International Journal of Sustainable Development and Planning
Vol. 15, No. 7, November, 2020, pp. 1113-1118
Journal homepage: http://iieta.org/journals/ijsdp

Prediction and Validation of the Promoting Effect of Technological Entrepreneurship on Sustainable Economic Growth

Erle Du[1,2*], Ji Meng[1]

[1] Strategic Development Department of Harbin Bank Co., LTD, Harbin 150010, China
[2] Postdoctoral Station of Harbin Institute of Technology, Harbin 150001, China

Corresponding Author Email: duerle@163.com

https://doi.org/10.18280/ijsdp.150715

Received: 8 May 2020
Accepted: 26 August 2020

Keywords:
technological entrepreneurship (TE), sustainable economic growth (SEG), data selection, prediction and validation

ABSTRACT

TE (TE) can potentially enhance the economic output of technological innovation, and thus promote sustainable economic growth (SEG). However, the TE-SEG relationship has been mainly analyzed subjectively through empirical analysis. This paper puts forward a novel strategy that automatically predict and validate the promoting effect of TE on SEG. Firstly, a multi-level analytical model of TE was constructed to automatically select the optimal sample subset from the original data, and eliminate noise and redundant data. Next, a multivariate linear regression model was adopted to analyze the TG-SEG relationship intelligently and intuitively. Finally, the proposed strategy was verified through experiments on the SEG data collected from 31 Chinese cities. The experimental results confirm that our strategy can effectively and reliably reflect the promoting effect of TE on SEG.

1. INTRODUCTION

Technological entrepreneurship (TE) is a potential driver of sustainable economic growth (SEG) [1]. The TE-SEG relationship has been mainly analyzed with varied indices, based on a set of effective samples selected from massive data on SEG [2-5]. However, the largely empirical method of sample selection cannot find a suitable sample subset that properly reflect the promoting effect of TE, given that the numerous independent SEG variables differ in impact and advantage. Even if the original dataset are the same, different researchers often select varied sample subsets, resulting in fluctuations in the analysis results. What is worse, some important data that truly reflect the effect of TE are not selected. In addition, the accuracy and robustness of the analysis may be suppressed if the original dataset are noisy and redundant. Therefore, it is necessary to select a sample subset that reveals the promoting effect of TE on SEG in a comprehensive, objective, and automatic manner.

Feature selection is critical to data analysis, machine learning (ML), and pattern recognition [6]. With the aid of an evaluation function, the goal of feature selection is to choose a proper feature subset for the search strategy. To acquire a proper feature subsets, many feature selection methods have been developed based on feature distance, namely, G-flip [7], Simba [8], Relief [9], and E-relief [10]. However, these methods construct the loss function for classification solely based on the distance of different samples close to the decision boundary. Some useful information might get lost, exerting a negative impact on data analysis.

The above problem can be solved by a feature selection approach based on the loss of nearest neighbor classification [11]: Firstly, a classification loss function is designed for each neighborhood on the Euclidean distances of the samples in the same class and the loss of the samples in different classes; then, the feature weights are computed through gradient descent. This approach was later extended by reconstructing the classification loss function with a new weight factor, which is decomposed from the Mahalanobis distance, and solving the weight through linear programming [12]. These two approaches [13-20] greatly improve the analysis effect of neighborhood-based classifiers.

Inspired by the feature selection based on the large loss nearest neighbor strategy, this paper puts forward a novel strategy that characterizes the promoting effect of TE on SEG. Firstly, a multi-level analytical model of TE was constructed through improved k-means clustering (KMC). The model automatically selects the optimal sample subset from the original dataset, while eliminating noise and redundant data. Next, a multivariate linear regression model was adopted to analyze the TG-SEG relationship intelligently and intuitively. Finally, the proposed strategy was validated through experiments on the SEG data from 31 Chinese cities, which fall into different classes. The experimental results show that our strategy can effectively predict and validate the promoting effect of TE on SEG.

2. METHODOLOGY

Considering the lack of clear hierarchical division for entrepreneurship, this paper employs the KMC, an unsupervised learning method, to decompose TE into multiple states, because this method is flexible, and easy to implement and understand.

In traditional KMC, the clustering results might be unstable and poor, if the cluster heads are not properly initialized. Here, the KMC is improved to solve the problem. First, the hierarchical clustering was adopted to make an initial division. Then, the mean of each initial cluster was taken as an initial cluster head. By using the class information of the dataset, the spatial distribution of the initial cluster heads was kept consistent with the actual data distribution, which promotes the clustering quality.

As shown in Figure 1, the improved KMC is implemented in five steps:

Step 1. Adjust and verify the number of clusters k;

Step 2. Perform hierarchical clustering on the dataset;

Step 3. Count the number of initial clusters, and take the mean of each cluster as the initial cluster head;

Step 4. Calculate the distance between each of the remaining objects and each of the initial cluster head, and allocate each object into the cluster of the nearest cluster head;

Step 5. Recalculate the head of each cluster.

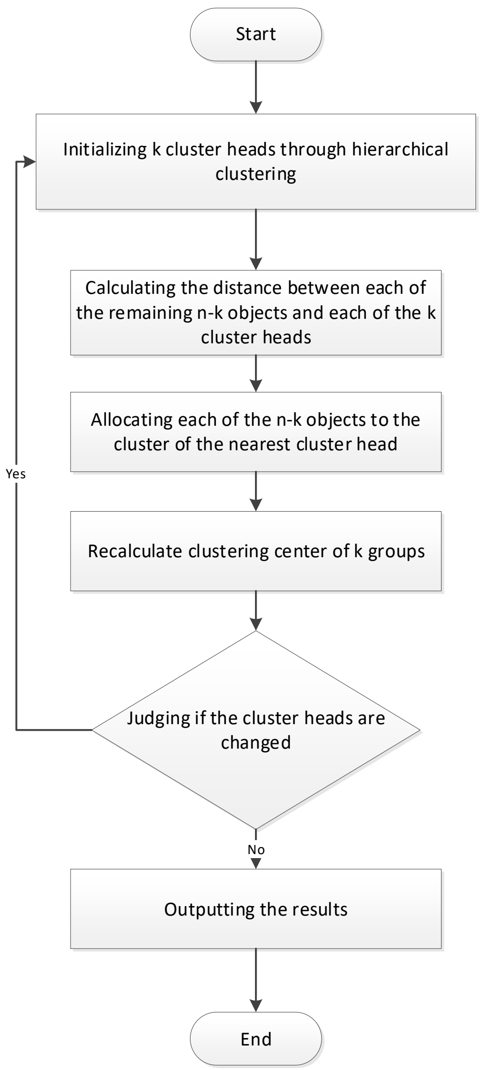

Figure 1. The flow chart of improved KMC

After data clustering, the optimal feature subset was derived from the preprocessed dataset by minimizing the difference between samples in the same class while maximizing the difference between samples in different classes.

Let DS: $\{x_i, y_i\}_{i=1}^{M} \in DS$ be the original dataset of m samples, where $y_i \in \{1,2,\ldots,l\}$ is the class label of sample x_i, and x_i is the sample with m indices $x_i=\{x_{i1}, x_{i2},\ldots, x_{im},\}$; and ε be the class discrimination matrix, in which each element $\varepsilon_{ij}=\{0,1\}$ indicates whether y_i matches with (i=1,2,…,l, j=1,2,…,l). If samples x_i and x_j have the same class label, then $y_i=y_j$ and ε_{ij}=1; otherwise, ε_{ij}=0.

Drawing on the k-nearest neighbors (kNN) algorithm, a neighborhood class record matrix R was established to store the class of the neighborhood composed of samples x_i and x_j with the same class label. In the matrix R, each element ω_{ij}:$\omega_{ij}\in\{0,1\}$ represents that whether x_i and x_j belong to the same neighbor. If yes, ω_{ij}=1; otherwise, ω_{ij}=0.

Following the large loss principle, the feature evaluation function G was modeled to detect the normal and abnormal samples of each index, and then minimize difference between samples in the same neighborhood while maximizing the difference between samples in different neighborhood:

$$G=\frac{\max(\sum_{ijp}^{N}\omega_{ij}(1-\varepsilon_{ip})\lambda_{ijp})}{\min(\sum_{ij}^{n}{}_{ij}\sum_{h=1}^{m}(x_{ih}-x_{jh})^{t}(x_{ih}-x_{jh})+\sum_{i,p}^{n}\theta_{ip})} \tag{1}$$

where,

$$\theta_{ip}=\min(\sum_{h=1}^{m}w_h(x_{ih}-x_{jh})^2-\sum_{h=1}^{n}w_h(x_{ih}-x_{ph})^2) \tag{2}$$

where, λ_{ijp} is the loss function; θ is a record function for noisy samples.

Two outputs were defined for our analysis framework: intermediate goods, and consumer goods. Labor, as a social product, was taken as an input. For simplicity, it is assumed that all residents have an infinitely long life, and the same preference in consumption throughout their life. This is a common assumption in microeconomic theories. Then, the marginal utility of individual consumption is a constant that can be expressed as an interest rate.

In each period, it is assumed that each individual supplies a unit of labor. For a society of L individuals, the total labor supply equals L. The utility difference in labor can be reflected by the intermediate goods. Thus, L can be defined as:

$$L=x+n \tag{3}$$

where, x is the skilled labor input in the production of intermediate goods.

The production of consumer goods y can be defined as:

$$y=Ax^{\alpha} \tag{4}$$

where, $y=Ax^{\alpha}$ is the product flexibility, representing the dependence of consumer goods on the input of intermediate goods x; A is the productivity of intermediate goods, i.e. the technical level of the production of consumer goods.

Taking the production process as a random innovation sequence, this paper defines the realization rate of innovation at any moment in the economy $\lambda\phi(n)$, where n is the investment of technical labor by research and development (R&D) department; λ is a constant for R&D productivity; $\phi(n)$ is a concave production function with constant returns to scale: ($\phi(0)=0$) and $\forall n\geq 0$.

In addition, it is assumed that the memory does not accumulate in the R&D process. That is, the realization probability of innovation only depends on the investment in the current period, and has nothing to do with that in the previous periods.

The innovation is mainly achieved in the R&D of intermediate goods. New intermediate goods can be used to produce more efficient consumer products. Using new intermediate goods, the productivity will be increased by a factor ratio of $A_t=\gamma A_{t-1}$. If there is no lag in technological diffusion, the factor ratio, i.e. rate of technological progress rate, satisfies:

$$A_t = A_0\gamma^t \quad t=1,2,3... \tag{5}$$

where, A_0 is the initial productivity.

The patented innovation can be utilized to monopolize the production of intermediate goods, throughout the long valid period of the patent. However, there is eventually an end to the monopoly of any manufacturer in intermediate goods. Once the next innovation emerges, the monopoly will cease to exist.

For a monopolist of intermediate goods, the goal is to maximize the present value of the expected profit during the monopoly period:

$$\pi_t = \max_x \left[p_t(x)x - w_t x \right] \tag{6}$$

where, w_t is the salary of the manufacturing industry; $p_t(x)$ is the price of intermediate goods sold by the second innovator at time t to the manufacturer, or set by the manufacturer through internal innovation.

In a competitive market, the optimal conditions for the production of consumer goods can be modeled as $p=MC$. Since the sector of intermediate goods is monopolized, the manufacturer will definitely set the price of intermediate goods $p_t(x)$ equal to that of consumer goods x to maximize his/her profit:

$$p_t(x) = A_t\alpha x^{\alpha-1} \tag{7}$$

Judging by whether the enterprise could produce new products, technologies, or services, entrepreneurship can be easily divided TE and general entrepreneurship (GE). An enterprise with TE can provide novel products/services, and boast strong pricing power and competitiveness. By contrast, an enterprise with GE only provides homogenous products, and passively accepts the price set by TE enterprises.

Without the loss of generality, it is assumed that the society has N1 TE enterprises and N2 GE enterprises, that is, a total of N=N1+N2 enterprises. The TE enterprises generally possess a group of R&D personnel, which are skilled labor force in technological R&D (hereinafter referred to as technical labor force), and have the ability to cultivate new technical labor force. On the contrary, GE enterprises are unable to cultivate technical labor force. Therefore, the technical labor force of the whole society comes from TE enterprises. Thus, the proportion of technical labor force in manufacturing industry can be described as:

$$n^* = \eta N_1 \tag{8}$$

where, η is the proportionality parameter.

When it comes to data availability, there is a severe lack of TE data or reports on TE quantification. But there is abundant literature on the quantification of entrepreneurship and the influence of entrepreneurship over SEG. In most studies, the entrepreneurship is measured by the number or proportion of private enterprises.

The previous studies have demonstrated the following advantages of measuring entrepreneurship with the number of enterprises: First, the number of enterprises is easy to obtain, and the number of the enterprises with property rights is a good indicator of entrepreneurship level in terms of SEG. Second, the entrepreneurship can be decomposed easily based on the ownership and other attributes of enterprises. Third, the number and proportion of enterprises can be monitored for a long time, reflecting the continuity of entrepreneurial activities.

Therefore, a similar method was developed to select TE indices. The most active areas of TE in China are high-tech zones in major cities. Under the guidance of national policies, many high-tech zones, especially early starters, have become major gathering places of TE enterprises. Suffice it to say that the high-tech zones are the synonym of TE in contemporary China. The high presence of TE in these high-tech zones is attributed to multiple factors: an institutional environment that encourages innovation, the complete infrastructure that supports the high-tech development, and the abundance of specialized talents that facilitates technical R&D.

As mentioned before, this paper aims to predict and validate the promoting effect of TE on SEG. Like any other country, China has been vigorously pursuing SEG, a key indicator of social welfare. Since the TE index is a scalar and a total factor index, the regional gross domestic product (RGDP), also a scalar and a total factor index, was selected to measure the SEG of each region. For convenience, the natural logarithm of RGDP was taken to highlight its sustainability in the face of TE. Instead of nominal GDP, the RGDP was calculated based on the actual GDP with 1994 as the base period, using the consumer price index (CPI).

As for the prediction model, regression analysis was adopted rather than black box-based methods like deep learning (DL). The regression analysis aims to verify the dependency between different variables by finding the optimal linear mapping from the feature space to the output space, such that the model can describe the TE-SEG relationship intelligently and intuitively. Hence, the regression model can be defined as:

$$\begin{aligned} \boldsymbol{lnGDP_{i,t}} &= \beta_0 + \beta_1 \boldsymbol{lnSent}_{i',t'} \\ &+ \gamma Z_{i',t'} + \lambda_{i'} + \eta_{t'} + \varepsilon_{i',t'} \end{aligned} \tag{9}$$

where, i and t are the serial numbers of city and period, respectively; lnGDP is the natural logarithm of GDP; *InSent* is the logarithm of TE *Sent*; $Z_{i',t'}$ is the set of control variables, including fixed asset investment, inflation rate, economic openness, financial development, etc.; $\lambda_{i'}$ is the individual-fixed effect; $\eta_{t'}$ is the time-fixed effect; β_1 is the impact of TE on SEG. If $\beta_1<0$, TE has a negative impact on SEG; if $\beta_1>0$, TE has a positive impact on SEG.

For the above econometric model, ordinary least squares (OLS) estimation will bring biased and inconsistent results, due to the problem of endogeneity. In our model, the endogeneity might arise from the influencing factors of SEG that are not covered by the control variables. The estimation might also be biased, owing to the reverse causal relationship between TE and control variables.

In addition, an instrumental variable related to TE but independent of SEG was designed for regression: the TE with

1-period lag was chosen as the instrumental variable of the current TE. This variable was designed based on the following considerations: Internally, TE is a gradual process to output technologies and products that satisfy the growing demand of residents. The current TE is strongly correlated with the TE in the previous periods. Externally, the current TE is not directly affected by the previous TE. The effect is mediated by many exogenous factors. Thus, it is to use the TE with 1-period lag as an instrumental variable of the current TE.

3. RESULTS ANALYSIS

To verify the effectiveness of our strategy, a panel dataset was established based on the SEG data from 31 Chinese cities, including provincial seats, municipalities directly under the central government, and the cities with independent planning status. The sample period was from 1997 to 2005.

The original data were collected from the following sources: *China City Statistical Yearbooks*, *China Statistical Yearbooks for Regional Economy*, *China Statistical Yearbooks*, *China Labor Statistical Yearbooks*, Wind database, CEInet Statistics Database, EPSDATA, as well as the statistical yearbooks and statistical communique on national economic and social development issued by the relevant provinces.

To eliminate the impact of inflation and facilitate the comparison of annual data, the data on some indices were deflated by the CPI with 1994 as the base period.

For comparison, two sample subsets were selected for our experiments. The first sample subset (our subset) was generated automatically by our strategy, while the other sample subset (contrastive subset) was constructed by an experienced economist. Both sample subsets contain the same core variable (*InSent*) and 4 control variables.

Before estimation on panel data, the regression models with fixed-effects and random-effects were subject to Hausman test. The results show that the null hypothesis was rejected. Thus, fixed-effects model is the better choice.

To clarify the effect of TE on SEG, the control variables were added to the regression model one by one. Tables 1 and 2 show the results on the contrastive subset and our subset, respectively.

Table 1. The results on the contrastive subset

	lnRGDP				
	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)
lnsent	1.018*** (32.72)	0.729*** (25.81)	0.708*** (25.13)	0.703*** (24.96)	0.102*** (7.22)
fixa		1.508*** (19.44)	1.436*** (18.44)	1.443*** (18.57)	0.049 (1.87)
cpi			2.900*** (4.50)	3.072*** (4.74)	-0.522 (-1.18)
open				0.156** (2.15)	-0.026 (-0.98)
fin					-0.115 (-11.65)
cons	10.43*** (53.92)	11.46*** (72.29)	11.57*** (73.32)	11.68*** (70.82)	15.399*** (187.54)
Time-fixed effect &individual-fixed effect	Yes	Yes	Yes	Yes	Yes
N	589	589	589	589	589
F-statistic	1,070.4***	1,086.1***	755.9***	571.8***	342.02***
R^2	0.658	0.796	0.803	0.805	0.806

Note: The bracketed values are t values of the regression coefficients; ***, **, and * are the significance levels at 1%, 5%, and 10%, respectively; lnGDP is the logarithm of the actual regional GDP; fixa, cpi, open, and fin are fixed asset investment, inflation, opening-up, and financial development, respectively; R^2 is the goodness of fit in regression.

Table 2. The results on our subset

	lnRGDP				
	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)
lnsent	1.018*** (32.72)	0.729*** (25.81)	0.051*** (3.48)	0.052*** (3.63)	0.074*** (5.23)
fixa		1.508*** (19.44)	0.032 (1.15)	0.087*** (2.99)	0.053*** (1.89)
pop			0.0009*** (7.89)	0.0009*** (8.90)	0.0007*** (7.01)
fer				-1.99** (-5.60)	-1.944*** (-5.64)
tind					-0.969*** (-6.24)
cons	10.43*** (53.92)	11.46*** (72.29)	14.801*** (143.16)	14.86*** (146.93)	15.292*** (127.71)
Time-fixed effect &individual-fixed effect	Yes	Yes	Yes	Yes	Yes
N	589	589	589	589	589
F-statistic	1,070.4***	1,086.1***	344.73***	363.77***	388.74***
R^2	0.658	0.796	0.976	0.974	0.977

Note: pop, fer, and tind are human capital, proportion of fiscal expenditure, and proportion of tertiary industry, respectively; the other symbols have the same meaning as those in Table 1.

As shown in Table 1, as the control variables were gradually added to the regression model, the R^2 increased continuously, and the F-statistics of all models passed the significance test, suggesting the reliability of our model. Moreover, the coefficient of *lnsent* was always positive at the 1% significance level, indicating that higher TE and more TE enterprises in a region promote the regional SEG.

After all control variables were added, the coefficient of TE stood at around 0.7. Since both TE and actual GDP are in the form of natural logarithm, it can be seen from the results with all control variables that each 1% growth of TE brings 0.68% increase of the actual GDP. This means the TE has a positive effect on SEG.

As shown in Table 2, human capital, proportion of fiscal expenditure, and proportion of tertiary industry were selected as control variables according to the feature evaluation function. It can be seen that, after the control variables were updated, the sign and significance of *Insent* remained unchanged, while the newly selected control variables were significant at least on the 5% level. Thus, the new variables enhance the explanatory power of the main explanatory variables and control variables. Compared with the results on the contrastive subset, the coefficients in our subset were highly significant, indicating that our model boosts the explanatory power while optimizing the goodness of fit.

To further verify its robustness, consistency, and correctness, our strategy was coupled with two popular estimation methods, namely, the OLS estimation, and system moment estimation (SYS-GMM). Before estimation, the SYS-GMM was subject to Sargan test and Arellano-Bond test. The comparison shows that TE exhibited a significant promoting effect on regional SEG at the 1% level, whether the estimation method was OLS or SYS-GMM. Hence, the promoting effect of TE on SEG is basically not affected by the estimation method, which confirms the robustness and goodness-of-fit of our method.

The SYS-GMM estimation also shows that the previous TE bolsters the inertia of SEG. In other words, the current SEG is promoted not only by the current TE, but also by the development of regional TE enterprises in the past. The previous TE exerts a pulling effect on SEG via the growth of TE enterprises. This is in line with the cumulative effect of TE development. Econometric theory proves that the SYS-GMM has advantages in eliminating endogeneity of econometric regression. In this paper, the SYS-GMM results also demonstrate that, even after the inclusion of endogeneity, the estimation results remained robust: TE can significantly promote SEG.

4. CONCLUSIONS

Empirical evidences show that TE has an obvious promoting effect on SEG. This paper proposes an effective strategy to predict and validate the TE-SEG relationship. Specifically, the KMC was improved to build a multi-level analytical model of TE, which automatically selects the optimal sample subset from the original dataset. Next, a multivariate linear regression model was adopted to analyze the TG-SEG relationship intelligently and intuitively. Finally, the proposed strategy was validated through experiments on the SEG data from 31 Chinese cities. The experimental results show that our strategy can effectively and reliably predict and validate the promoting effect of TE on SEG.

REFERENCES

[1] Acs, Z.J., Audretsch, D.B., Lehmann, E.E., Licht, G. (2016). National systems of entrepreneurship. Small Business Economics, 46(4): 527-535. https://doi.org/10.1007/s11187-016-9705-1

[2] Lazear, E.P. (2004). Balanced skills and entrepreneurship. American Economic Review, 94(2): 208-211.

[3] Kang, H.C., Anderson, R.M., Eom, K.S., Kang, S.K. (2017). Controlling shareholders' value, long-run firm value and short-term performance. Journal of Corporate Finance, 43: 340-353. https://doi.org/10.1016/j.jcorpfin.2017.01.013

[4] Miller, D., Le Breton–Miller, I. (2017). Sources of entrepreneurial courage and imagination: Three perspectives, three contexts. Entrepreneurship Theory & Practice, 41(5): 667-675. https://doi.org/10.1111/etap.12281

[5] Nason, R.S., Wiklund, J., McKelvie, A., Hitt, M., Yu, W. (2019). Orchestrating boundaries: The effect of R&D boundary permeability on new venture growth. Journal of Business Venturing, 34(1): 63-79. https://doi.org/10.1016/j.jbusvent.2018.05.003

[6] Hamedmoghadam, H., Jalili, M., Yu, X. (2018). An opinion formation based binary optimization approach for feature selection. Physica A: Statistical Mechanics and its Applications, 491: 142-152. https://doi.org/10.1016/j.physa.2017.08.048

[7] Gilad-Bachrach, R., Navot, A., Tishby, N. (2004). Margin based feature selection-theory and algorithms. In Proceedings of the Twenty-First International Conference on Machine Learning, 43. https://doi.org/10.1145/1015330.1015352

[8] Li, Y., Lu, B.L. (2009). Feature selection based on loss-margin of nearest neighbor classification. Pattern Recognition, 42(9): 1914-1921. https://doi.org/10.1016/j.patcog.2008.10.011

[9] Kononenko, I. (1994). Estimating attributes: analysis and extensions of RELIEF. In European Conference on Machine Learning, 784: 171-182. https://doi.org/10.1007/3-540-57868-4_57

[10] Sun, Y. (2007). Iterative RELIEF for feature weighting: algorithms, theories, and applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6): 1035-1051. https://doi.org/10.1109/TPAMI.2007.1093

[11] Weinberger, K.Q., Saul, L.K. (2009). Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research, 10(2): 207-244.

[12] Chen, B., Liu, H., Chai, J., Bao, Z. (2008). Large margin feature weighting method via linear programming. IEEE Transactions on Knowledge and Data Engineering, 21(10): 1475-1488. https://doi.org/10.1109/TKDE.2008.238

[13] Bailetti, T. (2012). Technology entrepreneurship: overview, definition, and distinctive aspects. Technology Innovation Management Review, 2(2): 5-12.

[14] Autio, E., Acs, Z. (2010). Intellectual property protection and the formation of entrepreneurial growth aspirations. Strategic Entrepreneurship Journal, 4(3): 234-251. https://doi.org/10.1002/sej.93

[15] Lancee, B. (2010). The economic returns of immigrants' bonding and bridging social capital: The case of the

Netherlands. International Migration Review, 44(1): 202-226. https://doi.org/10.1111/j.1747-7379.2009.00803.x
[16] García, J.A.T., Skotnicka, A.G., Maeso-González, E. (2015). The role of the entrepreneur in the new technology-based firm (NTBF). In Enhancing Synergies in a Collaborative Environment, 225-232. https://doi.org/10.1007/978-3-319-14078-0_26
[17] Kennedy, W.P. (1987). Industrial structure, capital markets and the origins of British economic decline. CUP Archive, 115-119.
[18] Huggins, R.A., Izushi, H. (2007). Competing for knowledge: creating, connecting and growing. Routledge. Economic Geography, 3: 45-49.
[19] Ojedokun, O. (2012). Role of perceived fair interpersonal treatment and organization-based self-esteem in innovative work behavior in a Nigerian bank. Psychological Thought, 5(2): 124-140. https://doi.org/10.5964/psyct.v5i2.33
[20] Angulo-Guerrero, M.J., Pérez-Moreno, S., Abad-Guerrero, I.M. (2017). How economic freedom affects opportunity and necessity entrepreneurship in the OECD countries. Journal of Business Research, 73: 30-37.