USING OF CANONICAL CORRELATION ANALYSIS IN MARKETING MANAGEMENT OF TOUR OPERATORS AND TRAVEL AGENCIES

Lukáš Volf, Ing., DiS., lecturer[150], Ph.D. student[151]

Abstract: *The document represents the theory of multivariate statistics that specifically discusses the canonical correlation analysis and explains the basic principles and possible outcomes. Work also performs its analysis of the population, based on a survey carried out in the framework of the preliminary research related to the topic of the dissertation of the author on the topic of Quality management - the strategy of enterprises – SMEs.*

Key words: *Canonical correlation, correlation analyse, tour operators, travel agencies, SMEs, marketing management*

1. INTRODUCTION

In the case of most commercial marketing research are the subject of analysis of pre-defined elements (variables) that are generally consistent with the meaning of the questions, and are no longer monitored reciprocal links (relationships or dependencies) between these variables. The research also takes place purely in the context of the interpretation of the answers to that question. While determining the interdependencies between different quantities can lead to the better understanding of the merits of the research and the research questions. Especially type scientific disciplines of economics and sociology, these methods of multivariate statistics should take in its fulfillment.

Regression and correlation analysis are used for investigating the various relationships between variables. Both are fundamental methods of multivariate statistics. Many methods of multivariate data analysis are just an extension of one-dimensional and two-dimensional analysis [2]. It is necessary to define the basic difference between a one-dimensional and two-dimensional analyses. The relevant correlation analysis among the multivariate method that is an extension of one-dimensional and two-dimensional analyses.

Under the univariate analysis can imagine an analysis of the distribution of one variable and also an independent analysis of several variables in terms of one after another without any context. For two-dimensional data analysis tools are considered pairwise correlation coefficients further regression equation with a single response measurable and with one explanatory measurable variable or analysis of variance with one measurable response variable and one categorical explanatory variable, and the combination of two-dimensional tables, simultaneous frequencies between any two variants (mainly categorical) variables. [2]
The response variable is the dependent variable, respectively one variable (Y), in which we observed a relationship with one or more variables (X_1, ..., X_n). These variables are treated as independent; it means explanatory variables. These relations are concerned regression and

[150] University College of Business in Prague, Spálená 76/14, 110 00 PRAGUE, Czech Republic
[151] University of South Bohemia in České Budějovice, Branišovská 1645/31a, 370 05 ČESKÉ BUDĚJOVICE, Czech Republic

correlation analysis, the first mentioned examines dependency forms (one-sided dependence) and expresses them with appropriate mathematical functions, it means by regression functions. The correlation analysis determines the degree of force that a given dependency is used for expression in a particular environment of secondary elements (interdependence). [3] Any dependence of variables raise the natural question of whether it is substantial or not, how strong the relationship is. Mutually and linear relationship between variables is just statistics known as correlation, that is specifically selected symmetric dependence of some type. [6]

***Lukáš Volf** is receiving his Ph.D. degree in the University of South Bohemia in České Budějovice focused on management of quality in the tourism sector. During his studies, he worked as a senior product manager of an international insurance company where he was responsible for travel insurance development. He was a Chairman of the Working Group of travel insurance in Czech Insurance Association in 2015. Currently, he gives lectures and consultations specialized on tour-operators management and financial services in tourism at the University College of Business in Prague.*

Dependence operates with a degree of tightness of statistical dependence. Since the degree of tightness of statistical dependence is required for movement in a specific, tightly defined, interval, which has to raise its value with respect to increasing the degree of dependence and at the same time not to be dependent on the specific types of units at the investigated variables or of their size (the size of values). The impermeability dependence corresponds to the degree with which the given dependence approaching functional dependence. [1], [4]

To express a particular intensity level of reciprocal linear relationship random variable Y and the random variable X is used **simple (Pearson) population coefficient** $p\ (Y, X)$. Extended coefficient for measuring the intensity of the linear relationship of mutual random variable Y and linear functions $a_1\ X_1 + a_2X2 + \ldots.\ a_p\ X_p$ random vector x is known as **multiple correlation coefficient**: $p\ (Y, a^T x)$. It is the highest correlation coefficient between Y and linear combinations of $a^T x$. In the case of group correlation coefficient $p\ (y, x)$ which presenting degree of correlation between two random vectors x and y may be considered as a kind of superstructure maximum correlation coefficient. [2] The relationship between these two linear combinations of variables is engaged in a process called canonical correlation analysis (known as CCA - Canonical Correlation Analysis).

Canonical Correlation Analysis is a generalization of multiple correlation analysis used in the regression jobs and the goal is to determine the appropriate degree of correlation (correlation) between the response variable and any linear combination of the explanatory variables. [2] It was designed in 1935 by Hotelling in connection with the search for more linear combination one group of characters $x = (x_1, \ldots, X_q)$ that best correlates with the linear combination of the second group of symbols $y = (y_1, \ldots, y_p)$. Based on the assumption of a joint distribution of the two groups of characters. Canonical correlation analysis are closely related to the behavior of multiple correlation coefficient R between one random variable and a linear combination of

other variables. Interpretation of canonical variables is more problematic than that of factor analysis. Therefore, they serve in many cases as pre-processing data for further analysis. Canonical correlation analysis is often utilized in the situations in which form regression models and where there is more than one dependent variable. It is particularly useful in situations where the dependent variables are intrinsically correlated, so it is not worth evaluated separately because they are neglecting their mutual internal correlation. A useful feature is the ability canonical correlation verify independence among groups of characters *x* and *y*. [5] When creating and working with canonical analysis it is necessary to take into account:
- means and standard deviations,
- chi-square test,
- variances/dissipation (proportions),
- canonical weight.

Determining the standard deviation is for a canonical correlation very important since the value testifies how much individual cases in the population differ from each other (small deviation indicates similarities, a large deviation testifies large mutual differences). **Pearson chi-square test** determines, respectively verifies, distribution predetermined probability for random variables. With this assay can also be encountered in connection with PivotTables and verifying hypotheses, equivalently is called as a test of good conformity. Distractions (volatility, variability) of random variable values around the central value express its **dissipation**. It is a measure of a random variable around its mean value (in relation to the topic, that measure is significant since it determines how much the respondents agree or disagree regarding the issue). **Canonical weights** represent one of the methods for the interpretation of canonical variables and determine the relative importance of all of the original characters in the canonical variables. [2], [5], [7] Other possible methods are canonical loads and canonical cross burden. Canonical correlation analysis include these graphic outputs:
- box plot showing e.g. median, min. and max., average and standard deviations;
- graph of individual correlations;
- graph of canonical variables and
- individual histograms.

2. METHODOLOGY

The presented work is used for the preparation of the canonical correlation analysis of data from the questionnaire which dealt with added value of offers of tour operators in relation to consumers. The questionnaire took place from 1st February to 29th February 2016. The survey was conducted in the Czech language and was approached by a random sample of persons, the return of the questionnaire reached almost 76% with a total number of 453 responses. The questionnaire has 11 questions, while the last two are identifying and pursuing gender of the respondent and his age. A total of 4 questions carry the possibility answer yes / no; in 4 questions respondents chose from a list of answers, while for two of them additionally determined using a Likert scale of levels of agreement with the choice. The questionnaire also includes one open question, by which it assessed the extent to what respondents understand the topic and what do not.

The population in this case is 100 observations (i.e. N = 100, 11 variables). Given that within the survey occurs subdivision of a questionnaire based on the positive or negative response, the file decreases gradually (this does not impinge on the analysis as such as missing data the

program STATISTICA [8] reported as missing value and the program it does not take into account/ individual calculations).

3. RESULTS

Custom analysis was performed in STATISTICA and the primary output, thus individual correlation recorded in the table of variables and observations are placed in Annex no. 1. The table contains a numeric value between -1 and 1, with the more positive value closer to 1, the more positive relationship to each of the variables have. Conversely, values close to -1, the relationship is strongly negative. The value oscillating around 0 state generally weaker link of the variables between themselves. The plaque can be traced diagonal of values 1 (pictured gray), which indicates that the variables are identical, in other words, the variable predicts the same variable (which is why this relationship is always the same combination of variables, i.e. q_1 *vs.* q_1, q_2 *vs.* q_2 ... q_{11} *vs* q_{11}).

At the outputs of the data you can notice such as the fact that q_2 and q_1 and between themselves have no correlation (the correlation is 0), in other words, that the form in which the population familiar with the offer of some tourism subject does not affect whether they realized their trip through the tour operator or travel agency.

The above table output may be replaced by a matrix of point diagrams, see Figure no. 1, its graphic form can be enjoyable form of expression of the results. In the above table are tinged with red random values, which are further discussed and suitably complements very well figure no. 1.

For the correlation between the *q5* and *q3* can be concluded that although impressed on consumer marketing campaign (recognized as a distinct marketing concept) were slightly disappointed - the results were not satisfactory - because of declining trend line. Unlike the correlation of variables $q6^a$ and $q6^b$, which can be expressed in a positive relationship with respect to clearly growing line. De facto it tells us that the more one variable (in this case, service having a high standard), the more even the second variable (the width of the services offered was sufficient). Other selected correlations between variables *q5* and $q8^a$ expresses negative trend, which can be interpreted so that the more failed bid submitted tours consumer expectations, the more they tended to compare the standard of services offered in case of ask trip with another subject of tourism. Conversely, a clear positive correlation can be found between high standard of service and fulfilling customer expectations (variable $q6^a$ and $q8^a$). The last selected correlation engaged in sex. The first one took into account gender when asked whether they were satisfied expectations when submitting bids tour (*q5*), which positively expressed more women (*q10*) and in the second case on the question of dealing with age (*q11*), when the result shows that on average were older men.

Figure 1: Matrix diagrams point of correlation analysis

Individual correlations can also display specific to the selected variables, such variables were chosen values $q6^{a-e}$ to ourselves, see table no. 1.

Kořen odstraněn	Korelace, levá sada (List1 v data_volf_chD)				
	q6a	q6b	q6c	q6d	q6e
q6a	1,000000	0,733429	0,190277	0,072888	0,062326
q6b	0,733429	1,000000	0,267570	0,019017	-0,062855
q6c	0,190277	0,267570	1,000000	0,120747	0,141736
q6d	0,072888	0,019017	0,120747	1,000000	0,741716
q6e	0,062326	-0,062855	0,141736	0,741716	1,000000

Table 1: Correlation of values $q6^{a-e}$

For example, this output can evaluate the correlation, that materially affects the (red-colored) and for variables $q6^{b}$ and $q6^{a}$ say that the rate of customer fulfillment is affected by another standard of service and sufficient breadth of services offered. In the second case ($q6^{e}$ and $q6^{d}$) then filling based on mutual intelligibility affecting supply and nature / purpose of the journey of the trip.

Correlation can of course also expressed to a single variable where, for the purposes of the work has been taken as the relationship between sex (*q10*) and the above-mentioned factors ($q6^{a-e}$), see table no. 2. From this it follows that these variables do not affect each other.

Kořen odstraněn	Korelace, levá i pravá sada q10
q6a	-0,044561
q6b	0,056089
q6c	0,071981
q6d	-0,075033
q6e	-0,103109

Table 2: Correlation of values *q6^{a-e} and q10*

Among other results of the analysis include the aforementioned averages and standard deviations, respectively for the purpose of evaluating these and standard errors, more or less we tell us how much we can be certain diameter. This output is presented in Table no. 3 and figure no. 2.

	Průměry a směrodatné odchylky (List1 v data_volf_chD)	
proměnná	Průměry	Sm.odch.
q1	1,160000	0,368453
q2	3,011905	1,514202
q3	1,119048	0,298304
q5	1,523810	0,460044
q6a	3,369048	1,461836
q6b	3,428571	1,485864
q6c	3,107143	1,477220
q6d	3,416667	1,429900
q6e	3,392857	1,421465
q7	1,654762	0,437949
q8a	2,862069	0,849529
q8b	3,137931	0,813077
q8c	2,793103	0,936136
q8d	3,482759	0,883298
q8e	3,000000	0,828775
q8e	3,000000	0,828775
q9	1,172414	0,204444
q10	1,520000	0,502117
q11	2,900000	1,480513

Table 3: Means and standard deviations

Regarding context, mean and standard errors are just better graphical representation, because we can express, how confident we can be given diameter. It must be demonstrated variable *q10*, the graph shows that both sexes were investigated balanced and due to the fact that the standard error of nearly diameter bounding area where probably average is diminished to a minimum, the same goes for other variables (*q5, q7*) and closest to the *q1* and *q3*. Another interpretation example can be variable *q9* when considering that the average value approaches 1, so we understand that the majority of the population buying trip with the first subject of tourism.

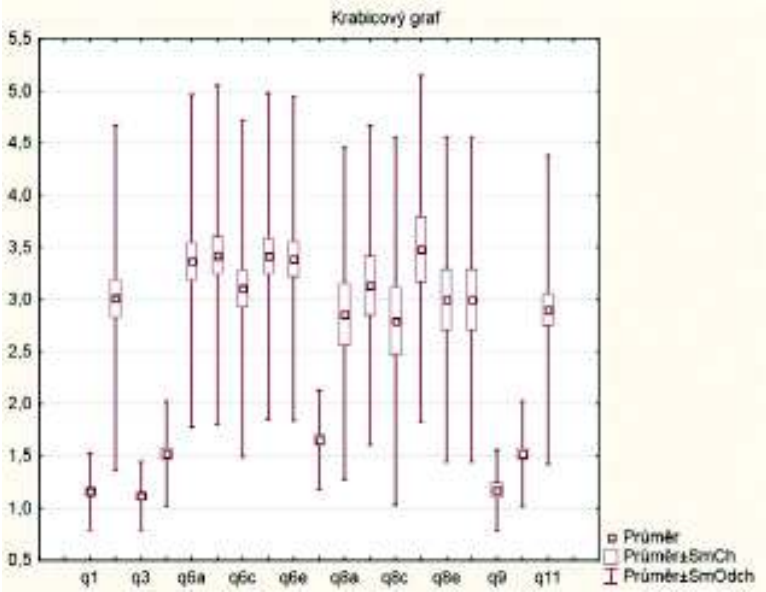

Figure 2: Box plot of average and standard deviations

Another graphical output of the box chart can be expressed median, percentiles and minimum and maximum values, see figure no. 3.

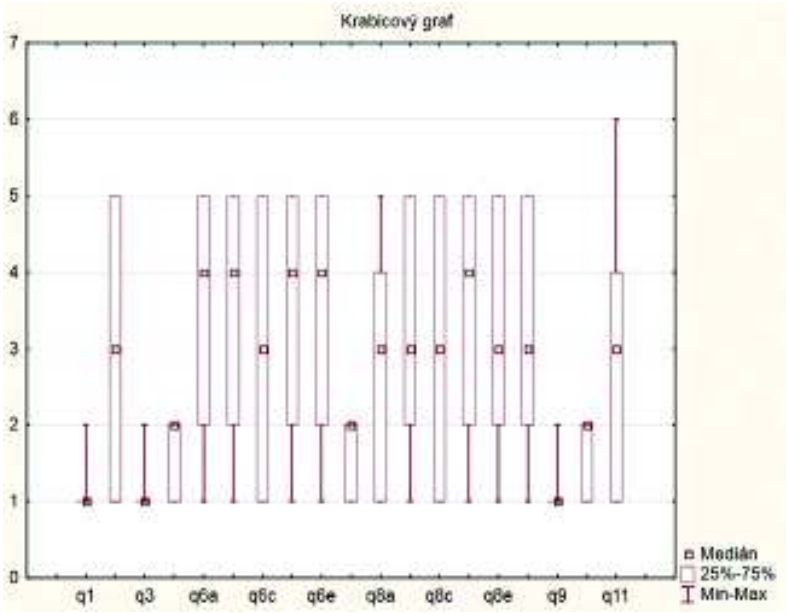

Figure 3: Box plot of median percentiles, min. and max. values

The above figure no. 3 notices several interesting facts. For example, variables *q2* and $q6^c$ were evaluated for their entire range, the median is the diameter of the consumers and thus are not satisfied. Another example may be mentioned by variables $q6^a$, $q6^b$, $q6^d$ and $q6^e$ when less than a quarter of respondents selected 1 (this follows from quartile range, which is between 2 and 5, or less than 25 % has been chosen value of 1.

The summary of canonical analysis allows expressing the overall redundancy in percentages and also R factor, a measure of how much the selected variables can influence each other. For purposes of summarizing the analysis of canonical variables were chosen which were

expressed using a Likert scale (already mentioned several times $q6^a$-$q6^e$), see table no. 4. The result is that 74 % of these variables influence each other, in other words the correlation of three-quarters thick.

	Souhrn kanonické analýzy (List1 v data_volf_chD) Kanonické R: ,73997	
N=100	L	P
Počet proměnných	5	5
Získaný rozptyl	100,000%	100,000%
Celková redundance	35,2057%	37,0351%
Proměnné: 1	q6a	q8a
2	q6b	q8b
3	q6c	q8c
4	q6d	q8d
5	q6e	q8e

Table 4: Summary of canonical analysis

As the control mechanism can be applied Chi-square test. Its output includes the value of the conical R (% of explanatory variables indicative of the response variable) but also the value of *p*, which represents a value 0 to 1, and is an indicator for us when we can reject the null hypothesis. This condition occurs when the result is 0 or very close to it (e.g. 0,1). Conversely, if we have the result of 0,7, as in the case below, see table no. 5, to rejection in relation to the null hypothesis does not occur.

	Test chí-kvadrát po odstranění post. kořenů (List1 v data_volf_chD)					
Kořen odstraněný	Kanonic. R	Kanonic. R-kvad.	Chí-kv.	sv	p	První
0	0,175729	0,030881	2,995596	5	0,700665	0,969119

Table 5: Chi-square test

Finally, we should mention the histograms as one of the outcomes of the analysis. For the purpose of the work was created on the issue of dealing with the evaluation factors and their influence on request tour or individual tourism services at a competitive subject, namely the factor of comparison standard of services offered, see figure no 4. For it is clear that this factor was put consumers the highest weight in spite of the general population (red line).

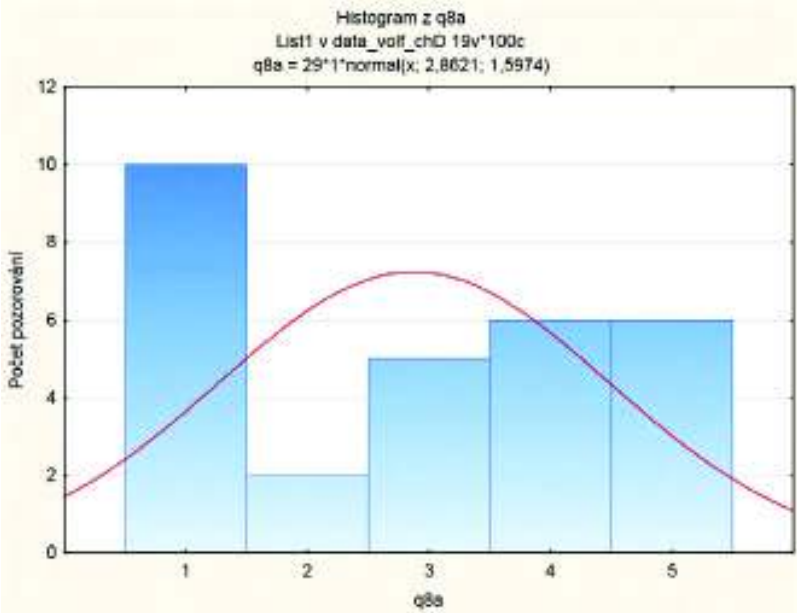

Figure 4: Histogram of $q8^a$

4. CONCLUSION

Statistical correlations are an important element, should be used for routine evaluation of research activities, socio-economic fields. Their application to the results of the investigation will ensure a proper understanding of the interrelationship between the observed variables. Correlations are a tool for complex data analysis.

One of the programs that can help with own correlation analysis and subsequent interpretation is the STATISTICA that offers relatively easy, after some time and intuitive, operating environment. Thanks to the program can perform own analysis, input data as needed to adjust, but also can in various forms results graphically customize to subsequent needs of interpretation.

Statistical – an analytical – tool of correlations can, after the experience gained in working on this subject, evaluated as useful not only for study purposes but also for regular commercial and other practical applications.

REFERENCES

[1] BUDÍKOVÁ, M., KRÁLOVÁ, M., MAROŠ, B. *Průvodce základními statistickými metodami.* Praha: Grada Publishing, 2010, 225 pp. 225.

[2] HEBÁK, P. a kol. *Statistické myšlení a nástroje analýzy dat.* Praha: Informatorium, 2013. pp. 25-411.

[3] KÁBA, B., SVATOŠOVÁ, L. *Statistické nástroje ekonomického významu.* Plzeň: Aleš Čeněk, 2012. pp. 93.

[4] KAŇKA, M., MALEC, L. *Učebnice kvantitativních metod.* Praha: Idea Servis, 2012. pp.231.

[5] MELOUN, M., MILITKÝ, J., HILL, M. *Statistická analýza vícerozměrných dat v příkladech.* Praha: Academia, 2012. pp 161-162.

[6] PECÁKOVÁ, I. *Statistika v terénních průzkumech.* Praha: Professional Publishing, 2011. pp 181.

[7] PAVLÍK, J. a kol. *Aplikovaná statistika.* Praha: VŠCHT v Praze, 2005, 2005. p 36.

[8] Program STATISTICA. StatSoft CR, s.r.o. Czech Republic.

Annex no. 1: Table of variables and observations

Korelace (List1 v data_volf_chD)																					
proměnná	**q1**	**q2**	**q3**	**q5**	**q6a**	**q6b**	**q6c**	**q6d**	**q6e**	**q7**	**q8a**	**q8b**	**q8c**	**q8d**	**q8e**	**q9**	**q10**	**q11**
q1	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	-0,017	-0,100
q2	0,000	1,000	-0,025	0,094	-0,025	0,056	0,018	-0,109	0,012	-0,101	0,070	0,099	-0,040	-0,124	0,089	-0,105	-0,034	0,063
q3	0,000	-0,025	1,000	-0,165	-0,062	-0,098	-0,139	-0,051	-0,141	-0,042	-0,137	-0,148	-0,042	0,003	-0,123	-0,114	-0,016	-0,106
q5	0,000	0,094	-0,165	1,000	-0,004	0,150	0,093	0,087	0,243	0,010	-0,179	0,025	0,096	0,168	0,291	-0,063	0,129	0,112
q6a	0,000	-0,025	-0,062	-0,004	1,000	0,733	0,190	0,073	0,062	0,232	0,581	0,353	0,090	0,007	0,008	-0,112	-0,045	0,024
q6b	0,000	0,056	-0,098	0,150	0,733	1,000	0,268	0,019	-0,063	0,131	0,333	0,547	0,202	-0,100	-0,066	-0,056	0,056	0,052
q6c	0,000	0,018	-0,139	0,093	0,190	0,268	1,000	0,121	0,142	0,142	0,098	0,234	0,634	0,038	0,050	-0,099	0,072	0,200
q6d	0,000	-0,109	-0,051	0,087	0,073	0,019	0,121	1,000	0,742	-0,031	0,008	-0,112	0,037	0,618	0,477	-0,083	-0,075	0,111
q6e	0,000	0,012	-0,141	0,243	0,062	-0,063	0,142	0,742	1,000	0,185	0,008	-0,070	0,046	0,451	0,583	-0,070	-0,103	0,183
q7	0,000	-0,101	-0,042	0,010	0,232	0,131	0,142	-0,031	0,185	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,055	0,015
q8a	0,000	0,070	-0,137	-0,179	0,581	0,333	0,098	0,008	0,008	0,000	1,000	0,608	0,155	0,013	0,014	0,014	-0,193	-0,120	0,182
q8b	0,000	0,099	-0,148	0,025	0,353	0,547	0,234	-0,112	-0,070	0,000	0,608	1,000	0,369	-0,182	-0,120	-0,120	-0,103	-0,048	0,120
q8c	0,000	-0,040	-0,042	0,096	0,090	0,202	0,634	0,037	0,046	0,000	0,155	0,369	1,000	0,060	0,078	0,078	-0,157	0,127	0,252
q8d	0,000	-0,124	0,003	0,168	0,007	-0,100	0,038	0,618	0,451	0,000	0,013	-0,182	0,060	1,000	0,773	0,773	-0,135	-0,131	0,154
q8e	0,000	0,089	-0,123	0,291	0,008	-0,066	0,050	0,477	0,583	0,000	0,014	-0,120	0,078	0,773	1,000	1,000	-0,119	-0,049	0,214
q8e	0,000	0,089	-0,123	0,291	0,008	-0,066	0,050	0,477	0,583	0,000	0,014	-0,120	0,078	0,773	1,000	1,000	-0,119	-0,049	0,214
q9	0,000	-0,105	-0,114	-0,063	-0,112	-0,056	-0,099	-0,083	-0,070	0,000	-0,193	-0,103	-0,157	-0,135	-0,119	-0,119	1,000	0,058	-0,189
q10	-0,017	-0,034	-0,016	0,129	-0,045	0,056	0,072	-0,075	-0,103	0,055	-0,120	-0,048	0,127	-0,131	-0,049	-0,049	0,058	1,000	-0,174
q11	-0,100	0,063	-0,106	0,112	0,024	0,052	0,200	0,111	0,183	0,015	0,182	0,120	0,252	0,154	0,214	0,214	-0,189	-0,174	1,000