On applications of the factor analysis in the agricultural research

Využití výsledků faktorové analýzy v zemědělském výzkumu

V. Brabenec, H. Nešetřilová

Czech University of Life Sciences, Prague, Czech Republic

Abstract: The authors give a brief overview of the outcomes of an application of the factor analysis and present results of two applications within the agricultural research. The first application is a study in which relations among 16 variables characterising production of fish in nearly 200 high production (mainly carp) fish ponds in the Czech Republic were explored using the factor analysis metod. In the second case, outcome of a household questionaire survey was analysed using factor analysis to shed light upon the relations among household annual income and expenditures. The paper was supported by the research project Informational and knowledge support of strategic management MSM 6046070904.

Key words: agricultural research, statistical data analysis, application of multivariate statistical methods, factor analysis

Abstrakt: Autoři podávají charakteristiku základních výstupů faktorové analýzy a uvádějí dva řešené příklady použití faktorové analýzy v zemědělském výzkumu. První z nich se týká studie, ve které byly metodou faktorové analýzy hledány vazby mezi produkčními ukazateli charakterizujícími intenzivní chov ryb (především kapra); do studie byly zahrnuty výsledky hospodaření z téměř dvouset produkčních rybníků v České republice. Druhý případ se týká hodnocení vazeb mezi příjmy a výdaji domácností v České republice, podklady pro tuto studii byly získány z dotazníkového šetření domácností. Příspěvek vznikl v rámci výzkumného záměru Informační a znalostní podpora strategického řízení MSM 6046070904.

Klíčová slova: zemědělský výzkum, statistická analýza dat, využití metod vícerozměrné statistické analýzy, faktorová analýza

The authors of this paper have often benefited from application of multivariate statistical methods when trying to address the problems of agricultural research. As applications of these methods have been relatively modest in this area of research untill now, the paper intends to present the basic information on applications of multivariate statistical analysis for decision making in the agrarian sector, namely the factor analysis, and also to present solutions of two practical problems.

Multivariate statistical methods are based on information on simultaneous measurements of several variables on a set of objects. Objectives of the multivariate statistical techniques are, very generally said, inferences on properties and relationships among such variables. Such data could be, in the technical sense, viewed as observations of one multidimensional variable. Suppose we measure k variables on a set of n objects. Such data can be represented by a matrix with n rows and k columns (each row represents an object, each column corresponds to one variable). To characterize this multidimensional variable, *mean vector* (vector of means of the considered variables) is used, *covariance matrix* (containing variances and covariances of the variables) and *correlation matrix* (consisting of correlation coefficients of all pairs of variables). The multivariate statistical methods can be classified into one of the following two groups:

(a) methods of multidimensional statistical sorting and grouping,

(b) methods of correlation structures analysis.

The methods of *multidimensional statistical sorting and grouping* are employed to split objects into homogeneous subgroups. The classification is technically

Supported by the Ministry of Education, Youth and Sports of the Czech Republic (Grant No. MSM 6046070904 – Informational and knowledge support of strategic management).

based on values of a suitable multivariate criterion for a given observation (k-tuples of observations of each considered variable). These techniques suit to situations in which the considered set of objects naturally tends to create distinct groups of similar objects. *Cluster analysis* and *discriminant analysis* are techniques in this group of methods.

The methods of *correlation structures analysis* are designed for studying relations among many variables. Some of the methods can be used as a tool for simplification of the data structure (reduction of the number of variables) without loss of substantial information. Into this group of multivariate methods, there belongs factor analysis, principal components analysis and less frequently used canonical correlation analysis. (Theoretically justified application of these methods imposes certain assumptions on properties of the variables.)

This paper contains two applications of factor analysis. These applications are based on the research carried out by the authors (Brabenec 1979; Brabenec, Šařecová 2001) and on publications of other authors on that topic (e.g. Hebák, Hustopecký 1987; Hebák et al. 2005 and Johnson, Wichternet 1998).

MATERIAL AND METHODS

Factor analysis is a method of variance-covariance analysis which can be successfully applied even in the situations of informationaly heterogeneous or little explored systems of variables with mutual correlations of varying strength. The factor analysis model is based on the assumption that it is possible to describe the variance-covariance structure of the system of variables by a few underlying, unobservable factors. The idea is to group original variables into several classes in such a manner that variables in the same class have strong correlations among themselves and much weaker correlations with variables in different classes. The underlying, unobservable factors are then represented by such classes of strongly intercorrelated observed variables (which, due to strong correlations, contain duplicate information).

Relations between such *common factors* and observed variables are described by *factor loadings* which, in fact, represent correlation coefficients between the original variable and the common factor. Generally, the factor solution is not unique. It is usually recommended to rotate the matrix of factor loadings (this corresponds to an orthogonal transformation) to simplify the resulting structure before interpretation of the common factors, because the orthogonal factors are mutually independent.

Let us consider a system of observed variables $X_1, X_2, \ldots, X_v$. A *factor analysis model* is a system of linear equations which present each of the observed variables as a linear combination of a few unobservable variables $F_1, F_2, \ldots, F_c$ called *common factors* and additional also unobservable variables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_v$ called *specific (or error) factors*

$$\begin{aligned} X_1 &= \mu_1 + a_{11} F_1 + a_{12} F_2 + \ldots + a_{1c} F_c + \varepsilon_1 \\ X_2 &= \mu_2 + a_{21} F_1 + a_{22} F_2 + \ldots + a_{2c} F_c + \varepsilon_2 \\ &\ldots \qquad\qquad \ldots \\ X_v &= \mu_v + a_{v1} F_1 + a_{v2} F_2 + \ldots + a_{vc} F_c + \varepsilon_v \end{aligned} \qquad (1)$$

The coefficient a_{jp} ($j = 1, 2, \ldots, v$; $p = 1, 2, \ldots, c$) is called the *factor loading* of the variable X_j on the factor F_p. The factor loadings a_{jp} are, under the assumptions given bellow, correlation coefficients $r_{X_jF_p}$ between the variable X_j and the factor F_p. We will consider all variables $X_1, X_2, \ldots, X_v$ including the factors $F_1, F_2, \ldots, F_c$ in their standardized form. Then, the covariance matrix which is to be reproduced by the model coincides with the correlation matrix. The *assumptions* imposed usually on the factor model are

$$\begin{aligned} &E(F_p) = 0,\ \operatorname{cov}(F_p, F_q) = 1 \quad \text{for } p = q \\ &\qquad\qquad \operatorname{cov}(F_p, F_q) = 0 \quad \text{for } p \neq q, \quad p, q = 1, 2, \ldots, c \\ &E(\varepsilon_j) = 0,\ \operatorname{cov}(\varepsilon_j, \varepsilon_k) = d_j^2 \ \text{for } j = k \\ &\qquad\qquad \operatorname{cov}(\varepsilon_j, \varepsilon_k) = 0 \quad \text{for } j \neq k, \quad j, k = 1, 2, \ldots, v \end{aligned}$$

and further that common factors $F_1, F_2, \ldots, F_c$ and specific factors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_v$ are independent, i.e.

$$\operatorname{cov}(F_p, \varepsilon_j) = 0 \ \text{ for } \ p = 1, 2, \ldots, c; \ \ j = 1, 2, \ldots, v$$

(orthogonal factor model).

It follows from the model (1) that the relationship between variables $X_1, X_2, \ldots, X_v$ and factors $F_1, F_2, \ldots, F_c$ is linear; this is an important assumption of the traditional factor analysis model. Further, in this model

a_{jp}^2 = the contribution of the factor F_p to the explanation of the variance of the variable X_j

h_j^2 = the *communality* of the variable X_j. This is the portion of the variance of X_j, which can be in the model attributed to the common factors $F_1, \ldots, F_c$, therefore

$h_j^2 = a_{j1}^2 + a_{j2}^2 + \ldots + a_{jc}^2$

for each row j ($j = 1, 2, \ldots, v$)

d_j^2 = the *specific variance* of the variable X_j, it is the proportion of variance of X_j, which is not explained by the common factors $F_1, F_2, \ldots, F_c$

v_p^2 = the contribution of the factor F_p to the explanation of the total variance of $X_1, X_2, \ldots, X_v$ in the model

$v_p^2 = a_{1p}^2 + a_{2p}^2 + \ldots + a_{vp}^2$

for each column p (p = 1, 2, ..., c)

h^2 = the *total communality* of the model. This is the portion of variance of all variables $X_1, X_2, \ldots, X_v$ which is explained by the factor analysis model

$h^2 = h_1^2 + h_2^2 + \ldots + h_v^2$

d^2 = the *total specific variance* of the model. This is the portion of the total variance of $X_1, X_2, \ldots, X_v$ which is not explained by the factor analysis model, $d^2 = d_1^2 + d_2^2 + \ldots + d_v^2$

There is an inherent ambiguity associated with the factor model. It can be shown that all factor loadings obtained from the initial loadings by an orthogonal transformation have the same ability to reproduce the covariance (or correlation) matrix, (see Johnson, Wichern 1998). Such an orthogonal transformation of the factor loadings corresponds to a rigid rotation of the coordinate axes rotation. Therefore, among the possible solutions of the factor model a solution with a simple structure is often seeked to simplify the interpretation; this solution obtained by an orthogonal transformation is usually called *a rotated solution.*

Factor analysis model is an attempt to characterize the structure of covariance/correlation matrix of the data. The model is build in such a way that the importance of the common factors in the model is decreasing. This is to say that the first factor F_1 explains the largest portion of the total observed variance of the variables $X_1, X_2, \ldots, X_v$ and represents the most important class of intercorrelated variables. The second factor F_2 explains the largest portion of the total observed variance not explained by the factor F_1. The last factor F_c included in the final factor analysis model should associate into one class at least 2 variables with significant factor loadings. Significance of a factor loading can be tested analogous to statistical significance of a coefficient of correlation. For the sake of comparability of factor analysis models for samples of different sizes, the limit of significance of a factor loading is often set subjectively (frequently to $|a_{jp}| = 0.5$). Then the class of variables associated by a common factor is constituted by those variables with significant factor loadings.

RESULTS AND DISCUSSION

Factor analysis method can be used as a tool for the evaluation of multiple mutual relations in systems of variables with a little known structure. The authors present results and evaluation of two factor analysis models which are based on their research.

Model Fish Ponds in the Czech Republic

The first model of the factor analysis was based on the data on approx. 200 production fish ponds of the Fish Farmers Association in České Budějovice. For the purpose of building a factor analysis model, 16 variables were considered. Here is the list of variables:

X_1 – location (based on the average annual temperature, scale 1–8)
X_2 – distance between the fish pond and the location of a fishery centre
X_3 – cadastral area (ha)
X_4 – production (carps /ha)
X_5 – carp weight gains (kg/ha)
X_6 – dose of lime and calcite (kg/ha)
X_7 – dose of nutrients P_2O_5 and N per 1 ha
X_8 – dose of organic fertiliser per 1 ha
X_9 – dose of total feedstuff per 1 ha
X_{10} – costs of current repairs per 1 ha
X_{11} – depreciations per 1 ha
X_{12} – labour costs per 1 ha
X_{13} – primary costs per 1 ha
X_{14} – intra-enterprise costs per 1 ha
X_{15} – overhead costs of a fishery centre
X_{16} – total gain of other fish (except the carp) per 1 ha.

The factor analysis for the data was performed and the solution with four factors was considered as adequate for the construction of the factor analysis model. More details are presented in Table 1.

Interpretation of the factor structure in the "Fish Ponds in the Czech Republic" model

Factor model presented in Table 1 contains four common factors F_1, F_2, F_3, F_4 which account for 68.375% of the observed variability of the data. The common factors indicate correlation structure of the set of variables. Factor loadings a_{jp} are correlations between the variables and the factors. In the same class represented by a common factor, there appear variables with significant factor loadings ($|a_{jp}| > 0.50$, significant loadings are in Table 1 printed in bold figures). Coincidence of the factor loadings signs signifies a positive relation between the corresponding pair of variables, when the factor loadings differ in sign, the relation between the corresponding pair of variables is negative. The

Table 1. Factor analysis model for the set of variables "Fish ponds in the Czech Republic"

Variable X_j	Factor loadings a_{jp} of the factor F_p				Communality	Specific variance
	F_1	F_2	F_3	F_4	h_j^2	d_j^2
X_1	–0.17	**0.58**	–0.33	0.21	0.5183	0.4817
X_2	–0.16	–0.27	0.44	0.14	0.3117	0.6883
X_3	–0.16	**0.65**	–0.25	0.03	0.5115	0.4885
X_4	**0.63**	0.30	–0.24	**0.53**	0.8254	0.1746
X_5	0.74	–0.18	–0.42	0.09	0.7645	0.2355
X_6	0.16	–0.35	0.19	0.71	0.6883	0.3117
X_7	0.45	0.25	**0.70**	0.20	0.7950	0.2050
X_8	0.17	**0.74**	–0.03	–0.32	0.6798	0.3202
X_9	**0.67**	–0.39	0.10	–0.42	0.7874	0.2126
X_{10}	–0.21	0.27	**0.59**	–0.42	0.6415	0.3585
X_{11}	–0.16	**–0.58**	–0.25	0.10	0.4345	0.5655
X_{12}	–0.06	–0.34	**–0.70**	–0.29	0.6933	0.3067
X_{13}	**0.71**	**–0.52**	0.07	–0.34	0.8950	0.1050
X_{14}	**0.87**	0.09	0.18	–0.06	0.8010	0.1990
X_{15}	**0.93**	0.05	0.01	0.10	0.8775	0.1225
X_{16}	0.59	0.56	–0.23	–0.03	0.7155	0.2845
Contribution v_p^2	4.2778	2.9584	2.1269	1.5771	h^2 = 10.5402	d^2 = 5.0598
v_p^2 (%)	26.736	18.490	13.293	9.856	h^2 = 68.375	d^2 = 31.625

Source: Brabenec (1979)

importance of individual variables within the same class (factor) can be deduced from the absolute value of factor loadings. Common factors thus represent aggregates of the information contained in the original set of variables, the interpretation of those aggregates is based on the nature of variables appearing in the same class (in the same common factor).

In the discussed model, the strongest common factor F_1 aggregates into the same class seven variables which are positively correlated (signs of their factor loadings coincide), see below.

Factor F_1 Variable	Factor loading
X_{15} – overhead costs of a fishery centre	0.93
X_{14} – intra-enterprise costs per 1 ha	0.87
X_5 – carp weight gains (kg/ha)	0.74
X_{13} – primary costs per 1 ha	0.71
X_9 – dose of total feedstuff per 1 ha	0.67
X_4 – production (carps/ha)	0.63
X_{16} – total gain of other fish (except the carp) per 1 ha	0.59

Considering the factor structure, F_1 can be interpreted as the *factor of the intensity of fishpond farming*. Some of the relations between variables associated in the class of F_1 are more complex, nevertheless it is possible to guess that more intensive farming (i.e. higher fish gains per 1 ha) implies higher primary costs of farming (feedstuff) and higher intra-enterprise costs per 1 ha.

The second strongest common factor F_2 associates a class of six variables (with significant factor loadings a_{jp}) and represents the most significant residual group of mutually correlated variables. The factor F_2 structure is schematically recorded bellow.

Factor F_2 Variable	Factor loading
X_8 – dose of organic fertiliser per 1 ha	0.74
X_3 – cadastral area (ha)	0.65
X_1 – location (based on the temperature)	0.58
X_{11} – depreciations per 1 ha	–0.58
X_{16} – total gain of other fish	0.56
X_{13} – primary costs per 1 ha	–0.52

Within the class of the second common factor, those variables with the same sign of the factor loading are positively correlated and those with the opposite sign of the factor loading are negatively correlated. Thus, there are positive correlations between the dose of fertiliser, the total area of the pond and the location temperature unit (the range. 1 – the warmest and 8 – the coldest). It is difficult to find a simple name for the factor F_2 but the factor probably represents another criterion of the intensity of farming in the fish ponds. The dose of the organic fertiliser is positively related to the pond size because it is conditioned by the technical equipment for the transport and storage of the organic fertilisers and at the same time it is positively related to the variable X_{16} – total gain of other fish.

Less important classes of variables are associated in relatively week common factors F_3 (variables X_7, X_{10}, X_{12}) and F_4 (variables X_4, X_6).

The variable X_2 (the distance between the fish pond and the location of the fishery centre) does not appear significant in any of the common factors, thus there is no significant relation between this variable and other variables included in the model. Elimination of the variable from the set of variables, upon which the model is based, would not cause any informational loss.

Communalities give information on the proportion of variance of each variable which is explained by the factor analysis model. Thus the comparison of communalities h_j^2 of the variables X_j gives implicit information on the importance of each variable within the considered system (from the point of view of relations to other variables). The highest portion of explained variance is show in the final model variables with $h_j^2 > 0.800$, consequently variables X_{13} ($h^2_{13} = 0.8950$), X_{15} ($h^2_{15} = 0.8775$), X_4 ($h_4^2 = 0.8254$) and X_{14} ($h^2_{14} = 0.8010$). The lowest communalities in the extracted model have the variables X_2 ($h_2^2 = 0.3117$) and X_{11} ($h^2_{11} = 0.4345$). The total communality of the considered factor analysis model is $h^2 = 10.5402$, thus the model explains 68.375% of the total variability of all variables (the unexplained part d^2 is 31.625%).

The declining importance of the common factors F_1 through F_4 can be understood by inspection of the contributions v_p^2 of individual factors to the total observed variability, the contribution of the factor F_1 is 26.736%, the contribution of the factor F_2 is 18.490%, the contribution of the factor F_3 is 13.293% and the contribution of the factor F_4 is 9.856%.

The extracted model of the factor analysis contains aggregate information on the structure of the relations between variables representing various aspects of farming on fish production ponds in the Czech Republic and can be used as an objective tool in the fish production management.

Factor analysis model households questionnaire

A survey of households was carried out as a part of the *Statistical Methods in Marketing and Business* course. For the purpose of this survey, a questionnaire was designed containing 13 questions on income and expenditures of a household, 8 out of the13 ques-

Table 2. Factor analysis model for the set of variables "Households questionnaire"

Variable X_j	Factor loadings a_{jp} of the factor F_p		Communality h_j^2	Specific variance d_j^2
	F_1	F_2		
X_1	**0.85**	–0.11	0.735	0.265
X_2	0.40	**0.72**	0.678	0.322
X_3	**0.61**	–0.44	0.566	0.434
X_4	**0.80**	0.30	0.730	0.270
X_5	**0.55**	0.46	0.525	0.475
X_6	**0.59**	**0.66**	0.772	0.228
X_7	–0.33	**–0.69**	0.585	0.415
X_8	**0.73**	**0.56**	0.847	0.152
contribution v_p^2	3.187	2.251	$h^2 = 5.438$	$d^2 = 2.562$
v_p^2 (%)	39.8	28.1	$h^2 = 67.9$	$d^2 = 32.1$

Source: Brabenec, Šařecová (2001)

tions (of quantitative character) were used for the extraction of a factor analysis model. The model was based on 83 responding households with complete questionnaires.

The list of variables used for extraction of the model:

X_1 – annual net money income per household member

X_2 – annual expenditure on food, beverages and tobacco per household member

X_3 – annual expenditure on non-food products per household member

X_4 – annual expenditure on services (specified in the questionnaire) per household member

X_5 – annual expenditures and investments into business per household member

X_6 – size of the place of residence

X_7 – annual home food production (specified in the questionnaire) per household member

X_8 – annual expenditure on recreation and culture (specified in the questionnaire) per household member

A more detailed specification was provided in the questionnaire to clarify the questions to respondents. The two factors solution of the factor analysis model is presented in Table 2.

The factor loadings printed in bold characters are significant.

Interpretation of the factor structure in the "Households questionnaire" model

Factor model solution presented in the Table 2 classified the set of variables into two common factors which associate variables with significant factor loadings and clarify the nature of mutual relations.

The first common factor F_1 is schematically shown bellow, the variables with significant factor loadings are in order according to the magnitude of a_{jp}.

Factor F_1 Variable	Factor loading
X_1 – annual net money income	0.85
X_4 – annual expenditure on services	0.80
X_8 – annual expenditure on recreation and culture	0.73
X_3 – annual expenditure on non-food products	0.61
X_6 – size of the place of residence	0.59
X_5 – annual expenditures and investments into business	0.55

The second common factor F_2 associates into the same class four variables with significant factor loadings. Within this class, the triple of variables X_1, X_2 and X_3 is in negative correlation to the variable X_7 – annual home food production. The more detailed structure of the factor F_2 is presented bellow.

Factor F_2 Variable	Factor loading
X_2 – annual expenditure on food, beverages and tobacco	0.72
X_7 – annual home food production	–0.69
X_6 – size of the place of residence	0.66
X_8 – annual expenditure on recreation and culture	0.56

Inspecting the values of communalities h_j^2 of the variables X_j it is possible to conclude that the discussed factor model explains the highest portions of variance of the variables X_8 (with $h_8^2 = 0.847$), X_6 (with $h_6^2 = 0.772$) and X_1 (with $h_1^2 = 0.735$), while the lowest portion of explained variance has the variables X_5 (with $h_5^2 = 0.525$). The factor model explains 67.9% of the observed variances of the whole set of variables, the strongest common factor F_1 contributes to this by 39.8% (see v_1^2, Table 2) and the second common factor F_2 by 28.1% (see v_2^2, Table 2).

The extracted factor model helps to throw light upon the inner structure of eight variables contained in the household questionnaires, which considered the annual net money income and the annual household expenditures on the most important items in the Czech Republic. The results illustrate the inner relations between the considered variables associated into two common factors F_1 and F_2 fairly well in spite of the fact that the size of the sample was relatively small.

CONCLUSION

The complexity of problems studied to explain an economical or social phenomenon usually requires collecting data on many different aspects of the phenomenon. Methods of the multivariate statistical analysis represent a suitable analytical tool which helps to get an insight into the inner structure of sets of data with many recorded variables.

In particular, when the data are informationaly coherent, the factor analysis method could help to clarify the covariance structure of the set of variables by grouping the variables into classes of variables called common factors. Variables in the same class have high correlations among themselves while for

variables in different classes the mutual correlations are week.

REFEENCES

Brabenec V. (1979): Uplatnění faktorové analýzy v řešených příkladech (Application of the factor analysis in examples). VÚSEI Praha.

Brabenec V., Šařecová P. (2001): Statistické metody v marketingu a obchodu – vybrané přednášky a příklady (Statistical methods in marketing and business – selected lectures and examples). Credit Praha, Reprografické studio PEF ČZU v Praze.

Hebák P., Hustopecký J. (1987): Vícerozměrné statistické metody s aplikacemi (Multivariate statistical methods with applications). SNTL, Alfa, Praha, Bratislava.

Hebák P. a kol. (2005): Vícerozměrné statistické metody – 3 (Multivariate statistical methods – 3). Informatorium, Praha.

Johnson R.A., Wichern D.W. (1998): Applied Multivariate Statistical Analysis. Prentice Hall, Upper Sadlle River, New Persey.

Arrived on 1st March 2007

Contact address:

Vladimír Brabenec, Helena Nešetřilová, Czech University of Life Sciences, Kamýcká 129, 165 21 Prague-Suchdol, Czech Republic
e-mail: brabenec@pef.czu.cz; nesetrilova@pef.czu.cz