Minimizing the influence of coronavirus in a built environment

MICROBE

REGRESSION AND CORRELATION ANALYSIS, MULTICRITERIA CALCULATIONS, REGRESSION MODELS OF JOY, ANGER, SADNESS AND VALENCE

Project No: 2020-1-LT01-KA203-078100

Contents

1. Introducing lacking data 3
   1.1. Checking the internal compatibility of a set of variables 3
   1.2. Correlation analysis of the average Happiness index and related variables 6
   1.3. Average Happiness index and related variables regression-based analysis 7
2. The multiple linear regression model of valence 8
   2.1. Data transformation and normalisation 11
   2.2. The multiple linear regression model of arousal 14
References 15

1. Introducing lacking data

The reliability of the results of this analysis declines due to an insufficient set of data. Nonetheless, since the data under analysis represents economic and social categories, it is possible to ascertain their interrelationships. The lacking data relates to the characteristics of certain events that happened in the past; there are no records of their values (or weights). It is possible to supplement the lacking data by use of REGEM – the Regression-based imputation method (Junninen, Niska, Tuppurainen, Ruuskanen & Kolehmainen 2004).

In this case, there are n-1 x and y pairs of variables (x_i, y_i), where i=1, ..., n-1; the value of x_n is also known. The task here is to discover the value of y_n. It is expedient to employ the set of data {x_i; y_i} for the following regression-based equation:

$$y_n = \beta_0 + \beta_1 \cdot x_n + \varepsilon$$

where ε is a random number taken from very many values that distribute according to the Law of normal distribution, N (0, σ2), where 0 is the mean, and σ^2 is the residual dispersal of variable x, when there are n values of variable x. Pairs of (x_i, y_i) are also used for calculating the value (y_{n+1}) of the following variable; however, here i=1, ..., n). All the lacking values for all variable are also completed.

1.1. Checking the internal compatibility of a set of variables

The Cronbach's alpha coefficient is used to check the internal consistency of the variables used in the study. Its foundation is the correlation of the variables used in the study. This coefficient permits checking the number of needed variables and evaluating the suitability of very many variables for reflecting the size under analysis. As Boscarino, Figley and Adams (2004) have pointed out, when the dispersion of all the many values of the variables is close to the dispersion sum of each variable used in the study, the variables are not consistent and do not represent the same subject. In such an instance, the random variables make up the very many variables, and the Cronbach's alpha coefficient is close to 0. Whenever the dispersion sum of different variables is markedly lower than the dispersion of very many values used in the study, the different variables reflect the same subject. If the Cronbach's alpha scales under examination are greater than 0.7, the scale is considered reliable. Assuming the calculation of Cronbach's alpha for the variables used in the study gives a value of 0.792, it means all the variables must be used in the study to assure their maximal, internal compatibility except for the two mentioned previously. Then the variables reflect the constructs adequately. The Cronbach's alpha value lessens upon the elimination of any one of the variables from the study.

Data transformation

The data must be transformed, because the scale measuring the values of the applied variables differs. This way, their interrelationships can be expressed more accurately. The transformation is performed by applying maximal variable values, $x \rightarrow y = \frac{x}{x_{max}}$ (OECD 2008).

Correlation analysis

Table 1 displays the results of the correlation analysis.

Table 1. Correlation analysis results

	HAPPY	GDPC	GDPG	GDPCPPP	INFL_G	UNE	LABP	PD	EBDR
HAPPY	1								

GDPC	.729**	1							
GDPG	-0.029	-0.108	1						
GDPCPPP	-.188*	-0.072	0.046	1					
INFL_G	-0.071	-0.005	-.771**	-0.012	1				
UNE	-.276**	-.215*	-.447**	0.023	.402**	1			
LABP	.702**	.864**	-0.124	-0.131	-0.063	-.173*	1		
PD	-0.062	0.054	-.324**	-0.1	.291**	.279**	0.02	1	
EBDR	-.677**	-.568**	-0.116	.225**	.176*	.200*	-.628**	0.041	1
CPI	.738**	.823**	0.002	-0.125	-0.121	-.182*	.759**	0.093	-.720**
HDI	.795**	.713**	-0.14	-.255**	-0.011	-0.121	.803**	0.076	-.814**
GII	-.751**	-.706**	0.027	.209*	0.065	0.101	-.748**	-0.065	.807**
SPI	.799**	.712**	-0.105	-.292**	-0.012	-0.102	.735**	0.137	-.812**
EI	.739**	.638**	-0.147	-.174*	0.005	-0.086	.693**	0.021	-.805**
TV_SRV	.447**	.517**	0.022	-0.09	-0.118	-0.148	.521**	-0.009	-.664**
SV_SEV	.666**	.735**	-0.139	-0.053	0.067	-0.068	.605**	0.151	-.373**
EPI	.700**	.631**	-0.156	-.209*	0.034	-0.004	.603**	.231**	-.582**
EFPC	0.164	0.137	0.021	0.088	-0.004	-0.03	0.097	-0.129	-0.154
QOL	.754**	.760**	0.004	-0.137	-0.134	-.178*	.733**	0.126	-.702**
GNPC	.702**	.853**	-0.116	-0.107	-0.023	-.228**	.911**	-0.018	-.589**
GNPG	-0.166	-0.106	.401**	0.049	-0.15	-.285**	-.229**	-.251**	0.036
EMP	0.116	0.07	-0.1	-0.106	0.004	.226**	0.08	0.164	-0.093
FC	-.585**	-.489**	-0.051	.195*	0.09	0.028	-.362**	-0.15	.498**
RD	.206*	.320**	0.022	-0.094	-0.044	-.223**	.284**	0.109	-.292**
PS	.637**	.610**	0.061	-0.125	-0.099	-.266**	.565**	-0.009	-.639**
EFPC	.595**	.590**	.228**	-0.114	-.265**	-.234**	.552**	-0.118	-.723**
DI	.639**	.581**	0.068	-.177*	-0.105	-0.064	.427**	0.119	-.569**

	CPI	HDI	GII	SPI	EI	TV_SRV	SV_SEV	EPI	EFPC	QOL
CPI	1									
HDI	.766**	1								
GII	-.761**	-.884**	1							
SPI	.817**	.958**	-.871**	1						
EI	.721**	.955**	-.859**	.912**	1					
TV_SRV	.576**	.642**	-.711**	.632**	.665**	1				
SV_SEV	.652**	.507**	-.491**	.536**	.435**	.214*	1			
EPI	.634**	.738**	-.701**	.755**	.706**	.434**	.591**	1		
EFPC	0.149	0.13	-0.113	0.105	.177*	0.085	.181*	0.097	1	
QOL	.929**	.762**	-.767**	.810**	.716**	.572**	.619**	.704**	0.134	1
CPI	.725**	.755**	-.715**	.707**	.652**	.558**	.653**	.588**	0.109	.724**
HDI	-0.129	-0.158	0.129	-0.155	-0.114	-0.114	-.230**	-.201*	0.082	-0.148
GII	0.057	.179*	-0.088	.214*	0.138	-0.037	0.115	.248**	-0.059	0.071
SPI	-.685**	-.561**	.533**	-.706**	-.569**	-.346**	-.457**	-.570**	-0.143	-.718**
EI	.365**	.248**	-.382**	.244**	.231**	.330**	.243**	0.111	.195*	.312**
TV_SRV	.745**	.687**	-.659**	.720**	.665**	.437**	.452**	.548**	.230**	.718**
SV_SEV	.706**	.587**	-.600**	.605**	.549**	.446**	.449**	.406**	0.055	.640**
EPI	.748**	.622**	-.572**	.756**	.622**	.385**	.538**	.585**	0.149	.745**

	GNPC	GNPG	EMP	FC	RD	PS	EFPC	DI
GNPC	1							
GNPG	-.196*	1						
EMP	0.007	-0.168	1					
FC	-.377**	0.062	-.171*	1				
RD	.284**	-0.081	-0.069	-0.123	1			
PS	.553**	0.031	-0.025	-.603**	.330**	1		
EFPC	.530**	-0.045	0.091	-.497**	.279**	.595**	1	
DI	.431**	-0.05	.196*	-.956**	.173*	.619**	.566**	1

HAPPY - Happiness index, 2018

GDPC - GDP per capita, 2018 (or 2017, 2016, 2013, 2011, 2007, 2000)

GDPG - GDP growth (by annual %), 2019 (or 2018, 2017, 2011, 2009, 2007, 2000)

GDPCPPP - GDP per capita in PPP, 2019 (or 2010)
INFL_G - Inflation growth in 2018, 2019 (or 2010)
UNE - Unemployment rate in 2019 (or 2010, 2011, 2017, 2018)
LABP - Labor productivity in 2018
PD - Public debt in 2019 (or 2010, 2018)
EBDR - Ease of doing business ranking, 2018
CPI - Corruption perceptions index, 2018
HDI - Human development index, 2018
GII - Gender inequality index, 2018
SPI - Social progress index, 2019
EI - Education index, 2018
TV_SRV - Traditional values vs. secular-rational values
SV_SEV - Survival values vs. self-expression values
EPI - Environmental performance index, 2018
EFPC - Ecological footprint per capita, 2016
QOL - Quality of life index, 2019
GNPC - GNP per capita PPP in 2018
GNPG - GNP growth in 2017, 2018
EMP – Employment, 2019
FC - Freedom and control, 2017
RD - Religious Diversity Index, 2010
PS - Political stability in 2017
EFPC - Economic freedom in 2019
DI - Democracy Index, 2018
** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

The results of the performed correlation analysis permit concluding that the insignificant links of variable with Happy are GDP growth (GDPG); Inflation growth in 2018, 2019 or 2010 (GDPG); Public debt in 2019 (PD); Ecological footprint per capita in 2016 (EFPC); GNP growth in 2017 or 2018 (GNPG) and Employment in 2019 (EMP). The strongest positive and statistically significant link has been established between Happy and the Social progress index, 2019, (SPI, r=0.799, p<0.01), whereas the weakest – between Happy and the Religious Diversity index, 2010, (RD, r=0.206, p<0.05). The strongest negative relationship is between Happy and Freedom and control, 2017, (FC, r=-0.585, p<0.01), whereas the weakest – between Happy and GDP per capita in PPP, 2019, (GDPCPPP, r=0.188, p<0.05).

Regression-based analysis

Upon performing a regression-based analysis of all the variables that correlate statistically significantly with the dependent variable Happy, it is established that this model is suitable for deliberation (p<0.01). Meanwhile the dispersion of the dependent variable, Happiness index, values explains 79.7 percent of the changes in the independent variables. The compiled regression equation is

$$HAPPY = 0.494 + 0.109 \cdot GDPC + 0.002 \cdot GDPCPPP - 0.158 \cdot UNE - 0.082 \cdot LABP - 0.022 \cdot EBDR - 0.099 \cdot CPI + 0.439 \cdot HDI - 0.140 \cdot GII - 0.005 \cdot SPI - 0.040 \cdot EI - 0.036 \cdot TV_{SRV} + 0.88 \cdot SV_{SEV} + 0.022 \cdot EPI 0.142 \cdot QOL - 0.020 \cdot GNPC - 0.061 \cdot FC - 0.050 \cdot RD - 0.009 \cdot PS + 0.59 \cdot EFPC - 0.047 \cdot DI$$

The results of the performed regression-based analysis show that the greatest negative value the Happiness index has is with the Unemployment rate, whereas the greatest positive value—with the Quality of life index.

1.2. Correlation analysis of the average Happiness index and related variables

Table 2 displays the results of the correlation analysis.

Table 2. Correlation analysis results

	HAPPY_A	GDPC_A	GDPG_A	GDPCPPP_A	INFL_A	UNE_A	LABP_A	PD_A
HAPPY_A	1							
GDPC_A	.684**	1						
GDPG_A	0.05	0.039	1					
GDPCPPP_A	-0.048	-0.05	-0.022	1				
INFL_A	0	-0.053	0.076	-0.009	1			
UNE_A	-0.146	-.204*	-0.123	-0.088	0.101	1		
LABP_A	.622**	.829**	0.016	-0.05	-0.034	-0.157	1	
PD_A	-0.138	0.015	0.011	-0.039	-0.07	0.05	-0.057	1
EDBR_A	-.664**	-.648**	0.059	0.025	.190*	-0.003	-.603**	0.165
CPI_A	.673**	.840**	0.009	-0.016	-0.121	-0.083	.676**	0.009
HDI_A	.745**	.719**	-0.116	-0.023	-0.011	0.028	.723**	-0.08
GII_A	-.662**	-.668**	0.075	0.02	0.075	-0.038	-.562**	0.116
SPI_A	.665**	.679**	-0.036	-0.039	-0.01	-0.015	.596**	-0.01
EI_A	.601**	.557**	-.209*	0.02	-0.018	0.023	.460**	-0.046
EPI_A	.563**	.548**	-0.096	-0.053	0.014	0.044	.489**	-0.064
EFPC_A	0.12	0.096	0.012	-0.044	0.001	-0.007	0.082	-0.096
QOL_A	.676**	.751**	0.008	0.003	-.234**	-0.049	.658**	0.003
GNPC_A	.588**	.823**	0.02	-0.042	-0.035	-.212*	.948**	-0.055
GNPG_A	-.280**	-.253**	0.036	-0.075	-0.041	-0.091	-.185*	-0.074
EMP_A	.192*	0.136	0.028	-0.11	0.023	.236**	0.119	.174*
FC_A	-.606**	-.581**	-0.064	.173*	0.056	-0.076	-.378**	0.009
PS_A	.614**	.659**	0.03	0.031	-0.11	-0.115	.568**	-0.151
EF_A	.608**	.620**	0.01	-.263**	-0.166	-0.077	.539**	-0.108
DI_A	.594**	.577**	0.015	-0.117	-0.067	0.06	.321**	0.08

	EDBR_A	CPI_A	HDI_A	GII_A	SPI_A	EI_A	EPI_A	EFPC_A
EDBR_A	1							
CPI_A	-.745**	1						
HDI_A	-.822**	.734**	1					
GII_A	.762**	-.711**	-.855**	1				
SPI_A	-.791**	.722**	.863**	-.793**	1			
EI_A	-.680**	.599**	.818**	-.760**	.773**	1		
EPI_A	-.706**	.554**	.774**	-.666**	.709**	.619**	1	
EFPC_A	-0.139	0.103	0.123	-0.081	0.116	.187*	0.05	1
QOL_A	-.744**	.835**	.733**	-.679**	.695**	.581**	.551**	-0.008
GNPC_A	-.563**	.642**	.653**	-.528**	.552**	.423**	.423**	0.059
GNPG_A	.230**	-.285**	-.333**	.319**	-.276**	-.275**	-.335**	0.014
EMP_A	-.173*	0.151	.276**	-0.166	.238**	.180*	.363**	-0.011
FC_A	.606**	-.712**	-.596**	.569**	-.639**	-.513**	-.589**	-0.103
PS_A	-.693**	.762**	.689**	-.647**	.695**	.553**	.545**	.179*
EF_A	-.738**	.698**	.602**	-.548**	.599**	.505**	.479**	0.12
DI_A	-.620**	.702**	.618**	-.584**	.672**	.540**	.599**	0.073

	QOL_A	GNPC_A	GNPG_A	EMP_A	FC_A	PS_A	EF_A	DI_A
QOL_A	1							
GNPC_A	.609**	1						
GNPG_A	-.318**	-0.14	1					
EMP_A	0.095	0.042	-0.134	1				
FC_A	-.671**	-.344**	.314**	-.224**	1			
PS_A	.766**	.556**	-.200*	0.036	-.688**	1		
EF_A	.633**	.524**	-.201*	.184*	-.656**	.639**	1	.
DI_A	.655**	.279**	-.316**	.276**	-.885**	.630**	.614**	1

HAPPY_A - Average happiness Index (2014-2019)
GDPC_A - Average GDP per capita (1990-2018)
GDPG_A - Average GDP growth (by annual %) (1990-2019)
GDPCPPP_A - Average GDP per capita in PPP (1990-2019)
INFL_A - Average inflation growth, 1990-2019
UNE_A - Average unemployment rate (1990-2019)
LABP_A - Average labor productivity, 1990-2018
PD_A - Average public debt, 1990-2019
EDBR_A - Average ease of doing business ranking (2006-2019)
CPI_A - Average corruption perceptions index (1995-2018)
HDI_A - Average human development index, (1990-2018)
GII_A - Average Gender inequality index (1990-2018)
SPI_A - Average Social progress index (2014-2019
EI_A - Average Education index (1990-2018)
EPI_A - Average Environmental Performance index (2008, 2010, 2012, 2014, 2016, 2018)
EFPC_A - Average Ecological footprint per capita (1995-2016)
QOL_A - Quality of life index (2012-2019)
GNPC_A - GNP per capita PPP (1990-2018)
GNPG_A - GNP growth (1990-2018)
EMP_A - Employment (1991-2019)
FC_A - Freedom and control (1990-2017)
PS_A - Political stability (1996, 2000, 2005-2017)
EF_A - Economic freedom (1995-2019)
DI_A - Average Democracy Index (2006-2018)
** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

There are statistically insignificant relationships between the Average Happiness index and Average GDP growth, Average GDP per capita in PPP, Average inflation growth (1990-2019), Average unemployment rate (1990-2019), Average public debt (1990-2019) and Average Ecological footprint per capita (1995-2016). The strongest, positive link is with Average Happiness index, and there is a statistically significant relationship with the Average Human Development index ($r=0.745$, $p<0.01$), whereas the weakest – with Employment (1991-2019) ($r=0.192$, $p<0.05$). The strongest negative albeit statistically significant relationship is between Average Happiness index and Average ease of doing business ranking (2006-2019) (EDBR_A, $r=-0.664$, $p<0.01$), whereas the weakest—with GNP growth (1990-2018) (GNPG_A, $r=0.280$, $p<0.001$).

1.3. Average Happiness index and related variables regression-based analysis

Upon performing a regression-based analysis of all the variables correlating statistically significantly with the dependent variable, Average Happiness index, it was established that this model is suitable for deliberation ($p<0.01$). Meanwhile the dispersion of dependent variable, the Average Happiness index, explains 63.9 percent of the changes in the independent variables. The compiled regression equation is

$$\begin{aligned} HAPPY_A = {} & 0{,}451 + 0{,}147 \cdot GDPC\_A - 0{,}11 \cdot LABP\_A - 0{,}002 \cdot EDBR\_A - 0{,}091 \cdot CPI\_A \\ & + 0{,}435 \cdot HDI\_A - 0{,}026 \cdot GII\_A - 0{,}062 \cdot SPI\_A - 0{,}103 \cdot EPI\_A + 0{,}063 \\ & \cdot QOL\_A + 0{,}068 \cdot GNPC\_A - 0{,}011 \cdot GNPG\_A + 0{,}010 \cdot EMP\_A - 0{,}092 \cdot FC\_A \\ & - 0{,}003 \cdot PS\_A + 0{,}148 \cdot EF\_A + 0{,}039 \cdot DI\_A \end{aligned}$$

The Average Human Development index (1990-2018) has the greatest, positive influence on the Average Happiness Index, and Average Employment (1991-2019) has the least amount. The strongest negative influence comes from Average Environmental Performance index, and the weakest, from Political stability.

2. The multiple linear regression model of valence

Filling in missing data

The valence dataset (the dependent variable valence and the independent variables (pollution (SO_2, $KD_{2.5}$, KD_{10}, NO_2, CO, O_3), magnetic storm, average wind speed, interest, boredom, heart and breathing rates) is incomplete and the results of the analysis may, therefore, be distorted. In reference to OECD (2008), the missing data can be categorised as missing not at random (MNAR) data, which means that the missing values depend on the observed results. The missing values are, then, related to the available dataset. This means the values of certain pairs of variables have not been recorded. Polynomial dependence relationships between the observed data can only be established after the missing data has been added. Since the missing data are categorised as MNAR data, all necessary missing values have been added by means of regression imputation (OECD, 2008).

In this case we have $n-1$ pairs of the variables x and y (x_i, y_i), where $i=1$, ..., $n-1$ and the value of x_n is known. We have to find the value of y_n. The following regression equation of the dataset {x_i; y_i} can be used for that purpose:

$$y_n = \beta_0 + \beta_1 \cdot x_n + \varepsilon$$

where ε is a random number $N(0,\sigma^2)$ and σ^2 is the residual variance of the variable x, when the variable x has n values. Pairs (x_i, y_i) are also used to calculate the next value (y_{n+1}) of the variable y, but in this case $i=1$, ..., n. This way all missing values for all variables can be added. Upon performing the data supplementation, the result showed that 11 variables were used for the analysis and that each variable contained 1762 values.

Removing variables

After the valence dataset (the dependent variable valence and the independent variables (pollution (SO_2, $KD_{2.5}$, KD_{10}, NO_2, CO, O_3), magnetic storm, average wind speed, interest, boredom, heart and breathing rates) has been populated with the missing values of the variables, a statistical analysis of the dataset should be performed in order to remove any accidental variables with accidental links to the variables being analysed. For that purpose, factor analysis is used that combines information about the original set of variables to rule out any redundant variables with no negative impact on the quality of the theoretical model.

The results of Anti-image Correlations in Table 3 suggest that the correlations seen in the diagonal are the Measures of Sampling Adequacy (MSA) and are at least 0.5 or above. This means that the observations of all variables are suitable for factor analysis.

Table 3. Suitability of the values of the variables for factor analysis

	VAL	SO2	KD2.5	KD10	NO2	CO	O3	MS	WS	INT	BOR	HR	RPM
VAL	,761[a]												
SO2	,267	,880[a]											
KD2.5	-,026	-,136	,916[a]										
KD10	-,009	-,139	-,252	,888[a]									
NO2	-,004	,020	,010	-,039	,467[a]								
CO	,004	-,145	-,214	-,299	-,031	,892[a]							
O3	,020	-,090	-,172	-,286	,055	-,234	,913[a]						
MS	,323	,165	-,047	-,079	-,020	-,129	-,046	,802[a]					
WS	,286	-,028	-,080	-,014	-,050	-,025	-,018	-,176	,946[a]				
INT	,065	-,002	-,056	-,054	,029	,023	-,027	-,413	-,055	,938[a]			
BOR	,063	,042	,016	-,014	,002	,023	,019	,024	,074	,021	,895[a]		
HR	,051	,006	,014	,030	,003	-,069	,070	,032	,028	,006	,048	,464[a]	
RPM	,456	,443	-,120	-,135	-,012	-,152	-,077	,679	,144	,049	,410	,066	,705[a]

VAL – Valence, MS – Magnetic Storm, WS – Average Wind Speed, INT – Interest, BOR – Boredom, HR – Heart Rate, RPM – Breathing Rate

[a] Measures of Sampling Adequacy (MSA)

Table 4. KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy	0.858
Bartlett's Test of Sphericity	
Approx. Chi-Square	15999.091
df	78
Sig.	0.000

The value of the Kaiser-Meyer-Olkin Measure of Sampling Adequacy criterion in Table 4 must be >0.5 (the actual value is 0.917). The selected factors, then, can explain 91.7% of the variance of the variables. The significance of the Bartlett's Test of Sphericity must be <0.05 (the actual value is <0.0001). The next step is to verify the null hypothesis that the correlation matrix of the variables is an identity matrix, which would indicate that the variables are unrelated to each other. The zero hypothesis has been rejected.

Table 5. Variance explained

Component	Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %
1	5.286	40.659	40.659
2	1.006	29.006	69.665

Table 5 presents what share of the variance of all the variables is explained by the factors. The bigger the cumulative share of the variance of the variables is explained by the selected factors, the more successful is the factor analysis. The proper value of the first factor is 5.286 (this factor explains 40.659% of the variance of the variables) and the proper value of the second factor is 1.006 (this factor explains 29.006% of the variance of the variables), and together the two factors explain 69.665% of the variance of the variables.

Table 6. Composition of components

	Component	
	1	2
VAL	-.602	
SO2	.677	
KD2.5		.841
KD10		.875
NO2		.769
CO		.867
O3		.844
MS		.925
WS		.745
INT	.828	
BOR	.796	
HR	.626	
RPM	-.953	

VAL – Valence, MS – Magnetic Storm, WS – Average Wind Speed, INT – Interest, BOR – Boredom, HR – Heart Rate, RPM – Breathing Rate
Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
[a] Rotation converged in 4 iterations.

Table 6 is practically the main result of the analysis, if all other conditions (no unrelated variables, etc.) have been met. The factors have been rotated to make the smallest correlations between the variables and the unrotated factors become smaller and the biggest ones become bigger. The first factor comprises Valence, SO2, Interest, Boredom, Heart Rate and RPM, whereas the second factor comprises KD2.5, KD10, NO2, CO, O3, Magnetic Storm and Average Wind Speed. The first factor may be called the emotional factor and the second factor may be called the environmental factor.

2.1. Data transformation and normalisation

Different measuring scales of the values of the variables mean that the data need transformation to make the expression of their interrelationships more accurate. Another important step is data normalisation, that adjusts the range of variables values and the units of measurement, because the variables can be presented in different units or scales. Transformation of Johnson is the best choice, as suggested by García-Sánchez, das Neves Almeida and de Barros Camara (2015). Yeo and Johnson (2000) argue that this method is the most efficient data normalisation method and the only normalisation method that can deal with negative values. Another benefit is that this method makes normal distribution more symmetrical and, thus, improves the quality of analysis.

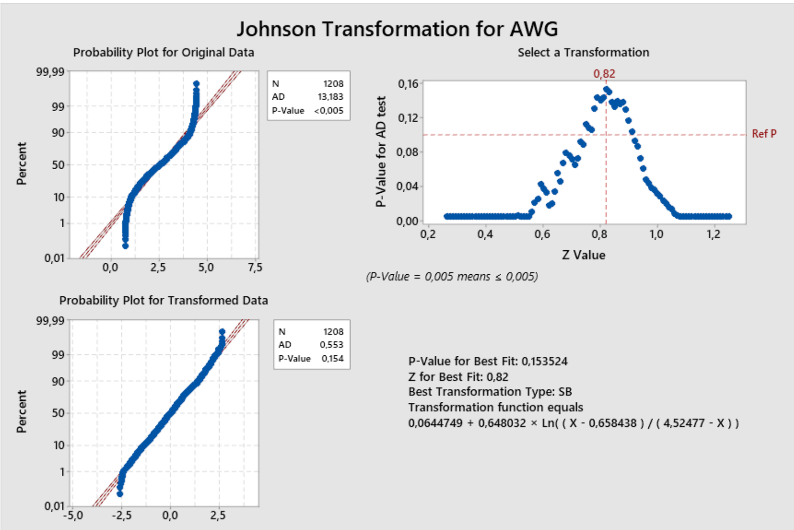

Fig. 1. The normal probability plot of Average Wind Speed before and after normalisation

We used Minitab v. 19.1.1 (64 bit) to normalise our data. The normality of the data was checked before their normalisation. The variable Average Wind Speed was our random choice to illustrate the normalisation results. Fig. 1 presents the distribution of the variable's values before and after normalisation.

Correlation

The next step after data normalisation is to make their correlation analysis and determine their Pearson correlation coefficient. Table 7 presents the results of the correlation analysis.

The results of the correlation analysis show a significant correlation between the variable Valence and all other variables being analysed. The variables Valence and KD10 show the strongest correlation ($r=-0,520$, $p<0,01$) and Valence and Heart Rate ($r=-0.287$, $p<0,01$) show the weakest correlation.

Regression analysis

The regression analysis shows that the dependence model of the dependent variable Valence on the independent variables is suitable for analysis ($p<0.05$) and the variations of all the independent variables included in the model explain 36.3% of the variance of the dependent variable. The following regression equation was generated:

$$VAL = 2{,}208 - 1{,}969 \cdot SO2 - 0.093 \cdot KD25 - 0.467 \cdot KD10 - 0.085 \cdot NO2 - 0.352 \cdot CO - 0.230 \cdot O3 - 0.068 \cdot MS - 0.084 \cdot WS + 0.050 \cdot INT - 0.027 \cdot BOR + 0.034 \cdot \mathrm{HR} - 0.024 \cdot RPM$$

The results of the regression analysis show that SO_2, $KD_{2.5}$, KD_{10}, NO_2, CO, O_3, Magnetic Storm and Interest are the variables with the variation of their values making the biggest impact on the variation of the variable Valence ($p<0.05$). Although Average Wind Speed, Boredom, Heart Rate and RPM make a certain impact on Valence, this impact is not significant ($p>0.05$).

Table 7. The results of the correlation analysis (N=1762)

	VAL	SO2	KD2.5	KD10	NO2	CO	O3	MS	WS	INT	BOR	HR	RPM
VAL	1												
SO2	-,519**	1											
KD2.5	-,435**	,615**	1										
KD10	-,520**	,732**	,586**	1									
NO2	-,327**	,420*	,063	,400**	1								
CO	-,507**	,719**	,576**	,698**	,403**	1							
O3	-,516**	,730**	,601**	,716**	,438**	,708**	1						
MS	-,388**	,493*	,397*	,498**	,001	,484**	,477**	1					
WS	-,484**	,682**	,581**	,675**	,424**	,655**	,676**	,490**	1				
INT	,436**	-,569**	-,457**	-,563**	-,300**	-,547**	-,554**	-,407**	-,515**	1			
BOR	-,390**	,553**	,454**	,536**	,332*	,524**	,531**	,436**	,436**	,499**	1		
HR	-,287**	,460**	,370*	,448**	,276	,440**	,452**	,019	,319**	,184*	,040	1	
RPM	-,344**	,474**	,386**	,469**	,299*	,472**	,468**	,313*	,313**	,458**	,347**	,202**	1

VAL – Valence, MS – Magnetic Storm, WS – Average Wind Speed, INT – Interest, BOR – Boredom, HR – Heart Rate

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

2.2. The multiple linear regression model of arousal

Fig. 1 shows that an employee's productivity depends on his or her arousal. The below arousal equation was obtained based on about 28 million depersonalised data points we have collected. The correlations between arousal and the independent variables were analysed and Table 8 presents the results of this analysis.

Table 8. The results of the correlation analysis of arousal and the independent variables

	Arousal	
	r	p
SO_2	0.698**	0.000
$KD_{2.5}$	0.614**	0.000
KD_{10}	0.566**	0.000
NO_2	0.669**	0.000
CO	0.444**	0.008
O_3	0.719**	0.000
Magnetic storms (MS)	-0.501**	0.000
Apparent temperature (AT)	-0.533**	0.000
Atmospheric pressure (AP)	0.689**	0.000

* Correlation is significant at the 0.05 level (2-tailed)

* Correlation is significant at the 0.01 level (2-tailed)

The results of the correlation analysis show that arousal correlates with all independent variables being analysed. All relationships demonstrate average strength and are statistically significant (p<0,05). The relationships with magnetic storms (MS) and apparent temperature (AT) are negative. Arousal, then, goes down as MS and AT values are increasing and vice versa. The strongest correlation links arousal to SO_2 concentration in air (r=0.698, p<0.01), while its correlation with CO concentration in air is the weakest (r=0.444, p<0.01).

A completed regression analysis allowed us to create the linear regression model ANOVA (1):

$$AROUSAL = 0{,}115 + 0{,}174 \cdot SO2 + 0{,}124 \cdot KD25 + 0{,}098 \cdot KD10 + \\ +0{,}325 \cdot NO2 + 0{,}058 \cdot CO + 0{,}208 \cdot O3 - 0{,}131 \cdot MS - 0{,}094 \cdot AT + 0{,}65 \cdot AP \quad (1)$$

where MS are magnetic storms, AT is the apparent temperature and AP is the atmospheric pressure.

Based on the linear regression model, we can state that it is suitable for analysis (p<0,05) and the independent variables included in the model explain 37.2% of the variance of the dependent variable (arousal). We can also state that the atmospheric pressure (AP) makes the biggest impact on arousal: an increase of 1% makes arousal go up by 0.65%. The CO concentration in air, in contrast, makes the lowest impact: an increase of 1% makes arousal go up by only 0.058%.

References

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, *38*(18), 2895-2907.

Boscarino, J., Figley, C., Adams, R. (2004), Compassion Fatigue following the September 11

Terrorist Attacks: A Study of Secondary Trauma among New York City Social Workers, *International Journal of Emergency Mental Health*, 6(2), 1-10.

OECD (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD: OECD Publishing, Paris