sim N(\mathbf{0}, \sigma^2 \mathrm{I}_{n_1+n_2}) \tag{22b}$$

The hypothesis to be tested is given as

$$H_0: \qquad \mathbf{HX} = \begin{pmatrix} \mathbf{I_p} & -\mathbf{I_p} \end{pmatrix} \begin{pmatrix} \mathbf{X_1} \\ \mathbf{X_2} \end{pmatrix} = \mathbf{0} \tag{23}$$

Substituting **Y, A** in (22a) and **H** in (23) into (10), (11) and (12), the F-statistic is obtained as follows

$$F = \frac{(SS_{He} - SS_e)/p}{SS_e/(n_1 + n_2 - 2p)} \sim F_{(p,n1+n2-2p)} \tag{24}$$

The testing statistic can be applied to the two-sample problems. Suppose there are two separate data samples, which have two independent sets of used data points $\{ y_i \}$ and $\{ z_i \}$. If $z_i \sim N(\mu_z, \sigma_z^2)$ for all i=1,2,…n_z and $y_i \sim N(\mu_y, \sigma_y^2)$ for all i=1,2, …,n_y are independent with the same variances and we define

$$\bar{z} = \frac{1}{n_z}\sum_{i=1}^{n_z} z_i, \quad \bar{y} = \frac{1}{n_y}\sum_{i=1}^{n_y} y_i \tag{25}$$

which are distributed independently according to $\bar{z} \sim N(\mu_z, \sigma_z^2/n_z)$ and $\bar{y} \sim N(\mu_y, \sigma_y^2/n_y)$, respectively. The variance estimates are

$$\hat{\sigma}_z^2 = \frac{1}{n_z - 1}\sum_{i=1}^{n_z} (z_i - \bar{z})^2 \quad , \tag{26a}$$

$$\hat{\sigma}_y^2 = \frac{1}{n_y - 1}\sum (y_i - \bar{y})^2 \tag{26b}$$

The difference of the two sample means $(\bar{z} - \bar{y})$

$$t = \sqrt{\frac{n_z n_y}{n_z + n_y}} \frac{\bar{z} - \bar{y}}{\hat{\sigma}_{zy}} \tag{27}$$

where (26) can be written as

$$\hat{\sigma}_{zy} = \sqrt{\frac{(n_z - 1)\hat{\sigma}_z^2 + (n_y - 1)\hat{\sigma}_y^2}{n_z + n_y - 2}}$$

2.2.3 Testing for univariate outliers

Now, we detect potential measurement outliers. The detection of univariate outliers in GNSS time series is relatively straightforward in the sense that outliers are generally observations that are somewhat distant from

the reminder of the data. The problem is how to avoid the effects of outliers on the regression models. With one outlier on the *i*th measurement, the regression equation is now expressed as

$$\mathbf{Y} = \mathbf{AX} + \mathbf{d_i b} + \mathbf{e}, \quad \mathbf{e} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}) \tag{28}$$

where b is the outlier and

$$\mathbf{d}_i = (0 \quad \dots \quad 0 \quad 1 \quad 0 \quad \dots \quad 0)^{\mathrm{T}}$$

We are now to test the hypothesis that the bias is zero, i.e.,

$$H_0: b=0 \tag{29}$$

The test statistic for the *j*th outlier is

$$F_j = \frac{(n-p-1)r_j^2}{n-p-r_j^2} > F_{1,n-p-1}(\alpha) \tag{30}$$

where

$$r_j = \frac{y_i - \mathbf{A_i}\hat{\mathbf{X}}}{\hat{\sigma}\sqrt{1-p_{jj}}} \tag{31}$$

where p_{jj} is the *j*th diagonal elements of the matrix

$$\mathbf{P} = \mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}$$

In the special case where

$$\mathbf{A} = (1 \quad 1 \quad \cdots \quad 1)^{\mathrm{T}}$$

$$\hat{\mathrm{X}} = \bar{\mathrm{y}} = \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{y_i}$$

the variance estimate is

$$\hat{\sigma}^2 = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - \bar{\mathrm{y}})^2 \tag{32}$$

The standardized residual is given by

$$\mathrm{r_j} = \frac{\mathrm{y_i} - \bar{\mathrm{y}}}{\hat{\sigma}\sqrt{\frac{\mathrm{n}-1}{\mathrm{n}}}} = \frac{\mathrm{y_i} - \bar{\mathrm{y}}}{\sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y_i} - \bar{\mathrm{y}})^2}} \tag{33}$$

The F statistic is

$$\mathrm{F_j} = \frac{(\mathrm{n}-2)\mathrm{r_j^2}}{\mathrm{n}-1-\mathrm{r_j^2}} > \mathrm{F}_{1,\mathrm{n}-2}(\alpha) \tag{34}$$

To avoid the effects of outliers, define the mean values and the standard deviation of $\hat{\sigma}$ with the jth measurement deleted,

$$\bar{\mathrm{y}}_{(\mathrm{i})} = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{j}=1,\mathrm{j}\neq 1}^{\mathrm{n}} \mathrm{y_j}, \; \hat{\sigma}^2_{(\mathrm{i})} = \frac{1}{\mathrm{n}-2}\sum_{\mathrm{j}=1,\mathrm{j}\neq 1}^{\mathrm{n}} (\mathrm{y_j} - \bar{\mathrm{y}}_{(\mathrm{i})})^2 \tag{35a}$$

The standardized residual is given by

$$\mathrm{r}_{(\mathrm{i})} = \frac{\mathrm{y_i} - \bar{\mathrm{y}}_{(\mathrm{i})}}{\hat{\sigma}_{(\mathrm{i})}\sqrt{\frac{\mathrm{n}-2}{\mathrm{n}-1}}} = \frac{\mathrm{y_i} - \bar{\mathrm{y}}_{(\mathrm{i})}}{\sqrt{\frac{1}{\mathrm{n}-1}\sum_{\mathrm{j}=1,\mathrm{j}\neq \mathrm{i}}^{\mathrm{n}} (\mathrm{y_j} - \bar{\mathrm{y}}_{(\mathrm{i})})^2}} \tag{35b}$$

3. Multivariate Analysis of GNSS Time Series: Models, Estimation and Hypothesis Testing

3.1 Linear models and least square estimation

The linear regression models for the multivariate GNSS time series have been expressed in a number of existing works as follows (eg. Amiri-Simkooei, 2009)

$$y_{i,j} = X_{0,j} + a_{i,1}X_{1,j} + a_{i,2}X_{2,j} + \dots + a_{i,p-1}X_{p-1,j} + e_{i,j}, \tag{36}$$

where i=1,…n, j=1,…,q. for instance, for each component, there is a equation (2). All the coordinates have the same independent variables and different coefficients. Using the matrix notations:

$$\mathbf{Y} = \begin{pmatrix} \mathrm{y_{11}} & \mathrm{y_{12}} & \cdots & \mathrm{y_{1q}} \\ \mathrm{y_{21}} & \mathrm{y_{22}} & \cdots & \mathrm{y_{2q}} \\ \vdots & \vdots & & \vdots \\ \mathrm{y_{n1}} & \mathrm{y_{n12}} & \cdots & \mathrm{y_{n1q}} \end{pmatrix} = \begin{pmatrix} \mathbf{y_1} \\ \mathbf{y_2} \\ \vdots \\ \mathbf{y_n} \end{pmatrix} = (\mathbf{Y_1}, \mathbf{Y_2}, \dots, \mathbf{Y_q})$$

$$\mathbf{X} = \begin{pmatrix} \mathrm{X_{01}} & \mathrm{X_{02}} & \cdots & \mathrm{X_{0q}} \\ \mathrm{X_{11}} & \mathrm{X_{12}} & \cdots & \mathrm{X_{1q}} \\ \vdots & \vdots & & \vdots \\ \mathrm{X_{p-1,1}} & \mathrm{X_{p-1,2}} & \cdots & \mathrm{X_{p-1,q}} \end{pmatrix} = \begin{pmatrix} \mathbf{x_0} \\ \mathbf{x_1} \\ \\ \mathbf{x_{p-1}} \end{pmatrix} = (\mathbf{X_1}.\mathbf{X_2}\dots, \mathbf{X_q}.)$$

$$\mathbf{A} = \begin{pmatrix} 1 & \mathrm{a_{1,1}} & \cdots & \mathrm{a_{1,p-1}} \\ 1 & \mathrm{a_{2,1}} & \cdots & \mathrm{a_{2,p-1}} \\ \vdots & \vdots & & \vdots \\ 1 & \mathrm{a_{n,1}} & \cdots & \mathrm{a_{n,p-1}} \end{pmatrix},$$

$$\mathbf{e} = \begin{pmatrix} \mathrm{e_{11}} & \mathrm{e_{12}} & \cdots & \mathrm{e_{1q}} \\ \mathrm{e_{21}} & \mathrm{e_{22}} & \cdots & \mathrm{e_{2q}} \\ \vdots & \vdots & & \vdots \\ \mathrm{e_{n1}} & \mathrm{e_{n12}} & \cdots & \mathrm{e_{n1q}} \end{pmatrix} = (\mathbf{e_1}, \mathbf{e_2}, \dots, \mathbf{e_q})$$

the linear model is expressed as follow

$$\begin{aligned} &\mathbf{Y} = \mathbf{AX} + \mathbf{e} \\ &E(\mathbf{e_i}) = \mathbf{0}; \mathrm{Cov}(\mathbf{e_i}, \mathbf{e_j}) = \sigma_{i,j}\mathbf{I_n}; \mathrm{Cov}(\mathbf{e_i}, \mathbf{e_i}) = \sigma_i^2\mathbf{I_n} \end{aligned} \tag{37}$$

To obtain the estimate of the matrix **X,** the basic principle is to vectorise the multivariate linear model (37) to univariate linear model (Jobson, 1992)

$$\begin{aligned} &\mathrm{Vec}(\mathbf{Y}) = (\mathbf{I} \otimes \mathbf{A})\mathrm{Vec}(\mathbf{X}) + \mathrm{Vec}(\mathbf{e}) \\ &\mathrm{E}[\mathrm{Vec}(\mathbf{e})] = \mathbf{0} \\ &\mathrm{Cov}[\mathrm{Vec}(\mathbf{e})] = \boldsymbol{\Sigma} \otimes \mathbf{I} \end{aligned} \tag{38}$$

where $\boldsymbol{\Sigma}$ is the q-by-q variance matrix. Using least square estimation and the properties of Kronecker products denoted by $\otimes$, the vector of **X** is estimated as

$$\mathrm{Vec}(\tilde{\mathbf{X}}) = \mathrm{Vec}\left[(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{Y}\right] \tag{39}$$

The estimate of matrix X is given as follows

$$\tilde{\mathbf{X}} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{Y} \tag{40}$$

For each vector of **X**, we have

$$\tilde{\mathbf{X}}_{\mathbf{i}} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{Y}_{\mathbf{i}}$$

The covariance matrix of Vec(**X**)

$$\mathrm{Cov}[\mathrm{Vec}(\tilde{\mathbf{X}})] = \boldsymbol{\Sigma} \otimes (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1} \tag{41}$$

where the variance matrix $\boldsymbol{\Sigma}$ is estimated by

$$\tilde{\boldsymbol{\Sigma}} = \frac{\tilde{\mathbf{e}}^{\mathrm{T}}\tilde{\mathbf{e}}}{\mathrm{n}-\mathrm{p}} = \frac{\mathbf{Y}^{\mathrm{T}}[\mathbf{I} - \mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}]\mathbf{Y}}{\mathrm{n}-\mathrm{p}} \tag{42}$$

where

$$\tilde{\mathbf{e}} = \mathbf{Y} - \tilde{\mathbf{Y}} = [\mathbf{I} - \mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}]\mathbf{Y} \tag{43}$$

when p=1, the solutions (40) and (43) are reduced to

$$\tilde{\mathbf{x}}_{\mathbf{0}} = \bar{\mathbf{y}} = \frac{1}{\mathrm{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}\mathbf{y}_{\mathbf{i}} \tag{44}$$

and

$$\tilde{\boldsymbol{\Sigma}} = \frac{\mathbf{1}}{\mathrm{n}-\mathbf{1}}\sum_{\mathbf{i=1}}^{\mathrm{n}}(\mathbf{y}_{\mathbf{i}} - \bar{\mathbf{y}})^{\mathrm{T}}(\mathbf{y}_{\mathbf{i}} - \bar{\mathbf{y}}) \tag{45}$$

respectively.

Similarly, for the statistic analysis, we assume the normal distribution for the noise vector:

$$\mathrm{Vec}(\mathbf{Y}) \sim \mathrm{N}_{\mathrm{np}}[(\mathbf{I} \otimes \mathbf{A})\mathrm{Vec}(\mathbf{X}), \boldsymbol{\Sigma} \otimes \mathbf{I}] \tag{46}$$

The most important distinction between the sets of univariate regression and multivariate regression is that in the multivariate regression model there are nonzero correlation among noise terms from different multiple univariate regression models. If joint inferences are required involving two or more of the multiple regression models, these correlation must be taken into consideration. These join inference procedures are discussed in the next section

3.2 Hypotheses testings in multivariate time series

The cases of hypothesis testings in multivariate regression analysis are similar to these in univariate regression analysis as outlined in section 2.2. For instance, it may be useful to be able to test the null hypothesis that a subset of the columns of the matrix A is superfluous, or some specific variable vectors of X are statistically insignificantly related to Y. In case of the regression model is given, it is useful to test the significance of the effect of the noise, or the mean values of residuals are zero. In addition, one may question whether the measurements have any outliers.

3.2.1 General formation of testing for multivariate regression models

In general, various hypotheses tests can be expressed as the test of the null hypotheses

$$\mathrm{H}_{\mathbf{0}}: \quad \mathbf{HX=B} \tag{47}$$

where **H** is a m-by-p matrix of known constants of rank m and **B** is a m-y-q matrix of given constants. The restricted least squares estimator of **X** subject to **HX=B** is given by (eg. Jobson, 1992).

$$\tilde{\mathrm{X}}_{\mathrm{H}} = \tilde{\mathrm{X}} - (\mathrm{A}^{\mathrm{T}}\mathrm{A})^{-1}\mathrm{H}^{\mathrm{T}}[\mathrm{H}(\mathrm{A}^{\mathrm{T}}\mathrm{A})^{-1}\mathrm{H}^{\mathrm{T}}]^{-1}(\mathrm{H}\tilde{\mathrm{X}} - \mathrm{B}) \tag{48}$$

The likelihood ratio test of the hypothesis (47) is carried out using the Wilk's Lambda statistic

$$\Lambda = \frac{\left|(\mathbf{Y} - \mathbf{A}\tilde{\mathbf{X}})^{\mathrm{T}}(\mathbf{Y} - \mathbf{A}\tilde{\mathbf{X}})\right|}{\left|(\mathbf{Y} - \mathbf{A}\tilde{\mathbf{X}}_{\mathbf{H}})^{\mathrm{T}}(\mathbf{Y} - \mathbf{A}\tilde{\mathbf{X}}_{\mathbf{H}})\right|} \tag{49}$$

If the H_0 is true, in large samples the distribution of Λ is approximated by the statistic which has an F distribution with m_1 and m_2 degrees of freedom:

$$\frac{1-\Lambda^{1/\nu}}{\Lambda^{1/\nu}}\frac{m_2}{m_1} \sim F_{m1,m2} \tag{50}$$

where,

$$m_1 = mq \tag{51a}$$

$$m_2 = \nu[n - p - \frac{1}{2}(q - m + 1)] - \frac{qm}{2} + 1 \tag{51b}$$

$$\nu = \sqrt{\frac{\mathrm{q}^2\mathrm{m}^2 - 4}{\mathrm{q}^2 + \mathrm{m}^2 - 5}} \tag{52}$$

If q=1, or 2, or m=1, or 2, this F-distribution is exact. When p=1, v=1. m_1=m, m_2 =(n-p), the F-statistic (50) is reduced to (13).

Similarly, to test different hypotheses, we only need to define the matrices **H** and **B** of known constants. Specifically, one can test the hypothesis that some coefficients are zero.

$$\begin{aligned} \mathbf{H}_{(mxp)} &= \begin{pmatrix} 0 & \mathbf{I}_m \end{pmatrix} \\ \mathbf{B}_{mxq} &= \mathbf{0} \end{aligned} \tag{53}$$

This is to test the hypothesis that the last m variables are superfluous. Following the same procedure to define **H** and **B** matrices, it is also possible to testing the hypotheses that two identical multivariate regression models are the same using the F- statistic (50).

3.2.2 T-squared statistics for testing of multivariate mean vector

In analysis of repeated multivariate GNSS state time series, we can test the hypothesis that the mean vector of the residuals is equal to a specific vector. We proceed by using Hotelling's T^2 statistics (Hotelling, 1931, Bowker, 1960). The null hypothesis

$$H_0: \mathbf{x_0} = \boldsymbol{\mu}_y \tag{54}$$

The T-squared statistic reads as follows (Anderson (1984), p156):

$$\mathbf{T^2} = n(\bar{\mathbf{y}} - \boldsymbol{\mu}_y)\mathbf{S}_y^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}_y)^T = (\bar{\mathbf{y}} - \boldsymbol{\mu}_y)\mathbf{S}_{\bar{y}}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}_y)^T \tag{55}$$

where

$$\mathbf{S}_{\bar{y}} = \frac{\mathbf{1}}{n}\mathbf{S}_y \tag{56}$$

and

$$S_y = \tilde{\Sigma} = \frac{\mathbf{1}}{n-\mathbf{1}}\sum_{i=1}^{n} \mathbf{y_i} - \bar{\mathbf{y}})^T(\mathbf{y_i} - \bar{\mathbf{y}}) \tag{57}$$

The distribution for the above statistics under the null hypothesis is the central F-distribution with p and n-p degrees of freedoms as follows

$$T^2\frac{(n-p)}{p(n-1)} = \frac{n(n-p)}{p(n-1)}(\bar{\mathbf{y}} - \boldsymbol{\mu}_y)\mathbf{S}_y^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}_y) \sim F_{p,n-p} \tag{58}$$

The ellipsoid $T^2 = n(\bar{\mathbf{y}} - \boldsymbol{\mu}_y)\mathbf{S}_y^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}_y)$ provides a confidence ellipsoid for μ_y. The critical region for testing the null hypothesis $\mu = \mu_y$ is

$$T^2\frac{(n-p)}{p(n-1)} > F_{p,n-p}(\alpha) \tag{59}$$

3.2.3 T-squared statistics for testing of multivariate two-sample problems

Similarly we can consider two-sample problems, to test the null hypothesis that mean of one normal population is equal to the mean of other where the covariance matrices are assumed equal but unknown. The null hypothesis is expressed as

$$H_0: \boldsymbol{\mu}_z = \boldsymbol{\mu}_y \tag{60}$$

Using (45) as the definition of one sample mean, and the 1-by-q vector

$$\bar{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i \tag{61}$$

as the second sample mean, and

$$\mathbf{S_{xy}} = \frac{1}{n_z+n_y-2}[\sum_{i=1}^{n_z}(\mathbf{z_i} - \bar{\mathbf{z}})^T(\mathbf{z_i} - \bar{\mathbf{z}}) + \sum_{i=1}^{n_y}(\mathbf{y_i} - \bar{\mathbf{y}})^T(\mathbf{y_i} - \bar{\mathbf{y}})] \tag{62}$$

as the unbiased pooled covariance matrix estimate, then the two-sample T-squared statistic is

$$T^2 = \frac{n_z n_y}{n_z + n_y}(\mathbf{z} - \bar{\mathbf{y}})\mathbf{S}_{zy}^{-1}(\mathbf{z} - \bar{\mathbf{y}})^T \tag{63}$$

which can be related to the F-distribution by

$$\frac{n_z + n_y - p - 1}{p(n_z + n_y - 2)}T^2 \sim F(p, n_z + n_y - p - 1) \tag{64}$$

The critical region is

$$\frac{n_z + n_y - p - 1}{p(n_z + n_y - 2)}T^2 > F(p, n_z + n_y - p - 1)(\alpha) \tag{65}$$

3.2.4 T-squared statistics for testing of multivariate outliers

We now turn attention to testing for multivariate outliers. The procedures commonly used for detecting outliers in univariate and bivariate distributions should be used as a preliminary step to identifying potential outliers for multivariate data. However, it is possible for a case of a multivariate outlier not to be an outlier with respect to any one of the underlying univariate distributions, the detection of extreme measurements in multivariate distributions is more difficult.

Following Jobson (1992), we outline the outlier detection procedures based on Hotelling's T^2 statistic.

One way of detecting multivariate outliers is to measure the distance of each repeated measurement from the centre of the data using the Mahalanobis distance. Each sample $\mathbf{y_i}$ can be ordered or ranked in terms of its value of

$$m_i^2 = (\mathbf{y_i} - \bar{\mathbf{y}})\mathbf{S_y^{-1}}(\mathbf{y_i} - \bar{\mathbf{y}})^{\mathbf{T}} \quad (66)$$

which is the equation of p-dimensional ellipsoid. A relative large value of m_i^2 would indicate that $\mathbf{y_i}$ is potential outlier. In practice, m_i^2 is related to the measure

$$b_i^2 = (\mathbf{y_i} - \bar{\mathbf{y}}_{(i)})\mathbf{S}_{\mathbf{y(i)}}^{\mathbf{-1}}(\mathbf{y_i} - \bar{\mathbf{y}}_{(i)})^{\mathbf{T}}$$

where $\bar{\mathbf{y}}_{(i)}$ denotes the sample mean vector $\bar{\mathbf{y}}$ with $\mathbf{y_i}$ omitted, $\bar{\mathbf{y}}_{(i)} = \frac{\mathbf{1}}{\mathbf{n-1}} \sum_{\mathbf{j=1, j\neq i}}^{\mathbf{n}} \mathbf{y_j}$ and

$$\mathbf{S}_{\mathbf{y(i)}} = \frac{1}{n-2} \sum_{\mathbf{j=1, j\neq i}}^{\mathbf{n}} (\mathbf{y_j} - \bar{\mathbf{y}}_{(i)})^{\mathbf{T}} (\mathbf{y_j} - \bar{\mathbf{y}}_{(i)})$$

The relationship between m_i^2 and b_i^2 is given by

$$m_i^2 = (n-1)^3 b_i^2 / [n^2(n-2) + (n-1)b_i^2] \quad (67a)$$

which is rewritten as

Table 1 Summary of three hypothesis tests for both univariate and multivariate analysis

	Univariate	Multivariate
One-sample problem Test 1: H_0: $\mu = \mu_y$ with unknown variance or covariance matrix	$s_y = \frac{1}{n-1}\sum_{i=1}^{n}[y_i - \bar{y}]^2$ $t = \frac{(\bar{y} - \mu_y)}{s_y / \sqrt{n}} \sim t(n-1)$ Two-tailed test: Given the confidence value α, if P-value is below threshold, H_0 is rejected	$S = \frac{1}{n-1}\sum_{i=1}^{n}[y_i - \bar{y}][y_i - \bar{y}]^T$ $T^2 = n(\bar{y} - \mu_y)^T S^{-1}(\bar{y} - \mu_y)$ $T^2 \frac{(n-p)}{p(n-1)} = \frac{n(n-p)}{p(n-1)}(\bar{y} - \mu_y)^T S^{-1}(\bar{y} - \mu_y) \sim F_{p,n-p}$ Given the confidence value α, if $T^2 \frac{(n-p)}{p(n-1)} > F_{p,n-p}(\alpha)$ The H_0 is rejected
Two-sample problem Test 2: H_0: $\mu_x = \mu_y$ with unknown covariance matrix	$s_x^2 = \frac{1}{n_x - 1}\sum_{i=1}^{n}[x_i - \bar{x}]^2$ $s_y^2 = \frac{1}{n_y - 1}\sum_{i=1}^{n}[y_i - \bar{y}]^2$ $s_{xy} = \sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}}$ $t = \frac{(\bar{y} - \bar{x})}{s_{xy} / \sqrt{n}} \sim t(n_x + n_y - 2)$ Two-tailed test: Given confidence value α, if the p-value is below threshold, the H_0 is rejected	$S_{xy} = \frac{1}{n_x + n_y - 2}[\sum_{i=1}^{n_x}(x_i - \bar{x})(x_i - \bar{x})^T + \sum_{i=1}^{n_y}(y_i - \bar{y})(y_i - \bar{y})^T]$ $T^2 = \frac{n_x n_y}{n_x + n_y}(\bar{y} - \mu_y)^T S^{-1}(\bar{y} - \mu_y)$ $\frac{n_x + n_y - p - 1}{p(n_x + n_y - 2)} T^2 \sim F(p, n_x + n_y - p - 1)$ Given confidence value α, if $\frac{n_x + n_y - p - 1}{p(n_x + n_y - 2)} T^2 > F(p, n_x + n_y - p - 1)(\alpha)$ the H_0 is rejected
Outlier detection problem In the linear mode $\mathbf{Y = AX + d_i b + e}$ $\mathbf{d}_i = (0 \ \ldots \ 0 \ 1 \ 0 \ \ldots \ 0)^T$ Test 3: H_0: b=0	$r_j = \frac{y_i - \bar{y}}{\hat{\sigma}\sqrt{\frac{n-1}{n}}}$ $F_j = \frac{(n-2)r_j^2}{n-1-r_j^2} \sim F_{1,n-p-1}$ Given confidence value α, if $\frac{(n-2)r_j^2}{n-1-r_j^2} > F_{1,n-p-1}(\alpha)$ the H_0 is rejected	$m_i^2 = (\mathbf{y_i} - \bar{\mathbf{y}})\mathbf{S_y^{-1}}(\mathbf{y_i} - \bar{\mathbf{y}})^{\mathbf{T}}$ $b_i^2 = (\mathbf{y_i} - \bar{\mathbf{y}}_{(i)})\mathbf{S}_{\mathbf{y(i)}}^{\mathbf{-1}}(\mathbf{y_i} - \bar{\mathbf{y}}_{(i)})^{\mathbf{T}}$ $T_i^2 = \frac{n-1}{n} b_i^2$ $\frac{(n-p-1)}{(n-2)p} T_{max}^2 \sim F_{p,n-p}$ Given confidence value α, if $\frac{(n-p-1)}{(n-2)p} T_{max}^2 \geq F_{p,n-p}(\alpha)$ then H_0 is rejected

$$b_i^2 = m_i^2 n^2 (n-2)/[(n-1)^3 - m_i^2 (n-1)] \qquad (67b)$$

Ordering based on m_i^2 is therefore equivalent to ordering based on b_i^2. An equivalent procedure is to compute the ratio of the variance

$$\gamma_i^2 = \frac{|\mathbf{S}_{y(i)}|}{|\mathbf{S}_y|} = 1 - n m_i^2/(n-1) \qquad (68)$$

A relative small value of γ_i^2 would indicate that xi is a potential outlier. The methods (68) and (67 of ordering is equivalent.

Under the multivariate normality and the null hypothesis that $\mathbf{y_i} \sim N_p(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}), i = 1,...,n$. The Hotelling T^2 statistic is given as

$$T_i^2 = \frac{n-1}{n} b_i^2 \qquad (69)$$

The largest value of T_i^2 over the sample, $T_{\max}^2$, is used to test for the presence of a single outlier. From (58), we form the F-statistic as follows

$$\frac{(n-p-1)}{(n-2)p} T_{\max}^2 \sim F_{p,n-p} \qquad (70)$$

Considering the relationship between b_i^2 and m_i^2, the F-statistic can also be written as

$$\frac{m_i^2 (n-p-1)}{p[(n-1) - m_i^2]} \sim F_{p,n-p} \qquad (71)$$

The identification of a subset of outliers is a more difficult problem. But, the F-test statistic based on the Mahalanobis distance given as above can be used to detect multiple outliers. The idea is to begin with the entire sample, the data point yielding the largest value of m_i^2 is removed from the sample if the corresponding F-statistic is considered significant. The value of m_i^2 are then recomputed and a new maximum value of m_i^2 is compared to F. The procedures will be demonstrated in the next section. In addition, the measure γ_i^2 introduced above for single outliers can be extended for multiple outliers. We denote the covariance matrix by $\mathbf{S}_{(I)}$ with the t observations $\mathbf{y}_{i_1}, \mathbf{y}_{i_2}, ..., \mathbf{y}_{i_t}$ removed, where $\mathbf{I}$ denotes the vector of subscripts $(i_1, i_2, \ldots, i_t)$. The critical ratio is given by $\gamma_i^2 = |\mathbf{S}_{y(I)}|/|\mathbf{S}_y|$. A subset of observations with a smaller γ_i^2 is a indication that outliers may be present.

For the sake of convenience, Table 1 provides a summary of test statistics for three special testing problems in univariate and multivariate analysis: one-sample problem, two-sample problem and outlier detection problem.

4. Numerical Analysis of IGS Daily Station Solutions

The IGS community has been generating daily station state solutions for hundreds of CORS stations worldwide since their installations, or their data sets have been reprocessed with the more advanced software editions. Data analysis to the station time series aims to extract useful signals, such as crustal deformation, seasonal variations of station dynamics etc. Essentially, knowing the station dynamics has to rely on physical knowledge. Suitable statistical methods, such multivariate data analysis methods, are not intended to replace physical analysis: these should be seen as complementary, and statistical methods can effectively be used to run a prior investigations, to sort out ideas, to put a new light on a problem, or to point out aspects which would not come out in a classical approach. Physical analysis is necessary subsequently to refine and interpret the results. Alternatively, the statistical analysis may be run as a posterior investigation, to give ideas whether the physical models have effectively extracted the dynamic information, or to detect the significance of effects of residual signals. The results may be useful to refine the physical analysis subsequently.

4.1 Daily state time series and correlation

The daily solutions of the IGS site COCO since mid 1996 were obtained with permission from SOPAC for analysis in this paper. Figure 1 plots the East-North-Up (ENU) coordinates biases with respect to ITRF2005 against the modelling results. Figure 2 shows the ENU coordinate residuals after removing the modeled values from the ENU daily solutions. We now perform the testing in both 1 D and 3 D coordinate domains for the residuals between the observed and modeled sequences. At the beginning, we examine the correlation coefficients of three residual components against the simulated three white noise time series, as shown in Figure 3. Their correlation coefficients mostly vary between ± 0.2 and ± ~0.4, while the white noise correlation coefficients fall within ±0.1~0.2, as a comparison.

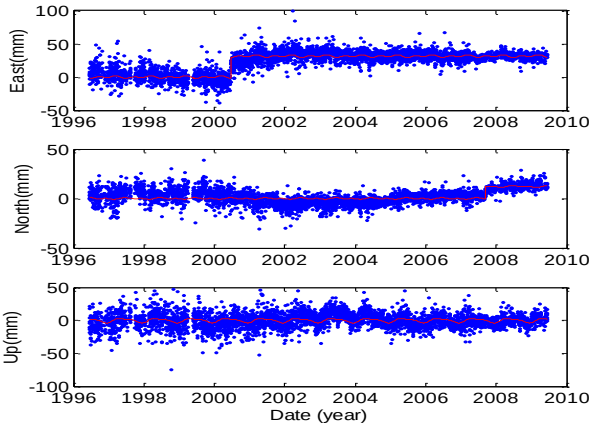

Figure 1: Illustration of the IGS station COCO daily solutions plotted against the physical models that reflect the half-yearly variations of the station solutions and solution jumps

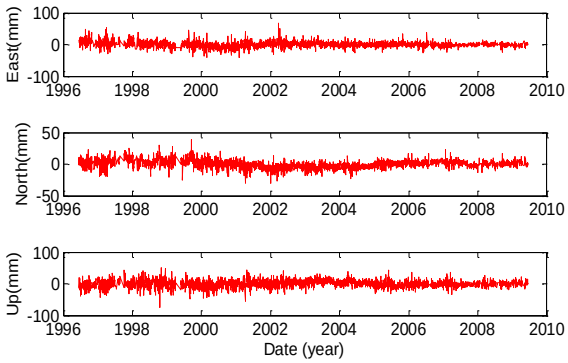

Figure 2: Illustrations of residuals of ENU components after removing the modelled values

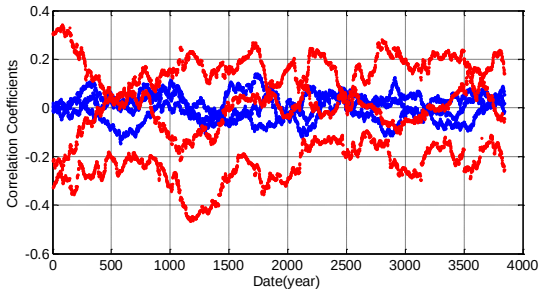

Figure 3: The correlation coefficients of three residual components against the simulated three white noise time sequences, showing the correlation between ENU components do exist.

4.2 Testing results for mean shits: univariate vs multivariate analysis

Next, using the statistics listed in Table 1 for both univariate and multivariate analysis, we perform Test 1 for the sample sizes of n=30 and n=91, respectively. For n=30, Figure 4 plots the t-statistic values in ENU components against the t-critical value at the confidence level of 0.001. For different components, their t-tests are rejected at different data points. Figure 5 shows the F-statistics derived from T-squared values of the 3D ENU time series against the F-statistics derived from 3 white noise time series, and the F-critical value. The rejected sample points decided by the multivariate tests are mostly different from the rejected points decided by the three individual t-tests. For n=91, univariate t-statistics and multivariate F-statistics are shown in Figure 6 and Figure 7 respectively. It appears that the rejected sample points decided by the multivariate tests are mostly the same as the rejected points decided by the univariate t-tests for each component. In general, testing for multivariate mean shifts tends to reject less data samples than testing for univariate mean shifts under the same confidence level.

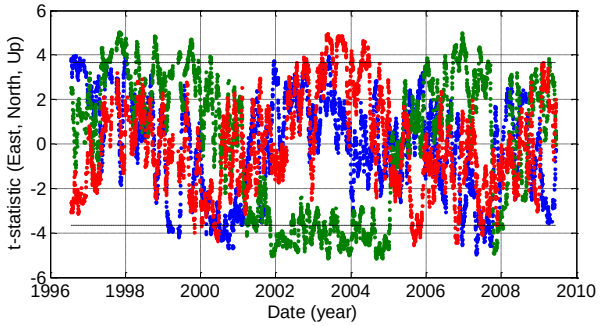

Figure 4: The t-statistic values in ENU components against their critical value (black), respectively, at the confidence level of 0.001. The sample size is n=30.

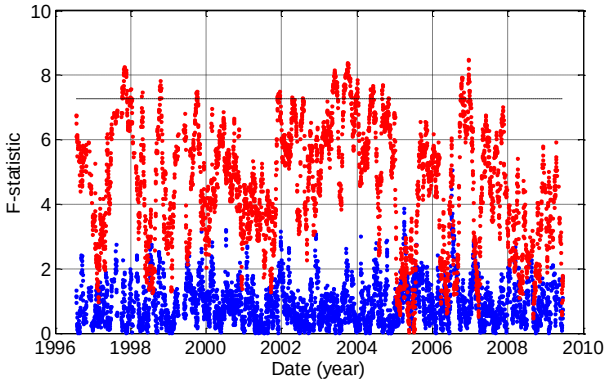

Figure 5: F-statistics (red) derived from T-squared values oof the 3D ENU time series against the F-statistics (blue) derived from 3 white noise time series, and the F-critical value (black). The sample size n is 30.

The results from Test 1 are summarized in Table 2. For Test 1, the mean covariance matrix over all the samples is used as the known covariance matrix. From Table 2, one may observe that the smaller the data sample size, the less the data samples are rejected in both tests for the hypothesis that the mean vector is equal to the given mean vector. This simply indicates that the assumption that the physical model established already fits into the

samples of daily coordinate solutions is more acceptable over shorter periods. Over longer terms, such as quarterly, the same assumption would be less acceptable.

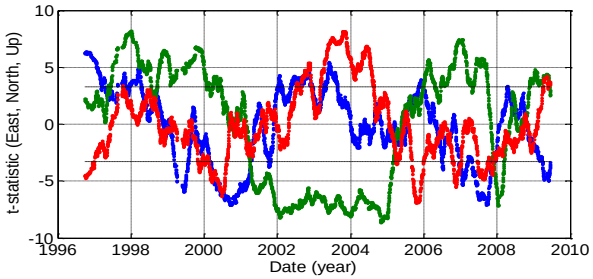

Figure 6: The t-statistic values in ENU components against their critical value (black), respectively, at the confidence level of 0.001. The sample size is n=91.

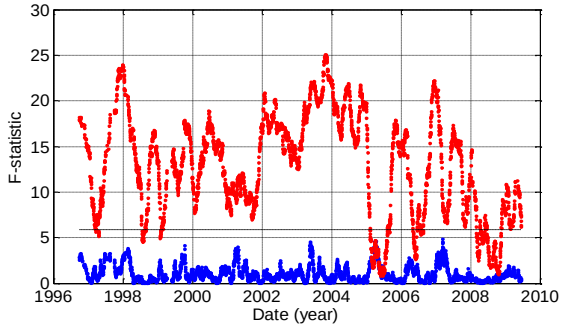

Figure 7: F-statistics (red) derived from the T-squared values of the 3D ENU time series against the F statistics (blue) derived from 3 white noise time series, and the F-critical value (black) for the confidence level of 0.001 and degrees of freedom (3,89) is 5.91.

Table 2. Comparison of Test 1 results between univariate and multivariate analyses with the sample sizes of 30 and 91, respectively

Scheme	Multivariate ENU		Univariate			Univariate		
			E	N	U	E	N	U
	$F_{p,n-p}$	$F_{p,n-p}$	t(n-1)	t(n-1)	t(n-1)	t(n-1)	t(n-1)	t(n-1)
p	3	3	3	3	3	3	3	3
n	30	91	30	30	30	91	91	91
DOF	27	88	29	29	29	90	90	90
α=0.001	7.27	5.91	±3.66	±3.66	±3.66	±3.29	±3.29	±3.29
Reject Rate %	7.06	88.22	8.07	32.32	9.53	36.35	70.87	31.12
			48.06			86.76		

4.3 Testing results for univariate and multivariate outliers

Similar results can be obtained from analysis of Test 2. To avoid repeating, we proceed to Test 3 directly, to detect outliers from based on univariate analysis from each ENU series individually and multivariate analysis from the ENU series together. Figure 8 shows the F statistics from the univariate analysis for each ENU component respectively, over the period from the day 239 to day 330 in 1998. The F-critical value for the confidence level of 0.01 is 6.93. Four data points were identified with outliers, 1 in each of the E and N components and 2 in the Up component. For the same data period, the results of multivariate analysis for the ENU components are shown in Figure 9. The F-critical value for the confidence level of 0.01 is 4.00. Only two data points were identified with outliers. This example shows that different testing conclusions were drawn using univariate and multivariate outlier testing statistics with the same data and under the same confidence level. In general, testing for multivariate outliers tends to reject less data samples than testing for univariate outliers under the same confidence level.

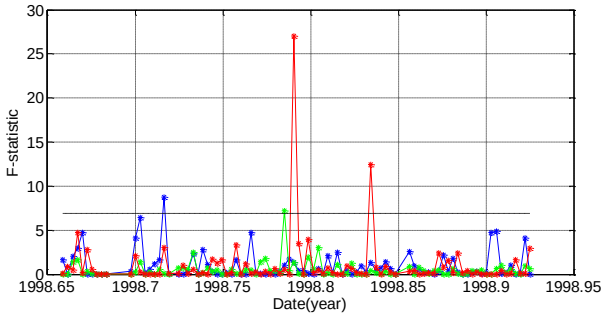

Figure 8: Univariate F-statistics for ENU components respectively, over the epochs from 600 to 690. The F-critical value for the confidence level of 0.01 is 6.93. Four data points were identified with outliers, 1 in each of E and N components and 2 in the Up component.

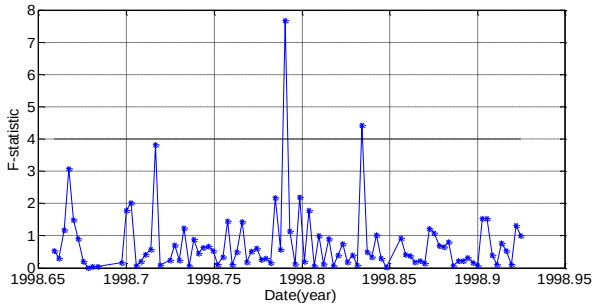

Figure 9: Multivariate F-statistics for ENU over the epochs from 600 to 690. The F-critical value for the confidence level of 0.01 is 4.00. Two data points were identified with outliers

5. Concluding Remarks

To analyse various state vector time series from GPS observation systems in the coordinate domain, such as 3 dimensional (3D) IGS station daily and weekly combined solutions, epoch by epoch real time kinematic positions, this paper has presented a number of T-squared statistics from the context of multivariate

analysis for use in the analysis of 3D GNSS station time series. Based on these T-squared statistics, F testing statistics for multivariate one-sample problem, two-sample problems and outliers are provided in comparison with test statistics for the same univariate analysis problems. These test problems are considered as the multivariate generalisation of univariate testing problems. T-squared statistics have taken the correlation between coordinate components into account, which is neglected in the univariate analysis.

A multi-year time series of an IGS station, COCO, has been analysed using some of the proposed tests to detect possible 3D mean shifts and outliers in the 3D residuals obtained by removing the modeled values from the observed coordinates. The man shits reflect the unmodelled biases in the residuals. The results have shown that in general, testing for multivariate mean shifts tends to reject less data samples than testing for univariate mean shifts under the same confidence level**.** Similarly, testing for multivariate outliers tends to reject less data samples than testing for univariate outliers under the same confidence level. Both the univariate and multivariate tests have shown that the assumption that the physical model established already fits into the samples of daily coordinate solutions is more acceptable from the shorter period perspective. From the longer term perspective, such as a quarter instead of a month, this assumption would be much less assumption.

It must be noted that data analysis for the station time series aims to extract useful signals, such as crustal deformation, seasonal variations of station dynamics etc. Essentially, knowing the station dynamics relies on physical knowledge. Suitable statistical methods, such multivariate data analysis methods, cannot replace physical analysis: these should be seen as complementary and statistical methods that can effectively be used to run a prior investigation, to sort out ideas, to put a new light on a problem, or to point out aspects which would not come out in a classical approach. Physical analysis is necessary subsequently to refine and interpret the results. Alternatively, the statistical analysis may be run as a posterior investigation, to give ideas whether the physical models have effectively extracted the dynamic information, or to detect the significance of effects of residual signals. The results may be useful to refine the physical analysis subsequently.

Acknowledgements

This work was carried out under the support of Cooperative Research Centre for Spatial Information (2010-2018) through the project 1.03 “Regionally enhanced orbits and clocks to support multi-GNSS real-time positioning”. The author also acknowledges the comments from the reviewers which have been beneficial for the improvement of the paper.

References

Altamimi, Z X. Collilieux (2008), ***IGS contribution to the ITRF***, Journal of Geodesy, Vol. 83, No. 3. (1 March 2008), pp. 375-383.

Amiri-Simkooei, A. (2009), ***Noise in multivariate GPS position time-series***, J. Geodesy, 83(2), 175-187

Anderson, T.W (2003) ***An Introduction to Multivariate Statistic Analysis***, Willey Seriers in probability and mathematical statistics, Jonhn Wiley and Sons, 2003, 752 pages.

Borghi, A., A. Aoudia, R.E.M Riva, R. Barzaghi (2009) ***GPS monitoring and earthquake prediction: A success story towards a useful integration***, Tectonophysics, Volume 465, Issues 1-4, 20 February 2009, Pages 177-189.

Caissy M, L Agrotis, G Weber, M Hernandez-Pajares, and U Hugentoble(2012), ***Coming soon: the Coming soon: the international GNSS real-time service***, *GPS World*, June 2012, pp. 52-8.

Ferland R, M. Piraszewski, (2009) ***The IGS-combined station coordinates, earth rotation parameters and apparent geocenter***, J Geod (2009) 83:385–392

Ferland R (2006) ***Proposed IGS05 Realization***, in IGS Mail. http://www.igs.org/mail/igsmail/2006/msg00170.html

Gnanadesikan, R, (1977), ***Methods of Statistic Data Analysis of Multivariate Observations***. Wiley, New York.

Hotelling. H (1931), ***The generalisation of Student’s ratio***, Annals of Mathematical Statistics, (2) 360-378.

Jobson J D (1992) ***Applied Multivariate Data Analysis: Regression and experimental design***, Volume 2, Springer.

Kouba, J (2009) ***A guide to using International GNSS Service (IGS) Products*** http://igscb.jpl.nasa.gov/igscb/resource/pubs/UsingIGSProductsVer21.pdf.

Mardia, K. V (1980) ***Tests of Univariate and Multivariate Normality***, Handbook of Statistics 1: Analysis of Variance (edited by P R Krishnaiah): p279-320. North-Holland

Nikolaidis, R (2002), ***Observation of Geodetic and Seismic Deformation with the Global Positioning System***, Ph.D Thesis, University of California, San Diego, 2002.

Williams, S. D. P (2003) ***The effect of coloured noise on the uncertainties of rates estimated from geodetic time serie***s, J. Geodesy, 76 (9-10), 483-494

Williams, S. D. P., Y. Bock, P. Fang, P. Jamason, R. M. Nikolaidis, L. Prawirodirdjo, M. Miller, and D. J. Johnson (2004) ***Error analysis of continuous GPS position time series***, J. Geophys. Res., 109, B03412, doi:10.1029/2003JB002741.

Yang, Y X (2006) ***Adaptive Navigation and Kinematic Positioning***, Publishing House of Surveying and Mapping, 234 pages.

Biography

Yanming Feng received his PhD degree in satellite geodesy from Wuhan Technical University of Surveying and Mapping (merged into Wuhan University in 2000), China. He is currently a Professor in Global Navigation Satellite Systems at School of Electrical Engineering and Computer Science, Queensland University of Technology, Australia. He has served as a project leader within Cooperative Research Centre for Spatial Information and Cooperative Research Centre for Automotive Technologies. His active research interests have included satellite orbit determination, wide area GNSS positioning, GNSS integrity determination, multiple GNSS data processing algorithms, and Dedicated Short-Range Communications for road safety applications. He is the Editor-in-Chief for The Journal of Global Positioning Systems.