Examining the Supply Chain Integration Impact on Economy by Regression Model

L.O. Babeshko [#1], I.V. Orlova [#2]

[#1, #2] *Financial University under the Government of the Russian Federation, Moscow, Russia,*

***Abstract*— In this study, we examine the current state of supply chain integration, estimate the economic impact of inadequate integration, and identify opportunities for governmental organizations to provide critical standards infrastructures that will improve the efficiency of supply chain communications. The development of methods to reduce the impact of multicollinearity in the construction of a linear regression model is an urgent task of applied econometrics. The article proposes a method for reducing multicollinearity in the construction of a linear regression model for evaluating the supply chain impact on economy. In the case of non-stationary multidimensional time series, it is assumed that all variables have a polynomial trend. Each predictor $x_j(t)$ is decomposed into a trend and a remainder $u_j(t)$, and then a regression *y(t)* is constructed for time *t* and the remainder $u_j(t)$. In this case, the regression coefficients for $u_j(t)$ are equal to the regression coefficients for $x_j(t)$, but they are estimated using less correlated regressors SCM. The article gives a quantitative assessment of the increase in the accuracy of the forecast of the considered model in comparison with other models. In the case of spatial variables, the proposed approach is that some X_j regressors SCM correlated with others are replaced by the sum of two summands. One of them is the predicted value of X_j obtained from the regression equation X_j on the predictor correlated with it; the other is the remainder of this regression U_j. As a result, we get a new set of regressors SCM that are much less correlated with each other. The new regressors - the remnants of U_j - are susceptible to meaningful interpretation. However, the new regression equation changes the regression coefficients only for variables that act as dependent variables in auxiliary regressions. The application of the proposed method is illustrated by examples through the supply chain process. Calculations are performed in the R software environment.**

***Keywords*—** *supply chain, economy development, block matrices, multicollinearity, spatial variables, multi-dimensional non-stationary time series.*

1. Introduction

The supply chain is how a company turns raw materials into finished goods and services for the customer. One of the conditions for correct application of regression analysis in supply chain process is to identify dependencies in empirical data is the absence of multicollinearity of regressors SCM. The negative consequences of multicollinearity are well known. These include an increase in the variance of estimates, which affects negatively the study of the degree and direction of the regressor's action on the endogenous variable, the complexity of determining the contribution of each of the explanatory variables to the variance of the dependent variable explained by the regression equation, and a decrease in the accuracy of predicting the response from the regression equation [1, 2]. Therefore, multicollinearity diagnostics is an integral part of regression analysis. Methods of diagnostics of multicollinearity are considered in numerous works [3-7]. Various multicollinearity tests sometimes lead to conflicting conclusions about the multicollinearity of regressors SCM. Therefore, to make an informed decision about the multicollinearity of regressors SCM, we have to use a set of tests. Tests for the presence of multicollinearity of variables are presented in the R software package quite fully.

If there is multicollinearity in the data, there is a need to either get rid of it, or at least weaken the degree of multicollinearity of regressors SCM [2, 3]. Possible approaches to statistical estimation of regression dependencies in multicollinearity conditions are as follows.

1) The simplest method for eliminating multicollinearity is to exclude one of the predictors that correlates most strongly with the others from the model. However, an important predictor for the model may be deleted, and deleting it will distort the essence of the regression model.

2) Data standardization. Data standardization improves the conditionality of linear regression computational algorithms, but it is of little use for weak conditionality of the sample correlation matrix. Standardization is particularly useful in the case of polynomial regression models.

International Journal of Supply Chain Management
IJSCM, ISSN: 2050-7399 (Online), 2051-3771 (Print)

3) Building a regression model under multicollinearity conditions without changing the composition of explanatory variables (ridge regression and Lasso regression). These methods reduce the variance of estimates of regression coefficients, and the estimates are biased.
4) Multicollinearity can be ignored if the regression model is intended only for forecasting.
5) Orthogonalization of regressors SCM.
Two methods are used for orthogonalization of variables – the principal component method and the Gram-Schmidt sequential orthogonalization method. A significant disadvantage of both methods is that the new variables obtained as a result of orthogonalization are almost impossible to be meaningfully interpreted. The regression equation for new variables is suitable for forecasting, but it is practically not suitable for meaningful regression analysis, for evaluating the impact of the original regressors SCM on the result.
Estimation of regression parameters under conditions of multicollinearity still remains an important problem in applied econometrics.
The purpose of this work is to develop new methods aimed at reducing the impact of multicollinearity on regression estimates and facilitating the interpretation of modeling results. The paper considers an approach associated with incomplete orthogonalization of variables, with the transition from the original regressors SCM to new, less correlated and susceptible to meaningful interpretation variables.
In this paper, we propose two methods for reducing the effect of multicollinearity, connected by a single approach, applicable for multidimensional non-stationary time series and spatial variables. The effectiveness of the methods is illustrated by examples based on real statistical data.
If the source variables are non-stationary multidimensional time series that have a trend, then a false correlation between the variables appears. Then the regressors SCM become correlated, even if they are not related to each other in meaning. To reduce the degree of multicollinearity, we divide each of the original regressors SCM $x_j(t)$ into two parts – the trend $\hat{x}_j$ and the remainder $u_j(t)$ – and then build a regression model of the dependence of the endogenous variable $y(t)$ on the regressors SCM *t*, $u_1(t),\ldots,u_m(t)$. The residuals $u_j(t)$are free from correlation caused by the trends of variables, and the regression coefficients for them are equal to the regression coefficients for the original regressors SCM $x_j(t)$. The error of the proposed trend-factor model is less than the errors of both the $y(t)$ regression model on t and the $y(t)$ regression model over the original time series. The paper provides a quantitative assessment of the improvement in the quality of the forecast by the trend-factor model in comparison with other models. The application of the method is illustrated by an example.
In the case of spatial variables, the proposed approach leads to incomplete orthogonalization of the original regressors SCM using a linear transformation. The linear transformation of the original regressors SCM consists in replacing part of the correlated regressors SCM X_j with the remainder U_j from their regression with the closely related regressor X_k. This is similar to the first step of the Gram-Schmidt orthogonalization algorithm, but it is applied not to a single variable, but to a group of highly correlated regressors SCM, and the new U_j variables do not participate further in auxiliary regressions of some regressors SCM on others. The remainder of U_j is the difference between X_j and the predicted value of $\hat{X}_j$ obtained from the linear regression equation X_j on the correlated regressor X_k. The new variables U_j are linear combinations of X_k and a constant, they are susceptible to meaningful interpretation and are less correlated than the original variables. In the new regression equation, only the regression coefficients change for variables that act as dependent variables in auxiliary regressions. The article presents formulas for the relationship of estimates of regression coefficients for old and new variables and their covariance matrices. The application of the proposed method is illustrated by the example of constructing a regression model of the volume of innovative goods, works and services for the subjects of the Russian Federation [21 – 25]. The source of statistical information is the official website of the Federal State Statistics Service (https://www.gks.ru).

2. Method

Firms engaged in supply-chain relationships, as customers, suppliers, or providers of services, need to share a great deal of information in the course of their interactions. Over the years, companies have managed these information flows in a number of ways, including telephone calls, letters, telex, faxes, and electronic data interchange. More recently, firms have

begun using the power of the Internet to create more effective and open transmission protocols for machine-to-machine communication of the same high-frequency data now handled by traditional electronic data interchange.

• **The original regressors SCM are multidimensional time series**

In practice, most often the trends of both the dependent variable $y(t)$ and the regressors SCM $x_j(t)$ *(j=1,…,m)* are polynomials of no higher than the second degree.

The linear regression model $y(t)$ on $x_j(t)$ has the form:

$y(t) = \beta_0 + \sum_{j=1}^{m} \beta_j x_j(t) + \varepsilon_t, t = 1, \dots, n$ (1)

The values of $x_j(t)$ *(j=1,…,m)* are represented as the sum of the trend and deviations from the trend

$x_j(t) = a_{0,j} + a_{1,j}t + a_{2,j}t^2 + u_j(t)$, (2)

along with this, no assumptions are made about the remainder $u_j(t)$. Estimates of coefficients $a_{k,j}$ will be obtained using the least squares method and, consequently, $\sum_{t=1}^{n} u_j(t) = 0$. Substituting (2) in (1), we get:

$y(t) = \gamma_0 + \gamma_1 t + \gamma_2 t^2 + \sum_{j=1}^{m} \beta_j u_j(t) + \varepsilon_t$,

where $\gamma_0 = \beta_0 + \sum_{j=1}^{m} \beta_j a_{0,j}$, $\gamma_1 = \sum_{j=1}^{m} \beta_j a_{1,j}$, $\gamma_2 = \sum_{j=1}^{m} \beta_j a_{2,j}$. (3)

To avoid a large coefficient, the correlation coefficient between t and t^2 is replaced in (3) *t* by $t - \bar{t}$, where $\bar{t}$ is the average value of *t*. So, the specification of the trend-factor model has the form:

$$y(t) = \gamma_0 + \gamma_1 t + \gamma_2 t^2 + \sum_{j=1}^{m} \beta_j u_j(t) + \varepsilon_t \text{ , (4)}$$

where the model parameters are estimated using the least squares method rather than using formulas (3). Let us note that the regression coefficients β_j for $u_j(t)$ in (4) are the same as the regression coefficients β_j for x_j in (1), but their estimates are obtained from other predictors than in (1). The regressors SCM of the model (4) are less correlated compared with the regressors SCM of model (1).

To predict the level of the dependent variable at t=L, you must have the forecast values $u_j(L)$. As the average value of $u_j(t)$ is zero, it is natural to put $u_j(L)$ equal to zero. Then the predicted value of $y(L)$ is $y(L) = \gamma_0 + \gamma_1 L + \gamma_2 L^2$. The obtained values of $y(L)$ differ from the forecast obtained from the trend model, as the resulting regression equation contains, in addition to t, regressors SCM $u_j(t)$. Therefore, the confidence intervals of forecasts will be less than the forecast intervals for the trend model, which only takes into account time t.

Quantifying improvements of forecast accuracy

Let *U* be a block matrix of regressors SCM values in model (4), $\boldsymbol{U} = (\boldsymbol{T}, \boldsymbol{D})$.

$$\boldsymbol{U} = \begin{pmatrix} 1, t_1, t_1^2, & u_1(t_1), u_2(t_1), \dots, & u_m(t_1) \\ \dots & \dots & \dots \\ 1, t_n, t_n^2, & u_1(t_n), u_2(t_n), \dots, & u_m(t_n) \end{pmatrix},$$

$$\boldsymbol{T} = \begin{pmatrix} 1 & t_1 & t_1^2 \\ . & . & . \\ 1 & t_n & t_n^2 \end{pmatrix},$$

$$\boldsymbol{D} = \begin{pmatrix} u_1(t_1), & u_2(t_1), \dots, & u_m(t_1) \\ \dots & \dots & \dots \\ u_1(t_n), & u_2(t_n), \dots, & u_m(t_n) \end{pmatrix}$$

The standard forecast error $\hat{\sigma}_p$ for $t = L$ is

$$\hat{\sigma}_p = \hat{\sigma}_e \sqrt{1 + \left(\boldsymbol{v}'(L), \boldsymbol{u}'(L)\right) (\boldsymbol{U}'\boldsymbol{U})^{-1} (\boldsymbol{v}'(L), \boldsymbol{u}'(L))'}, \quad (5)$$

where $\hat{\sigma}_e$ – standard error of model residuals, the $'$ sign indicates transposition, $\boldsymbol{v}'(L) = (1, L, L^2)$, $\boldsymbol{u}'(L) = (u_1(L), u_2(L), \dots, u_m(L))$,

Matrix $(\boldsymbol{U}' \cdot \boldsymbol{U})$ has the form $(\boldsymbol{U}' \cdot \boldsymbol{U}) = \begin{pmatrix} \boldsymbol{T}'\boldsymbol{T} & \boldsymbol{T}'\boldsymbol{D} \\ \boldsymbol{D}'\boldsymbol{T} & \boldsymbol{D}'\boldsymbol{D} \end{pmatrix}$

.

Let us denote $\boldsymbol{V_{u,t}} = \boldsymbol{D}'\boldsymbol{T}$, $\boldsymbol{V_{t,t}} = \boldsymbol{T}'\boldsymbol{T}$, $\mathbf{V_{u,u}} = \mathbf{D}'\mathbf{D}$, $\boldsymbol{B_{t,u}} = (\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{T}$, $\boldsymbol{B_{u,t}} = (\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{D}$. As it can be seen from the above formulas, the matrices $\boldsymbol{B_{t,u}}$ and $\boldsymbol{B_{u,t}}$ are matrices of estimates of regression coefficients ***v(t)*** on ***u(t)*** and ***u(t)*** on ***v(t),*** respectively. It is natural to expect that they are close to zero matrices. The matrix inverse to the block matrix $(\boldsymbol{U}' \cdot \boldsymbol{U})$is equal to [8]

$(\boldsymbol{U}' \cdot \boldsymbol{U})^{-1} = \begin{pmatrix} \boldsymbol{U_{1,1}} & \boldsymbol{U_{1,2}} \\ \boldsymbol{U_{2,1}} & \boldsymbol{U_{2,2}} \end{pmatrix}$, where $\boldsymbol{U_{1,1}} = \mathbf{V_{t,t}^{-1}}\text{-}\mathbf{B_{u,t}}(\mathbf{V_{u,t}}\mathbf{B_{u,t}}\text{-}\mathbf{V_{u,u}})^{-1}\mathbf{B}'_{\mathbf{u,t}}$, $\mathbf{U_{1,2}} = (\mathbf{V}'_{\mathbf{u,t}}\mathbf{B_{t,u}}\text{-}\boldsymbol{V_{t,t}})^{-1}\boldsymbol{B}'_{\boldsymbol{t,u}}$, $\boldsymbol{U_{2,1}} = (\boldsymbol{V_{u,t}}\boldsymbol{B_{u,t}}\text{-}\boldsymbol{V_{u,u}})^{-1}\boldsymbol{B}'_{\boldsymbol{u,t}}$, $\boldsymbol{U_{2,2}} = \boldsymbol{V_{u,u}^{-1}} - \boldsymbol{B_{tu}}$ $(\boldsymbol{V}'_{\boldsymbol{u,t}}\boldsymbol{B_{t,u}}\text{-}\boldsymbol{V_{t,t}})^{-1}\boldsymbol{B}'_{\boldsymbol{t,u}}$.

Let $\mathrm{C} = (\boldsymbol{V_{u,t}}\boldsymbol{B_{u,t}} - \boldsymbol{V_{u,u}})^{-1}\boldsymbol{B_{u,t}}$. As the matrix $(\boldsymbol{U}'\mathrm{U})^{-1}$ is symmetric, so $(\boldsymbol{V}'_{\boldsymbol{u,t}}\boldsymbol{B_{t,u}} - \boldsymbol{V_{t,t}})^{-1}\boldsymbol{B}'_{\boldsymbol{t,u}} = \mathrm{C}'$. Then the matrix $(\boldsymbol{U}'\boldsymbol{U})^{-1}$ can be written in the form

$$(\boldsymbol{U}'\boldsymbol{U})^{-1} = \begin{pmatrix} \boldsymbol{V_{t,t}^{-1}} - \boldsymbol{B_{u,t}}\boldsymbol{C} & \boldsymbol{C}' \\ \boldsymbol{C} & \boldsymbol{V_{u,u}^{-1}} - \boldsymbol{B_{t,u}}\boldsymbol{C}' \end{pmatrix}. \quad (6)$$

Substituting the predicted values of the remainder $\boldsymbol{u}(L) = \mathbf{0}$ in the formula (5), taking into account (6), we get

$$\hat{\sigma}_p = \hat{\sigma}_e \sqrt{1 + \boldsymbol{v}'(L)\left(\boldsymbol{V_{t,t}^{-1}} - \mathrm{B}_{u,\mathrm{t}}\mathbf{C}\right)\boldsymbol{v}(L)} \quad (7)$$

If the forecast was made only by the trend, without taking into account $u_j(t)$, the forecast error would be calculated using the formula

$$\hat{\sigma}_{p,trend} = \hat{\sigma}_{e,trend}\sqrt{1 + \boldsymbol{v}'(L)\boldsymbol{V}_{t,t}^{-1}\boldsymbol{v}(L)}\,, \quad (8)$$

where $\hat{\sigma}_{e,trend}$ – estimation of the standard deviation of the trend model. If the regression model specifications $x_j(t)$ from t are selected correctly, then $\boldsymbol{B}_{t,u} \approx 0$, $\boldsymbol{B}_{u,t} \approx \boldsymbol{0}$, and hence $\boldsymbol{C} \approx 0$. As the matrix $\boldsymbol{B}_{u,t}\boldsymbol{C}$ is close to zero, the forecast errors (7) and (8) of the two models actually differ only by a multiplier equal to the standard error of the model residuals. As $\hat{\sigma}_e < \hat{\sigma}_{e,trend}$, the forecast error of the trend-factor model (4) will be less than the forecast error of the trend model by about as many times as $\hat{\sigma}_e$ is less than $\hat{\sigma}_{e,trend}$.

As ≈ 0 , from (6) we get

$$(\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{U})^{-1} \approx \begin{pmatrix} \boldsymbol{V}_{t,t}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_{u,u}^{-1} \end{pmatrix}. \quad (9)$$

Substituting (9) in (5) and considering that $\boldsymbol{V}_{t,t} = \boldsymbol{T}'\boldsymbol{T}$ и $\boldsymbol{V}_{u,u} = \boldsymbol{D}'\boldsymbol{D}$, we get:

$$\hat{\sigma}_p \approx \hat{\sigma}_e\sqrt{1 + \boldsymbol{v}'(L)\,(\boldsymbol{T}'\,\boldsymbol{T})^{-1}\boldsymbol{v}(L) + \boldsymbol{u}'(L)(\boldsymbol{D}'\boldsymbol{D})^{-1}\,\boldsymbol{u}(L)}$$

If we square the last equality, we see that $\hat{\sigma}_p^2$ splits into two summands $\hat{\sigma}_p^2 \approx \hat{\sigma}_e^2\,\varphi(\boldsymbol{v}(L)) + \hat{\sigma}_e^2 c(\boldsymbol{u}(L))$ where $\varphi(L) = 1 + \boldsymbol{v}'(L)(\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{v}(L)$ does not depend on the remainder $\boldsymbol{u}(L)$, and $c(\boldsymbol{u}) = \boldsymbol{u}'(L)(\boldsymbol{D}'\boldsymbol{D})^{-1}\,\boldsymbol{u}(L)$ depends only on the remainder.

• *Discussion. Let us compare the simulation results for different models.*

We study the dependence of the money supply Y (billion dollars) from GNP X1 (billion dollars) and the interest rate on 6-month US government bonds X2 (%) [9]. The number of observations is n=24. Graphs of the source data are shown in Fig.1, 2, 3.

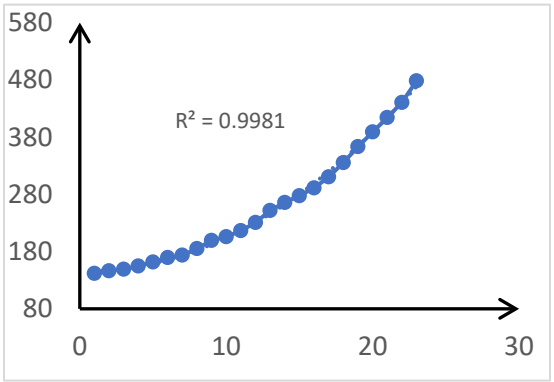

Figure 1. Graph of the Y dynamics

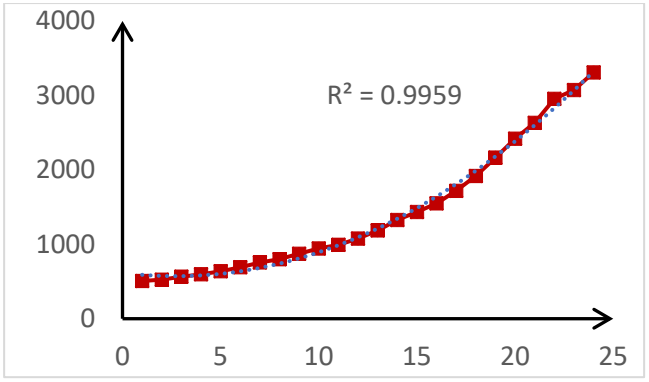

Figure 2. Graph of the X_1 dynamics

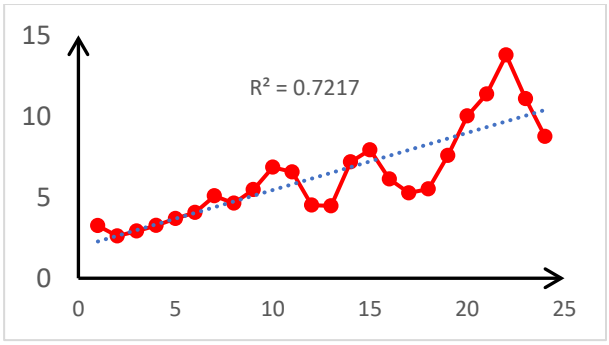

Figure 3. Graph of the X_2 dynamics

Preliminary visual analysis shows that $y(t)$ and $x_1(t)$ have parabolic trends, while $x_2(t)$ has a linear trend. We get the estimates $\hat{y}(t)$ of the three models y(t) for comparison.

Model 1- trend model of dependency of $y(t)$ from t and t^2: to get rid of the correlation between t and t^2, we replace t and t^2 with t- $\bar{t}$ and $(t-\bar{t})^2$, where $\bar{t}$ is the average value of t. the The estimate of the first model $\hat{y}_1(t)$ has the form: $\hat{y}_1(t) = 237{,}09 + 15{,}50t + 0{,}68t^2$,

The standard deviation of the remainder $\hat{\sigma}_{e1}$ is equal to $\hat{\sigma}_{e1} = 6{,}21$; the coefficient of determination R^2=0.997; the standard forecast error for a step forward is equal to $\hat{\sigma}_p$=7.46.

Model 2 regression $y(t)$on the source variables $x_1(t)$ and $x_2(t)$.

All regression model coefficients are significant at the 5% level. The coefficient of determination R^2 is equal to 0.995; the standard error of the model $\hat{\sigma}_{e2}$ = 8.13, the standard forecast error for a step forward $\hat{\sigma}_p$=10.21. The correlation coefficient of $x_1(t)$ with $x_2(t)$ is 0.874, so there is reason to consider the regressors SCM multicollinear. The correlation coefficients y(t) with $x_1(t)$ and $x_2(t)$ are 0.997 and 0.856, respectively, and the regression coefficient for $x_2(t)$ is negative, which is a manifestation of multicollinearity.

Model 3 regression y(t) on t, t^2 and $u_1(t)$, $u_2(t)$ – deviations from their trends $x_1(t)$ and $x_2(t)$. The estimate $\hat{y}(t)$ has the form :

$$\hat{y}(t) = 235{,}99 + 15{,}50t + 0{,}70t^2 + 0{,}086u_1(t) - 2{,}33u_2(t)$$

All model coefficients are significant at the 5% level. The coefficients for t and t^2 of model 3 are naturally close to the coefficients of model 1. The theoretical regression coefficients at $u_1(t)$ and $u_2(t)$ of model 3 coincide with the coefficients for $x_1(t)$ and $x_2(t)$ of model 2. In our case, the original regressors SCM have multicollinearity, the sample size is small, and the estimates of the coefficients of model 3 are 61% and 90% of the estimates of model 2.

The correlation coefficient $u_1(t)$ with $u_2(t)$ is 0.649 and then we can assume that there is no multicollinearity. Testing the presence of multicollinearity using the method of inflationary factors confirms our conclusion, as all VIFj were small less than 2.1 (table1).

Table 1. Method of inflationary factors

regressors SCM	VIFj
t	1,000
t^2	1,155
$u_1(t)$	1,877
$u_2(t)$	2,032

The determination coefficient of model 3 is 0.998; the standard error of the model $\hat{\sigma}_{e3}$= 5.23; the forecast error for a step forward is $\hat{\sigma}_p$= 6.37. The standard error of model 3 is 1.2 times less than the standard error of model 1, $\hat{\sigma}_{e1}/\hat{\sigma}_{e3}$=1.2. The matrix *C* in (6) is equal to $\boldsymbol{C} = \begin{pmatrix} -0{,}00014 & -7\text{E}-19 & 2{,}6\text{E}-06 \\ 0{,}0078 & 1{,}4\text{E}-17 & -0{,}00015 \end{pmatrix}$. In accordance with the above, *C* was close to the zero matrix. The simulation results are shown in table **2.**

Table 2. Simulation results

model	Model specification	R^2	$\hat{\sigma}_e$	$\hat{\sigma}_p$
1	$y(t) = a_0 + a_1 t + a_2 t^2 + \varepsilon_{1t}$	0,997	6,21	7,46
2	$y(t) = d_0 + d_1 x_1(t) + d_2 x_2(t) + \varepsilon_{2t}$	0,995	8,13	10,21
3	$y(t) = \gamma_0 + \gamma_1 t + \gamma_2 t^2 + \beta_1 u_1(t) + \beta_2 u_2(t) + \varepsilon_{3t}$	0,998	5,23	6,37

The standard error $\hat{\sigma}_e$ of model 3 is less than the standard errors of models 1 and 2, and the forecast errors for one step forward $\hat{\sigma}_p$ of model 3 are also less than the forecast errors of models 1 and 2.

Thus, the modified trend-factor model allows to calculate estimates of regression coefficients y(t) for the original regressors SCM $x_j(t)$ for significantly less correlated new regressors SCM that have the same regression coefficients as in the original model. However, the standard forecast error for model 3 is less than for models 1 and 2.

• **The original regressors SCM are spatial variables**

If the original regressors SCM are spatial variables, then the approach under consideration undergoes some changes. The set of regressors SCM is divided into disjoint groups of strongly correlated predictors. Predictors that do not have a high correlation with any of the regressors SCM are not included in any of the groups. Then, in each group, one predictor X_k is selected, called "selected", and auxiliary regression equations are constructed for X_j regressors SCM in the rest of the group on the selected X_k regressor. The coefficients of auxiliary regressions will be calculated using the least squares method. No assumptions are made about the distribution of regression residuals and their relationship to other residuals and regressors SCM. The residuals from regressing X_j on the selected regressor we denote using U_j. Further in the regression of the endogenous variable Y on the explanatory variables instead of X_j new variables U_j are involved. Variables U_j are equal to the difference between the replaced variable and the predicted variable by the regression equation values of this variable. Thus, U_j is a linear combination of the original X_j regressors SCM. Let $x_j^{(i)} = a_0^j + a_k^j x_k^{(i)} + u_j^{(i)}$, then

$$u_j^{(i)} = x_j^{(i)} - a_0^j - a_k^j x_k^{(i)}, \qquad (10)$$

where i – observation number. The variables X_k and U_j are uncorrelated, while X_k and X_j can be correlated in any way. Thus, by replacing part of the X_j regressors SCM with U_j, we get rid of some

correlation dependencies and thereby reduce the level of multicollinearity of the variables involved in the regression. Substituting (10) into the regression equation for the original regressors SCM $Y = b_0 + \mathrm{b}_1 X_1 + \cdots + \mathrm{b}_m X_m + e$, we note that the regression coefficients $\mathrm{b_j}$ for all X_j, except for those j for which X_j are "selected", are equal to the regression coefficients g_{j} for the new regressors SCM U_j. The regression coefficients of g_{k} for such k for which X_k are "selected" are equal to $g_k = b_k + \sum_j b_j a_k^j$. Summation is performed for all such j, for which X_j is a dependent variable, and X_k is independent in additional regressions. From the last formula, we get that the coefficients g_k of the regression Y for the new variables have a meaningful interpretation. When X_k is increased by one, Y changes to b_k by changing X_k, but, in addition, if the correlations are unchanged, X_j regressors SCM that are in the same group as X_k will change on average by a_k^j. This entails changing Y to $\sum_j b_j a_k^j$. As you can see, the change Y is equal to g_k. Thus, the regression coefficient g_k can be interpreted as an increment of Y when X_k changes by one, taking into account the corresponding changes of those X_j that participated in additional regressions X_j on X_k, that is, taking into account the correlations of X_k with other regressors SCM.

Explanatory variables X_j, which did not act as dependent variables in auxiliary regressions and were not replaced by U_j, will be denoted as U_j for convenience of writing. We denote with **X** and **U** the matrices of the values of the original regressors SCM $X_1, X_2, \dots, X_m$ and the new regressors SCM $U_1, U_2, \dots, U_m$. The matrices X and **U** have dimension $(m+1) \times n$, where n is the number of observations, and the first column of the matrices **X** and **U** consists of units. By ***Y*** we denote the vector of values of explanatory variables, by **b, g** we denote the vectors of regression coefficients Y on $X_1, X_2, \dots, X_m$ and $U_1, U_2, \dots, U_m$, respectively.

We denote with ***B*** the matrix of the variable transformation X_j, represented by the formula (10). Then

$$\boldsymbol{U}=\boldsymbol{X}\cdot\boldsymbol{B}. \quad (11)$$

The diagonal elements of the matrix ***B*** are equal to 1, the other elements are equal to 0, except for those columns j for which X_j act as independent variables in additional regressions. These columns are filled in according to the formula (10). The matrix ***B*** is non-degenerate, so the linear transformation given by the matrix ***B*** is one-to-one, and it follows that the residuals from the regressions for the original and new variables coincide.

Regression equations **Y** on $X_1, X_2, \dots, X_m$ and $U_1, U_2, \dots, U_m$ can be written in the form:

$$\boldsymbol{Y}=\boldsymbol{X}\cdot\boldsymbol{b}+\boldsymbol{e}, \quad (12)$$

$$\boldsymbol{Y}=\boldsymbol{U}\cdot\boldsymbol{g}+\boldsymbol{e} \quad (13)$$

Substituting (11) in (12), we get

$$\boldsymbol{Y}=\boldsymbol{X}\cdot\boldsymbol{B}\cdot\boldsymbol{g}+\boldsymbol{e} \quad (14)$$

Comparing (14) with (12), we obtain a formula that reflects the relationship of regression coefficients

$$\boldsymbol{b}=\boldsymbol{B}\cdot\boldsymbol{g}. \quad (15)$$

Taking into account (15), we obtain a formula for the relationship of covariance matrices of estimates of regression coefficients for the original and new variables $cov(\boldsymbol{b}) = \boldsymbol{B}\cdot cov(\boldsymbol{g})\cdot\boldsymbol{B}'$.

3. Results discussion

In a dynamic marketplace and a changing economic environment, the supply chain management system must coordinate the revision of plans/schedules across supply chain functions. The efficiency of the production system is ultimately determined by the agility with which the supply chain is managed at the tactical and operational levels to enable timely dissemination of information, accurate coordination of decisions, and management of actions among people.

Let us illustrate the proposed method of modeling spatial variables by building a regression model of the volume of innovative goods, works, and services for the subjects of the Russian Federation, Y (million rubles). The explanatory variables are: X_1 − gross regional product for the subjects of the Russian Federation, million rubles; X_2 − the number of workers of 15 years old and older, thousand people; X_3 − the average monthly salary for the subjects of the Russian Federation for 2018, rubles; X_4 − capital expenditures for research and development, million rubles; X_5 - innovative activity of organizations (the share of organizations that carried out technological, organizational, marketing innovations in the reporting year, in the total number of surveyed organizations), %; X_6 − number of researchers with a scientific degree, people. [10], [11], [12].

All calculations were performed in a freely distributed R environment [13].

Four indicators X_1, X_2, X_4, X_6 form a group of highly correlated indicators, all the correlation coefficients between them are greater than 0.8 (Fig. 7). This indicates multicollinearity of regressors SCM.

Multicollinearity testing is performed using the **mctest** package [14, 15] for diagnostics of general and individual multicollinearity of data.

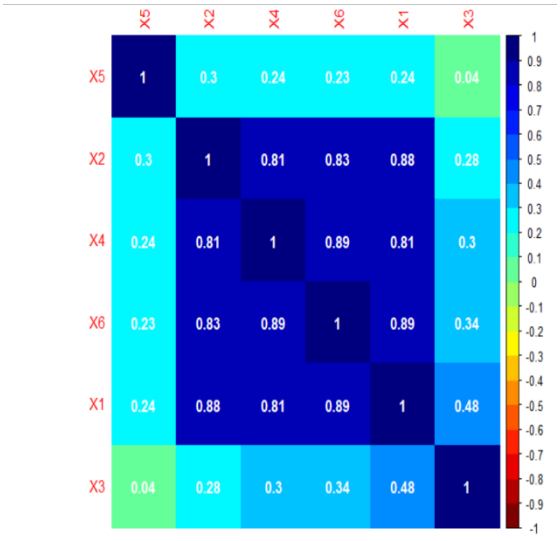

Figure 7. A graph of the correlation coefficient matrix reflecting groups of closely related factors.

Five out of six tests showed that the data is redundant and multicollinearity is present. An individual check of multicollinearity of regressors SCM is performed using the **vif()** function from the **car** package.

> fit<-lm(Y~X1+X2+X3+X4+X5+X6,data=tab1)
> vif(fit)

X1	X2	X3	X4	X5	X6
9.341974	5.570682	1.534861	5.278449	1.100184	**8.356084**

Much supply chain integration literature tends to be biased towards its positive impact on operational performance. However, inconclusive results demand investigation of the mechanisms through which supply chain integration can lead to superior operational performance. Large values of the VIF_j variance inflation factors, two of which are equal to 9.3 and 8.4, and two more than 5, confirm the presence of multicollinearity.

When applying the step-by-step regression procedure to the model of dependence Y on $X_1,\ldots, X_6$, factors X_3 and X_2 are removed from the model, the regression equation has the form:

$\hat{Y} = -32565 + 0{,}036\, X_1 + 24{,}9\, X_4 + 7679{,}8\, X_5 - 20{,}9\, X_6$. (16)

All regression coefficients are significant; the P-values of all coefficients for variables do not exceed 0.0002. According to the regression equation, the relationship of Y to X_6 is negative, as the regression coefficient for X_6 (the number of researchers with a scientific degree) is negative, while the correlation coefficient of Y to X_6 is positive and equal to 0.41. This is a manifestation of multicollinearity. The factors of VIF_j variance inflation are equal to 4.9; 5.0; 1.1; 8.2, which indicates the multicollinearity of the regressors SCM in the built model.

Let us use the approach of variable transformation proposed in this paper.

One of the four SCM – X_6 is called "selected" and in the regression model we replace X_1, X_2, X_4 with U_1, U_2, U_4 - the remainder of the regression X_1, X_2, X_4 with X_6. The regression equations X_1 and X_4 have the form: $X_1 = 522495 + 384{,}55X_6 + U_1$, $X_4 = 110{,}5 + 0{,}59X_6 + U_4$.

Now let us make the regression equation Y for U_1 , U_2 , U_4 , X_3 , X_5 , X_6 . The factors of inflation variance VIF_j are equal 1.9; 1.7; 1.1; 1.5; 1.1; 1.2, therefore, there is reason to believe that multicollinearity is absent. After excluding insignificant factors X_3 and U_2, we get the regression equation:

$\hat{Y}$ = -10862+0,036 U_1+ 24,9 U_4 + 7679,8 X_5 + 7,71 X_6. (17)

The standard error and the coefficient of determination of models (16) and (17) are the same. The regression coefficients for U_1 , U_4, X_5 in (17) are equal to the regression coefficients for X_1, X_4, X_5 in equation (16). Regression coefficients for the "selected" variable. X_6 in (16) and (17) differ not only in size, but also in sign. All regression coefficients in (17) are significant, the regression coefficient for U_1 has a P-value equal to 0.0002, the remaining coefficients have P-values less than 0.0001.

Testing multicollinearity in the new data set (U_1, U_4, X_5 , X_6) using the **imcdiag()** function of the **mctest** package showed its complete absence.

> omcdiag (x1 = XX, y = Y)

Call:
omcdiag(x = XX, y = Y)
Overall Multicollinearity Diagnostics
MC Results detection

Deter6minant |X'X|:	0.9276	0
Farrar Chi-Square:	6.0010	0
Red Indicator:	0.1109	0
Sum of Lambda Inverse:	4.1541	0
Theil's Method	-1.4018	0
Condition Number:	1.2994	0

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

The imcdiag () function implements tests for individually checking regressors for multicollinearity [16-20]. The result of executing imcdiag()function showed that none of the regressors SCM U_1, U_4, X_5 , X_6 can be the cause of multicollinearity.

```
> imcdiag(x = XX, y = Y)
Call:
imcdiag(x = XX, y = Y)
All Individual Multicollinearity Diagnostics Result
VIF TOL Wi Fi Leamer CVIF Klein IND1 IND2
U1 1.0121 0.9880 0.3187 0.4842 0.9940 1.3690 0 0.0375 0.3275
U4 1.0144 0.9858 0.3782 0.5745 0.9929 1.3720 0 0.0374 0.3877
X5 1.0689 0.9355 1.8145 2.7562 0.9672 1.4458 0 0.0355 1.7653
X6 1.0587 0.9445 1.5470 2.3499 0.9719 1.4321 0 0.0359 1.5195
1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
* all coefficients have significant t-ratios
R-square of y on all x: 0.516
* use method argument to check which regressors SCM may be the reason of collinearity

All Individual Multicollinearity Diagnostics in 0 or 1
VIF TOL Wi Fi Leamer CVIF Klein IND1 IND2
U1  0  0 0 0    0   0   0   0   0
U4  0  0 0 0    0   0   0   0   0
X5  0  0 0 1    0   0   0   0   0
X6  0  0 0 1    0   0   0   0   0
1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
* all coefficients have significant t-ratios
R-square of y on all x: 0.516
* use method argument to check which regressors SCM may be the reason of collinearity
```

SCM: 1 − Collinearity is determined by the test; 0 − Collinearity is not detected by the test.

Inflation factors of the variance VIF_j (Fig. 11) are equal to 1.01; 1.01; 1.07; 1.06, therefore, the model regressors SCM are almost orthogonal and, therefore, the regression coefficients have a variance close to minimal.

A change in X_6 by one with constant correlations results in an average change in X_1 by the value a_1^6=384 and X_4 by the value a_4^6 = 0.59, and this, in turn, leads to a change in Y by the value g_6, equal to b_6+ b_1·384+b_4·0.59 (7.71= - 20.9+ 0.036·384+24.9 ·0.59). Thus, in equation (17), the regression coefficient g_6=7.71 for the "selected" variable X_6 it takes into account the correlation relationship of X_6 with variables belonging to the same group of highly correlated variables, which are replaced by U_j. The linear transformation matrix of variables $\boldsymbol{B}$ and the estimates of regression vectors $\boldsymbol{b}$ and $\boldsymbol{g}$ in our example are equal

$$B = \begin{pmatrix} 1 & -522495 & -110{,}5 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & -384{,}55 & -0{,}5908 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} -32565 \\ 0{,}036 \\ 24{,}9 \\ 7679{,}8 \\ -20{,}9 \end{pmatrix}, \boldsymbol{g} = \begin{pmatrix} -10862 \\ 0{,}036 \\ 24{,}9 \\ 7679{,}8 \\ 7{,}71 \end{pmatrix}$$

The vectors $\boldsymbol{b}$ and $\boldsymbol{g}$ satisfy the equation $\boldsymbol{b} = \boldsymbol{B}\cdot\boldsymbol{g}$.

4. Conclusion

The proposed method of incomplete orthogonalization of spatial source variables allows obtaining meaningful interpretation of modeling results. Formulas are obtained that allow, if necessary, going to the regression equation for the initial variables and get all the characteristics of this equation. If the original SCM are non-stationary time series with a polynomial trend, the approach considered in this paper allows obtaining estimates of linear regression coefficients for the original variables using a modified regression model with much less correlated regressors SCM. In this case the standard forecast error for the modified model is less than the forecast errors of other models. In the future, the proposed methods are supposed to be implemented in the R software environment.

References

[1] Draper, Norman; Smith, Harry "*Applied regression analysis"* [Text], 3rd ed.: Trans. from English. – Moscow: Williams Publishing house,. 912 p. 2007.

[2] IRINA, IOUDINA VERA ORLOVA, and IOUDINA VERA. *"Analysis of information content of metric data when constructing models of linear regression."* In C40 System analysis in economics–2018: Proceedings of the V International research and practice conference–biennale (21–23 november 2018).—Moscow, p.

196. 2018.

[3] Orlova I.V. “Approach to solving the multicollinearity problem when analyzing the influence of factors on the resulting variable in regression models” // Fundamental study — 2018. — № 3. C. 58—63. DOI 10.17513/fr.42103.

[4] Gordinsky, Anatoly. *"New Facts in Regression Estimation under Conditions of Multicollinearity."* Open Journal of Statistics 6, no. 5 (2016): 842-861.doi: 10.4236/ojs.2016.65070

[5] Diagnostics, Regression. “*Identifying Influential Data and Sources of Collinearity* (David A. Belsley, Edwin Kuh, Roy E. Welsch) John Wiley & Sons, New York..P. 297. 1980.

[6] Lindner, Thomas, Jonas Puck, and Alain Verbeke. *"Misconceptions about multicollinearity in international business research: Identification, consequences, and remedies."* (2019). https://doi.org/10.1057/s41267-019-00257-1

[7] Yoo, Wonsuk, Robert Mayberry, Sejong Bae, Karan Singh, Qinghua Peter He, and James W. Lillard Jr. *"A study of effects of multicollinearity in the multivariable analysis."* International journal of applied science and technology 4, no. 5 (2014): 9.

[8] Gantmacher F.R. Matrix theory — 5th ed. — Moscow: Fizmatlit,. 560 p. 2010

[9] https://fraser.stlouisfed.org/title/45/item/8155/toc/317713 (access date 20.11.2017)

[10] http://old.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/science_and_innovations/science/ (access date 3.02.2019)

[11] https://www.gks.ru/free_doc/new_site/population/trud/tab_trud1.htm (access date 3.02.2019)

[12] https://www.gks.ru/labor_market_employment_salaries?print=1 (access date 3.02.2019)

[13] Hadley, Wickham,. Garrett ,Grolemund R ,. “*Language for tasks of data science: the import, preparation, processing, visualization and data modeling”*: Trans. from English.−SPb.: DIALEKTIKA LLC, 592 p., ISBN 978-5-9909446-8-8. 2019.

[14] Imdadullah, Muhammad, Muhammad Aslam, and Saima Altaf. "mctest: *An R Package for Detection of Collinearity among Regressors." R J.* 8, no. 2 (2016): 495.

[15] Muhammad, Imdad Ullah,. Aslam Muhammad.,*Multicollinearity Diagnostic Measures.* Package ‘mctest’ https: //cran.r-project.org/web/packages/mctest/mctest.pdf (access date 30.11.2019)

[16] O’brien, Robert M. *"A caution regarding rules of thumb for variance inflation factors."* Quality & quantity 41, no. 5 (2007): 673-690.

[17] Curto, José Dias, and José Castro Pinto. *"The corrected vif (cvif)."* Journal of Applied Statistics 38, no. 7 (2011): 1499-1507.

[18] Salmerón, Román & Pérez, Jose & Garcia, Catalina & López Martín, María.. “*A note about the corrected VIF”*. Statistical Papers. 58. 10.1007/s00362-015-0732-9. (2015)

[19] Kovács, Péter, Tibor Petres, and László Tóth. "*A new measure of multicollinearity in linear regression models."* International Statistical Review 73, no. 3 (2005): 405-412.

[20] Curto, José Dias, and José Castro Pinto. *"New multicollinearity indicators in linear regression models."* International Statistical Review/Revue Internationale de Statistique (2007): 114-121.

[21] Moiseenko, Zh N. "*state support of small forms of management in agriculture: status and directions of development."* Modern economy success (2500-3747) (2017).

[22] Komarova, S. L. *"The assessment of the consumer basket for the analysis of the region competitiveness." Russian Economic Bulletin 1, no. 2 (2018): 19.*

[23] Kobets, E. A. *"The implementation of import substitution programme in the agricultural sector."* Современный ученый (2541-8459) (2017).

[24] Kupryushin P.A., Chernyatina G.N. ”*Economic and environmental aspects of rational nature management and optimization of the process of import substitution in the agro-industrial complex”.* Modern Economy Success. № 3. P. 44 – 48. (2017)

[25] Narkevich L.V. *“Analysis of industrial capacity and break-even production in the crisis management system”.* Russian Economic Bulletin.. Vol. 1. Issue 3. P. 28 – 41. (2018).