Economic Development, Technological Change and Growth

Studying Covariance and Variance Components in the Czech Regions Arrival Tourism Data

Lukáš Malec[1], Antonín Pavlíček[2], Jaroslav Poživil[3]

Abstract: Lots of tourism data vary over time for that reason the quality processing techniques are demanded to relate such a series. This study concerns the covariance, resp. variance analysis of arrivals to regions of the Czech Republic using static multivariate methods of partial least squares (PLS), resp. principal component analysis (PCA). In addition, an analysis based on eigenvalue decomposition of Euclidean similarity matrix is employed. The relation of profiles and the direction of processes using linear trends in annual data are studied between arrivals of non-residents and residents. The results are attempted to connect them with some economic parameters and global events.

Keywords: partial least squares; principal component analysis; relation of non-residents and residents; analysis of trends; arrival data

JEL Classification: L83; C33

1. Introduction

Tourism is a very important part of economy, but also has a direct impact on the social, cultural and educational sectors of the society. Despite of some travelling and in-place tourism negative effects, the Tourism Satellite Account in 2011 indicates that the direct impact of tourism on Gross Domestic Product (GDP)[4] in the Czech Republic was approximately 2.7% (Main indicators of the national

[1] Department of Information Technologies and Analytical Methods, University of Business in Prague, Czech Republic, Address: Spálená 76/14, 110 00, Prague 1, Czech Republic, Corresponding author: Lukas.Malec@vso-praha.eu.
[2] Department of Information Technologies and Analytical Methods, University of Business in Prague. Czech Republic, Address: Spálená 76/14, 110 00, Prague 1, Czech Republic, E-mail: Antonin.Pavlicek@vso-praha.eu.
[3] Department of Information Technologies and Analytical Methods, University of Business in Prague. Czech Republic, Address: Spálená 76/14, 110 00, Prague 1, Czech Republic, E-mail: Jaroslav.Pozivil@vso-praha.eu.
[4] The GDP at purchaser prices is estimated by production method, i.e. based on the sum of the value added and net taxes on products.

economy and tourism in the Czech Republic), and the tourism ratio on total employment was 4.5% (Tourism employment module). The Czech tourism is strongly influenced by its position in the Central Europe and suffers from typical spatial and temporal problems. Firstly, the tourism is very influenced by economic crisis from the second half of 2008 through the end of 2009. After a worldwide increase of about 7% for international tourism arrivals in 2007, the negative trends were registered in the second half of 2008 resulting to an only 2% increase in 2008 (International tourism challenged by deteriorating global economy). The negative effect is also intensified in 2009 due to the outbreak of the H1N1 influenza virus resulting in a worldwide decrease of about 4% in arrivals for the year (World Tourism Barometer interior update). Until now, the international travel demand has been recovering from the late-2000s recession. Secondly, although the arrivals of non-residents and residents are almost equal to the Czech Republic, on average 57% of non-residents at present visit only Prague (Praha), the Czech capital, despite the fact that there are many assumptions to develop tourism in other regions as well, e.g. considering the term of cultural tourism throughout the country (Exploring historical towns of Bohemia, Moravia and Silesia).

The evaluation of the tourism sector potential over individual territories and over time (repeated measures data or panel data) is a difficult task. Many conclusions are in large measure intuitive and sometimes of unpredictable nature. To understand the tourism leading processes, good knowledge is required from the fields such as the volume of tourism, the characteristics of tourism trips, the structure of the tourists, the tourism expenditure and the financial benefits for the territory (Methodological manual for tourism statistics). Also seasonality is a much discussed topic in the Central European tourism (Bender *et al.*, 2005). Although publicly available statistical databases in Europe (e.g. Eurostat, Czech Statistical Office – CSO) cannot provide a complete picture of tourism, they still serve as useful instrument to model situation, relations and trends.

In this study are applied multivariate statistical methods of PLS and principal component analysis to the Czech Republic arrival tourism data in the static sense. The time series data are considered as non-residents and residents and the covariance (using PLS), resp. variance (using PCA) components between, resp. within both data sets are analyzed using various direction vectors, i.e. latent variables – linear combinations of original variables. The detailed study of relations between non-residents and residents sets is a new contribution in tourism. For studying the multivariate linear trends by principal component analysis, a time variable is inserted to both sets. The economic parameters or various events are compared with the processes *a posteriori* to explain many relations and trends. Also the method based on eigenvalue decomposition of Euclidean similarity matrix is introduced (Elmore & Richman, 2001). Instead of similarity of profiles, this method reveals the profiles which are close to each other in a Euclidean sense. The

last mentioned technique belongs to a relatively novel method with a possible wide spectrum of applications.

Knowing the exact multivariate relations in profiles of arrival tourism data as a basic indicator of tourism intensity is a critical task for policy-makers and the whole range of tourism industry. Particularly, the similarity in the progression of arrivals of non-residents and residents for the individual regions illustrate stable conditions in tourism market, especially for hospitality and accommodation services. On the other hand, the different processes denote change in such conditions. In the case of significantly unequal progression in the numbers of non-residents and residents, the market should move to adapt the required concerns, or on the contrary make an effort to change the marketing strategy and gain the desired composition of guests.

Although the basic concept of PLS method originates in the 1975 work of H. Wold, until now, partial least squares methodologies are negligibly adopted in econometrics with seldom publications in tourism. The tourism aimed publications concentrate predominantly on one specific field called structural equation modelling (or its form – PLS path modelling) such as (Assaker *et al.*, in press; Yuksel & Yuksel, 2008; Gardiner *et al.*, 2013; Arteaga *et al.*, 2010). While Assaker *et al.* (in press) studied a countries' destination competitiveness and examined the causal relationships among various parameters such as the economy, infrastructure, environment and tourism based on the set of countries from various databases, the others apply this method to service marketing, economic and behavioural sectors using the data from questionnaire surveys. Those approaches are conceptually distinct from the technique and methodology used in this study. On the other hand, extensive publications are known to apply PCA and other multivariate approaches, separately to various data sets (see e.g. Vajčerová *et al.*, 2012; Bender *et al.*, 2005; Yuksel & Yuksel, 2008).

Partial least squares method (described and applied below) is one of the ranges of approaches to partial least squares techniques, abbreviated PLS-SVD, and called robust canonical analysis or canonical covariance. PLS-SVD approach is closely related to more familiar canonical correlation analysis. The reason why to use PLS method instead of canonical correlation is the singularity of tourism data in this study caused by a small-sample problem (less observations than variables), and also due to collinearity within sets. Although the singularity of data caused by a small-sample problem and consequent computational issues can sufficiently be overcome by the regularization approach originated by Vinod (1976), the collinearity of data violates the statistical assumptions for applying canonical correlation analysis in a similar way as in the multiple regression. According to Wegelin (2000), PLS-SVD method is robust to the collinearity within sets, and also the singularity due to small-sample problem does not make computational issues.

In her work, Lee (2007) demonstrates the case of canonical correlation analysis and PCA first directions are nearly orthogonal to each other. The significant vectors are very different in such a case. We wish to find an interpretation of direction vectors in which the covariance components between sets and variance components within sets are close. In such a particular case, the PLS analysis not only describes the covariance structure between sets, but also the structure which expresses simultaneously the main information (structure of data) within the sets. Those considerations are intended from the point of view of linear relations. In our particular case, the covariance and variance relations are very similar for the first two latent variables based on arrivals to Czech regions.

2. Experimental

2.1. Tourism Data

In this study we investigate data gathered by the Czech Statistical Office covering the arrivals (number of guests in collective accommodation establishments) of non-residents and residents according to NUTS III regions of the Czech Republic (Franke, 2012; CSO database). Note that the children are also included to arrivals of guests. We are interested in annual data despite the fact that all seasonal or monthly data would increase the number of repeated observations. At considering fundamental annual data, the idea of tourism processes has some influence on smoothness in the year cycle and provides a better understanding of the linear trends on this scale. The data studied covers a relatively short time series from 2003 to 2011 (inclusive). At CSO all collective tourist accommodation establishments using the report of questionnaire survey are investigated from the year 2003. Investigation is divided into a monthly examination comprising sample establishments, and a quarterly examination comprising the rest of accommodation establishments. The resulting data are a summary of the submitted questionnaire outcomes and an imputed missing data, i.e. corresponding to collective tourist accommodation establishments that failed to report (Methodology of Czech Statistical Office). Until 2002 (inclusive) the collective accommodation establishments were examined on the basis of a 30% random sample from the Register of accommodation establishments.

Within the EU, effective Directive 95/57/EC of 23 November 1995 on the collection of statistical information is valid in the field of tourism (Council Directive 95/57/EC) for the gathering of statistical data. According to Directive 95/57/EC the EU member states are obligated to offer statistical data covering the capacity and occupancy of collective accommodation establishments and also data on domestic and outbound tourism. On the base of this Directive the tourism data are of high quality and internationally comparable. Council Directive 95/57/EC is revised and updated by Regulation 692/2011 of the European Parliament and of the

Council of 6 July 2011 concerning European statistics on tourism which also takes into account the internationally recommended methodology described in International recommendations for tourism statistics (2008). This Regulation is currently being implemented by EU member states.

2.2. Methodology

The main idea of this study is to reveal relations of profiles between arrivals of non-residents and residents on an annual scale and to find multivariate linear trends by inserting a time variable to PCA. Because the input data are standardized, the relations of profiles cover more specifically the characteristics of together varying without considering the distance term between the original time series (Pinto da Costa *et al.*, 2007). Studied tourism data vary over time and over individual territories, thus they can be considered as repeated measures or panel data. Those data are of a complex nature because they suffer from the case of small-sample problem and also from the collinearity within sets. Our attention is concentrated on PLS, resp. principal component analysis approaches producing latent variables for studying covariance, resp. variance components. The PLS method, instead of the more familiar canonical correlation analysis in the econometric literature, is used because of collinearity within sets (Wegelin, 2000). In our study, regions that are important for explaining the covariances between the two data sets (of non-residents and residents) show higher values of coefficients in the PLS analysis; on the other hand, regions important in within-set relations reveal higher values of coefficients in the principal component analysis. The latent variables which express relations between sets of data (PLS results), and which are simultaneously close to the corresponding latent variables covering relations within sets (PCA results), have in some sense higher weight because in addition to the basic relations between the sets, they also cover principal information of the relations carried by both sets, separately. Also the technique of eigenvalue decomposition of Euclidean similarity matrix is applied to reveal the overlying processes in arrivals to individual regions of the Czech Republic.

The programs are written by the author using MATLAB 7.1 (Mathworks, Natick, MA, USA) software platform. Many built-in functions are used, especially singular and spectral decompositions from the branch of linear algebra.

3. Algorithms

Notation: X and Y are data matrices standardized by columns and measured on n observations of types (n, p) and (n, q), respectively. Furthermore, $rank(X) = \min(n, p)$ and $rank(Y) = \min(n, q)$. The products $X^T X = R_{xx}$

and $Y^TY = R_{yy}$ denote within sets (symmetrical, positive-definite) sample correlation matrices. $X^TY = R_{xy}$ is the between sets sample correlation matrix. $E = (e_{ij})$ denotes (symmetric) Euclidean similarity matrix where all the elements e_{ij} are nonnegative. Vectors are boldface lowercase; scalars and variables are lowercase. The sets (groups) are denoted with uppercase.

The methods used in this study involve a form of optimization tasks and can be expressed as a Rayleigh quotient, or their formulas are very close to this quotient with similar properties of critical points given by eigensystem of a generalized eigenvalue problem (Borga *et al.*, 1997). Various approaches of hypothesis testing on such objects can be implemented for the data. However, since our approach is descriptive in sense of time series analysis, no statistical hypotheses are to be examined here.

Based on what reason the analysis is performed, various similarity matrices and statistical techniques are employed in current multivariate statistical literature other than standard covariance or correlation metrics. We also use one of them, based on eigenvalue decomposition of Euclidean similarity matrix E, easily derived from the standardized Euclidean distance matrix (Elmore & Richman, 2001).[1] All the elements of first eigenvector $\boldsymbol{u}_{\max}$ are nonnegative (event. all nonpositive). Instead of finding direction vectors which study the similar profiles, i.e. relations in the time series (the topic of PLS and principal component analyses), the original variables which are close to each other (overlie) in a Euclidean sense are demonstrated by using the eigenvalues λ and corresponding eigenvectors $\boldsymbol{u}$ of Euclidean similarity matrix. This solves the problem (Harville, 1997, p. 516; Hasan, 2004)

$$\max_{\boldsymbol{u}\neq\boldsymbol{0}} \frac{\boldsymbol{u}^T E\boldsymbol{u}}{\boldsymbol{u}^T\boldsymbol{u}} \quad (1)$$

which corresponds to the eigenvalue decomposition of matrix E. The interpretation of elements of eigenvectors can be considered as taking only the information about magnitudes of grouped similarities (Elmore & Richman, 2001). The positive eigenvalue numbers indicate the significance of corresponding eigenvectors. Since here is no way to introduce the term latent variable by using Euclidean distance metric in eigenvalue decomposition and the corresponding

[1] During the computation of similarities from the standardized Euclidean distances, scaling by maximum Euclidean distance is used. This is not an issue in practical problems where the complete descriptive picture is revealed.

theory, this approach was mentioned before PLS and principal component analysis methods.

We algebraically derive the two-set partial least squares and one-set principal component analysis methods. The PLS is derived in a similar way to the well-known canonical correlation analysis (Krzanowski, 2000, p. 436). Contrary to canonical correlation analysis, in partial least squares, instead of correlations, the covariances of individual latent variables are maximized. Principal component analysis is the basic technique in multivariate processing methods which is based on maximization of the variances of individual latent variables.

3.1. Partial Least Squares

Partial least squares (PLS-SVD) technique is explained as finding vectors (Wegelin, 2000; De Bie *et al.*, 2005), coefficients of linear combinations $(\boldsymbol{u},\boldsymbol{v}) \in R^p \times R^q$, by solving problem

$$\max_{\boldsymbol{u},\boldsymbol{v}\neq\boldsymbol{0}} \frac{\boldsymbol{u}^T R_{xy} \boldsymbol{v}}{\sqrt{\boldsymbol{u}^T \boldsymbol{u}}\sqrt{\boldsymbol{v}^T \boldsymbol{v}}}. \quad (2)$$

Because (2) is a homogenous function in $\boldsymbol{u}$ and $\boldsymbol{v}$, this task is equivalent to finding local extrema of the constrained optimization problem

$$\begin{aligned} &\max_{\boldsymbol{u},\boldsymbol{v}\neq\boldsymbol{0}} \boldsymbol{u}^T R_{xy} \boldsymbol{v} \\ &s.t. \boldsymbol{u}^T \boldsymbol{u} = 1, \boldsymbol{v}^T \boldsymbol{v} = 1 \end{aligned} \quad (3)$$

where the combinations $a = X\boldsymbol{u}$ and $b = Y\boldsymbol{v}$ are called the latent variables – LVs (also called factors or components). Such variables can be used as a dimension reduction tool and as a graphical display.

According to the Kuhn-Tucker theorem (Kuhn & Tucker, 1951), there exist numbers λ_1 and λ_2 (Lagrange multipliers) in such a way that the solution is a stationary point of the corresponding Lagrangian

$$\boldsymbol{u}^T R_{xy} \boldsymbol{v} - \frac{\lambda_1}{2}\left(\boldsymbol{u}^T \boldsymbol{u} - 1\right) - \frac{\lambda_2}{2}\left(\boldsymbol{v}^T \boldsymbol{v} - 1\right). \quad (4)$$

Solving derivatives with respect to $\boldsymbol{u}$ and $\boldsymbol{v}$, the maximization leads to the system

$$\begin{aligned} R_{xy} \boldsymbol{v} &= \lambda_1 \boldsymbol{u} \\ R_{yx} \boldsymbol{u} &= \lambda_2 \boldsymbol{v}. \end{aligned} \quad (5, 6)$$

Since $\boldsymbol{u}^T R_{xy}\boldsymbol{v}$ is a scalar, then according to expression (4), the following is valid: $\lambda_1 = \lambda_2 = \lambda = \boldsymbol{u}^T R_{xy}\boldsymbol{v}$. From Eq. (6) we have

$$\boldsymbol{v} = \frac{1}{\lambda} R_{yx}\boldsymbol{u} \quad (7)$$

and the substitution to Eq. (5) gives

$$R_{xy}R_{yx}\boldsymbol{u} = \lambda^2\boldsymbol{u}. \quad (8)$$

Thus, the solution of optimization task (2) is turned into a generalized eigenvalue problem to find eigenvalue λ and corresponding eigenvector $\boldsymbol{u}$. Because of positive-semidefinite matrix $R_{xy}R_{yx}$, this problem can be solved by singular or spectral decompositions (Harville, 1997, p. 555). Using Eq. (7), we find the vector $\boldsymbol{v}$. Note, the latent variables covariance can be expressed such that $\lambda = \mathrm{cov}(a,b)$.

The higher-order latent variables and covariances are defined identically as in expression (2), but now under additional restriction so that a latent variable of order k, with $1 < k \leq \min(p,q)$, should be uncorrelated with all the lower-order latent variables of the other set.

According to De Bie *et al.* (2005), the PLS-SVD problem (contrary to other PLS methods) can be solved directly by using singular value decomposition of matrix R_{xy}, i.e. $R_{xy} = UDV^T$ where U and V are orthogonal matrices of types (p,p) and (q,q), respectively, and D is a diagonal matrix of not necessarily distinct singular values.

3.2. Principal Component Analysis

Principal component analysis (developed as a one-set method) is explained (Krzanowski, 2000, p. 60) as solving the following term

$$\max_{\boldsymbol{u}\neq\boldsymbol{0}} \frac{\boldsymbol{u}^T R_{xx}\boldsymbol{u}}{\boldsymbol{u}^T\boldsymbol{u}}. \quad (9)$$

Because (9) is a homogenous function in $\boldsymbol{u} \in R^p$, this task is equivalent to finding local extrema of the optimization problem

$$\max_{\boldsymbol{u}\neq\boldsymbol{0}} \boldsymbol{u}^T R_{xx}\boldsymbol{u} \quad (10)$$
$$s.t. \boldsymbol{u}^T\boldsymbol{u} = 1.$$

The combinations $a = Xu$ are called the latent variables (principal components). Equivalently to PLS method, the analytical solution according to the Kuhn-Tucker theorem (Kuhn & Tucker, 1951) can be defined as finding stationary points of the Lagrange function by eigenvalue decomposition of R_{xx} matrix. This problem is equivalent to computing the singular value decomposition of standardized matrix X.

4. Results and Discussion

This study covers the time series data of arrivals to fourteen individual Czech regions (corresponding to the NUTS3 classification used by Eurostat and CSO).[1] The annual profiles are graphically represented in Figure 1 (for non-residents) and Figure 2 (for residents). Those figures can serve as a visual interpretation of the basic similarity in standardized time series (of variables which vary together) between and within sets. The most important results, particularly of PLS analysis, dealing with similarity, are also expressed in both figures by various line types and marks. By visual inspection, a fall in the years 2008 and 2009 is evident (in particular for non-residents) and was influenced by the worldwide economic crisis and outbreak of the H1N1 influenza virus. This situation continued for resident arrivals during the year 2010. The other significant fall in arrivals to the Czech Republic is visible mainly for residents from 2004 to 2006 period. The arrivals are probably influenced by the year 2004 when the Czech Republic joined the European Union. The number of non-residents may have been reduced particularly by global events, such as terroristic attacks (Beslan school, Madrid and London), Indian ocean earthquake (in 2004) and the outbreak of the H5N1 influenza virus (also in 2004). On the other hand, the number of residents was probably influenced by local meteorological events, e.g. 2004 tornado in Litovel (Olomoucký region) and floods in the Czech Republic and a wider part of Central Europe in 2006.

[1] The abbreviations considered here are according to the Czech classification system, i.e. PHA – Praha, STC – Středočeský region, JHC – Jihočeský region, PLK – Plzeňský region, KVK – Karlovarský region, ULK – Ústecký region, LBK – Liberecký region, HKK – Královéhradecký region, PAK – Pardubický region, VYS – Vysočina, JHM – Jihomoravský region, OLK – Olomoucký region, ZLK – Zlínský region, MSK – Moravskoslezský region.

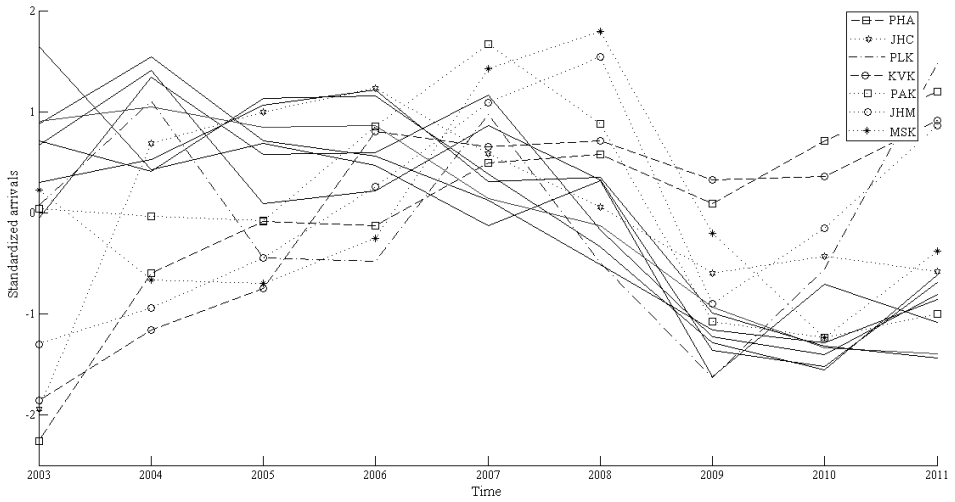

Figure 1. Arrivals of non-residents

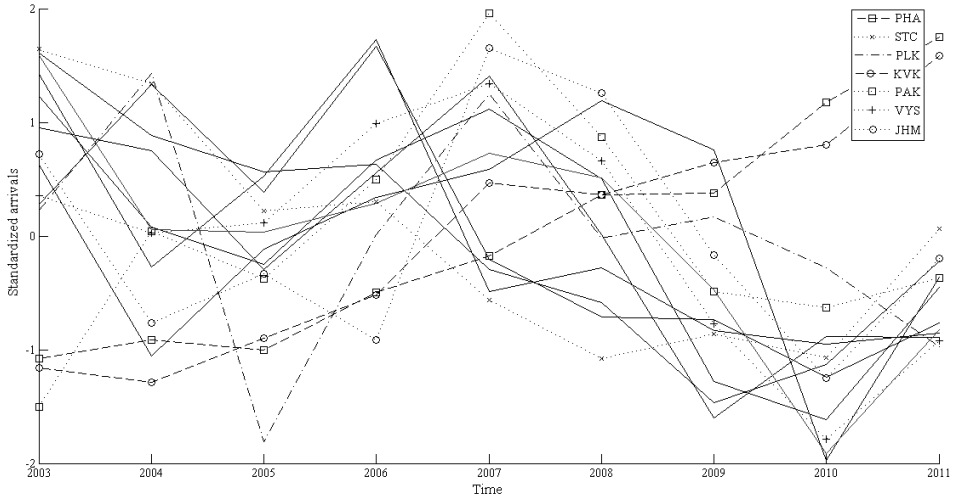

Figure 2. Arrivals of residents

It is important to note that the worldwide economic crisis did not have such a deep negative effect on the tourism economy in comparison to other industry sectors, because, after few years' fall, in 2009 the small increase to 2.9% in tourism direct impact on GDP[1] took place in the Czech Republic relative to a 2.7% impact on GDP in 2011 (Main indicators of the national economy and tourism in the Czech Republic). Also the tourism ratio on total employment (jobs in %) growths during

[1] The GDP at purchaser prices is estimated by production method.

the three consecutive years 2008 (4.48), 2009 (4.56%) and 2010 (4.62), see (Tourism employment module).

As in many real data processing situations, the original data are collinear within sets. The values of Pearson correlation coefficient higher than 0.9 are considered as an indication of collinearity (Griffith & Amrhein, 1997). In the case of non-residents, there occur such magnitudes of correlations at STC and LBK (0.95), ULK and LBK (0.91), ULK and HKK (0.91), LBK and HKK (0.93), and VYS and ZLK (0.92). In the case of residents, we identified correlations exceeding 0.9 at PHA and KVK (0.95), STC and HKK (0.92), VYS and OLK (0.90), and OLK and MSK (0.91). The extent of collinearity within sets of non-residents and residents can also be identified by the values of determinant of the sample correlation matrices, both of orders 10^{-98}. For comparison purposes, the determinant of between set sample correlation matrix is also determined, with a value of order 10^{-99}.

In Table 1, the elements of individual eigenvectors (event. coefficients of latent variables) are introduced together with various measures of their significance for PLS, PCA and analysis using eigenvalue decomposition of Euclidean similarity matrix E. It is important to note that sum of all three LVs eigenvalue ratios (%) for PLS gains a value 85.82, for principal component analysis of non-residents it is 86.84 and of residents it is 85.63. In the one-set case (originated as a union of both sets), the significance of only the first eigenvector is studied by the eigenvalue decomposition of matrix E. The corresponding eigenvalue ratio is 90.58%. The elements of eigenvector are scaled in this case to have unit norms in both sets and the results from the parts are considered as distinct. Because PCA is processed separately for both of the sets, the covariance of combinations of the various latent variables (of maximally 3rd order) is presented in Table 2. It is evident that the directions of maximum covariance differ only at LVs#3. In the following, we discuss particularly the elements of eigenvectors of a magnitude greater than or close to 0.3. These are descriptively marked as significant.

Table 1. Elements of eigenvectors and their significance

	Two-set cases*						One-set case
	PLS			PCA			Euclidean similarity
	LVs#1	LVs#2	LVs#3	LVs#1	LVs#2	LVs#3	
Eigenvalues ratio (%)/ Covariance	54.84/ 7.142	21.78/ 2.836	9.20/ 1.199	53.23(nr),** 52.64(r)/ 7.072	23.81(nr), 23.08(r)/ 2.483	9.80(nr), 9.91(r)/ 0.455	90.58
Non-residents							
PHA	-0.274	0.250	-0.334	-0.232	0.376	-0.245	0.022
STC	0.340	-0.146	-0.127	0.350	-0.098	-0.119	0.280
JHC	0.071	0.117	-0.687	0.129	0.317	-0.614	0.281
PLK	0.090	0.069	-0.053	0.104	0.143	0.109	0.278
KVK	-0.206	0.357	-0.230	-0.207	0.421	-0.110	0.278
ULK	0.341	-0.021	-0.080	0.338	0.039	-0.122	0.277
LBK	0.357	-0.074	-0.072	0.358	-0.065	-0.085	0.280
HKK	0.325	0.051	-0.153	0.334	0.088	-0.153	0.281
PAK	0.268	0.423	0.041	0.255	0.340	0.198	0.272
VYS	0.309	0.127	-0.230	0.330	0.163	-0.139	0.273
JHM	-0.074	0.483	-0.020	-0.066	0.510	0.166	0.280
OLK	0.325	-0.099	0.158	0.318	-0.113	0.187	0.275
ZLK	0.335	0.127	0.030	0.337	0.098	0.144	0.274
MSK	0.121	0.555	0.478	0.087	0.337	0.582	0.276
Residents							
PHA	-0.345	0.169	0.063	-0.316	0.137	0.015	0.266
STC	0.267	-0.332	0.121	0.256	-0.321	-0.011	0.265
JHC	0.307	-0.098	-0.474	0.284	-0.218	-0.221	0.256
PLK	0.113	0.113	0.108	0.131	0.178	-0.594	0.273
KVK	-0.330	0.274	0.054	-0.291	0.266	0.059	0.271
ULK	0.106	0.273	0.330	0.166	0.341	0.458	0.272
LBK	0.274	-0.124	0.015	0.288	-0.181	0.255	0.267
HKK	0.326	-0.198	0.039	0.310	-0.246	0.061	0.259
PAK	0.070	0.556	-0.520	0.065	0.425	-0.377	0.273
VYS	0.300	0.326	-0.164	0.313	0.256	-0.013	0.273
JHM	0.100	0.367	0.450	0.137	0.455	0.209	0.255
OLK	0.310	0.225	0.132	0.333	0.183	0.044	0.273
ZLK	0.325	0.162	-0.047	0.319	0.105	-0.285	0.270
MSK	0.307	0.091	0.336	0.335	0.145	0.218	0.268

* note that two-set cases are solved exclusively using sample correlation matrix

** nr means non-residents, r means residents

Table 2. Covariance matrix of latent variables

Residents / Non-residents	LVs#1	LVs#2	LVs#3
LVs#1	7.072	-0.544	-0.387
LVs#2	-0.059	2.483	-0.478
LVs#3	0.629	1.083	0.455

The analysis of first latent variables of PLS reveals that the Praha, Středočeský, Karlovarský, Liberecký, Královéhradecký, Vysočina, Olomoucký and Zlínský regions prove the related profiles for non-residents and for residents (to its counterparts from the other sets), although Praha and Karlovarský regions demonstrate opposite profiles to the others and have lower values of coefficients for non-residents (see Figs. 1 and 2). On the other hand, the Ústecký region (non-residents) and the Jihočeský together with Moravskoslezský regions (residents) have similar profiles to the others from the second sets. At second LVs are significant the Karlovarský, Pardubický and Jihomoravský regions (for both non-residents and residents), as well as the Moravskoslezský region (non-residents), and the Ústecký together with Vysočina regions (residents), which in some sense prove similar profiles to the others from the second sets. The Středočeský region (residents) proves an opposite profile to the others mentioned. Third LVs describe the similar profiles between sets of non-residents and residents for the Jihočeský and Moravskoslezský regions. The Praha region (non-residents) and also the Ústecký, Pardubický and Jihomoravský regions (residents) prove significantly similar (event. opposite) profiles to the others from the second sets. The Moravskoslezský region (at both sets) as well as the Ústecký and Jihomoravský regions (residents) prove opposite profiles to the others mentioned at LVs#3.

The PCA analysis in comparison to partial least squares reveals a very important property of data as the variance directions within both sets, separately, are almost of the same pattern as the direct covariance directions between sets.[1] This is valid especially for the first and second latent variables. For the LVs#1, again the Praha and Karlovarský regions prove opposite behaviour to others and demonstrate lower values of coefficients for non-residents, which case is intensified in the Praha region contrary to PLS results. Second LVs can be described by the similar profiles of the Pardubický and Jihomoravský regions (as well as in case of PLS) for non-residents and residents. Also the similarity is revealed in the Praha, Jihočeský, Karlovarský and Moravskoslezský regions (non-residents) as well as the Ústecký region (residents) to the others from the second sets. The Středočeský region (residents) proves the opposite behaviour to the others in accordance to the PLS results. LVs#3 describe the similar (event. opposite) profiles of the Jihočeský and Moravskoslezský regions (non-residents) as well as the Plzeňský, Ústecký, Pardubický and Zlínský regions (residents) to the others from the second sets.

The similarity pattern between PLS and principal component analysis is not the case also in eigenvalue decomposition of Euclidean similarity matrix. Here, the similarity metric indicate the closeness of variables in space made from a distance

[1] The values of Pearson correlation coefficient between corresponding first-order LVs of PLS and principal component analysis are close to the value 0.999 for both sets. At the second LVs, the correlations for non-residents is approximately 0.974 and for residents is 0.987; at the third LVs, the approximate values of correlations are 0.853 for non-residents and 0.647 for residents.

standardized by corresponding standard deviations of individual observations. Because in this approach the variables are left unstandardized, the relations are dominated by the large values of such variables. In this case, the Praha region (non-residents) only appears dissimilar to the others in the one-set case.

Latent variables are a time series just like the original time series they are obtained from. They can be considered as directions that concentrate information contained in all the $(p+q)$ time series (specifically, in our case $p=q$). We can see from Figures 1 and 2 that the first LVs of PLS and also of PCA separate only increasing profiles from the decreasing ones according to the signs of elements of the eigenvectors. On the contrary, the second and third latent variables approximately demonstrate first the roughly increasing trends, and then the decreasing ones in profiles (the more complicated processes in time), thus regions greatly influenced by various events and the worldwide economic crisis. In some sense, a nonstandard profile is revealed in the Plzeňský region, which proves very low values of coefficients in the first three LVs of PLS and also of the principal component analysis. The only exception is the case of residents, where by inspection of Table 1 third latent variable at PCA proves similarity only within sets.

The linear trends in data by PCA are also studied using a time variable inserted to the sets of non-residents as well as the residents, separately (see Table 3). Also the influence of the economic crisis is emphasized by the independent analyses of whole range data and the data dealing with the period from 2003 to 2008. Only the first LVs are studied which prove, in 2003 – 2011, a percent proportion of 55.51 and a time coefficient of 0.328 for non-residents, and for residents a percent proportion of 54.61 and a time coefficient of 0.322. In the case of the 2003 – 2008 period, a percent proportion of 51.19 and a time coefficient of 0.357 is revealed for non-residents, and for residents the percent proportion is 50.62 and time coefficient is 0.334.

Table 3. Linear trends – elements of eigenvectors

PHA	STC	JHC	PLK	KVK	ULK	LBK	HKK	PAK	VYS	JHM	OLK	ZLK	MSK
Non-residents (2003-2011)													
0.236	-0.333	-0.107	-0.088	0.214	-0.317	-0.341	-0.312	-0.227	-0.303	0.085	-0.305	-0.314	-0.070
Residents													
0.309	-0.253	-0.275	-0.120	0.290	-0.144	-0.276	-0.301	-0.044	-0.285	-0.115	-0.306	-0.298	-0.312
Non-residents (2003-2008)													
0.307	-0.322	0.129	-0.053	0.325	-0.207	-0.334	-0.060	0.303	-0.033	0.360	-0.272	-0.099	0.294
Residents													
0.337	-0.331	-0.252	0.070	0.361	0.251	-0.192	-0.332	0.300	0.272	0.275	0.122	0.045	0.015

We can see that the Praha and Karlovarský regions predominantly grew regarding number of guests. Particularly from 2003 to 2011 for non-residents, the Středočeský, Ústecký, Liberecký, Královéhradecký, Vysočina, Olomoucký and Zlínský regions decreased in arrivals. In the case of residents, the increase in arrivals to the Praha and Karlovarský regions is more significant in comparison to non-residents and the significant decrease is visible in the Jihočeský, Liberecký, Královéhradecký, Vysočina, Olomoucký, Zlínský and Moravskoslezský regions. In the case of data dealing with the 2003 – 2008 period, greater increases are detected in arrivals in addition to the Praha and Karlovarský regions. Those increases are achieved in the Pardubický, Jihomoravský and Moravskoslezský regions for non-residents. In the case of residents, the Pardubický, Vysočina and Jihomoravský regions also reveal a greater increase of arrivals.

Figures 3 and 4 demonstrate the processes of the first two latent variables in time. Lines are made in the sequence as maintain the time axis direction, i.e. roughly from left to the right in the course of nine consecutive years, i.e. from 3003 to 2011 for both, PLS and PCA methods. In accordance with the property that strong covariance directions between sets are almost of the same pattern as variance directions within sets on first LVs, but also on second latent variables, the data show proximity of latent variables for both sets (non-residents and residents), separately. In the case of PLS as well as in the case of principal component analysis, the doubles of points corresponding to years 2007 and 2008, and also 2010 and 2011, separately for both analyses, are close. For the other years, the results are more distinct for PCA method. In the case of close points in PLS and also in principal component analysis, mutually more similar processes are between sets of non-residents and residents with the processes within those sets, separately. Those years form main base of the relations (processes) between and within both sets of data. In PLS and principal component analysis the year 2011 reveals the recovery from the worldwide economic crisis which began in 2008.

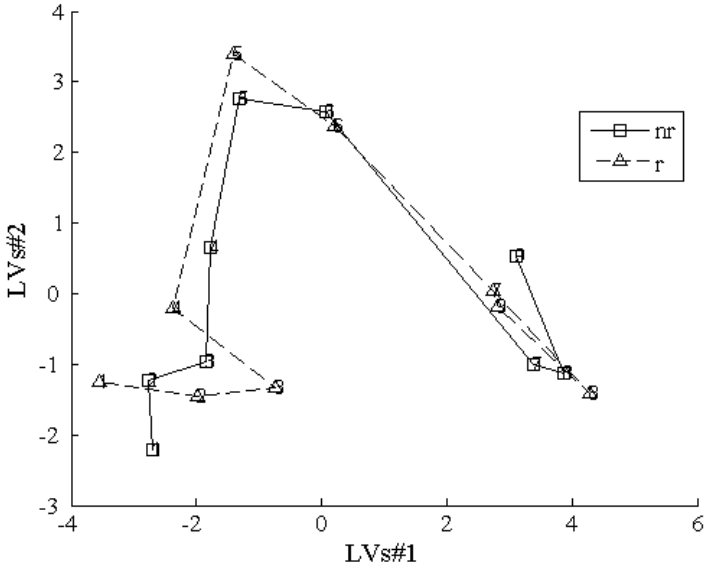

Figure 3. PLS score plot

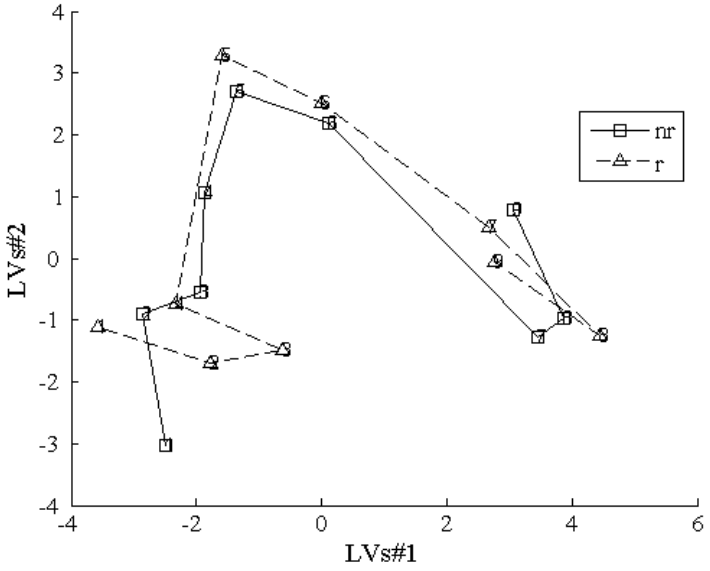

Figure 4. PCA score plot

5. Conclusion

To evaluate the contribution of tourism sector in the economy and in other related branches, high quality procedures for statistical data processing are required. In this study, PLS and principal component analysis methods prove their importance in Czech arrival data. Because the covariance directions (using PLS) between sets of non-residents and residents are simultaneously close to the variance directions (using PCA) within sets, particularly at first two LVs, the covariance directions also carry a great amount of principal information within both sets. This in some sense demonstrates a higher significance of such covariances. It has been found in both analyses that first latent variables describe approximately only increasing and decreasing profiles, while second and third LVs describe the global economic crisis and other events connected to arrivals of guests. The technique of eigenvalue decomposition of Euclidean similarity matrix reveals the Praha region as the most distinct for non-residents. This corresponds to the above mentioned prevalent share of non-residents visiting the Praha region, and also to the ratio of non-residents and residents in total.

From the range of PLS and PCA results, the most important are summarized as:

- By analysis of first LVs of PLS, the Praha, Středočeský, Karlovarský, Liberecký, Královéhradecký, Vysočina, Olomoucký and Zlínský regions have been identified as providing the related profiles for non-residents and residents, although the Praha and Karlovarský regions demonstrate opposite profiles to the others and reveal lower values of coefficients for non-residents. Moreover, the Ústecký region (non-residents) and the Jihočeský together with Moravskoslezský regions (residents) have similar profiles to the others from the second sets.

- The Praha and Karlovarský regions are characterized by a predominant growth in arrivals in the whole data set and the data excluding the late-2000s recession. In comparison to the prevalent directions of decrease in whole period studied, in the Pardubický, Jihomoravský and Moravskoslezský regions (non-residents), and the Pardubický, Vysočina and Jihomoravský regions (residents), changes in direction are identified at the 2003 – 2008 period, which indicates the great impact of the worldwide economic crisis and consequent related processes.

- Inspecting points of the first two LVs of PLS and principal component analysis, the years 2007 and 2008, and also 2010 and 2011 are close where the mutually more similar processes between sets of non-residents and residents with the processes within those sets, separately, are identified. The year 2011 reveals the recovery from the worldwide economic crisis.

6. Acknowledgement

The financial support of the University of Business in Prague internal grant FRV No. 3/2013 is being acknowledged.

7. References

Arteaga, F., Gallarza, M.G. & Gil, I. (2010). A new multiblock PLS based method to estimate causal models: Application to the post-consumption behavior in tourism. In Vinzi, V.E., Chin, W.W., Henseler, J. & Wang, H. (Eds). *Handbook of Partial Least Squares: Concepts, Methods, and Applications* (pp. 141-69). Springer, Berlin.

Assaker, G., Hallak, R., Vinzi, V.E. & O'Connor, P (in press). An empirical operationalization of countries' destination competitiveness using partial least squares modelling. *Journal of Travel Research*.

Bender, O., Schumacher, K.P. & Stein, D. (2005). Measuring seasonality in Central Europe's tourism – how and for what?. In Schrenk, M. (Ed.). *CORP 2005 & Geomultimedia05, 10th International Conference on Information & Communication Technologies (ICT) in Urban Planning and Spatial Development and Impacts of ICT on Physical Space* (pp. 303-9). Wien.

Borga, M., Landelius, T. & Knutsson, H. (1997). A unified approach to PCA, PLS, MLR and CCA. *Report LiTH-ISY-R-1992, ISY, SE-581 83*. Sweden, Linkoping.

Council Directive 95/57/EC of 23 November 1995 on the collection of statistical information in the field of tourism. URL:

http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:1995L0057:20081211:EN:PDF. Retrieved 11.8.2013.

CSO database. URL:

http://www.czso.cz/eng/redakce.nsf/i/cru_ts. Retrieved 9.8.2013.

De Bie, T., Cristianini, N. & Rosipal, R. (2005). Eigenproblems in pattern recognition. In Bayro-Corrochano, E. (Ed.). *Handbook of Geometric Computing: Applications in Pattern Recognition, Computer Vision, Neural Computing, and Robotics* (pp. 1-39). Springer-Verlag, Heidelberg.

Elmore, K.L. & Richman, R. (2001). Euclidean distance as a similarity metric for principal component analysis. *Monthly Weather Review*, 129(3): 540-9.

Exploring historical towns of Bohemia, Moravia and Silesia. The Association of Historical Settlements in Bohemia, Moravia and Silesia, EU project. URL:

http://www.shscms.cz/en/eu-project-uexploring-historical-towns-of-bohemia-moravia-and-silesiau-2232.html. Retrieved 17.8.2013.

Franke, A. (2012). *Statistiky cestovního ruchu*. Praha: Wolters Kluwer.

Gardiner, S., King, C. & Grace, D. (2013). Travel decision making: An empirical examination of generational values, attitudes, and intentions. *Journal of Travel Research*, 52(3): 310-24.

Griffith, D.A. & Amrhein, C.G. (1997). *Multivariate statistical analysis for geographers*. NJ: Prentice Hall.

Harville, D.A. (1997). *Matrix algebra from a statistician's perspective*. NY: Springer-Verlag.

Hasan, M.A. (2004). Information criteria for reduced rank canonical correlation analysis. In *Proceedings of the IEEE International Joint Conference on Neural Networks*. Vol 3 (pp. 2215-20). Piscataway, NJ.

International recommendations for tourism statistics (New York/ Madrid, 2008). Series M No 83/ Rev 1. URL:

http://unstats.un.org/unsd/trade/IRTS/IRTS%202008%20unedited.pdf. Retrieved 11.8.2013.

International tourism challenged by deteriorating global economy. UNWTO World Tourism Barometer (World Tourism Organization). January 2009.

Krzanowski, W.J. (2000). *Principles of multivariate analysis: A user's perspective* (2nd ed.). Oxford: University Press.

Kuhn, H.W. & Tucker, A.W. (1951). Nonlinear programming. In Neyman, J. (Ed.). *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (pp. 481-92). Berkeley: University of California Press.

Lee, M.H. (2007). *Continuum direction vectors in high dimensional low sample size data*. Ph.D. thesis. Chapel Hill, US: University of North Carolina.

Main indicators of the national economy and tourism in the Czech Republic. Czech Statistical Office. URL:

http://www.czso.cz/eng/redakce.nsf/i/tsa_main_indicators_of_the_national_economy_and_tourism_in_the_czech_republic. Retrieved 9.8.2013.

Methodological manual for tourism statistics (Version 1.2). EUROSTAT Methodologies and Working Papers. URL:

http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-11-021/EN/KS-RA-11-021-EN.PDF. Retrieved 17.8.2013.

Methodology of Czech Statistical Office. URL:

http://www.czso.cz/eng/redakce.nsf/i/methodology_time_series_tourism. Retrieved 9.8.2013.

Pinto da Costa, J., Silva, I. & Silva, M.E. (2007). Time dependent clustering of time series. *Bulletin of the 56th Session of the International Statistical Institute*. Lisboa, Portugal.

Regulation (EU) No. 692/2011 of the European Parliament and of the Council of 6 July 2011 concerning European statistics on tourism and repealing Council Directive 95/57/EC. URL:

http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:192:0017:0032:EN:PDF. Retrieved 11.8.2013.

Tourism employment module, Czech Statistical Office. URL:

http://www.czso.cz/eng/redakce.nsf/i/tourism_employment_module. Retrieved 9.8.2013.

Vajčerová, I., Šácha, J. & Ryglová, K. (2012). Using principal component analysis for evaluating the quality of a tourist destination. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*. 60(2): 449-58.

Vinod, H.D. (1976). Canonical ridge and econometrics of joint production. *Journal of Econometrics*. 4(2): 147-66.

Wegelin, J.A. (2000). A survey of partial least squares (PLS) methods, with emphasis on two-block case. *Technical Report*. Seattle, US: University of Washington.

Wold, H. (1975). Path models with latent variables: The NIPALS approach. In Blalock, H.M. *et al.* (Eds.). *Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building* (pp. 307-57). NY: Academic.

World Tourism Barometer interior update, UNWTO World Tourism Barometer (World Tourism Organization). August 2010.

Yuksel, F. & Yuksel, A. (2008). Perceived clientelism: Effects on residents' evaluation of municipal services and their intentions for participation in tourism development projects. *Journal of Hospitality & Tourism Research*. 32(2): 187-208.