Some Observations on Assessment of Gaussianity for Correlated Profiles

M. Uzielli[1], M.ASCE and K.-K. Phoon[2], M.ASCE

[1] Independent consultant, Piazza degli Unganelli, 3; 50125 Firenze, Italy; PH +39-328-6549144; FAX +39-055-2341232; email: marco@marcouzi.net
[2] Associate Professor, Department of Civil Engineering, National University of Singapore; Blk E1A, #07-03, 1 Engineering Drive 2; Singapore 117576; PH +65-68746783; FAX +65-67791635; email: cvepkk@nus.edu.sg

Abstract
The classical Anderson-Darling test for Gaussianity cannot be applied on correlated data. It is possible to circumvent correlation by resampling the data at wider spacing. A simple resampling criterion is developed using digitally simulated standard Gaussian realizations following a single exponential correlation structure. The effects of resampling on Gaussianity assessment are assessed on actual normalized cone tip resistance profiles from highly homogeneous soil units.

Introduction

The spatial variations of soil measurements in any direction are generally correlated with each other. This is most readily observed in the case of cone penetration test soundings, because of the small measurement intervals. Statistical inference from correlated data is very complicated. Invariably, classical tests are based on the important assumption that the data are independent. When they are applied indiscriminately to correlated data, large bias will appear in the evaluation of the test statistics. Phoon et al. (2003) demonstrated this bias quantitatively for the case of the Bartlett test. This paper shows that correlation introduces significant bias, even when the object of interest is the marginal information, such as the ubiquitous Gaussian goodness-of-fit test. Intuitively, a simple method to circumvent this correlation problem is to resample the data at wider spacings. However, the power of a test reduces with sample size. Hence, the practical challenge is to determine the minimum spacing beyond which a classical goodness-of-fit test can be applied. A simple resampling criterion would be developed based on simulation.

Simulation of Correlated Standard Gaussian Profiles

Realizations belonging to a zero-mean stationary Gaussian process X(t) can be generated using the well established spectral approach as follows:

$$X(t) = \sum_{k=1}^{M} \sigma_k (U_k \sin 2\pi f_k t + V_k \cos 2\pi f_k t) \tag{1}$$

in which $\sigma_k = \sqrt{2S(f_k)\Delta f}$, Δf is the interval over which S(f) is discretized, f_k = (2k-1)Δf/2, and U_k and V_k are standard normal variates. The following target autocorrelation function is selected:

$$R(\tau) = \exp(-2|\tau|/\delta) \tag{2}$$

in which τ is the spatial separation distance and δ is the scale of fluctuation, in the same units as τ. This single exponential autocorrelation function (SNX) is quite commonly used to model spatial variability in geostatistics. A first-order Markov or AR(1) process would also produce this autocorrelation function. Fig. 1 shows examples of simulated profiles for $k_\delta = \delta/(\Delta z)$=10 and 40, in which Δz=measurement interval. If the random seed used in Eq. (1) is

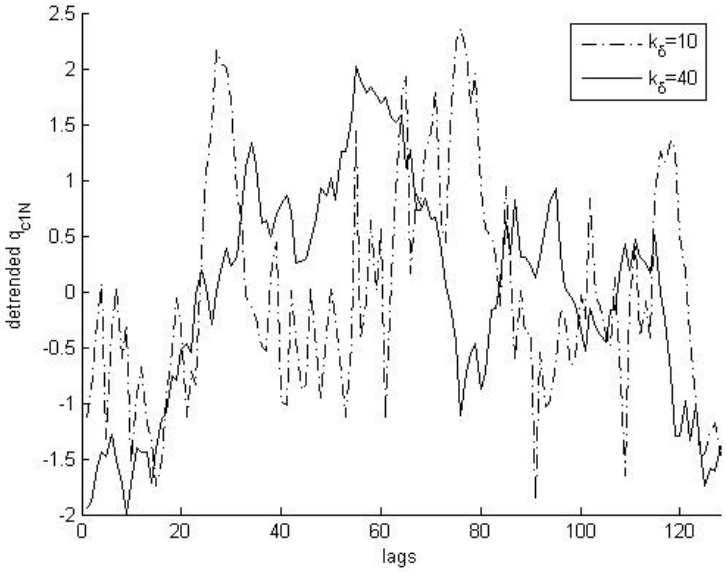

Figure 1. Examples of simulated profiles of detrended q_{c1N} for k_δ=10 and k_δ=40

maintained constant (as in the present study), it can be shown that the same dimensionless parameter k_δ (i.e. the number of measurements in one scale of fluctuation) would produce the same set of realizations.

Resampling

The aim of resampling in the context of the present paper is to remove the correlation in residuals of q_{c1N} by increasing the separation distance between consecutive data values. The physical motivation is that it will be less likely for measurements to be related if their separation distance is greater. However, it is desirable to limit the resampling interval to a minimum, as resampling results in a reduced sample size. The resampling interval can be conveniently expressed as a multiple of the scale of fluctuation, i.e. $k_n\delta$.

The scale of fluctuation is a concise indicator of the spatial extent of strong correlation. Hence, if data are resampled at a separation distance corresponding to δ (i.e. k_n=1), they should be uncorrelated. As the meaning of "strong" is not unequivocally defined and possibly a function of context, it is of pragmatic interest to see whether data resampled at smaller separation distances (i.e. k_n<1) could be considered "uncorrelated" from the viewpoint of a classical goodness-of-fit test. This would reduce the test bias caused by correlation while minimizing degradation in test power caused by inevitable reduction in resampled sample size.

Based on results obtained by Uzielli et al. (2005), typical values of k_δ for normalized cone penetration test parameters range from 10 to 40. Sets of 10000 realizations of different lengths (n_d=64, 128, 512 and 1024 data points) were generated for k_δ =10, 20, 30, 40. All realizations were resampled at 10 values of separation distances $k_n\delta$, with k_n increasing from 0.1 to 1.0 in steps of 0.1.

Normality Testing for Optimum Resampling

If data are uncorrelated and Gaussian, a normality test should reject the null hypothesis of Gaussianity at a rate equal to the prescribed level of significance (say 0.05). If, on the contrary, the Gaussian data are correlated, the null hypothesis would be rejected too frequently. The basic idea is to resample until the resulting data achieve the correct rejection rate.

The Anderson-Darling test [AD] (Anderson and Darling 1954) as modified by Stephens (1986) consists of the test statistic A^{2*} and the subsequent comparison with a critical value for a given level of significance. The AD test was performed on resampled realizations for each combination of k_δ and k_n. The ratio of rejections of the null hypothesis of normality to the total number of realizations (10000) was recorded.

The optimum value of the resampling parameter k_n was taken as the minimum value after which the rejection ratio for the resampled correlated simulations was definitively below 0.05, the confidence level adopted in the AD test. Fig. 2 shows the rejection ratios obtained for k_δ = 10, 20, 30, 40 and for k_n=0.1 to 1.0 in steps of 0.1 (only the range 0.1-0.6 is shown as the rejection rate is below 0.05 for k_n>0.6).

Fig. 3 shows the optimum resampling coefficient, k_n*, as a function of the dimensionless parameter $I_1=n_d/k_\delta$ (Phoon et al. 2003), which expresses the number of scales of fluctuation in the spatial extension of a realization (or, in case of real data, the width of a soil layer under investigation). Two distinct zones are observed in Fig. 3: for small values of I_1, k_n* increases with increasing I_1; for larger values, k_n* is essentially independent of I_1. The proposed relations for k_n*, obtained by empirical fitting, are

$$k_n^* = 0.118\ln(I_1) + 0.100 \qquad 1.5 \le I_1 \le 15 \tag{3a}$$

$$k_n^* = 0.42 \qquad 15 < I_1 \le 100 \tag{3b}$$

Curves deriving from Eq. (3a) and Eq. (3b) are shown in Fig. 3, superimposed to the rejection rate data obtained from the AD test. Hence, for data whose autocorrelation structure is well described by the single exponential model, correlation may be sufficiently reduced, insofar as the AD test is concerned, by resampling data using an interval of $k_n^* \cdot \delta$ (rounded up to the nearest multiple of Δz). It is also interesting to observe that Gaussianity is rejected too frequently when $k_n < k_n$*, clearly indicating that indiscriminate application of classical tests would lead to erroneous conclusions.

The AD test was also performed on control sets of uncorrelated standard Gaussian data of the same length as the correlated resampled simulations (i.e. varying according to k_n). The rejection ratio for such sets was always constant around 0.045.

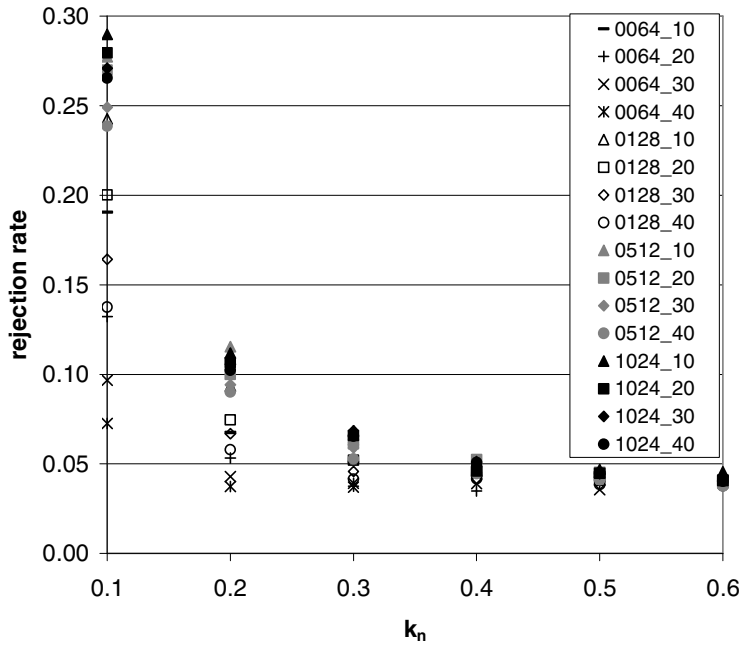

Figure 2. Rejection ratios (for sets of 10000 simulated realizations) for the estimation of the optimum resampling coefficient k_n^*. Sets are denoted by XXXX_YY where XXXX=number of elements in each realization; YY=k_δ.

Application to Real Data

To apply the procedure to a real data set: (a) obtain a zero-mean random field by subtracting a deterministic function [e.g. a linear trend]; (b) calculate the sample autocorrelation function [ACF] of the residuals; (c) evaluate the goodness of fit of the SNX model to the ACF by regression [e.g. verify that $R^2 \ge 0.9$]; (d) evaluate the scale of fluctuation; (e) calculate k_δ and check that $10 \le k_\delta \le 40$; (f) calculate I_1 and check that $1.50 \le I_1 \le 100$; (g) resample the data using k_n calculated from Eq. (3a) or (3b). The paper by Uzielli et al. (2005) can be referred to for detailed description of steps (a)-(d). Once correlation has been removed, it is possible to apply the classical AD test for the assessment of Gaussianity.

The criterion was applied to one profile of normalized cone tip resistance q_{c1N} (calculated according to the revised Robertson [1999] algorithm) whose SNX autocorrelation structure and weak stationarity of residuals were assessed rigorously by Uzielli et al. (2005). The profile comprises n_d=555 measurements, taken at interval Δz =0.025m.

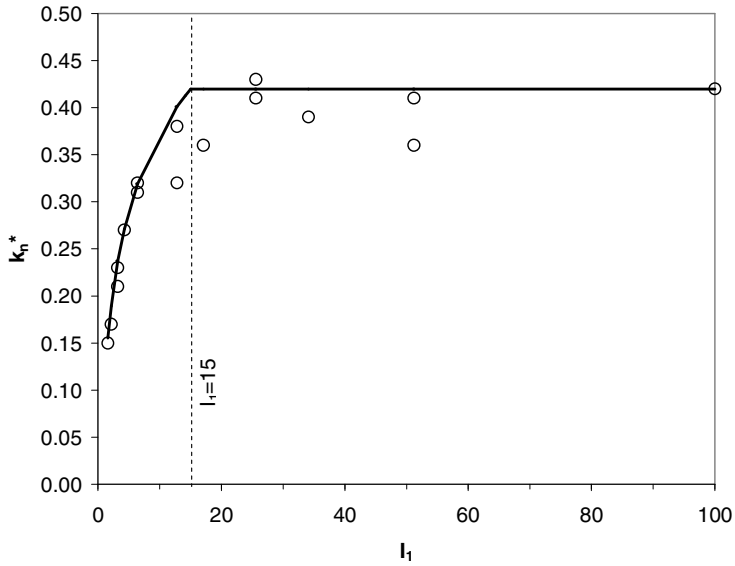

Figure 3. Optimum values of the resampling coefficient k_n^* versus I_1, with proposed curves for practical application

A linear trend is removed from the original data set. The fit of the SNX model to the initial part of the autocorrelation function whose coefficients exceed Bartlett's limit $r_b=1.96/(n_d^{0.5})$ is given by the determination coefficient $R^2=0.99$; the resulting scale of fluctuation is $\delta = 0.98$m. Hence, $I_1=(555\cdot0.025)/0.98=14.16$; Eq. (3a) yields $k_n^*=0.41$. The resampling thus should occur at a separation distance at least $0.41\delta=0.40$m, i.e. every 16 measurements. The original set of residuals [NONRES] and 5 of the possible 16 resampled sets [RES1–RES5] are shown in Fig. 4. Sets RES1-RES4 contain 35 measurements, while RES5 contains 34 measurements. Note the drastic reduction in sample size from 555 measurements to 35 measurements. If $k_n=1$, then the separation distance is 0.98m, i.e. every 40 points. The resulting resampled sample only consists of a meagre 14 measurements!

To evaluate the effect of resampling in terms of Gaussianity, the AD test is performed on the non-resampled set and RES1-RES5. The variation in the assessment in Gaussianity by the AD test due to the reduction of correlation is reported in Table 1. The critical value of the A statistic for a significance level of 0.05 is 0.787; it is seen that applying the AD test to NONRES provides a biased assessment of Gaussianity, as the A^{2*} statistic is very close to the critical value. The A^{2*} statistics for [RES1-RES5] are significantly smaller, ranging from 0.168 to 0.290. Such values show that in general the differences in the A^{2*} statistic between resampled sets are small; hence, in practice, it is not paramount to perform the AD test on more than one resampled set. Fig. 5 shows normal probability plots for NONRES and RES1.

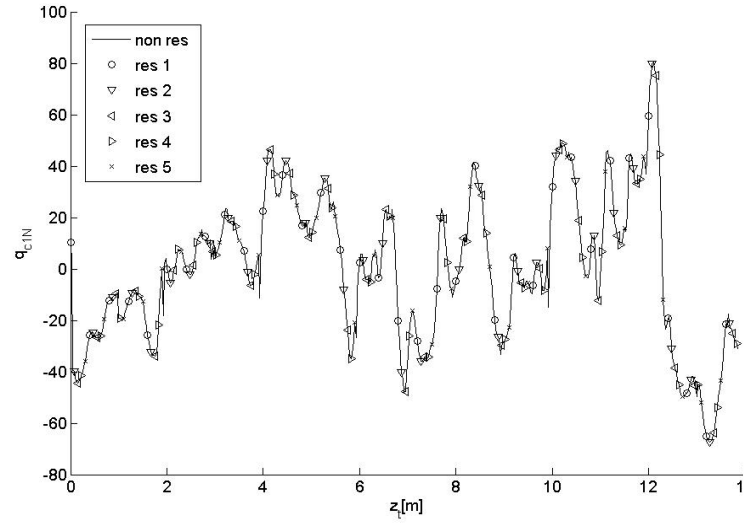

Figure 4. Residuals from real q_{c1N} profile [NONRES] and 5 resampled sets [RES1-RES5] (Δ=0.025m; δ=0.98m; k_n^*=0.41)

Conclusions

This paper shows that the classical AD test for Gaussianity cannot be applied on correlated data. It is possible to circumvent correlation by resampling the data at wider spacing, although this is accompanied by a significant loss in power. While the procedure can be extended to any type of spatially correlated data and to any correlation structure, the results of this paper strictly apply to data sets with an SNX-type correlation structure, with $1.5\leq I_1\leq100$ and with $10\leq k_\delta\leq40$.

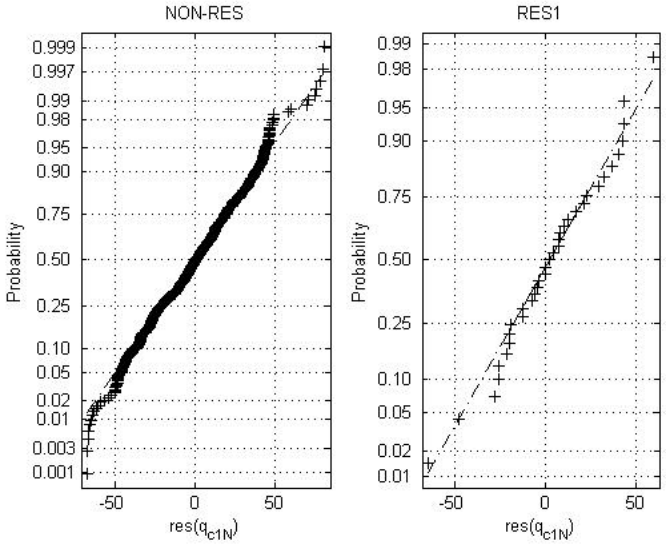

Figure 5. Normal probability plots for non-resampled residuals and resampled set RES1

Table 1. Anderson-Darling test statistic A^{2*} for non-resampled and resampled q_{c1N} residuals

set	A^{2*}
NONRES	0.714
RES1	0.228
RES2	0.214
RES3	0.170
RES4	0.290
RES5	0.168

References

Anderson, T.W and Darling, D.A. (1954). "A test of goodness of fit." Journal of the American Statistical Association, 49, 765-769.

Phoon, K.K., Quek, S.T. and An, P. (2003). "Identification of statistically homogeneous soil layers using modified Bartlett statistics." Journal of Geotechnical and Geoenvironmental Engineering, 2003, 129, No. 7, 649-659.

Robertson, P.K. (1999). "Estimation of minimum shear strength for flow liquefaction using the CPT." *Earthquake geotechnical engineering*, Balkema, Rotterdam, 1021-1028.

Stephens, M.A. (1986). "Tests based on EDF statistics." *Goodness of fit techniques*, R.B. D'Agostino and M.A. Stephens, eds., Marcel Dekker, New York.

Uzielli, M., Vannucchi, G. and Phoon, K.K. (2005). "Random field modeling of normalized cone penetration testing parameters." Géotechnique, 55(1), 3-20.