AUSTRIAN JOURNAL OF STATISTICS
VOLUME 31 (2002), NUMBER 2&3, 189-200

Retropolation Models of Economical Aggregates

Josef Kozák, Richard Hindls, and Stanislava Hronová
The University of Economics, Prague

Abstract: The contemporary practice used for determination of quarterly values of GDP comes out to prevailing extent from exhaustive surveys. This approach is demanding from the time, labour and financial points of view. Therefore, so-called indirect methods become increasingly important at present. Here the quarterly value of certain aggregates is estimated using a regression model of time series, in which a short-timely determined indicator enters as explaining variable. As for explained variable, there are only yearly values at our disposal in this model. The contribution deals with a specific approach to parameter estimation of such a regression model and with possibilities of its use under conditions of Czech statistics.

Zusammenfassung: Die Methode der Retropolation ermöglicht die Quartalwerte der Kennziffern mit Hilfe der Jahreswerten zu schätzen. Zur Schätzung benutzt man eine Regresionsfunktion. Als Beispiel wurde die Applikation dieser Methode auf die Zeitreihen der Bauwirtschaft der Tschechischen Republik illustriert.

Keywords: National Accounting, Quarterly Values of GDP, Indirect Methods.

1 Introduction

This paper follows up with general methodology of the allocation of yearly data into seasons worked out by the authors of this contribution in French version Kozák, Hindls, Hronová (2000). The methodology of the allocation of yearly values of economical indicators contained here was gradually corrected according to experience with official data produced by Czech Statistical Office. The result of this collation of theory and practice represents the content of this paper.

2 The Allocation Problem

1. Let us start with the necessary summary of basic premises. For the years $i = 1, 2, ..., m$ and seasons $j = 1, 2, ..., n$ with $N = mn$, let us define N-dimensional vector of explained variable

$$\mathbf{y} = [y_{11}, y_{12}, \dots, y_{1n}, y_{21}, y_{22}, \dots, y_{2n}, \dots, y_{m1}, y_{m2}, \dots, y_{mn}]'$$

with the values ordered in years according to seasons. Let us suppose, it can be described using a classical linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \ E(\boldsymbol{\varepsilon}) = \mathbf{0}_N \text{ and } E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \ \mathbf{I}_N,$$

where **X** denotes a non-stochastic matrix of known explaining variables with dimensions Np, $1 < p < N$, and with rank p, $\boldsymbol{\beta}$ denotes the p-dimensional vector of unknown parameters and $\boldsymbol{\varepsilon}$ is an unobservable N-dimensional vector of mutually independent random disturbances with zero mean values and constant variances σ^2.

2. We shall work under conditions, when the vector y is unknown; only yearly sums are considered as known and they are ordered into m-dimensional vector

$$\mathbf{Y} = [Y_1, Y_2, \ldots, Y_m]'$$

using denotation

$$\mathrm{Y}_i = \sum_{j=1}^{n} y_{ij} \text{ for } i = 1, 2, ..., m$$

for yearly sums. There is linear relation between **y** and **Y**

$$\mathbf{Q}\mathbf{y} = \mathbf{Y},$$

where

$$\mathbf{Q} = \begin{bmatrix} 1, & 1, & \ldots, & 1, & 0, & 0, & \ldots, & 0, & 0, & 0, & \ldots, & 0, & \ldots, & 0, & 0, & \ldots, & 0 \\ 0, & 0, & \ldots, & 0, & 1, & 1, & \ldots, & 1, & 0, & 0, & \ldots, & 0, & \ldots, & 0, & 0, & \ldots, & 0 \\ 0, & 0, & \ldots, & 0, & 0, & 0, & \ldots, & 0, & 1, & 1, & \ldots, & 1, & \ldots, & 0, & 0, & \ldots, & 0 \\ . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ 0, & 0, & \ldots, & 0, & 0, & 0, & \ldots, & 0, & 0, & 0, & \ldots, & 0, & \ldots, & 1, & 1, & \ldots, & 1 \end{bmatrix}$$

denotes the non-stochastic matrix with dimension mN. Therefore, introducing the non-stochastic matrix with dimension mp

$$\mathbf{W} = \mathbf{Q}\mathbf{X},$$

with denotation

$$\boldsymbol{\xi} = \mathbf{Q}\boldsymbol{\varepsilon}$$

for m-dimensional vector of unobservable random disturbances and, with respect to identity

$$\mathbf{Q}\mathbf{Q}' = n\ \mathbf{I}_m$$

using above mentioned model for **y** , the vector **Y** will have form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi}, \ E(\boldsymbol{\xi}) = \mathbf{0}_m \text{ and } E(\boldsymbol{\xi}\boldsymbol{\xi}') = n\sigma^2 \ \mathbf{I}_m.$$

3. In the introduction of the paper quoted, we appreciate namely the fact, it contains the proposal of unique method of the allocation of yearly data into seasons

$$\mathbf{u} = n^{-1}\mathbf{Q}'\mathbf{Y} + (\mathbf{X} - n^{-1}\mathbf{Q}'\mathbf{W})\mathbf{B},$$

where N-dimensional vector **u** can be interpreted as an estimator of unknown vector **y**, with **B** as the estimator of unknown vector $\boldsymbol{\beta}$. Above all, the attention deserves the specific property of the estimation proposed, because, for arbitrary vector **B,** it holds

$$\mathbf{Q}\mathbf{u} = \mathbf{Y},$$

and the estimation of class proposed preserves known yearly values.

4. But, the allocation method proposed generates the new problem, which was mentioned in the material quoted only as marginal one: how reasonably select the vector B representing the estimator of unknown vector β?

a) The problem would not exist in the case of known vector y. In this case, the estimator of β can be obtained with the use of common least squares method

$$\mathbf{B}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

with matrix **X** of dimensions *Np***,** which represents the estimator with optimum properties in double sense. Taking into consideration the identity

$$E(\mathrm{B}_0) = \beta,$$

it is an unbiased estimator providing minimum square error

$$R(\mathbf{B}_0,\mathbf{f}) = \mathrm{E}(\mathbf{f}'(\mathbf{B}_0 - \boldsymbol{\beta})^2) = \sigma^2\mathbf{f}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{f} \ ,$$

where f denotes given *p*-dimensional non-zero vector.

In our situation, we know the vector y, because we „think out“ the methodology of the allocation (actual values serve as an criterion of estimators quality) and, therefore, we can try to obtain β vector estimation with respect to mentioned optimum properties. The solution found will be only a starting point for comparisons from the point of view of the allocation itself, and, namely, for the possible specification of „losses“, which must be admitted during parameters estimation process.

b) In the paper quoted, only the estimator based on the model of yearly sums has been briefly mentioned

$$\mathbf{B}_1 = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y} \ .$$

With respect to the structure of the vector of yearly sums, this estimator is unbiased. In comparison with an ideal estimator B_0 (not realizable with respect to the fact the vector y is unknown) this is a case of estimator reduced from *N* items contained in rows of X matrix to $m = N/n$ known rows contained in the W matrix and, thus, intentionally ignoring information contained in rows of X, i.e. not taking into consideration known values of explaining variables in the seasons of individual years.

3 Simulation

1. Under the described circumstances, a feasible way to improve statistical properties of the estimation of the vector β is a simulation of elements of unknown vector y. Let us mention some possibilities in this direction under assumption of two starting points.

a) Let us mention, the knowledge of yearly sums of explained variable Y_i, $i = 1, 2, ..., m$ is supposed. Therefore, we are able to determine average values of the explained variable corresponding to one season as $y_{0,ij} = n^{-1} Y_i$ constant for $j = 1, 2, ..., n$ in the year $i = 1, 2, ..., m$, which can be in the first approximation considered as trend component of time series of explained variable under investigation.

b) Further, let us suppose, the matrix of explaining variables **X** with dimensions 2*N* is known. After transposition, it will contain ones in the first row and the values of single explaining variable in the second row

$$x_{11}, x_{12}, ..., x_{1n}, x_{21}, x_{22}, ..., x_{2n}, ..., x_{m1}, x_{m2}, ..., x_{mn}.$$

Therefore, the vector of parameters $\boldsymbol{\beta}$ will contain two scalars: free parameter β_0 and the slope β_1.

2. As for the defined sequence of yearly values of explained variable, one can reproach for the fact, it does not contain stochastic component in comparison with real case. This shortage can be easily rectified by simulation. For given variance comparable e.g. with the variance S^2 of the sequence $y_{0,ij}$, $i = 1, 2, ..., m$, $j = 1, 2, ..., n$, we can simulate the sequence of random quantities and to create a new sequence

$$y_{1,ij} = y_{0,ij} + \psi_{ij}, \qquad i = 1, 2,..., m,\ j = 1, 2, ..., n,$$

where ψ_{ij} for $i = 1, 2,..., m$, $j = 1, 2, ..., n$ are simulated values of stochastic component arisen by „generating“ of random sample from e.g. normal distribution with zero mean value and the variance S^2.

3. In this connection, the possibility is offered for the estimation of parameters of the vector $\boldsymbol{\beta}$ containing scalars β_0 and β_1 to use the model

$$y_{1,ij} = \beta_0 + \beta_1\, {}_a x_{ij} + \eta_{ij}, \qquad i = 1, 2, ..., m,\ j = 1, 2, ..., n,$$

where besides the quantities introduced for $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$, the additional two sequences play the role: the sequence ${}_a x_{ij}$ containing the estimations of seasonally adjusted time series of used explaining variable x_{ij} and the sequence of random disturbances η_{ij}. Thus, the estimations within given model can be obtained with the use of common least squares, when the estimation b_1 of an unknown parameter β_1 is calculated as

$$b_1 = \frac{\operatorname{cov}(y_1, {}_a x)}{\operatorname{var}({}_a x)}.$$

4. As acceptable can be considered simulation reflecting the starting idea, the simulated explained variable considered „rather“ similar in comparison with explaining variable is comparable even in the nature of seasonal component. Under validity of such a hypothesis, then, according to our opinion, it is possible to „graft“ the seasonal component contained in the sequence of explaining variable on simulated explained variable and to specify it for our purposes. This idea can be put into concrete terms e.g. by this way.

Let us suppose, during the years $i = 1, 2, ..., m$ and seasons $j = 1, 2, ..., n$, the explaining variable x_{ij} can be described using a multiplicative model, i.e. by a set of seasonal indices I_{ij} systematically changing their values during the seasons $j = 1, 2, ..., n$ and constantly repeating during the years with the condition

$$n^{-1} \sum_{j=1}^{n} I_{ij} = 1,$$

i.e. under additional assumption, the average of seasonal indices within given year equals one. If we can justify the assumption, the action of seasonal influences is equally intensive on all variables, we can simulate explained variable in the form

$$y_{3,ij} = y_{2,ij} + \eta_{ij}, \quad i = 1, 2, ..., m, \; j = 1, 2, ..., n,$$

where

$$y_{2,ij} = I_{ij}\, y_{0,ij}, \quad i = 1, 2, ..., m, \; j = 1, 2, ..., n,$$

represents the estimation of „mixed" model of regular components (both trend and seasonal one) of time series under investigation. Thus, using such a model, we suppose, for $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$ the deterministic component of time series of explained variable is equal to the constant $I_{ij}\, y_{0,ij}$ and the corresponding stochastic component is equal to simulated value ψ_{ij} in above mentioned sense of word. In this connection, it is useful to mention two facts.

a) For i-th year, $i = 1, 2, ..., m$, let us consider the sums

$$\sum_{j=1}^{n} y_{2,ij} = \sum_{j=1}^{n} I_{ij}\, y_{0,ij} \; .$$

With respect to above given definition of the quantities $y_{0,ij}$ for $i = 1, 2, ..., m$, $j = 1, 2, ..., n$, we can write

$$\sum_{j=1}^{n} y_{2,ij} = n^{-1} Y_i \sum_{j=1}^{n} I_{ij} \; .$$

But, using the relation

$$\sum_{j=1}^{n} I_{ij} = n \, ,$$

we obtain the identities

$$\sum_{j=1}^{n} y_{2,ij} = Y_i \, , \quad i = 1, 2, ..., m,$$

which can be interpreted by this manner: the quantities $y_{1,ij}$ can be considered within given year i, $i = 1, 2, ..., m$, as the rough allocation of the values of unknown values of explained variable y_{ij} in seasons $j = 1, 2, ..., n$.

b) If the hypotheses considered within this article and concerning with simulated values of explained variable are justified, we can use the linear model

$$y_{3,ij} = \beta_0 + \beta_1 x_{ij} + \omega_{ij}, \; i = 1, 2, ..., m, \; j = 1, 2, ..., n$$

and to obtain the estimation b_2 of the parameter β_1 using least squares method and the formula

$$b_2 = \frac{\operatorname{cov}(y_1, x)}{\operatorname{var}(x)} \, .$$

4 Illustration Using the Building Industry Data

The above given procedure of the allocation of yearly values into seasons is illustrated using official data from Czech Statistical Office with the goal to manifest, even in conditions of economical transformation (i.e. in conditions of relative instability of short time series) is possible to use this approach. After consultations with experts from the area of branch statistics, we decided to perform a simulation using the data from the building industry, where did not occur substantial methodological changes of the content of indicators and the way of their survey during the last period. As the explained variable *y the volume of building works* has been selected (in CZK-Czech Crowns-million in fixed prices 1994), as the explaining variable *x* natural indicator – *the production of cement* in thousands of tons. The reason for the selection of natural indicator as explaining variable is its easy accessibility, the rate of its survey and the relative stability in time. These properties concern not only the indicator of cement production, but of all natural indicators investigated by Czech Statistical Office, which, due to these attributes, manifest themselves as suitable explaining variables in contemporary state of matters.

As for the modelling itself, four different possibilities have been used and we describe them now in detail.

1. Let us start with somewhat not realistic model assuming the knowledge of explained variable and based on linear regression function

$$y_{ij} = \alpha + \beta x_{ij} + e_{ij} \quad \text{for } i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n.$$

As for practical applicability of this model, it is suitable to take into consideration the first row in Table 1, where coefficient denotes the adjusted coefficient of determination, significancy expresses, whether the estimations of both parameters are sufficiently statistically significant (the answer „yes“) or not (the answer „no“), „estimation“ denotes the estimation of β parameter obtained by common least squares method and „lower bound“ and „upper bound“ denote lower and upper bound of 95% two-sided confidence interval of given estimation.

Table 1: Characteristics of individual models

Model No.	Variables	Coefficient	Significance	Estimation	Lower bound	Upper bound	Figure No.
1	Y, x	0.357 9	yes	12.422	6.222	18.223	1
2	$Y_1, {}_a x$	0.000 0	no	0.487	–6.375	6.472	2
3	y_2, x	0.613 9	yes	21.784	15.762	27.807	3
4	y_3, x	0.588 6	no	21.939	15.544	28.333	4

As mentioned above, the model given has no practical meaning for the allocation itself with a view to the fact, the *y* variable will be unknown. The data in the first row of table 1 indicate, the measure of determination of a given model is not too high. But the positive value of estimation indicates good interpretability of direct dependence between both variables.

The situation can be better understood with the use of Figure 1 showing the time course of both variables, with the scale of y variable „building works in mil. CZK in fixed prices“ and the values of x variable (the cement production in thousands of tons) are created by multiplying of y values by certain constant. At the abscissa, the time variable is depicted (the order of quarters during the years 1990–1998).

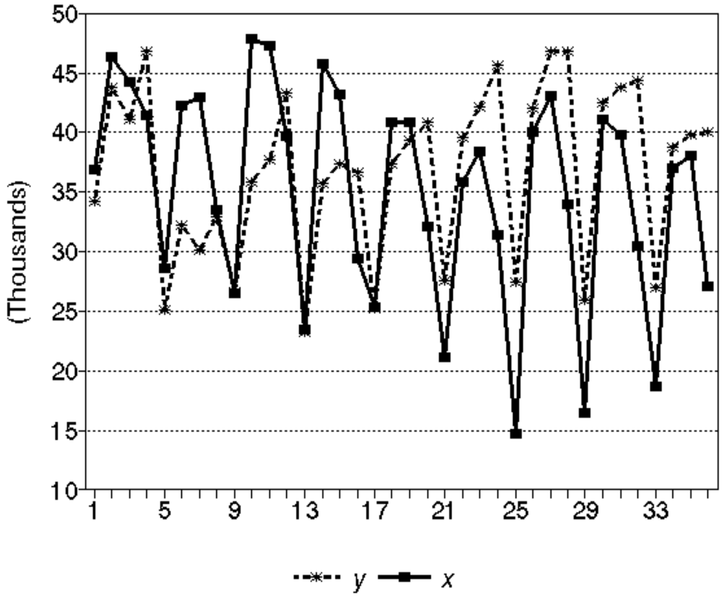

Figure 1

2. From the point of view of practical application of the allocation problem, the additional model

$$y_{1,ij} = \alpha + \beta\, {}_a x_{ij} + e_{ij} \text{ for } i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n\,,$$

based on the search for the relation between known yearly averages of explained variable with added simulated random disturbances (i.e. variable y_1 differing from variable y_0 by simulated values generated from normal distribution with zero mean value and standard deviation 3000) on the one hand and the variable ${}_a x$ representing seasonally adjusted values of x variable could be the most simple multiplicative method. The second row in Table 1 manifests the „futileness“ of such an attempt from all three points of view considered – the measure of determination, significance and bad interpretability of the estimation.

State of matters can be seen from Figure 2, where time variable is again depicted at the abscissa, the axis of ordinates denotes the variable y_1 (expressed in mil. CZK in fixed prices) and the second variable ${}_a x$ (again in thousands of tons) denotes seasonally adjusted time series of explaining variable scaled in the same way as at the Figure 1. The picture indicates that seasonal adjustment of explaining variable cannot be considered as good „step“ during simplifying of the relation between both variables.

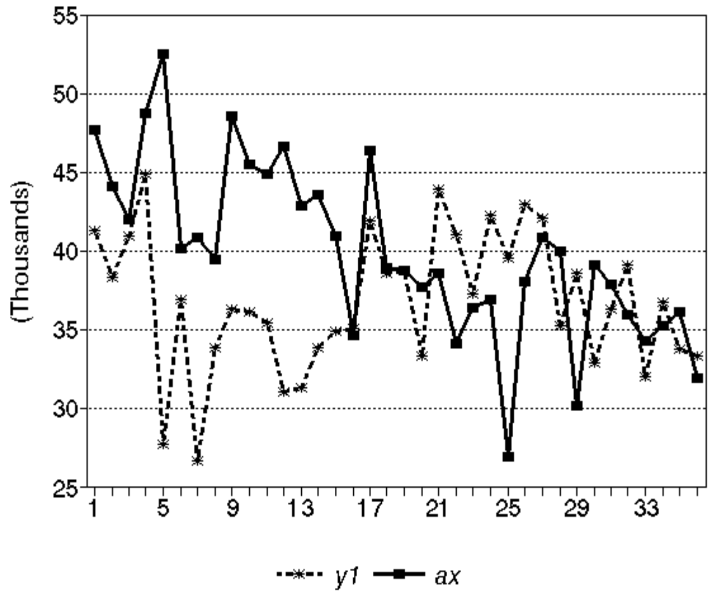

Figure 2

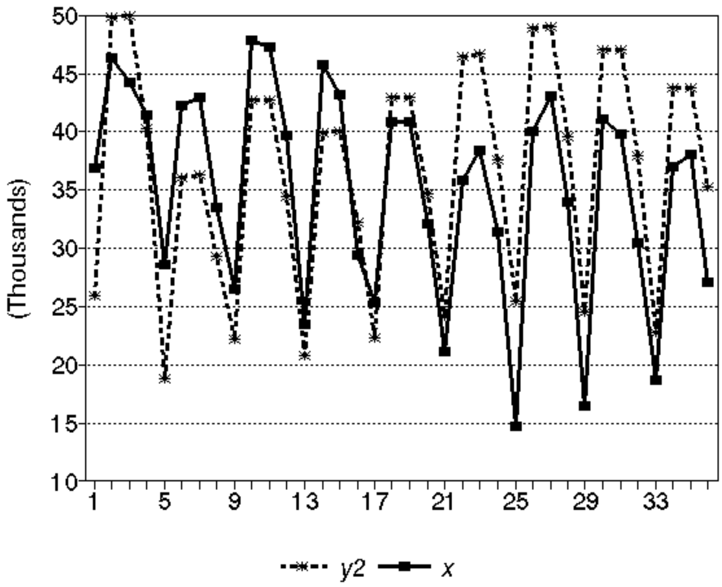

Figure 3

3. The preceding attempt indicates, it is necessary to employ either the model

$$y_{2,ij} = \alpha + \beta x_{ij} + e_{2,ij} \quad \text{for } i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n,$$

or the model

$$y_{3,ij} = \alpha + \beta x_{ij} + e_{3,ij} \quad \text{for } i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n,$$

i.e. the models similar to theoretical model 1 with the difference, its left side differs in comparison with „ideal“ model in simulated values of explained variable. Table 1 shows in the third and fourth row good interpretability of results from all three judged points of view. The state of matters is illustrated in Figures 3 and 4 having similar character like in Figure 1 with the only difference, namely, at the ordinate axis is given the variable y_2 in Figure 3 and the variable y_3 in Figure 4 and the second variable x is scaled in the same manner like in preceding two Figures 1 and 2.

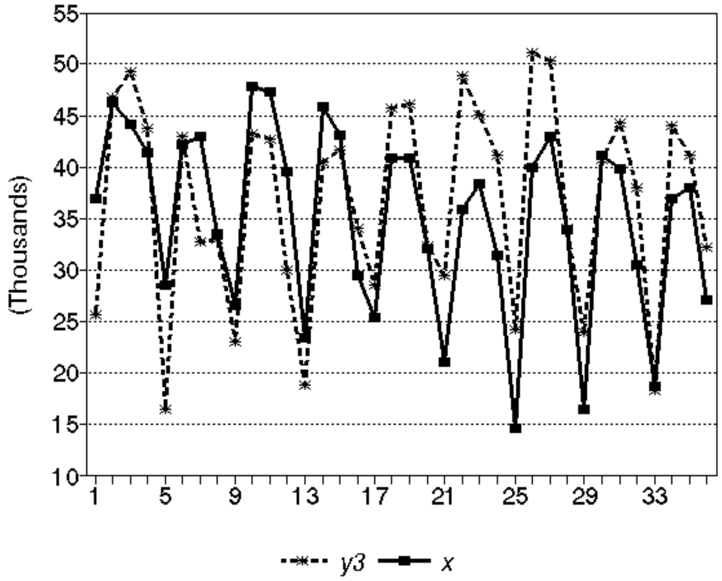

Figure 4

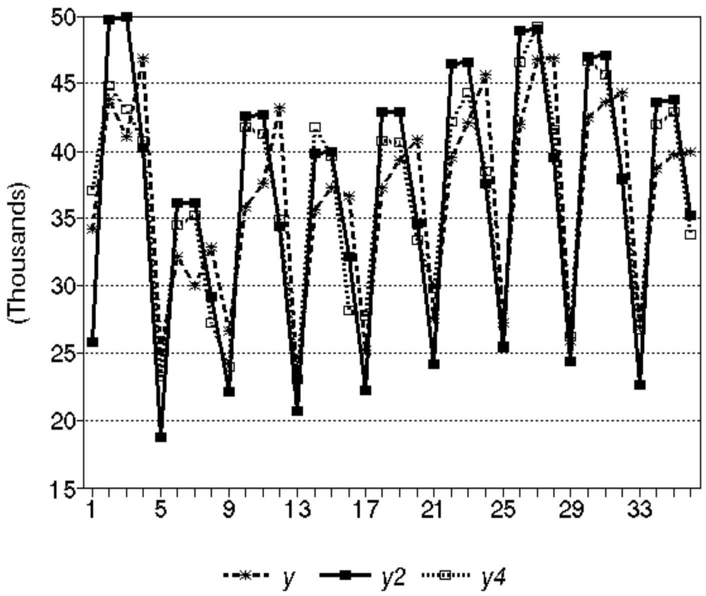

Figure 5

Now, it is suitable to close the explanation by own considerations based on three types of allocation of the indicator of building works.

a) The first one is theoretical allocation represented by time series of explained variable, i.e. of empirical values of indicator, which naturally meets the condition needed on prescribed yearly sums. This allocation can serve only as a basis for the comparation for additional practically usable allocation, of which we use only two.
b) The example of practically usable allocation is simple allocation represented by the values of indicator of dependent variable adjusted according to the above mentioned pattern of y_2, i.e. of rough estimation of theoretical values of this time series constructed for i-th year and j-th quarter from the multiple of known yearly average and the value of seasonal index.
c) The additional allocation was obtained according to original allocation formula suggested by authors with the use off the estimation of parameters of the fourth model. The calculation was performed using the formulae

$$y_{4,ij} = y_{0,ij} + b(x_{ij} - x_{0,ij}) \quad \text{for } i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n ,$$

$$x_{0,ij} = \frac{1}{n}\sum_{j=1}^{n} x_{ij} \quad \text{for } i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n ,$$

where for given i and $j = 1, 2, \ldots, n$ denotes yearly average of explaining variable and $b = 21.9$, according to the last row of Table 1 denotes the estimation of β contained in above formulated model 4. The values of the allocations mentioned are given in Table 2 for given years and quarters. They are depicted in Figure 5.

Table 2: Recapitulation

Year	Time	y	y_2	y_4	Year	Time	y	y_2	y_4
1990	1	34 183	25 872.85	37 036.95	1994	19	39 340	42 939.44	40 727.20
	2	43 769	49 785.54	44 877.15		20	40 890	34 628.35	33 412.60
	3	41 018	49 899.52	43 125.15	1995	21	27 597	24 157.87	29 849.23
	4	46 829	40 241.28	40 759.95		22	39 536	46 485.51	42 222.73
1991	5	25 080	18 769.49	23 247.95		23	42 048	46 591.93	44 281.33
	6	32 223	36 116.97	34 548.35		24	45 628	37 573.89	38 455.93
	7	30 035	36 199.66	35 205.35	1996	25	27 333	25 420.43	25 526.40
	8	32 941	29 193.08	27 277.55		26	42 008	48 914.98	46 594.20
1992	9	26 634	22 148.27	23 957.93		27	46 739	49 026.97	49 200.30
	10	35 842	42 618.55	41 762.63		28	46 820	39 537.61	41 579.10
	11	37 654	42 716.12	41 302.73	1997	29	25 911	24 407.99	26 171.05
	12	43 241	34 448.26	34 907.93		30	42 515	46 966.79	46 691.35
1993	13	23 138	20 717.30	23 149.15		31	43 657	47 074.32	45 662.05
	14	35 614	39 865.02	41 829.85		32	44 329	37 962.90	37 887.55
	15	37 302	39 956.29	39 617.95	1998	33	27 007	22 691.63	26 761.10
	16	36 707	32 222.60	28 164.25		34	38 712	43 664.11	42 025.40
1994	17	25 189	22 264.06	27 762.40		35	39 738	43 764.08	42 879.50
	18	37 254	42 841.36	40 771.00		36	39 956	35 293.38	33 747.20

In this connection, natural question arises, which of mentioned practical allocations is „better“ in the sense, it is closer to the theoretical sequence *y*. For *h*-th allocation, one can use the measure defined as the square root from average square error of the values allocated, i.e. the square root of the characteristics

$$ASE(y_h) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(y_{h,ij} - y_{ij})^2}{mn}\ .$$

Using in this formula the data from Table 2 and calculating square roots, we get for h = 2 and 4 the values of square roots of average square errors as

$$\sqrt{ASE(y_2)} = 5\,419.17 \ \text{ and } \ \sqrt{ASE(y_4)} = 4\,365.30\,,$$

of which follows:

a) because the square root of average square error can be considered as the average square deviation of the value allocated from ideal allocation, we can say, this value is, regardless of the sign, equal to about 5.4 mil. CZK in fixed prices in case of the second allocation and about 4.4 mil. CZK in fixed prices in case of the improved fourth allocation.

b) therefore, the improved allocation y_4 is roughly by 20 % better in comparison with simple allocation y_2 (measured by the square root of average square error).

Basic References

J. Bournay and G. Laroque. Réflexions sur la méthode d´élaboration des comptes trimestriels. *Annales de l´INSEE* n° 36, Paris, 1979.

J. Kozák. Možnosti využití expertních informací při konstrukci předpovědí na bázi klasického lineárního modelu časové řady. *Acta Oeconomia Pragensia*, VŠE, pages 117-142, Praha, 1995.

J. Kozák, R. Hindls, and S. Hronová. Les expériences de la République tchèque dans l´utilisation des méthodes indirectes d´estimation des agrégats trimestriels. 8ème colloque de l´Association de Comptabilité Nationale, Paris, 2000.

Other References

J.C.C. Boots, W. Feibes, and H.H.C. Lisman. Further methods of derivation of quarterly figures from annual data. *Applied Statistics*, pages 65-75, vol. 16, 1967.

G. Chow and A.L. Lin. Best linear unbiased interpolation, distribution and extrapolation of time series by related series. *Review of Economics and Statistics*, vol. 43, pages 372-375, 1971.

G. Dureau. *Les comptes nationaux trimestriels*. INSEE-Méthodes, n^o 13. Paris 1991.

M. Friedman. The interpolation of time series by related series. *JASA*, vol. 57, pages 729-757, 1962.

S. Hronová. *Čtvrtletní národní účty ve Francii*. VŠE, Praha 1994.

Authors' addresses:

Univ.-Prof. Ing. Dr. Josef Kozák
Univ.-Prof. Ing. Dr. Richard Hindls
Univ.-Prof. Ing. Dr. Stanislava Hronová

University of Economics
Dep. of Statistics
nám. W. Churchilla 4
130 67 Prague 3
Czech Republic

Tel. +42 02 24-095-410, +42 02 24-095-484
Fax +42 02 24-22-03-62
E-mail: kozak@vse.cz, hindls@vse.cz, hronova@vse.cz
Internet: http://www.vse.cz