A study of relationship between socio-economic factors and R&D funds based on spatial analysis

Dingyuan Xue*
Beijing National Day School, Beijing, China

*Corresponding author e-mail: hzhe_wjdi@163.com

Abstract. In this essay, we are going to study the relationship between socio-economic factors and R&D funds based on Spatial Autoregression. To start with, we introduce Moran’s I to describe a relationship between a variable and its spatial lag; After that, Principle Component Analysis is manipulated to operate a dimension reduction by orthogonal transformation; Ultimately, we use the Spatial Autoregression model to reflect the relationship between a province’s R&D funds and these five components as well as spatial lag of this province. According to the coefficients of the five principles, we find out that education and development of high-tech companies have positive impact on R&D funds while the population structure and employment structure also significantly influence the amount of funds.

1. Introduction
This paper will mainly discuss the relationship between socio-economic variables and the amount of R&D funds of the corresponding region. These variables include education, development of high-tech companies, and structure of employment. These data come from the National Bureau of Statistics. Principal component analysis is a standard tool in modern data analysis - in diverse fields from neuroscience to computer graphics - because it is a simple, non-parametric method for extracting relevant information from confusing data sets. With minimal effort PCA provides a roadmap for how to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified structures that often underlie it. In terms of the data we have selected, they may closely relate to each other. In order to reduce multiple collinearity, we decide to utilize PCA to reduce dimension and provide a highly simplified form of independent variables.

In this case, we apply Moran’s I to justify that there the R&D funds have a spillover effect within 31 provinces of China. Subsequently, the next step is to regress five components and spatial lag on R&D funds. Conventional models commonly used to analyze cross-section and panel data assume that variables are independent of one another. It is commonly observed that sample data collected for regions or points in space are not independent, but rather spatially dependent, which means that observations from one location tend to exhibit values similar to those from nearby locations. Ignoring this violation of independence between observations will produce estimates that are biased and inconsistent. Spatial regression methods allow us to account for dependence between observations, which often arises when observations are collected from points or regions located in space. To sum up, SAR is able to combine the results of both PCA and Moran’s I and provide us with an estimated linear relationship between these factors and R&D funds from the perspective of generalizing all provinces.
2. Characteristics of the distribution of R&D funds within 31 provinces

2.1. Spatial analysis

2.1.1. Introduction to Moran’s I model. The Moran scatter plot the relation between the commuting times dependent variable vector \( y \) and the average values of neighboring observations in the spatial lag vector \( W_y \), where we relied on the ten nearest neighboring provinces to form the matrix \( W \).

Hot spots of R&D funds can be either clustered or stay individually. In this study, a clustered distribution would be provinces with high R&D funds surrounded by provinces with high R&D funds; a discrete distribution would be provinces with high R&D funds surrounded by provinces with normal or low R&D funds. They can be identified using the global Moran’s I index:

\[
I = \frac{\sum_i \sum_j w_{ij} z_i z_j / S_0}{\sum_i z_i^2 / n}
\]

2.1.2. Results and explanation. According to the above formula, the global Moran’s I index and its significance are shown as follow:

Table 1. Results of Moran’s I.

| Year | Moran’s I | S.D. | z-score | p-value | Spatial Autocorrelation | Spatial distribution |
|------|-----------|------|---------|---------|-------------------------|---------------------|
| 2012 | 0.2592    | 0.109| 3.0531  | 0.01    | Positive                | Aggregation         |
| 2013 | 0.2933    | 0.1091| 3.0322  | 0.01    | Positive                | Aggregation         |
| 2014 | 0.2996    | 0.1092| 3.0839  | 0.01    | Positive                | Aggregation         |
| 2015 | 0.3015    | 0.1091| 3.1002  | 0.01    | Positive                | Aggregation         |
| 2016 | 0.3021    | 0.109 | 3.1082  | 0.01    | Positive                | Aggregation         |
| 2017 | 0.3181    | 0.109 | 3.2526  | 0.009   | Positive                | Aggregation         |

For statistical hypothesis testing, Moran’s I can be transformed to Z-scores. Given a set of factors and related attributes, this tool evaluates whether the examined pattern is clustering, discrete, or random. With z-score and p-value is used to indicate statistical significance, the result can be interpreted as follow: if Moran’s I value is positive, it shows a clustering trend while if it is negative, it indicates a discrete trend. Our model represents that the 31 provinces’ spatial distribution is classified as aggregation.

2.2. Clustering trend identifying model

2.2.1. Overview of the model. As the conclusion drawn by global Moran’s I index only reflects the general trend, which is not specific enough, we decide to use local Moran’s I statistics to better judge the clustering characteristics of our R&D funds.

2.2.2. Results and explanation. Using Geoda spatial analysis software, we can get the Moran scatter plot as below:
Figure 1. Local Moran’s I.

The scatter plot is then easily decomposed into four quadrants; these four quadrants reflect:

Quadrant I (upper right) provinces that have R&D funds above the average, where the average of neighboring provinces R&D funds is also above the average.

Quadrant II (upper left) provinces that have R&D funds below the average, but the average of neighboring provinces R&D funds is above the mean.

Quadrant III (lower left) provinces that have R&D funds below the average, where the average of neighboring provinces R&D funds is also below the mean.

Quadrant IV (lower right) provinces that have R&D funds above the average, but the average of neighboring provinces R&D funds is below the mean.

From the scatter plot, we see a positive association between points y associated with the horizontal axis and points Wy from the vertical axis, suggesting positive spatial dependence in province-level R&D funds. In fact, the magnitude of the slope from a line fitted through the points in the Moran scatter plot would equal Moran’s I. (The alternative hypothesis in this test is that the slope equals zero, indicating no spatial dependence.)

3. Factor Analysis model

3.1. Overview of the method

In the above model, we only conclude that there is a spatial correlation in R&D funds within 31 provinces, but we do not carry a specific analysis on the key factors of R&D funds based on the 16 sets of data we collect. Therefore, we decide to adopt PCA model to apply a dimensionality reduction on the socio-economic data of 31 provinces.

In principal component analysis, variables are often scaled. This is particularly recommended when variables are measured in different scales. When scaling variables, the data can be transformed as follow

\[
\frac{x_m - \text{mean}(x)}{sd(x)}
\]

After data standardization,

\[
P_{C_t} = \sum_{i=1}^{N} c_m x_m, \quad t = 1, 2, 3, \ldots, M
\]
In the above equation, we consider $PC_z$ as a principal component after dimension reduction, while $x_m$ as a socio-economic factor and $c_m$ as its coefficient. $N$ is known as the total number of socio-economic factors, and $M$ represents the number of principal components selected.

3.2. Results and explanation

![Scree plot](image)

**Figure 2.** Contribution of each dimension

From the plot above, we might want to stop at the fifth principal component. 89% of the information (variances) contained in the data are retained by the first five principal components.

![Variables - PCA](image)

**Figure 3.** Coordinates of variables

The plot above allows us to visualize variables and draw conclusion about their correlations. We also highlight the variables according to their quality of representation ($cos2$).
Positively correlated variables are grouped together. Note that factors relating to development of high-tech companies aggregate at the lower-right quadrant, just like factors correlating to education and employment.

Negatively correlated variables are positioned on opposite sides of the plot origin (opposed quadrants). Note that ratio of minorities and education are negatively correlated with each other.

The cos2 is represented by the distance between variables and the origin, which measures the quality of the variables on the factor map. Variables that are away from the origin are well represented on the factor map. Note that factors correlating to education, employment, and development of high-tech companies contribute greatly to dimension one and two.

We can visualize the cos2 of variables on all the dimensions:

![Figure 4. Cos2 of variables](image)

It can be seen that number of high-tech companies, total revenue of high-tech companies, and educational fund contribute the most to principal component one and two.
Figure 5. Individual’s PC values

Note that, similar individuals are grouped together on the plot. For instance, underdeveloped provinces like Xinjiang, Gansu, Ningxia etc. are grouped at the lower left quadrant and note that Beijing, Jiangsu, and Guangdong contribute greatly to either dimension one or two, suggesting they are well-developed.

4. Spatial Autoregression model

We can use the spatial autoregressive process to construct an extension of the conventional regression model. The model has been labeled the spatial autoregressive (SAR) model. The dependent variable vector $y$ is of dimension $n$ by 1, containing (logged) R&D funds for each region/observation. The $n$ by $k$ matrix $X$ contains exogenous explanatory variables possibly including a constant term vector, and the $k$ by 1 vector $\beta$ are associated regression parameters. The $n$ by 1 spatial lag vector $Wy$ reflects an average of (log) R&D funds from neighboring regions specified by the matrix $W$, and the associated scalar parameter $\rho$ reflects the strength of spatial dependence. When the scalar parameter $\rho$ takes on a value of zero, the model in (3) simplifies to the conventional linear regression model. Finally, we assume the $n$ by 1 disturbance vector $\epsilon$ contains independent, normally distributed terms with a vector mean zero $(0_{nx1})$, constant variance, $(\sigma^2)$.

$$y = \rho Wy + X\beta + \epsilon$$

$$y = (I_n - \rho W)^{-1}X\beta + (I_n - \rho W)^{-1}\epsilon$$

$$\epsilon \sim N(0_{nx1},\sigma^2)$$

The spatial regression model adds a spatial lag vector, which takes the form of the average of the neighboring regions, reflecting the average commuting times from neighboring regions to help explain variation in commuting times across the regions.

4.1. Results and explanation

Using the Geoda software, we can get the final result as below:
Table 2. Results of SAR

| Variable     | Coefficient | Std.Error | z-value | Probability |
|--------------|-------------|-----------|---------|-------------|
| W_RD2016     | 0.1909      | 0.1011    | 1.8892  | 0.0589      |
| CONSTANT     | -0.0064     | 0.0646    | -0.0985 | 0.9215      |
| PC1          | 0.2978      | 0.0262    | 11.3581 | 0.0000      |
| PC2          | -0.0874     | 0.0402    | -2.1706 | 0.0300      |
| PC3          | -0.0040     | 0.0444    | -0.0897 | 0.9285      |
| PC4          | 0.1746      | 0.0568    | 3.0758  | 0.0021      |
| PC5          | 0.2602      | 0.0655    | 3.9719  | 0.0001      |

Additionally, we note that the value of R2 is 0.87.

Table 2 reports the result of SAR. The p-value illustrates that PC3 does not significantly contribute to R&D funds. We notify that the coefficient of is 0.1909, suggesting that the investment of R&D funds of a province has a significant spillover effect. It is also worth emphasizing that coefficients of PCs indicate its contribution to the ultimate R&D funds, which can be clearly seen that PC1, mainly including information of developing companies, has largest effect on R&D funds of a region. Additionally, the positive sign suggests a positive relationship between a PC and R&D funds and vice versa, indicating that PC2, mainly containing employment information, negatively influences R&D funds.

Table 3. Results of SAR

| Variable     | Coefficient | Std.Error | z-value | Probability |
|--------------|-------------|-----------|---------|-------------|
| W_RD2016     | 0.3123      | 0.2168    | 1.4404  | 0.1498      |
| CONSTANT     | 0.2218      | 0.1946    | 1.1398  | 0.2544      |
| Ratio_rity   | -1.5526     | 0.7737    | -2.0068 | 0.0448      |

As the ratio of minorities is not significantly represented in each of PCs, we simulate this variable separately. The coefficient clearly reports the negative contribution it gives to R&D funds. compared with the coefficient of PC1, which is 0.2978, the absolute value also suggests the significance of the minority’s effect.

5. Conclusion

The paper attempts to shed some light on the spatial dimension and explanatory variables of local R&D funds. This analysis is applied to a sample of 31 regions in China from 2012 to 2017, a developing country. Based on three models, this work tries to examine whether regional R&D funds depend on local education, technology, or employment and whether regional R&D funds have a spillover effect. We ultimately obtain a formula and substantiate the spillover effect.

In conclusion, both technology and education have great impact on regional R&D funds while population dedicating in traditional industries may drag the local R&D funds. Moreover, the sign of the coefficient of minorities and large absolute value of this coefficient suggest that regions with large population of Han nationality enjoy better research and development funds. This trend greatly reminds us to reflect and to pay effort on development of minority-prevailing regions.

References

[1] Carboni, & Oliviero, A. . (2013). A spatial analysis of r&d: the role of industry proximity. Economics of Innovation and New Technology, 22 (8), 820-839.

[2] Audretsch, D. . (1996). R & d spillovers and the geography of innovation and production. American Economic Review, 86 (3), 630-640.

[3] Shlens, J. (2014). A Tutorial on Principal Component Analysis. ArXiv, abs/1404.1100.
[4] Bernardí Cabrer-Borrás, & Serrano-Domingo, G. . (2007). Innovation and r&d spillover effects in spanish regions: a spatial approach. Research Policy, 36 (9), 0-1371.

[5] Arbia, & Giuseppe. (2014). A primer for spatial econometrics ||. , 10.1057/9781137317940.

[6] Fingleton, & Bernard. (2008). A generalized method of moments estimator for a spatial panel model with an endogenous spatial lag and spatial moving average errors. Spatial Economic Analysis, 3 (1), 27-44.