Spatial bootstrapping using deep clustering methods: spatial machine learning applied to Lombardy high-tech businesses

Alessio Bumbea, Giuseppe Espa and Andrea Mazzitelli

1. Introduction

The productive structure of the high-tech sectors is rich and articulated: it is a world characterised by high heterogeneity in terms of the type of productive specialisations, in terms of different relevance on the economic activities and in terms of the average size of the players. Innovation is considered a central driver of economic development and the growth and competitiveness of firms (Schumpeter, 1934; Romer, 1986).

Thanks to this approach to promoting creative entrepreneurship, technology transfer, and intensifying market competition, policymakers' long-term goal is to speed up the industrial revolution and increase production levels, competitiveness, and efficiency in the whole economy (Giuliani et al., 2023).

The determinants of location in high-tech industries are market size, availability of high-quality resources such as scientific infrastructure and supply of skilled labour, and agglomeration effects from the proximity of other companies and public knowledge centers.

The key technical innovations that propel regional economic growth include agglomeration and spillovers. Agglomeration effects and spillovers are recognized to dominate the economic geography of innovation, where geographical closeness and localised learning processes (Krugman, 1991; Saxenian, 1996) promote innovation and economic growth (Ellison and Glaeser, 1997; Glaeser, 1999). Through information spillover processes of localization and urbanisation, agglomeration leads to the concentration of diversity and a different partitioning of geographic areas (Porter, 1990; Duranton and Overman, 2005). Nevertheless, understanding the role that spatial partitioning plays concerning the diversification of technology and tacit knowledge (Polanyi, 1966) has always been a critical aspect of statistical analysis.

Several methodologies have been proposed in recent years to assess the presence of clusters of high-tech firms. Distance-based methods are frequently used to analyse spatial structures in economics, namely, which geographic distances firms produce specific goods are more spatially clustered regarding their technological content at any scale. (Arbia et al., 2012, Kopczewska, 2021; Marcon and Puech, 2023). We refer to this phenomenon as the clustering of firms and the clustering of economic activities. For this purpose, marked point processes are used in the specialised literature and Ripley's K function is an estimator used to characterise the correlation of such spatial point processes.

2. Objective

To investigate how firms are spatially clustered, we introduce a spatial machine learning algorithm (Kopczewska, 2022) that utilises the geographical coordinates of firms but also takes into account the other attributes and characteristics of the firms.

The source of information that we used is the ASIA-ISTAT (Registro Statistico delle Imprese Attive) database by tax code[1]. The analysis is conducted using micro-data including different characteristics of firms as the sector of activity of each enterprises using a 5-digit code called ATECO, the local version of the European NACE code. The available data are updated to 2020 but we restricted the dataset to firms that have been active in the years 2017, 2018, and 2019 to avoid distorting effects with the outbreak of the pandemic from Covid 19. Finally, the cleaned dataset had a total of 109 features for 24976 businesses.

We focused on high-tech enterprises from the following sectors:

1. Pharmaceuticals
2. ICT Manufacturing: semiconductor, computer and TLC hardware
3. Aerospace: aircraft manufacturing, spacecraft and related devices, satellite communications
4. Biomedical: electromedical and medical devices
5. ICT Trade: wholesale and retail trade of ICT equipment
6. IT Services: software publishing and production, IT consulting, database management
7. Telecommunications: fixed and mobile telecommunications.

2.1 Methods

The method we propose is to calculate quantities of interest, such as the correlation between two variables. Specifically, we want to calculate productivity at two different temporal instants, measured as value added per employee, using a stratified bootstrap technique. We briefly review what stratified bootstrap is and then introduce our proposed method for creating an efficient division of the original dataset into layers of firms with similar characteristics using a deep clustering algorithm called Deep Embedded Clustering (DEC).

2.2 Stratified bootstrap

This method assumes that the entire population is divided into distinct subgroups or 'strata,' formed based on shared features. Within each stratum, multiple samples are drawn with replacements. This means that the same individual or element can be chosen more than once in each sample. This process ensures that the variability within each stratum is captured. The sampling process is repeated many times to create several bootstrap samples. Each sample is a mini-representation of the population,

[1] The ASIA-ISTAT contains information about the firms' structure, financial situation, and whether and what they export. Further information is available at https://www.istat.it/it/archivio/216767 (last access 17/01/2024)

respecting the stratified structure. The correlation coefficient is computed for each bootstrap sample, and therefore, a distribution for this variable is obtained.

2.2 Deep clustering

The main issue now is how these strata can be created. Firms should be grouped into homogeneous clusters concerning specific characteristics and as different as possible from those in the other clusters. The proposed architecture is based on the DEC algorithm presented in (Xie et al. 2016), one of the most well-known models in deep clustering (Gonzales, 2022). The entire proposed architecture is presented in Figure 1.

First, the dataset has been pre-processed by removing the target variables that will be used in the computation of correlation coefficients; in this way, we ensure that they don't influence the process that constructs the strata. The ASIA-ISTAT database has no missing data by construction; therefore, no imputation technique needs to be implemented. On the other hand, several variables are nominal categorical variables incompatible with neural networks that require entirely numerical variables as inputs. This problem can be solved by implementing an embedding technique that maps the original dataset to a higher dimensional but entirely numerical space. The embedding procedure has been implemented using a Tensorflow Stringlookup[2]. Once a numerical representation of the dataset becomes available, using a traditional clustering algorithm like k-means to produce the clusters and use them directly as the strata for the stratified bootstrap would be tempting.

Unfortunately, these predictions would not be reliable, given that the embedded space is highly dimensional and most traditional clustering algorithms suffer from the curse of dimensionality. Still, they can be used to initialize more sophisticated algorithms.

The DEC algorithm nonetheless requires an initial "naive" guess for the clusters to improve upon. Therefore, a MiniBatchKMeans model, a lighter version of KMeans adapted for extensive data settings, has been trained on the embedded data and fed to DEC.

The DEC model also requires training an autoencoder to work. An autoencoder is a Neural Network designed to take a starting high-dimensional dataset and project it into a low-dimensional "latent space" (encoding) and then back to its original dimension (decoding). This model is a nonlinear generalization of dimensionality reduction techniques like PCA for our intended purposes. The encoding part of the autoencoder and the simple clustering guess are combined into the DEC.

This model then iteratively adjusts the "latent space" according to the points assigned to clusters with high reliability and updates the groups according to this new feature

[2] Further information about the StringLookup layer and the embedding process can be found in the official Tensorflow documentation https://www.tensorflow.org/api_docs/python/tf/keras/layers/StringLookup (Last access 26/01/2024)

representation. This procedure is iteratively repeated to improve the clustering procedure's performance.

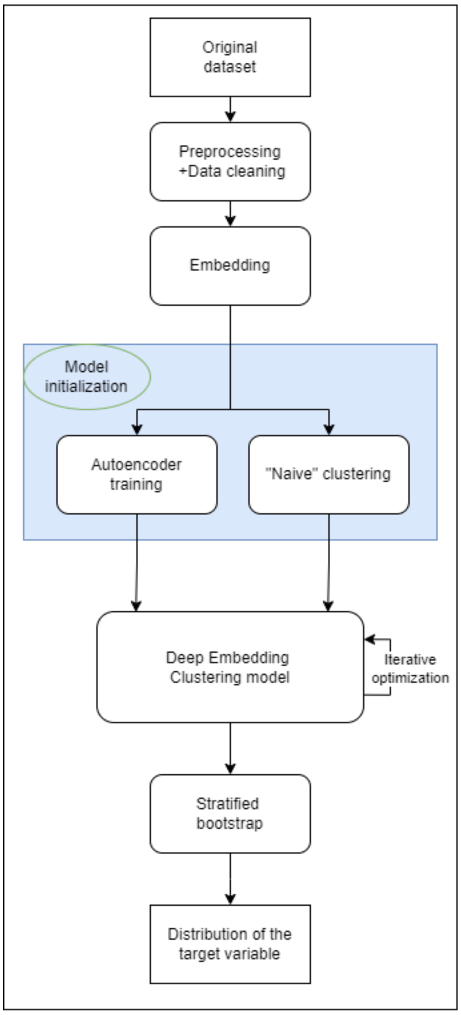


Figure 1 Summary of the applied method: first we prepare the data and embed them, then we initialise the model's latent space structure and clusters, subsequently we ran the DEC model and then use its clusters as strata for the bootstrap.

3. Results

Once the algorithm converges, the clusters can be used as strata for the Stratified bootstrap algorithm (Kopczewska, 2021). The number of clusters found is 15, with the larger clusters comprising 6,598 businesses and the smaller ones comprising 15 firms.

In the present study, we will compute the correlation between business productivity between 2018 and 2019. The results are presented in Figure 2. The correlation coefficient is 0.815, which is lower than the correlation on the full sample (0.872) due to the heavy left tail of the distribution. Figure 3 presents a map of the Lombardy region with the high-tech firms colored according to their cluster. The algorithm is not affected by spatial inhomogeneity and can assign neighboring firms to different clusters.

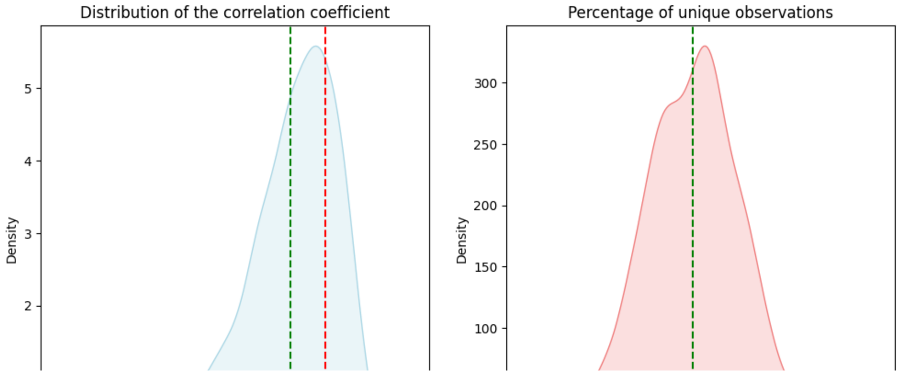
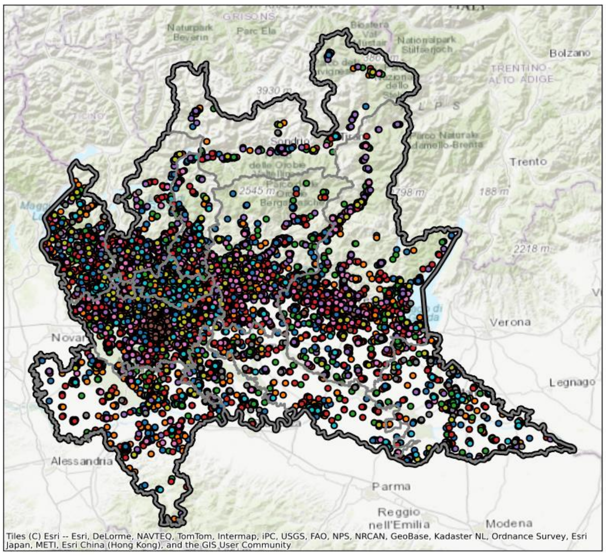


Figure 2 The correlation between the productivity between the years 2018 and 2019. The average coverage of each bootstrap replica (in the right figure) is low because the clusters have different sizes



Tiles (C) Esri -- Esri, DeLorme, NAVTEQ, TomTom, Intermap, iPC, USGS, FAO, NPS, NRCAN, GeoBase, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), and the GIS User Community

Figure 3 Image of the Lombardy region where each business is coloured according to its cluster. Note that even though the latitude and longitude are two considered variables, the clusters are mixed.

What the identified clusters share is that they use the same or similar production technology. nonetheless, “if [the production of] two goods...require similar institutions, infrastructure, physical factors, technology, or some combination thereof, they will tend to be produced [in the same location],” (Hidalgo et al., 2007, 484). Moreover, the spatial co-production of products indirectly catches their production technology similarity (Neal et al., 2022). Therefore, the clusters determined considered some characteristics of the firms such as economic activity, technology, turnover, number of employees, geographic localization, etc. It implies that two geographically close companies may belong to distinct clusters because they differ in these attributes. These differences result in large distances in latent space that the clustering algorithm can capture.

4. Conclusions

In the present paper we applied a spatial bootstrapping algorithm to the high-tech businesses of the Italian Lombardy region. The spatial bootstrapping algorithm proposed is the stratified clustering algorithm. This algorithm assumes that the data are separated in strata of similar data points from which sample with repetition. These strata have been created using a deep clustering algorithm called Deep Embedding Clustering which is based on an Autoencoder Neural Network. The algorithm was able to cluster the data successfully by using the characteristics of the businesses and execute the bootstrap to compute the distribution of the correlation coefficient.

References

1. Arbia G, Espa G, Giuliani D, Mazzitelli A (2012) Clusters of firms in an inhomogeneous space: the high-tech industries in Milan. Economic Modelling 29(1):3–11

2. Duranton, G., and Overman, H. G. (2005) Testing for Localization Using Microgeographic Data. Rev. Econ. Stud. 72, 1077–1106.

3. Ellison, G., and Glaeser, E. L. (1997). Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach. J. Polit. Economy 105, 889–927.

4. Giuliani, G., Toffoli, D., Dickson, M.M., Mazzitelli, A., Espa, G. (2023) Assessing the role of spatial externalities in the survival of Italian innovative startups, *Regional Science Policy and Practice,* DOI: 10.1111/rsp3.12653

5. Glaeser, E. L. (1999). Learning in Cities. Journal of. Urban Economics. 46, 254–277.

6. Gonzales, F.F. (2022) State of the Art on: Deep Clustering. Politecnico di Milano, Honours Programme, March, CSE Track

7. Hidalgo, C. A., Klinger, B., Barabási, A.-L., & Hausmann, R. (2007). The product space conditions the development of nations. Science, 317(5837), 482–487

8. Kopczewska, K. (2021) Applied Spatial Statistics and Econometrics, 448-455. Routledge

9. Kopczewska, K. (2022) Spatial machine learning: new opportunities for regional science, The Annals of Regional Science (2022) 68:713–755

10. Krugman, P. R. (1991). Geography and Trade. Cambridge: MIT press.

11. Marcon, E., Puech, F. (2023) Mapping distributions in non-homogeneous space with distance-based methods, Journal of Spatial Econometrics, 4-13

12. Neal, Z.P., Domagalski, R., and Bruce Sagan (2022) Analysis of spatial networks from bipartite projections using the R backbone package, Geographical analysis, 54(3), 623-647

13. Polanyi M. (1966) The Tacit Dimension, Anchor Books, New York

14. Porter, M. E. (1990). The Comparative Advantage of Nations. New York: Free Press.

15. Romer, P.M. (1986) Increasing Returns and Long-Run Growth, *The Journal of Political Economy*, 94(5), 1002-1037

16. Saxenian, A. (1996). Regional Advantage: Culture and Competition in Silicon Valley and Route 128, with a New Preface by the Author. Cambridge: Harvard University Press.

17. Schumpeter, J. A (1934) The theory of economic development. Cambridge, MA: Harvard University Press

18. Xie, J., Girshick, R., and Farhadi, A. (2016) Unsupervised deep embedding for clustering analysis. International conference on machine learning, PMLR, 478–487.