A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 269, 2012

Jacek Stelmach[*]

USING PERMUTATION TESTS IN MULTIPLE CORRELATION INVESTIGATIONS

Abstract. An indication of correlation between dependent variable and predictors is a crucial point in building statistical regression model. The test of Pearson correlation coefficient – with relatively good power – needs to fulfill the assumption about normal distribution. In other cases only non-parametric tests can be used. This article presents a possibility and advantages of permutation tests with the discussion about proposed test statistics. The power of proposed tests was estimated on the basis of Monte Carlo experiments. The investigations were carried out for real data – a sample of refinery process parameters, where the indication of changes in correlation, even for sample with small size is very important. It creates an opportunity to react to changes and update statistical models quickly and keep acceptable quality of prediction.

Key words: permutation tests, Data Mining, correlation analysis, batch process, Monte Carlo.

I. INTRODUCTION

Paraffin wax – white, odorless, tasteless hydrocarbon solid in normal conditions, is produced in a process of deoiling – run in the so-called "sweating chambers" from raw material – slack wax. This material is a by-product of refinery processes (deparafination of base oils). As a by-product, nobody cares about quality of its parameters – and as a result - parameters of slack waxes are mostly unpredictable with high variation. Three parameters have the greatest influence on production of paraffin waxes: congealing temperature, viscosity and oil content.

The value of slack waxes parameters decides about the efficiency and economics of paraffin production. Unfortunately, the parameters known before the purchasing decision come only from a producer (manufacturer certificate), next package of information about parameters comes from a border quality control. The most reliable parameters are known only in a plant – after final quality control. And these packages of data are different because of:

- Different standards (ISO, DIN, GOST).
- The difference in quality of laboratory staff and equipment.
- A way of sample collecting and others.

[*] Ph.D. student, Department of Statistics, Katowice University of Economics.

It is possible to optimize the efficiency of whole logistic chain (including production) using statistical models. But these models are valid until any major change (i.e. new supplier, a change of laboratory staff, new reloading place) occurs. It is very important to recognize such a change and update the model as soon as possible.

II. RESEARCH PROBLEM DESCRIPTION

Both: parametric and non-parametric statistical models need proper predictors' packages. The independent variables should describe the examined process as wide as possible – they should be correlated with a dependent variable and rather weakly correlated with each other. The changes of multiple correlation coefficients can be indicated as the result of investigation of:

- The function of partial correlations: $f(r_{xi,y})$.
- The canonical correlation (especially first pair of canonical variables).
- The maximum of partial correlation - see Blackford J.U. *et al.* (2010).

The most known, usually used tests of correlation are carried out with certain limitations:

1. Test of Pearson correlation coefficient (parametric) with the statistics:

$$t = \frac{r\sqrt{n-1}}{\sqrt{1-r^2}} \tag{1}$$

where:

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \bullet \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} \tag{2}$$

2. Non-parametric tests (for Spearman's and Kendall's rank correlation coefficient) with the coefficients:
   Spearman:

$$r_s = 1 - \frac{6 \bullet \sum_{i=1}^{n}\left(Rx_i - Ry_i\right)^2}{n\left(n^2 - 1\right)} \tag{3}$$

Kendall:

$$\tau = \frac{2U}{n(n-1)} \tag{4}$$

Test of Pearson correlation coefficient can be used only when the assumption about normal distributions of examined data is fulfilled. Distribution-free nature of non-parametric tests enables avoiding the limitation but the power of these tests is usually lower. Additional important issue is a need to indicate a change of correlation already for small amount of data – to keep the model valid. The method that is discussed in this article is a permutation test for correlation coefficient. And the formulated hypothesis tested here is:

It is possible to test the null hypothesis about lack of correlation using the permutation test, even for the sample with small amount of observations; H_0: $\rho(x, y) = 0$.

The hypothesis is tested here for the real data sample of 36 observations, collected from January to July 2010 with the predictors:

- Congealing temperature – T_k, T_g, T_p
- Viscosity – L_k, L_g, L_p
- Oil content – O_k, O_g, O_p

Where 'k' index represents plant quality control, 'g' index represents border control and 'p' index represents manufacturer certificate results.

III. PERMUTATION TEST FOR CORRELATION COEFFICIENT

The goal of the test is to reject the null hypothesis (i.e. to discover a correlation between variables) at certain level of confidence. The idea of permutation test was worked out by R. A. Fisher. This test doesn't need any knowledge about the distribution of test statistics because instead of using any theoretical distribution, ASL (Achieved Significance Level) is estimated by Monte Carlo sampling from permutation distribution. And the power of permutation test is similar to parametric test, see Good P. I. (1994). The permutation tests sequence used in the investigations is as below:

1. Choose the test statistics that can 'measure' the correlation coefficient. It was decided to carry out the investigations with statistics presented in table 1.
2. Calculate the value of statistics for tested sample – T^*.
3. Proceed a permutation (M times, it is recommended in most cases M to be greater than 1000) of data, that destroys existing dependencies between variables.
4. Calculate test statistics value for these permutations and create empirical distribution – T_i, where i=1, 2,…,M.

5. Locate calculated value of T^* on this distribution and estimate p-value as ASL:

$$ASL = \frac{card\left\{T^* > T_i\right\}}{M} \tag{5}$$

6. If received ASL value is less than assumed value of α level, null hypothesis cannot be rejected.

Table 1. The test statistics used in described permutation test

Name	Test statistics	Formula for calculation
perm1	Pearson's correlation coefficient	$r_{XY} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \bullet \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$
perm2	Kendall's correlation coefficient	$\tau = \frac{2U}{n(n-1)}$, where: $U = \sum_{i<j}^{n} \xi(y_i, y_j)$
perm3	Spearman's correlation coefficient	$r_s = 1 - \frac{6 \bullet \sum_{i=1}^{n}\left(Rx_i - Ry_i\right)^2}{n\left(n^2 - 1\right)}$
perm4	Chi-square statistics	$\chi^2 = \sum_{j=1}^{k}\sum_{i=1}^{r} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$, where: $\hat{n}_{ij} = \frac{\sum_{j=1}^{k} n_{ij} \bullet \sum_{i=1}^{r} n_{ij}}{n}$
perm5	Chi-square statistics with Yates correction	$\chi^2 = \sum_{j=1}^{k}\sum_{i=1}^{r} \frac{(\left\lvert n_{ij} - \hat{n}_{ij}\right\rvert - 0.5)^2}{\hat{n}_{ij}}$
perm6	F statistics	$F = \frac{b^2}{S_e^2}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$

The idea used in presented investigations is based on *MPT.Corr* package, see Blackford J.U. *et al.* (2010) with the changes in test statistics and implementing additional Monte Carlo simulation – to estimate the power of analyzed tests.

IV. DATA ANALYSIS

For majority of analyzed variables, a hypothesis about normal distribution must be rejected, see figure 1 with histograms and value of Shapiro-Wilk test results.

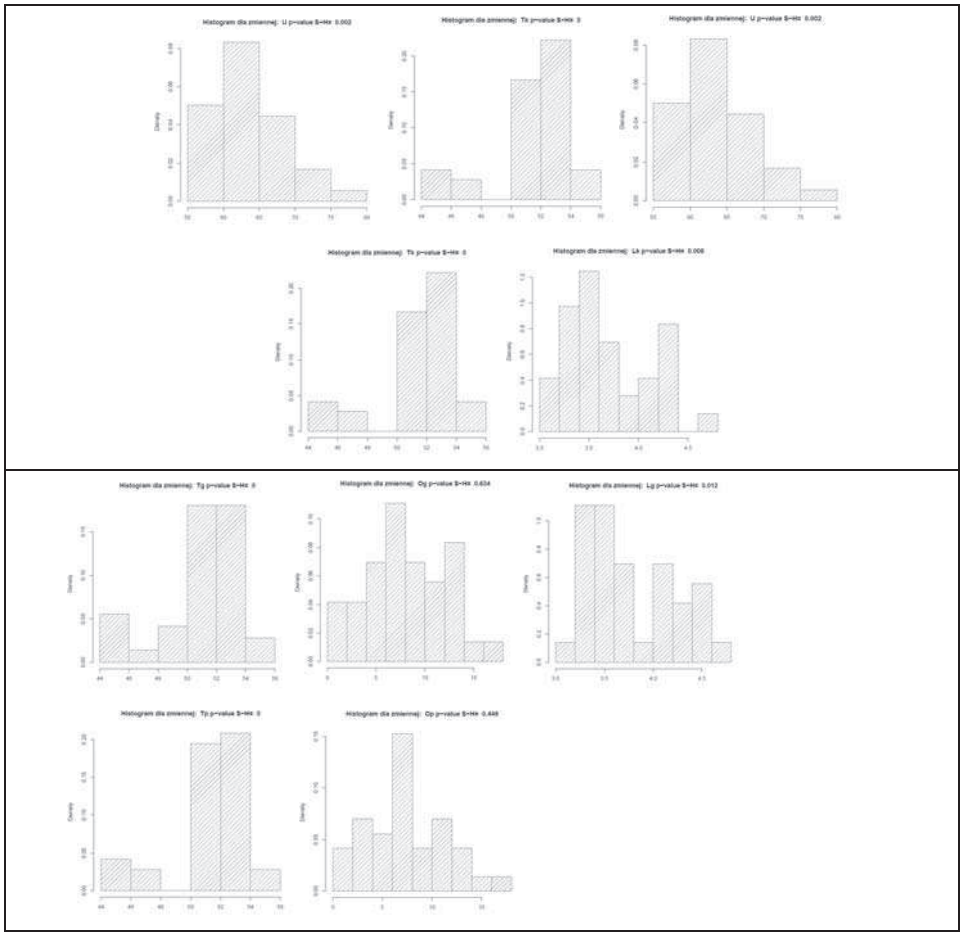

Figure 1. The histograms and p-values of Shapiro-Wilk test

It causes that the correlation coefficients for the whole sample were calculated only with non-parametric tests. For example, p-values of the tests for the correlation between L_k variable and the rest of known parameters are shown in table 2. The results of such a test are to be compared with Monte Carlo results of permutation tests.

Table 2
P-values of tests for tests of the Spearman's and Kendall's correlation coefficients for the whole sample

	Test of the Kendall's correlation coefficient					
predictor	Tg	Og	Lg	Tp	Op	Lp
correlation coef.	0.329	0.111	1	0.464	0.102	0.824
p-value	0.006	0.34	0	0	0.383	0
	Test of the Spearman's correlation coefficient					
predictor	Tg	Og	Lg	Tp	Op	Lp
correlation coef.	0.475	0.06	1	0.578	0.068	0.932
p-value	0.003	0.728	0	0	0.693	0

V. MONTE CARLO SIMULATION

The computer simulations were prepared using R CRAN software (www.r-project.org) in order to estimate the power and errors of permutation tests. The simulations were performed for subsamples with a size of: 5, 10, 15, 20 and 30 cases (100 subsamples). The power of the described tests was calculated during Monte Carlo experiment as a percentage of cases where null hypothesis was rejected (if such hypothesis was rejected for the whole sample). An error was calculated - as the percentage of I type error – number of cases where null hypothesis was rejected (if such a hypothesis was **not** rejected for the whole sample). Examples presented in figures 2 and 3 give different results. The case presented in figure 2 shows that the power of permutation test with *perm1* statistics was relatively poor in comparison with non-parametric tests or permutation test for other statistics (excluding chi-square types). It was caused by an outlier – that disrupted the shape of empirical histogram – see figure 4. Figure 3 shows the advantage of permutation tests, especially for statistics: *perm1* and *perm6*. Table 3 shows the total results: the number of cases (in Monte Carlo study) when a given test has got bigger power.

Table 3. Total results of Monte Carlo investigations for non-parametric and permutation tests

Test	Kendal	Sperman	perm1	perm2	perm3	perm4	perm5	perm6
sample size								
5	7	7	**11**	4	5	5	5	**9**
10	4	6	**13**	4	6	3	3	**10**
15	5	6	**14**	6	5	2	2	**12**
20	6	6	**17**	6	6	3	3	**15**
30	10	13	**17**	10	12	5	5	**16**
Sum	32	38	**72**	30	34	18	18	**62**

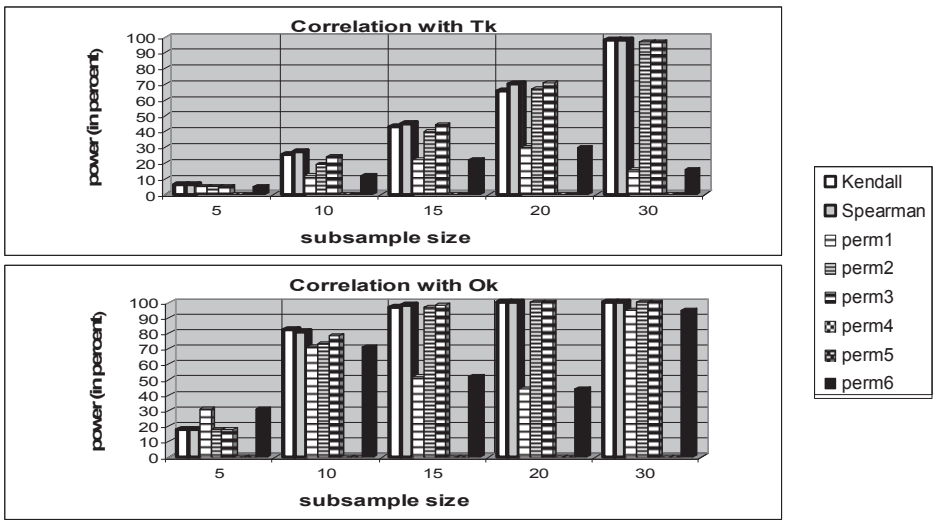

Figure 2. The comparison between non-parametric and permutation tests
(case with indicated outlier)

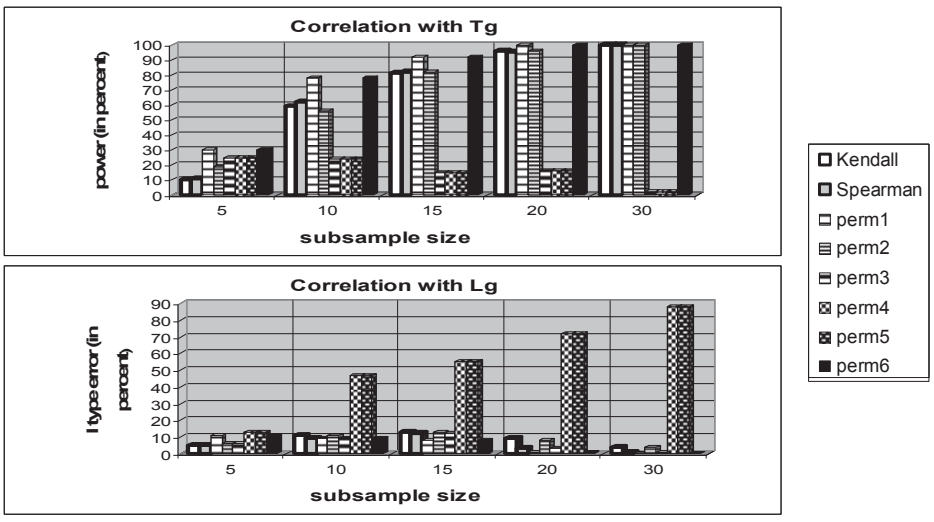

Figure 3. The comparison between non-parametric and permutation tests
(case with no indicated outlier)

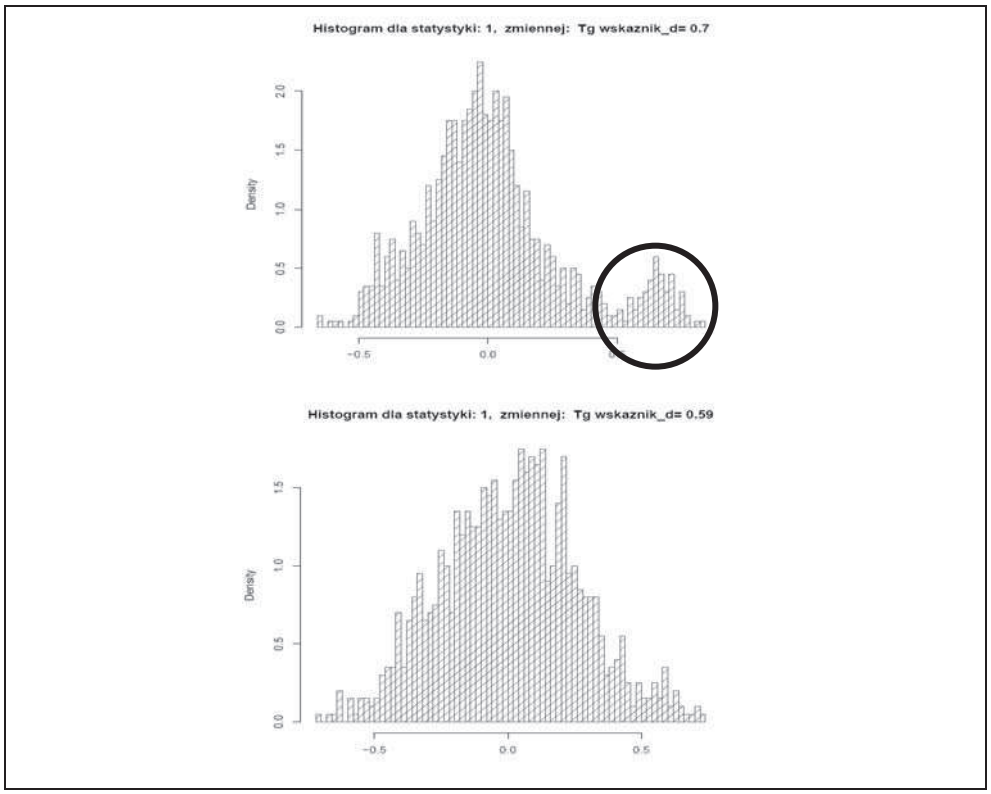

Figure 4. The histograms (empirical distributions) of the same statistics, on the left – the case with an outlier that created additional local maximum

VI. CONCLUDING REMARKS

Test permutation used for correlation purposes can have bigger power than standard non-parametric correlation tests (Spearman's and Kendall's correlation coefficients), especially for small amount of observations. The crucial point is to choose the proper statistics – adjusted to investigated data, and to observe a shape of empirical distribution (no significant local extremum). Such possibility gives an opportunity to discover the changes of conditions in examined process that needs the update of statistical models. There is another opportunity – to use permutation tests in canonical correlation, as well – to check the statistical significance. It will be the issue of next investigations.

REFERENCES

Blackford J.U., Kim G., Waller N., Koder P.: *A Manual for the Multivariate Permutation Test for Correlations* http://www.psych.umn.edu/faculty/waller/downloads/mpt/mptcorr.pdf [2010.05.31]

Efron B., Tibshirani R. (1993) *An Introduction to the Bootstrap*, Chapman & Hall. N.York

Good P. I. (1994) *Permutation Tests: A practical guide for testing Hypotheses*, Springer-Verlag, N. York

Kończak G. (2008) *O pewnym teście dla weryfikacji hipotezy o równości wartości przeciętnych w k populacjach*, [w:] Rola informatyki w naukach ekonomicznych i społecznych, Zeszyty Naukowe 8, tom 2, str. 337–344. WSH Kielce.

Odiase J.I., Ogbonmwan S.M. (2007) *Correlation Analysis: Exact Permutation Paradigm*, Matiematiczki Wiesnik, vol. 59, str. 161–170.

Jacek Stelmach

WYKORZYSTANIE TESTU PERMUTACYJNEGO W BADANIACH KORELACJI WIELOWYMIAROWEJ

Istotnym zagadnieniem w procesie tworzenia modeli statystycznych jest dobór predyktorów skorelowanych ze zmienną zależną. Test współczynnika korelacji Pearsona o stosunkowo dużej mocy wymaga spełnienia założenia o normalności rozkładu badanych danych. W innym przypadku wyłącznie mogą być wykorzystane testy nieparametryczne. Artykuł przedstawia zalety testów permutacyjnych wraz z propozycją zastosowania konkretnych statystyk testowych. Moc tych testów została oszacowana metodą Monte Carlo. Badania zostały przeprowadzone dla rzeczywistych danych reprezentujących parametry procesu rafineryjnego, w którym wykrycie zmian korelacji, nawet dla małych liczności jest bardzo ważne. Daje to możliwość reakcji na zmiany i szybkiego uaktualniania modeli statystycznych, utrzymując zadowalającą jakość prognoz.