Computers & Geosciences 40 (2012) 1–9

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Calculating a robust correlation coefficient and quantifying its uncertainty

Eric B. Niven, Clayton V. Deutsch *

Centre for Computational Geostatistics, School of Mining and Petroleum Engineering, 3-133 Markin/CNRL, Natural Resources Engineering Facility, University of Alberta, Edmonton, Alberta, Canada T6G 2W2

ARTICLE INFO

Article history:
Received 29 September 2009
Received in revised form
30 June 2011
Accepted 30 June 2011
Available online 3 August 2011

Keywords:
Pearson
Spearman
Robust
Least median of squares
Outlier
Cokriging

ABSTRACT

Relationships between primary and secondary data are frequently quantified using the correlation coefficient; however, traditional means of calculating experimental correlation coefficients are known to be adversely affected by outlier data. A new method for calculating a robust correlation coefficient is proposed based on a weighted average correlation calculated from different combinations or subsets of the original data. The proposed robust correlation coefficient is shown to have a higher breakdown point than either Pearson's or Spearman's correlation coefficients as well as two out of three other robust correlation coefficients. The least median of squares (LMS) correlation coefficient has the highest possible breakdown point; however, it also tends to give unrealistically high or low correlation coefficients. A simulation study demonstrates the differences between the proposed robust correlation coefficient and other robust correlation coefficients. When the sample size is small, the uncertainty in the measured correlation can be very large, especially when the measured correlation is low. The uncertainty in the correlation coefficient is calculated based on the measured correlation and the number of data. This sampling distribution for the correlation coefficient requires a number of independent data; however, earth sciences data are often spatially dependent. Thus, a method for calculating an effective number of independent data using the variogram is proposed. An example is presented that applies the developed techniques to a petroleum geostatistics problem. The methodologies presented in this paper are implemented in FORTRAN code made available as part of this paper.

1. Introduction

The relationship between bivariate data is frequently summarized by the correlation coefficient. The sign of the correlation coefficient is positive if the variables are directly related and negative if they are inversely related. The closeness to +1 or −1 measures the closeness to a linear relationship. In some instances a few outliers significantly decrease an otherwise high correlation. The traditional Pearson correlation coefficient is known to be highly affected by outlier data (Abdullah, 1990; Isaaks and Srivastava, 1989; Jeongtae and Fessler, 2004; Shevlyakov, 1997). The Spearman rank correlation coefficient is considered to be more resistant to outliers, although it is also adversely affected by outlier data.

Methods for detecting outlier data have been suggested (Barnett and Lewis, 1994; Davies and Gather, 1993; Johnson and Wichern, 2007; Penny and Jolliffe, 2001; Rousseeuw and Zomeren, 1990). Outliers could be trimmed from the data and the correlation of the remaining points can be calculated. However, in some cases, the outlier data may be reliable data and should not be excluded (Gideon and Hollister, 1987), especially when the sample size is small. However, the influence of such data should not be inordinately large.

Often, correlations are estimated from a small number of observations. When the sample size is small, the uncertainty about the value of the true correlation can be very large, particularly when the estimated correlation is low (Kalkomey, 1997). It is useful to quantify the uncertainty in the correlation coefficient to assess its significance and to perform sensitivity studies.

Many statistical and geostatistical models and techniques rely on the correlation between different data variables. This research establishes a procedure to calculate a robust correlation and quantifies the uncertainty in the correlation coefficient through its sampling distribution of the correlation coefficient.

The correlation coefficient is particularly important in cases with sparse primary data and exhaustive secondary data such as offshore petroleum well data and seismic data. In this case, there may be only five to eight wells that have been drilled for production potential and not statistical representivity. Each of these wells is expensive and important. The final geological models will be highly dependent on the correlation coefficient established by simple spreadsheet calculations. Making this

* Corresponding author. Tel.: +1 780 492 9916; fax: +1 780 492 0249.
E-mail addresses: eniven@ualberta.ca (E.R. Niven), cdeutsch@ualberta.ca (C.V. Deutsch).

0098-3004/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2011.06.021

correlation robust and understanding its uncertainty have a large practical impact.

2. Measures of correlation

2.1. Pearson's correlation coefficient

Let $(x_1,y_1),\ldots,(x_n,y_n)$ be n observations from a bivariate normal distribution with parameters $(\mu_x,\mu_y,\sigma_x^2,\sigma_y^2,\rho)$, where μ_x and σ_x^2 are the mean and variance of x, μ_y and σ_y^2 are the mean and variance of y, and ρ is the correlation coefficient between x and y given by $\rho=\beta\sigma_x/\sigma_y$ where β is the slope parameter of regression of y on x. The sample correlation coefficient commonly used for estimating ρ is the Pearson's product–moment correlation coefficient defined by (Pearson, 1920; Rodgers and Nicewander, 1988):

$$r_p=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{[\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2]^{1/2}}. \quad (1)$$

One problem with using Pearson's product–moment correlation coefficient is that the sample means for x and y are sensitive to outlier data. As a result, the correlation estimate r_p is also sensitive to outliers in x, y, or both variables (Abdullah, 1990; Jeongtae and Fessler, 2004). Even a few outliers can degrade the sample correlation coefficient.

2.2. Spearman's rank correlation coefficient

As an alternative to Pearson's correlation coefficient, the nonparametric Spearman's rank correlation coefficient, r_s, can be calculated as

$$r_s=1-\frac{6[\sum_{i=1}^{n}(r_{y_i}-r_{x_i})^2]^2}{n(n^2-1)} \quad (2)$$

where r_{xi} and r_{yi} are the ranks of x_i and y_i, respectively. Spearman's correlation coefficient does not require the assumption of a linear relationship between the variables and is generally more resistant to outliers than Pearson's coefficient. However, as is shown later in this paper, Spearman's rank correlation coefficient is still quite sensitive to outliers, particularly in the presence of sparse data.

2.3. Outlier data

Outliers can be loosely defined as observations, which appear to deviate markedly from the other members of the sample (Grubbs, 1969). Hawkins (1980) defines an outlier as "an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism". There is no mathematical definition for what constitutes an outlier (Davies and Gather, 1993) and determining which data (if any) are outliers remains subjective. Data quality control and checking should identify erroneous data for removal; our concern is with influential multivariate observations that influence our calculated statistics.

Outliers can occur by two main ways. They may occur due to random variability in the data. In this case, outliers would normally be generated from a heavily tailed distribution. The second way for outliers to occur is when the data arise from two different underlying distributions. The "good" data come from one distribution and the "bad" or "contaminated" data come from another distribution. In this case, the contaminated data could be due to experimental or measurement error or any number of other ways.

If the data come from a heavily tailed distribution, the outliers are valid. In this case we would want to keep and use those observations. When the outliers occur from another distribution, we would hope to be able to identify and discard those values or use statistical methods that are robust to outliers. Reasons for different distributions of data could be different geological structures or processes.

Outlier detection has been widely discussed in the literature. Barnett and Lewis (1994), in particular, provide extensive reviews on this topic giving over 100 discordancy tests for a number of distributions. Despite the number of options for detecting outliers, there is no guarantee of finding any because there may not be a test developed for a particular combination, or the data do not follow any standard distribution.

Outlier detection in the bivariate or multivariate case can be even more challenging than in the univariate case. If the bivariate dataset is large and highly correlated, detecting outliers may be relatively easy. However, if the dataset is small (say less than 20 paired data values) and the correlation is low to moderate (say less than 0.5), we may not be able tell whether or not suspicious data points are outliers or an example of lack of correlation between the two variables.

2.4. Robust estimates of correlation

The idea behind robust estimation of means or covariances (and hence correlation) is to reduce the effect of outlier samples either by weighting or removing them altogether (Campbell, 1980; Jeongtae and Fessler, 2004; Rousseeuw and Zomeren, 1990; Titterington, 1978).

One of the most popular robust methods for estimating correlation (and regression coefficients) is the least median of square (LMS) estimation (Rousseeuw, 1984). The LMS regression coefficients minimize the median of the squared residuals. One of the big advantages of the LMS estimators is their noted 50% breakdown point, which means that LMS regression can give reliable results up to the point where 50% of the data are outliers. The LMS algorithm is similar to the bootstrap in that it proceeds by repeatedly drawing subsamples of p different observations from the dataset. For each subsample, $J=\{i_1,\ldots,i_p\}$, a regression line is found for the p points. Each regression line is viewed as a trial estimate and denoted θ_J. For trial, θ_J, the residuals between the regression line and the full dataset are calculated. The LMS objective function is defined by the median of the residuals:

$$\underset{i=1,\ldots,n}{med}(y_i-\mathbf{x}_i\theta_J)^2. \quad (3)$$

The trial estimate, which gives the minimal median of the squared residuals, gives the LMS coefficients and the correlation.

Shevlyakov (1997) introduced a robust correlation coefficient that utilizes the Hampel medians of absolute deviations to obtain the median correlation coefficient.

$$r_{med}=\frac{\mathrm{med}^2|u|-\mathrm{med}^2|v|}{\mathrm{med}^2|u|+\mathrm{med}^2|v|} \quad (4)$$

$$u_i=\frac{x_i-\mathrm{med}\,x}{\mathrm{med}|x_i-\mathrm{med}\,x|}+\frac{y_i-\mathrm{med}\,y}{\mathrm{med}|y_i-\mathrm{med}\,y|},\quad i=1,\ldots,n$$
$$v_i=\frac{x_i-\mathrm{med}\,x}{\mathrm{med}|x_i-\mathrm{med}\,x|}-\frac{y_i-\mathrm{med}\,y}{\mathrm{med}|y_i-\mathrm{med}\,y|},\quad i=1,\ldots,n. \quad (5)$$

Gideon and Hollister (1987) approach robust correlation from another perspective by introducing a robust rank correlation coefficient based on the principle of maximum deviations.

$$r_g=(d(\boldsymbol{\varepsilon}\circ\mathbf{p}-d(\mathbf{p}))/[N/2], \quad (6)$$

where $\circ$ is a group operation that is a composition of mappings $(\boldsymbol{\varepsilon}\circ\mathbf{p}=(N+1-p_1,\ldots,N+1-p_N))$, and $\mathbf{p}=\mathbf{p}(x,y)$ is the permutation

determined by the sample and $\boldsymbol{\varepsilon}$ is the reverse permutation.

$$d_i(\mathbf{p}) = \sum_{j=1}^{N} I(r(x_j) \le i < r(y_i)), \tag{7}$$

where $r(x_j)$ and $r(y_i)$ are the ranks of x and y, respectively.

$$d_i(\boldsymbol{\varepsilon}\circ\mathbf{p}) = \sum_{j=1}^{i} I(i < N+1-p_j). \tag{8}$$

Another method for calculating a robust correlation coefficient involves calculating an ellipsoid and trimming any data that do not fall within the ellipsoid. This technique works best for data that follow a normal distribution and other swarms of data that are elliptical (Titterington, 1978).

Despite these methods aimed at calculating robust correlation coefficients, there is room for improvement. The key idea developed below is to isolate the influence of each individual data pair (and sets of data pairs) and to ensure that the correlation coefficient is robust, yet fairly considers all data. The performance of any correlation coefficient estimator can be checked by a simulation study.

3. Correlation in sparse datasets in the presence of outliers

Fig. 1 shows a scatter plot of the bivariate relationship between two variables, x and y. The scatter plot shows what appears to be a strong direct correlation between the two variables marred by one potential outlier data point at a location of (7,1). The Pearson and Spearman correlation coefficients for the data shown in Fig. 1 are 0.291 and 0.214, respectively. Note that, while the Spearman correlation coefficient is usually more resistant to the effects of outliers, in this case it is more strongly affected by the potential outlier data point. In general, the Spearman correlation will be more robust when the outliers appear as unusually low or high values and not as values within the data distribution.

Of course when a dataset such as the one shown in Fig. 1 is observed, it would be natural to think that the point at (7,1) is an outlier. It may be a sample that belongs to another statistical population or perhaps there is an error with one of the measurements.

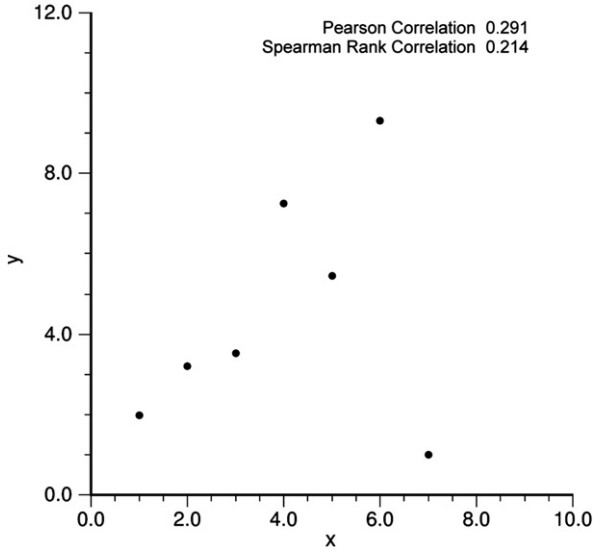

Fig. 1. An example dataset with one potential outlier data point. The potential outlier appears to be negatively affecting what would otherwise be a strong correlation between x and y. Note that, in this case, the rank correlation coefficient appears to be more strongly affected by the outlier than the Pearson's correlation coefficient.

If we could be reasonably sure that the point at (7,1) is an outlier, we could simply remove it from the dataset altogether. In this case, the Pearson and Spearman correlations would increase to 0.910 and 0.943, respectively. However, suppose we have carefully looked at all possibilities and have not found any reason to believe that the point at (7,1) belongs to another statistical population. In this case, the data should be considered in the calculations, but its importance should not be unreasonably large.

3.1. A weighted average correlation from a leave-one-out test

In order to arrive at a more robust correlation coefficient, we begin by introducing a "leave-one-out test" (LOOT), whereby a data point is removed from the dataset and the correlation is recalculated. This procedure can be repeated n times for a dataset with n points, leaving a different data point out each time. The result is n calculated Pearson correlation coefficients. A LOOT was conducted for the data shown in Fig. 1 and the results are shown in Table 1. For the first 6 leave-one-out tests, the resulting correlations are very low and unrepresentative of the obvious correlation in the data. However, the last test results in a correlation of 0.910.

The proposed robust correlation coefficient is based on the idea of a weighted average of the correlations calculated in the LOOT. The idea is to weight the correlations according to their difference from the actual correlation as follows.

$$w_i = \left|r_{Actual} - r_{i,LOOT}\right|^{\alpha}, \tag{9}$$

where:

- w_i is the resulting weight assigned to each correlation calculated in the leave-one-out test;
- r_{Actual} is the Pearson's correlation coefficient calculated using all of the original data;
- $r_{i,LOOT}$ is the ith Pearson's correlation coefficient calculated from leaving out the ith data point in the leave-one-out test;
- α is a weighting exponent and is a function of the number of data ($\alpha = 1 + n/12$). α is restricted to a maximum of 15 due to computational limitations and the observation that beyond a certain point a larger exponent is unnecessary.

The weighting coefficient is directly related to the sample size. As the sample size increases, the outlier data point effectively gains more influence on the calculated correlations in the LOOT since there are more combinations that use the outlier and only a few that do not. Thus, a larger exponent gives a larger weight (relative to the other weights) to the correlation that does not use the outlier data point.

In the above example the correlation obtained from removing the point at (7,1) is the most different from the actual correlation of 0.291 and thus receives the most weight. This makes sense

Table 1
Resulting correlations from a leave-one-out test (actual correlation is 0.29).

Coordinates of left out point (x,y)	Resulting correlation	Weight (w_i)
(1.00,1.98)	0.081	0.085
(2.00,3.20)	0.235	0.010
(3.00,3.53)	0.269	0.002
(4.00,7.25)	0.318	0.003
(5.00,5.44)	0.272	0.002
(6.00,9.31)	0.002	0.140
(7.00,1.00)	0.910	0.468

Updated correlation from leave-one-out test = 0.615.

since we want to somehow minimize the impact of this suspicious data point.

Then, the more robust correlation coefficient calculated from the LOOT for sparse datasets is defined as

$$r_{Robust,LOOT} = \frac{\sum_{i=1}^{n} w_i r_{i,LOOT}}{\sum_{i=1}^{n} w_i}. \quad (10)$$

The updated correlation coefficient is essentially a weighted average of the correlations calculated in the LOOT, where the weights are defined in (9). The weighting scheme in (9) assigns the greatest weights to the correlations that are the most different from the actual Pearson's correlation coefficient. The idea is that the data points that have the biggest impact on the correlation are the ones that are most likely to be outliers. For the data shown in Fig. 1, the LOOT correlation weights and updated correlation coefficient are as follows in Table 1. As is shown in the table, the updated correlation coefficient from the LOOT is 0.615, which seems reasonable given that we do not want to exclude the potential outlier.

3.2. A weighted average correlation from a leave-X-out test

The LOOT and weighted average correlation is effective for the case where there is one potential outlier data point. Of course, the idea can be extended to account for multiple potential outlier data points by considering a "leave-X-out test" (LXOT), where X varies from 1 to $n-3$ (correlations cannot be calculated with 1 data and correlations calculated using 2 data have little value). This calculation would yield $n-3$ weighted average correlations from each of the LXOTs. However, as the number of data increases the updated correlations calculated using only small amounts of data tend to be unreliable. For example, say we have 30 data and we want to conduct a L_{27}OT (where varying combinations of 27 data are left out and the remaining 3 data are used). The resulting correlations from the L_{27}OT would likely be very erratic depending on which data are selected. Even for highly positively correlated data there would be some subsets of 3 data that would form a strong negative correlation. Since the correlations are weighted by their difference to the actual correlation, those subsets would receive a lot of weight. Thus, we suggest constraining X from 1 to φ, as in (13). Thus, in a case with 100 data points, the proposed correlation considers LXOTs where X ranges from 1 to 77. There is no need to consider leaving out very large subsets of data anyway. If there are 100 data points, leaving out a maximum of 77 (a L_{77}OT) will already consider subsets with no outliers since there are normally far less than 77/100 outliers in the data. If there were 77 data that appeared to be from one distribution and 33 from another, one would normally call those 33 data the outliers.

Then, each of the weighted average correlations can be weighted again in a similar manner as follows:

$$w_{X,LXOT} = |r_{Actual} - r_{X,LXOT}|^{\alpha}, \quad (11)$$

where:

- r_{Actual} is the original data correlation
- $r_{X,LXOT}$ is the updated correlation calculated in each leave-"x-data"-out test
- $w_{X,LXOT}$ are the weights calculated for each updated correlation from the leave-"x-data"-out test
- α is the same weighting exponent as in (9) (i.e., $\alpha = 1 + n/12$).

The weighting exponent in (11) works similar to that of (9). We want to access the correlation from the LXOTs that are the most different from the original Pearson correlation. However, as the number of data points increases, the number of LXOTs that must be performed also increases. A larger exponent for larger sample sizes effectively gives more weight to the correlations from the LXOTs that are the most different from the original Pearson correlation.

Then a single robust correlation is calculated as follows:

$$r_{Robust} = \frac{\sum_{X=1}^{\varphi} w_{X,LXOT} r_{X,LXOT}}{\sum_{X=1}^{\varphi} w_{X,LXOT}}, \quad (12)$$

where

$$\varphi = 0.8n - 3, \quad (13)$$

where φ is rounded up to the nearest integer. Eq. (13) allows for a reasonable maximum number of data to leave out.

3.3. The distribution of r

It is also necessary to examine the uncertainty in the correlation coefficient. Naturally, since we have limited sample information, we do not know the true underlying data correlation. When the sample size is small, the uncertainty in the correlation coefficient can be very large, particularly when the measured correlation is low (Kalkomey, 1997).

The distribution of r (the sample correlation coefficient) as given in Johnson et al. (1995) is

$$p_R(r) = \frac{(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2}}{\sqrt{\pi}\Gamma((1/2)(n-1))\Gamma((1/2)n-1)} \times \sum_{j=0}^{\infty} \frac{[\Gamma((1/2)(n-1+j))]^2}{j!}(2\rho r)^j. \quad (14)$$

where $-1 \le r \le 1$.

Note that Eq. (14) also assumes that (X_i,Y_i) and (X_j,Y_j) are mutually independent if $i \neq j$. In formula (14), ρ is the estimated correlation, n is the number of *independent* data points, and Γ is the gamma function.

3.4. Calculating the number of independent data

Eq. (14) requires the number of independent data points. However, earth sciences data are rarely independent and are usually spatially related. We can, however, calculate an effective number of independent data.

Consider a number of observations X_i, where $i=1,\ldots,n$. The variance of the mean is given by

$$Var\{\overline{x}\} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}. \quad (15)$$

But, we also know that the variance of the mean can be calculated by (Edwards, 2006)

$$Var\{\overline{x}\} = \frac{\sigma^2_{data}}{N_{Independent}}, \quad (16)$$

where $N_{Independent}$ is the number of independent data.

The covariance of the data can be calculated from the variogram:

$$C_{ij} = \sigma^2_{data} - \gamma_{ij}. \quad (17)$$

Therefore,

$$N_{Independent} = \frac{\sigma^2_{data}}{Var\{\overline{x}\}} = \frac{\sigma^2_{data}}{(1/n^2)\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}} = \frac{\sigma^2_{data}}{(1/n^2)\sum_{i=1}^{n}\sum_{j=1}^{n} \sigma^2_{data} - \gamma_{ij}}. \quad (18)$$

Simplifying, we have

$$N_{Independent} = \frac{n^2\sigma^2_{data}}{\sum_{i=1}^{n}\sum_{j=1}^{n}(\sigma^2_{data} - \gamma_{ij})}. \quad (19)$$

Thus, the effective number of independent data can be calculated using only the number of data and the variogram. When the correlation between two variables is being considered, the variogram with the longest range should be used since it will yield a lower effective number of independent data.

Now that we have a method for calculating the effective number of independent data, the sampling distribution for the correlation coefficient (Eq. (13)) can be used to calculate the uncertainty in any measured correlation coefficient.

4. Computer codes

We created a suite of three FORTRAN codes to estimate a robust correlation coefficient and its associated uncertainty.

A FORTRAN code called ROBUSTCORRCO automatically calculates the updated correlation coefficients for each LXOT as well as an updated robust correlation. In cases where there are more than approximately 20 data points, the time to calculate the number of combinations of data in the LXOT becomes prohibitively large. As a result, a specified number (say 10,000) of data combinations are randomly sampled rather than calculating the correlation for every possible data combination. ROBUSTCORRCO also calculates the two traditional correlation coefficients as well as Shevlyakov's r_{med}, Gideon and Hollister's r_g, and Rousseeuw's r_{lms} for comparison purposes.

A FORTRAN code called NIND automatically calculates the effective number of independent data based on (19) and the input variogram model.

A FORTRAN code called SAMP_DIST_CORR calculates the sampling distribution for the correlation coefficient. The program uses the formula in (14) with the measured data correlation and the effective number of independent data as inputs.

4.1. Practical considerations

Although the summation in Eq. (14) is to infinity, it tends to converge rapidly except where the measured correlation is quite high (i.e., $\rho > 0.9$). Thus, an upper summation limit and a tolerance parameter are specified inputs into the SAMP_DIST_CORR program. The program calculates the percentage of instances where the summation parameter does not converge to a value smaller than the specified tolerance parameter. If the percentage of values not converging is too high, the summation parameter can be increased (or the tolerance can be increased).

4.2. Breakdown properties of the proposed robust correlation

A simulation study, similar to the one presented in Abdullah (1990), illustrates the breakdown properties of the proposed robust correlation coefficient (12) compared to the traditional Pearson and Spearman correlation coefficients, as well as the three robust correlation coefficients, r_{med}, r_g, and r_{lms}, proposed by Shevlyakov (1997), Gideon and Hollister (1987) and Rousseeuw (1984), respectively.

First, 100 good observations are generated according to the linear relation $y_i = 2 + x_i + u_i$, where x_i is drawn randomly from a normal distribution with a mean of 5.0 and a variance of 1.0. u_i is drawn from a normal distribution with a mean of 0 and a standard deviation of 0.2. The results were as follows: $r_{Pearson} = 0.974$, $r_{Spearman} = 0.969$ and $r_{Robust} = 0.906$. Note that the proposed correlation (r_{Robust}) is slightly lower than the Pearson and Spearman coefficients. The original Pearson correlation is quite high (0.974), so when the proposed algorithm leaves out data near the tips of the bivariate distribution, a slightly lower correlation is measured in the remaining data, which has the biggest impact on the proposed robust correlation. However, if the measured Pearson correlation was lower (say 0.5), this effect would be less pronounced.

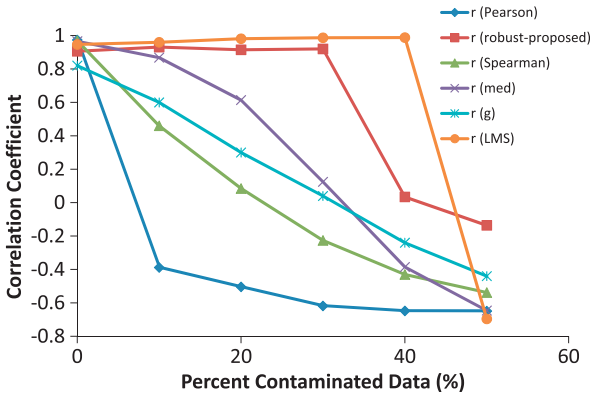

Fig. 2. Simulation study comparing the effect of contaminated data on the Pearson, Spearman, the proposed robust correlation coefficients and three other robust correlation coefficients.

Next, the data were slowly contaminated. In increments of 10 data points, the good data were replaced with bad data points. The contaminated data points were generated according to the linear relation where x_i is uniformly distributed on [5,10] and y_i is drawn from a normal distribution with a mean of 2 and a standard deviation of 0.2.

This was repeated until only 50 good observations remained. Fig. 2 shows the comparison of the proposed robust correlation coefficient against the traditional Pearson and Spearman correlation coefficients as well as three other robust correlation coefficients and serves to highlight the point at which the correlation coefficients begin to breakdown. In this study, Pearson's correlation coefficient breaks down with less than 10% contamination. Spearman's is more robust, as expected, but the proposed approach is significantly better than either of the two traditional correlation coefficients. Gideon and Hollister's r_g fares only slightly better than Spearman's correlation coefficient and its measured correlation with no contamination is much lower than the others. Shevlyakov's r_{med} exhibits reasonable resistance to data contamination until about 20% contamination, but by 30% contamination the correlation drops substantially. Rousseeuw's least median of square correlation, r_{lms}, is known to have a 50% breakdown point, as is shown in the figure. This is one of the main advantages of LMS regression.

5. Applications and discussion

Fig. 3 shows a porosity versus log permeability dataset with 12 paired points. Each point is labeled with an arbitrary number for reference purposes. Fig. 4 shows the location maps for the porosity and log permeability values. The left side of the circles indicate porosity (in %) and the right side of the circles indicates $\log_{10}$ permeability (in mD). The Pearson and Spearman correlation coefficients between the porosity and the log permeability data are 0.545 and 0.776, respectively. Since the Pearson correlation is lower than the Spearman rank correlation coefficient, the Pearson correlation may be affected by outlier data. Visual inspection of the scatter plot in Fig. 3 confirms that data point number one, in particular, and to a lesser extent two and three, appears to be "suspicious" or outliers. In this example, we assume that there are no known errors in the measurements. The program ROBUSTCORRCO calculates a robust updated correlation coefficient of 0.739, which agrees with the Spearman rank coefficient.

The program also calculates three other robust correlations, which are also noted on Fig. 3. In this case Shevlyakov's r_{med} and Gideon and Hollister's r_g are slightly lower than the proposed robust correlation and Spearman's rank correlation. However, here the LMS correlation coefficient is 0.957, which seems much too high based on visual inspection of the data and is much higher than any of the other correlations.

With knowledge of the estimated and proposed robust correlation coefficients, the uncertainty in the correlation coefficient can be calculated by utilizing the sampling distribution for the correlation coefficient. First, however, we need to know the number of independent data points. We can use the program NIND to calculate the effective number of independent data points. The data file and a variogram model are the only two inputs into the NIND program. In this case, a single-structure spherical variogram with a range of 4000 units in the horizontal plane and 10 units in the vertical direction was assumed since there is not enough data to calculate a reliable experimental variogram. Based on the data configuration, the number of data, and the assumed variogram, the effective number of independent data calculated by NIND is 10.8.

The program SAMP_DIST_CORR is used to calculate the sampling distribution for the correlation coefficient. The robust correlation (0.739) and the number of independent data (N_{ind}=10.8) are input into the program. The output is a probability density sampling distribution for the correlation coefficient, which is shown in Fig. 5. Note that the P50 for this distribution is approximately 0.76, which is different than the mean due to the asymmetric nature of the sampling distribution.

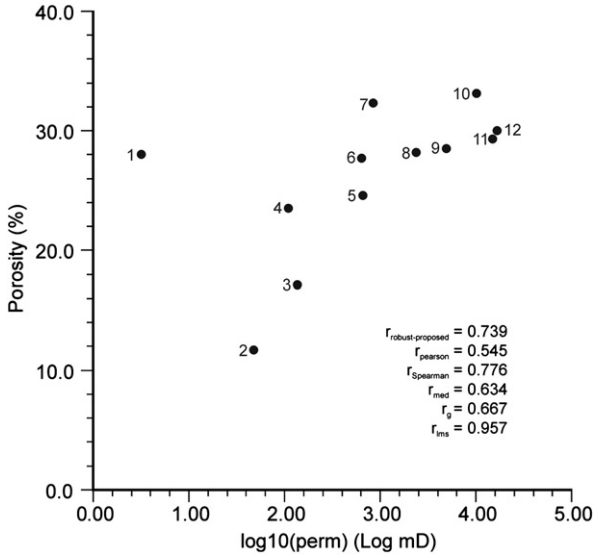

Fig. 3. Synthetic sparse dataset.

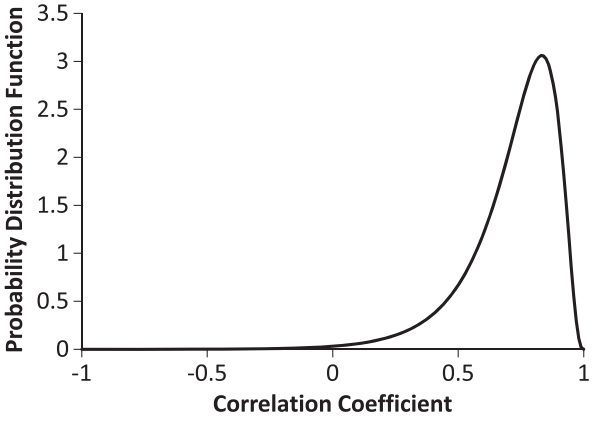

Fig. 5. Sampling distribution for the correlation coefficient for the synthetic core data shown in Fig. 3.

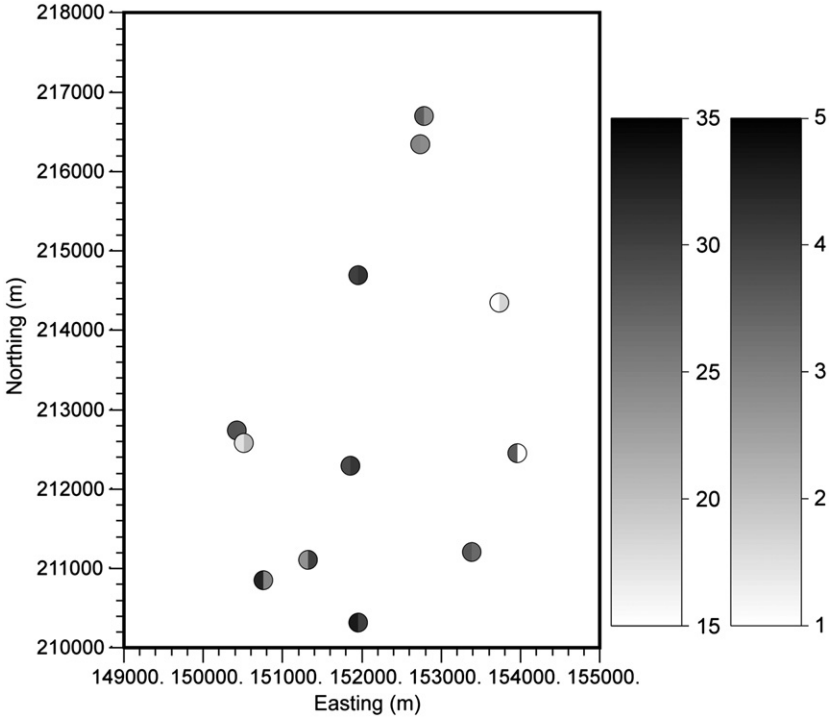

Fig. 4. Location map of 12 samples. The left scale and left side of the circles indicate porosity (in %). The right scale and right side of the circles represent Log_{10} permeability (in mD).

The 10th and 90th percentile (the P10 and P90) correlation values are approximately 0.11 and 0.90, respectively.

Note that if the measured correlation was lower, or if there were fewer data, the sampling distribution for the correlation coefficient would be even wider. For example, Fig. 6 shows the sampling distribution for the correlation coefficient for a measured correlation of 0.3 and 8 independent data points. The uncertainty in the correlation is much wider in this case and the P10 and P90 correlation values are –0.34 and 0.78, respectively.

For one final example, consider the data from an offshore reservoir shown in Fig. 9. The figure shows six paired points on a scatter plot between a seismic attribute and porosity. The data points have been labeled with arbitrary numbers for reference purposes. The Pearson correlation is –0.471 and the Spearman rank correlation is –0.771. On examination of the figure, data points one and six appear to be outliers or at least suspicious. Here, the program ROBUSTCORRCO calculates a robust updated correlation coefficient of –0.533, which is in good agreement with the traditional Pearson correlation coefficient. This makes sense when the results are examined in more detail. When point number one is left out of the calculation in a LOOT, the correlation between points two to six is –0.087. However, when point number six is left out of the calculation, the correlation of the remaining points is –0.927. Elimination of any of the other points makes little difference to the resulting correlation. Thus, the effects of point one and six roughly offset each other.

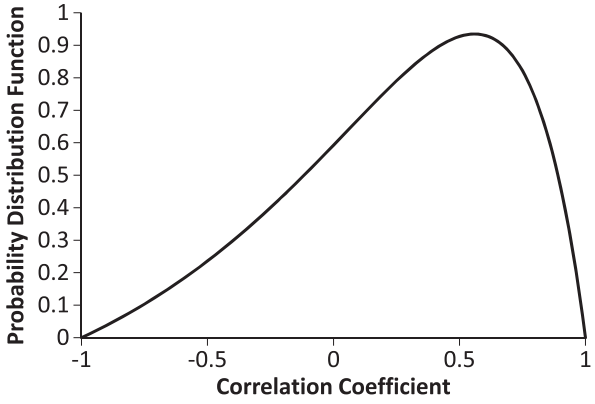

Fig. 6. Sampling distribution for the correlation coefficient for a measured correlation of 0.3 and 6 independent data points.

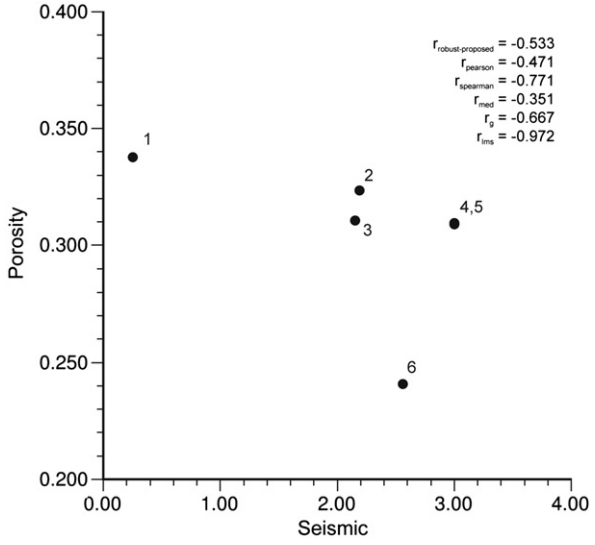

Fig. 7. A seismic attribute versus porosity from an offshore reservoir. Although only five points are visible, there are actually six points since there are two points very close together near (3, 0.31).

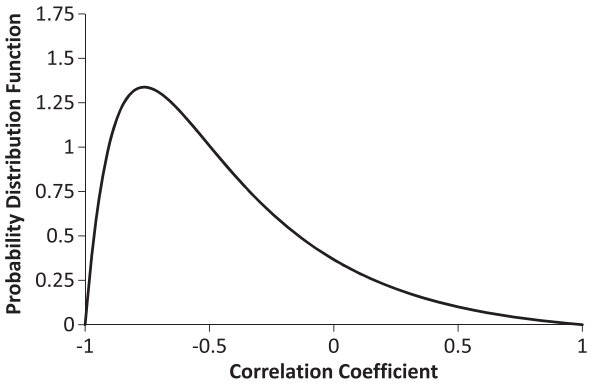

Fig. 8. Sampling distribution for the correlation coefficient for the data in Fig. 8.

The three robust correlations are also seen in Fig. 7. As shown, the robust correlation coefficients are $r_{med}=-0.351$, $r_g=-0.667$ and $r_{lms}=-0.972$. Just as in the last example the LMS correlation tends toward the extreme end of the correlation spectrum.

Fig. 8 shows the sampling distribution for the robust updated correlation coefficient for the offshore reservoir data. In this case, there are no spatial locations for the data so it is assumed that the data are independent of each other. The sampling distribution shows P10/P50/P90 correlation values of approximately –0.89/–0.60/0.00, respectively. Thus, according to its distribution, there is a 10% chance that the correlation is greater than zero, based on the number of data and the calculated robust correlation.

6. Simulation study

In an effort to compare the proposed robust correlation coefficient to the other robust correlation coefficients and to help explain the extreme r_{lms} values, a few small simulation studies were performed. In the first simulation study, 100 realizations of 10 data points (x,y) are generated by drawing x and y values randomly and independently from a uniform distribution between 0 and 10. Correlation coefficients are calculated for each 10 point realization. Since the x and y values are drawn randomly and independently from uniform distributions, the average correlation is expected to be 0.0. The results of the study are shown in Table 2. For each correlation coefficient, the average was very close to 0.0 as expected. More interesting is the standard deviation of correlation of the 100 realizations. The $\sigma_{Pearson}=\sigma_{Spearman}=0.34$ while $\sigma_{Robust\text{-}proposed}=0.29$ and is similar to $\sigma_g=0.25$. The lower standard deviation for $\sigma_{Robust\text{-}proposed}$ and σ_g makes sense since they should be less affected by outliers that give correlation to Pearson and Spearman's coefficients for some realizations. Interestingly, $\sigma_{lms}=0.711$, which is much larger than for any other correlation coefficient. Fig. 9 provides additional insight showing the relative frequency histograms of $r_{Pearson}$, $r_{Robust\text{-}proposed}$ and r_{lms}. The histograms of $r_{Pearson}$ and $r_{Robust\text{-}proposed}$ are symmetric and centered around 0.0, as expected. However, the histogram for r_{lms} shows a distinct tendency toward values near –1 and 1. Fig. 10 shows similar relative frequency histograms except that the number of data points per realization was increased from 10 to 50. When the number of data per realization is increased, the standard deviation of the correlations decreases as expected since the extra data points decrease the impact of a few outliers. However, even with 50 data points generated from two independent uniform distributions, the range of correlation calculated by the LMS algorithm remains very wide.

Table 2
Simulation study results for 100 realizations of 10 data points drawn from independent uniform distributions.

	Pearson's r_p	Spearman's r_s	Proposed robust r_{Robust}	Shevlyakov r_{med}	Gideon and Hollister, r_g	Rousseeuw r_{lms}
Average r	−0.040	−0.035	−0.039	−0.026	−0.044	−0.035
SD	0.340	0.340	0.289	0.444	0.247	0.711
Minimum r	−0.852	−0.952	−0.792	−0.907	−0.600	−0.991
Maximum r	0.711	0.842	0.666	0.873	0.400	0.989

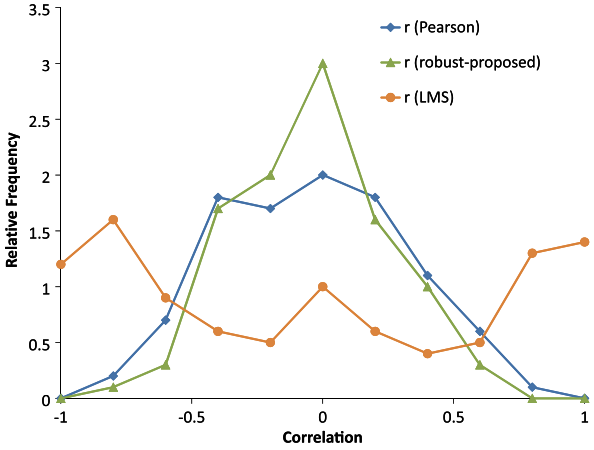

Fig. 9. Relative frequency histograms of correlation for 100 realizations of 10 data points measured by three correlation coefficients, $r_{Pearson}$, $r_{Robust\text{-}proposed}$, and r_{lms}.

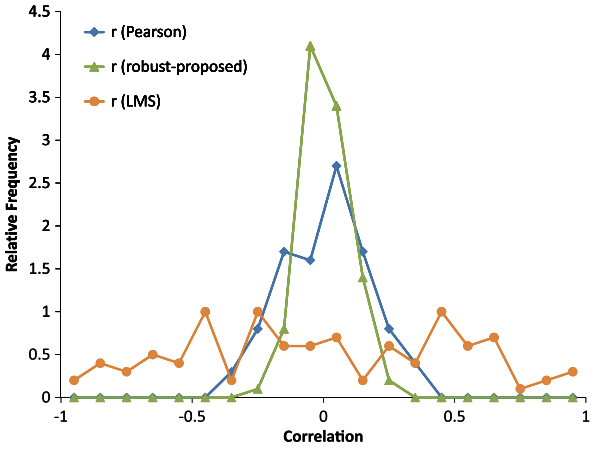

Fig. 10. Relative frequency histograms of correlation for 100 realizations of 50 data points measured by three correlation coefficients, $r_{Pearson}$, $r_{Robust\text{-}proposed}$, and r_{lms}.

The standard deviation for $r_{lms}=0.509$ with minimum and maximum correlations of –0.98 and 0.99.

For another simulation study, 100 realizations of 10 data points are generated. This time the x values were drawn from a normal distribution with $m_x=15$ and $\sigma_x=4$. The y values are drawn from a normal distribution with a mean conditional to x ($m_{y|x}$), conditional variance ($\sigma^2_{y|x}$), and a target correlation of 0.6 where

$$m_{y|x}=m_y+\rho\sigma_y\frac{(x-m_x)}{\sigma_x} \tag{20}$$

and

$$\sigma^2_{y|x}=\sigma^2_y(1-\rho^2) \tag{21}$$

Pearson's, the LMS and the proposed robust correlation coefficients are calculated for each of the 100 realizations. The average Pearson correlation is 0.634. The average proposed robust correlation is 0.524 and is lower than the target of 0.6. The average LMS correlation is 0.797, which is considerably higher than the target. The relative frequency histograms are shown in Fig. 11. The histograms for Pearson's and the proposed robust correlation coefficients are similar, although the proposed correlation tends to be slightly lower than Pearson's. However, similar to the previous examples, the LMS correlation has a strong tendency toward the extreme end of the correlation spectrum. In fact, the LMS estimator calculates a correlation greater than 0.9 in 37 of 100 realizations.

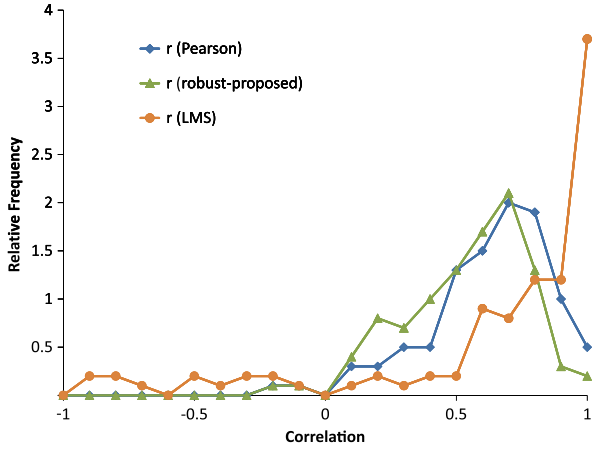

Fig. 11. Relative frequency histograms of correlation for 100 realizations of 10 data points (with expected correlation=0.60) measured by three correlation coefficients, $r_{Pearson}$, $r_{Robust\text{-}proposed}$, and r_{lms}.

Given that there is no formal mathematical definition for an outlier and that we may or may not want to exclude outliers depending on their origin and the number of data, it should be no surprise that there are a many ways to calculate a robust correlation coefficient. It appears that no correlation coefficient is perfect in every situation. The traditional correlation coefficients are particularly sensitive to outliers and two of the three robust coefficients (r_{med} and r_g) are slightly more robust, but still sensitive to outliers as shown by the breakdown test. The test also showed that the proposed correlation coefficient is quite resistant to outliers, though not as much as the LMS correlation coefficient. However, in examples with lower correlation and fewer data points the LMS correlation coefficient tends toward the extreme ends of the correlation spectrum. Moreover, for any particular underlying data distribution, the LMS estimator may calculate a correlation anywhere along the spectrum depending on the particular data configuration.

It would seem to be a good idea to calculate and compare several correlation coefficients for any particular dataset. If some of the coefficients agree with each other, it may be easier to trust one of those calculated values. On the other hand, visual inspection of a scatter plot of the data should not be overlooked. If the calculated correlation appears to disagree with visual inspection of the scatter plot, perhaps one of the other robust or traditional correlation coefficients makes more sense in that situation.

One example of application of this research is in collocated cosimulation. In collocated cosimulation a Markov-type assumption is made where collocated secondary information is assumed to screen further away data of the same type. This means that the available primary data and a single secondary datum at the estimation location are used in the calculation (Deutsch, 2002). The collocated cosimulation relies on the measured correlation between the primary and the secondary data. In some cases such as in off shore oil and gas reservoirs, there may be few wells or samples on which a correlation may be calculated. In these cases, the uncertainty in the correlation coefficient may be quite large.

It is useful and valuable to use the program ROBUSTCORRCO to obtain a more robust correlation coefficient. Then the program NIND can be used to calculate the effective number of independent data. Finally the program SAMP_DIST_CORR can be used to obtain the sampling distribution for the correlation coefficient.

The aim is to arrive at a more robust correlation coefficient, which minimizes the effect of outliers and is hopefully more reflective of the true correlation coefficient, which will never be known. By using a more robust correlation as input into the collocated cosimulation, the influence of secondary data can be more correctly scaled in the estimation and simulation of the variable of interest.

The use of the sampling distribution for the correlation coefficient allows one to examine the impact of the secondary data on our estimates. In this example, the geostatistician could perform three scenarios of collocated cosimulation with the P_{10}, P_{50}, and P_{90} correlations to observe the impact of the uncertainty in the correlation on the measured reserves. Or, a Monte Carlo simulation approach could be used to randomly sample the distribution of r as an input into the collocated cosimulation. Given the demonstrated uncertainty in the correlation coefficient, it is not difficult to imagine different scenarios having a major impact on estimated reserves.

7. Conclusions

The relationship between multiple variables in geostatistics is frequently estimated using the correlation coefficient. In cases where there are a small number of samples, the uncertainty in the correlation coefficient can be very large.

If a scatter plot of the data indicates the possibility for outlier data, the first step is to examine the data for any potential errors or inaccuracies. If no errors in the data are found and the sample is small, the engineer or geologist may not want to eliminate the suspicious data points from the dataset.

The sensitivity of Pearson's correlation coefficient to outliers is well known. An empirical study of the breakdown properties of the traditional and robust correlation coefficients indicates that Spearman's, Shevlyakov's, and Gideon and Hollister's correlation coefficients are also significantly affected by outliers, although they are more robust than Pearson's coefficient.

A new robust correlation coefficient was proposed, which showed a higher breakdown point than the traditional correlation coefficient and two other robust correlation coefficients. The LMS correlation coefficient has a breakdown point of 50%, which is the maximum possible. However, it was shown that the LMS correlation coefficient may exhibit a tendency toward calculating extreme correlation values. Moreover, even when realizations are drawn from some underlying distribution with a strong positive correlation, the LMS estimator may calculate low or negative correlations.

Care and judgment should be used in selecting a correlation coefficient to represent a dataset. It is fairly easy and quick to use ROBUSTCORRCO to calculate each of the robust correlation coefficients in addition to the traditional ones. Then the calculated values can be compared to each other and a scatter plot of the data to arrive at a reasonable value.

Regardless of the calculated value or choice of correlation between two variables, if the dataset is small, the uncertainty in the correlation can be very large. The sampling distribution for the correlation coefficient can be used to quantify its uncertainty based on the measured correlation value and the number of independent data.

Finally, the calculated uncertainty in the correlation coefficient should be propagated through geological modeling (or any further statistical or geostatistical analysis) to determine its impact on the resulting models. This can be easily achieved by running scenarios with different percentiles of the correlation coefficient or a Monte Carlo simulation approach.

Appendix. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.cageo.2011.06.021.

References

Abdullah, M.B., 1990. On a robust correlation coefficient. The Statistician 39, 455–460.
Barnett, V., Lewis, T., 1994. Outliers in Statistical Data, 3rd ed. Wiley, 604 pp.
Campbell, N.A., 1980. Robust procedures in multivariate analysis. I: robust covariance estimation. Applied Statistics 29 (3) 237–237.
Davies, L., Gather, U., 1993. The identification of multiple outliers. Journal of the American Statistical Association 88 (423), 782–792.
Deutsch, Clayton V., 2002. Geostatistical Reservoir Modeling. Oxford University Press, Inc., New York, NY, 376 pp.
Edwards, R.V., 2006. Processing Random Data: Statistics for Engineers and Scientists. World Scientific Publishing Co., Singapore, 152 pp.
Gideon, R.A., Hollister, R.A., 1987. A rank correlation coefficient resistant to outliers. Journal of the American Statistical Association 82 (398), 656–666.
Grubbs, J., 1969. Procedures for detecting outlying observations in samples. Technometrics 11, 1–21.
Hawkins, D.M., 1980. Identification of Outliers. Chapman and Hall, London 188 pp.
Isaaks, E.H., Srivastava, R.M., 1989. An Introduction to Applied Geostatistics. Oxford University Press, New York, NY, 561 pp.
Jeongtae, K., Fessler, J.A., 2004. Intensity-based image registration using robust correlation coefficients. IEEE Transactions on Medical Imaging 23 (11), 1430–1444.
Johnson, N.L., Kotz, S., Balakrishnan, N., 1995. Continuous Univariate Distributions, 2nd ed. John Wiley and Sons, Inc., New York, NY, 719 pp.
Johnson, R.A., Wichern, D.W., 2007. Applied Multivariate Statistical Analysis, 6th ed. Pearson Prentice Hall, Upper Saddle River, NJ, 773 pp.
Kalkomey, C.T., 1997. Potential risks when using seismic attributes as predictors of reservoir properties. The Leading Edge, 247–251.
Pearson, K., 1920. Notes on the history of correlation. Biometrika 13 (1), 25–45.
Penny, K.I., Jolliffe, I.T., 2001. A comparison of multivariate outlier detection methods for clinical laboratory safety data. The Statistician 50 (3), 295–308.
Rodgers, J.L., Nicewander, W.A., 1988. Thirteen ways to look at the correlation coefficient. The American Statistician 42 (1), 59–66.
Rousseeuw, P.J., 1984. Least median squares regression. Journal of the American Statistical Association 79 (388), 871–880.
Rousseeuw, P.J., Zomeren, B.C., 1990. Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association 85 (411), 633–639.
Shevlyakov, G.L., 1997. On robust estimation of a correlation coefficient. Journal of Mathematical Sciences 83 (3), 434–438.
Titterington, D.M., 1978. Estimation of correlation coefficients by ellipsoidal trimming. Applied Statistics 27 (3), 227–234.