CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17795

In What is Europe Investing? A Text Mining Approach on Cohesion Projects

Nicola Caravaggio, Giuseppe Di Renzo, Laura Fanelli, Giuliano Resce, Agapito Emanuele Santangelo
Department of Economic, University of Molise, Italy.

How to cite: Caravaggio, N.; Di Renzo, G.; Fanelli, L.; Santangelo, A. E. 2024. In What is Europe Investing? A Text Mining Approach on Cohesion Projects. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17795

Abstract

In this Paper we analyze the dynamics of the interventions of European cohesion policies in the Italian territory - according to a regional division - focusing on the topics financed, extracted through text mining, in particular with the term document matrix and term frequency – inverse document frequency. From the analysis conducted on 51.971 projects financed from 2000 to 2022, a clear preponderance of the topics of internships, civil service and more generally aspects relating to the world of work emerges. Although these are the most frequent topics, by weighing the frequency of words in the summary of the call with the total value of the financing, it emerges that the most costly projects in terms of resources are those linked to the development of infrastructure. This innovative approach allows an innovative understanding of the spending trends, useful to improve the action of the European political decision- maker.

Keywords: *Cohesion Policy; Text mining; Public investments; Term Document Matrix; Term frequency – inverse document frequency*

1. Introduction

The European Cohesion Policy is a series of interventions and measures aimed at reducing economic and social inequalities among territories, thus promoting homogeneous, inclusive, and long-term growth. The Cohesion Policy is provided for by art. 119 of the Italian Constitution and the Treaty on the Functioning of the European Union. Cohesion Policy essentially finds its genesis in the very definition of the European Union, or rather from the observation that the creation of a single market, as well as a regulation valid for all member states, would inevitably have aggravated the disparities and differences in terms of per capita income and wealth of the territories. Cohesion interventions involve central authorities and local administrations equally. When evaluating the impact of Cohesion Policies, it is necessary to focus more on the outputs of the interventions, in terms of reducing economic-social inequalities, rather than on the quantity of resources used, which represent only the amount of expense.

But, in the study of the long-term growth process it is believed that capital accumulation is not enough to guarantee continuous and sustainable growth in the long term. Neoclassical economic analysis believes that, alongside the accumulation of capital, a driver of growth is agglomeration, and therefore the movement of large masses of individuals from rural and/or peripheral areas towards metropolitan centers, favoring thus the contamination of knowledge, ideas, and therefore enrichment of human capital and growth of technological progress (Marshall, 1890) The empirical evidence found in recent years highlights a dichotomous picture: dynamic urban agglomerations and remote regions characterized by low growth rates and depopulation (Iammarino et al., 2019). From this, it follows, as a logical corollary, that the colorful European picture is composed on the one hand of prosperous nations (central-northern Europe), while other areas are characterized by low growth and/or low income (southern Europe, Eastern Europe). Considered analytically, even the leading States in terms of wealth and economic growth present internal conditions of imbalance also in terms of wealth distribution.

Therefore, in the analysis of growth processes it is impossible to ignore the consideration that in Europe many regions have very different growth paths, compared to the performances of the States to which they belong (Cuaresma et al., 2014). This means that a prosperity of a Nation does not necessarily determine homogeneous prosperity across its entire territory (Liberati et al., 2022). Rather, it is the local, atomistic dynamics that guide the progress (Petrakos et al., 2011), and therefore the analysis of regional gaps and internal dynamics constitutes a crucial point for adapting cohesion interventions to the emerging needs of the territories, which are increasingly attentive to attributes such as sustainability, respect for the environment, the protection of workers' rights.

However, if today we still continue to debate the effectiveness of the EU's cohesion policy – despite the fact that its importance has become clear recently, during the pandemic period (European Commission, 2022) – it means that, not considering merely ideological contestations, there are problems that do not allow cohesion interventions fully express their potential. On the one hand, its detractors point out that, despite the enormous amount of resources used in recently years, the disparities have worsened only modestly, and that in reality, the Cohesion Policy only serves to "keep public employees and resources busy", but at the same time preventing them from carrying out and carrying out more productive activities (Molle, 2007); others, however, believe that the cohesion interventions have actually achieved good results (Gagliardi & Percoco, 2016). The truth is probably somewhere "in the middle": Cohesion policy is important, but it still suffers from a certain delay in its implementation due to the combined effect of various factors, such as the inability to plan and spend these resources. Regardless of these considerations, one fact remains constant, and remains a topos in this context, at least as regards the levels of effectiveness of the Cohesion Policy and the investments related to it: the low quality of the institutional context, a determining factor in guaranteeing the operation of the convergence and growth mechanisms of the European Regions. In particular, the ability of local political decision-makers proves crucial (Fratesi & Whishlade, 2017) in ensuring that large investments produce the corresponding fruits (Arbolino & Boffardi, 2017). The latter, in particular, are crucial to allow the so-called less developed Regions to become Regions in transition and subsequently, hopefully, more developed. According to the data provided by OpenCoesione, for the 2014-2020 programming cycle, taking into account the Italian framework, only 13% of the projects were concluded. The data is striking, and prompts reflections especially regarding the implementation time of public works in Italy. In a specific report from the Agency for Territorial Cohesion, the figure of the slowness regarding the construction and completion of public works emerges.

The average implementation time of the infrastructure works is approximately 4.4 years, but progressively increases as the economic value of the projects increases. These timings are necessarily also influenced by the reference contexts.

This problem emerges, naturally, for those European areas, such as southern Italy, where in the last twenty years low growth rates have been recorded, if not close to zero or negative as in the case of real wages, also deriving from the fact that investments carried out have been victims of the so-called "distributive drift", with a consequent excessive fragmentation of the interventions and therefore also of their unitary dimension (Agrello, 2019). This favors a sort of sub-optimization phenomenon: the management of resources by local political decision-makers - from a realistic and rational point of view interested in their reconfirmation and therefore in maintaining their office - will inevitably be directed towards political goals, favoring only the preservation of the status quo, without any real prospects for change and modernization being revealed. From this perspective - while distancing ourselves from any form of demonization of the work of local politicians

- in order to bring about a real positive turning point, it would be appropriate to disconnect the operation of cohesion policies from the political cycle and from the purely electoral logic, through actions long-term (such as improvements in education) and the provision of tools to support the action of local public administrations.

Despite the increase in monitoring initiatives for cohesion policies, the majority of indicators primarily focus on expenditure, with limited evidence attempting to understand the actual impact of the projects. This paper aims to address this gap by employing text mining techniques to analyze all European projects financed in Italy from 2000 to 2022 Preliminary results offer fresh insights into the priorities targeted by cohesion funding, paving the way for a novel approach to policy analysis.

2. Methods

Advances in machine learning research offer great potential for international development agencies to leverage the vast information generated from accountability mechanisms to gain new insights, providing analytics that can improve decision-making. (Resce, Garbero & Carneiro, 2021)Within digitally based content analysis approaches, text Mining is broadly defined as an Artificial Intelligence (AI) technique that uses Natural Language Processing (NLP) to transform unstructured text of documents/databases such as web pages, newspaper articles, e-mails, fundings, press, posts/comments on social media, in structured and normalized data (Resce & Maynard 2018). Words, the carriers of meaning, are identified and transformed into a processable data structure.

Indeed, many studies focus on this technique to provide an aid to policy maker to optimally allocate resources: Resce et al, (2021) explores how text mining can harness existing project data to uncover latent information about food systems dynamics taking 900 projects of the International Fund for Agricultural Development (IFAD); Choudhary et al (2009), instead, identified text mining as a potential tool for addressing the identified problems of Post-Project Reviews.

In our study, text mining has been employed to analyze 50.971 projects implementing cohesion policies from 2000 to 2022, funded by Structural Funds, the National Fund for Development and Cohesion (FSC), and the Cohesion Action Plan (PAC). Our objective is to identify predominant themes in the projects and analyze their evolution over the years, with particular attention to the division into the macro-areas of Italy, namely "Centre-North" and "South". (Open Coesione, 2023)

The first step in the analysis involved retrieving data, including the project code, title, summary, destination macro-area, funding amount, and project start date. The text chosen for analysis was the summary of each project, which was extracted and aggregated based on the project's start year and month. Subsequently, the analysis corpus was prepared using functions from the R package "tm" (Feinerer & Hornik, 2018; Feinerer, Hornik & Meyer, 2008): punctuation, stop words (e.g. words like "the," "is," "of," etc.), and numbers were removed from the corpus. The words were then converted to lowercase and stemmed. Finally, a Term Document Matrix was produced, with the project's start year and month as columns (126) and words as rows (28 928 unique terms). The Term Document Matrix indicates the number of times each word appears in each project started in that month and year. It serves as the starting point for text mining by transforming unstructured text into numerical data.

A central question in text mining is how to quantify what a document is about. One measure of a word's importance is its term frequency (tf), which counts the occurrence of a word in a document. Another approach is to consider the inverse document frequency (idf), which decreases the weight of commonly used words and increases the weight of words that are not frequently used in a collection of documents. The two can be combined to calculate the tf-idf of a word (the product of the two quantities), which measures the frequency of a word adjusted for how rarely it is used (Silge & Robinson, 2017). Formally:

$$\mathrm{idf(term)} = \ln\left(\frac{\mathrm{n_{documents}}}{\mathrm{n_{documents\ containing\ term}}}\right) \tag{1}$$

The statistic tf-idf is widely used to measure how important a word is to a document in a collection of documents (Silge & Robinson, 2017). In our case, the tf-idf combines frequency - how many times a word is associated is associated to month and year of the project's start - and the inverse of ubiquity - how exclusive the association is between a word and month and year of the project's start. To this regard, it is worth stressing that more ubiquitous words are more likely to have less informative power than exclusive words.

The Term Document Matrix and tf-idf statistics were initially computed for all projects and later differentiated for projects targeted at the “Center-North”, composed by twelve regions (Liguria, Lombardia, Piemonte, Valle d'Aosta, Emilia-Romagna, Friuli Venezia Giulia, Trentino-Alto Adige, Veneto, Lazio, Marche, Toscana and Umbria) and “South” of Italy, so called “Mezzogiorno” composed by eight regions (Abruzzo, Basilicata, Calabria, Campania, Molise, Puglia, Sardegna and Sicilia). Using the functions from the R package “wordcloud2”, project keywords were incorporated into word clouds, visual representations where a word's size is proportional to its frequency (Silge & Robinson, 2017).

Subsequently, graphs were generated to illustrate the trend of words over time. Finally, to comprehend the economic value associated with the most frequent words in the projects, a new matrix was created by weighting the Term Document Matrix according to the project's funding value where the word is present.

3. Results

Among the 50.971 projects implemented under cohesion policies from 2000 to 2022, the findings in Figure 1 show that the most frequent words, using the Term Document Matrix and the Term Frequency-Inverse Document Frequency, are nearly identical

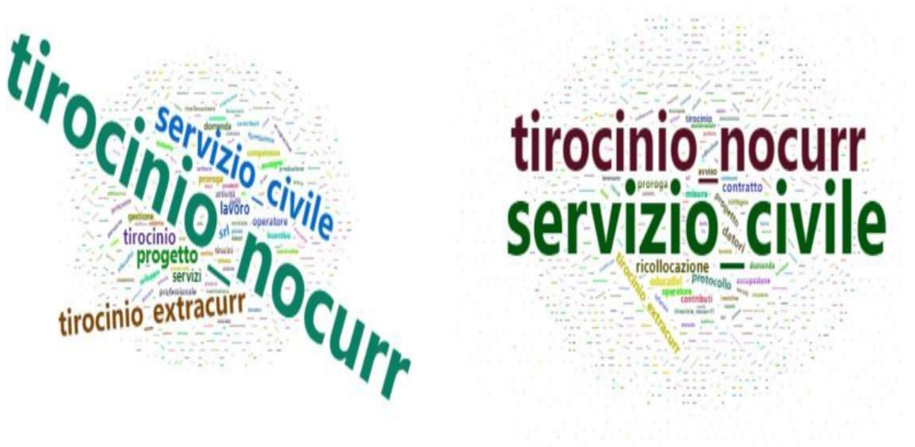

Figure 1. Comparison of the most frequent words in projects using the Term Document Matrix (wordcloud on the left) and the Term Frequency-Inverse Document Frequency (wordcloud on the right).

Most projects are dedicated to internships and civil service. Among the internships, non-curricular and extracurricular internships stand out, aimed at training young individuals for the world of work, followed by curricular internships targeting students to integrate professional training with academic education. Similarly, projects dedicated to civil service represent a significant opportunity for training and professional maturity. This highlights how cohesion policies prioritize youth growth and labour.

Examining Figure 2, the trend over time of the frequency of the top 10 words in the calls for proposals is evident, using both Term Frequency and Term Frequency-Inverse Document Frequency.

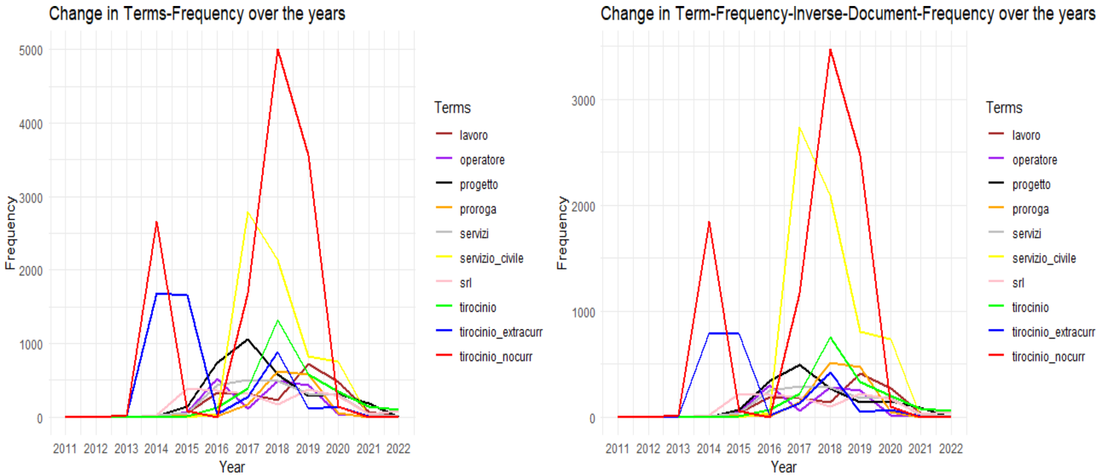

Figure 2. Variation in the frequency of the top 10 words over time from 2011 to 2022 using the Term Document Matrix (left plot) and the Term Frequency-Inverse Document Frequency (right plot).

The x-axis represents the project initiation year, while the y-axis represents the frequency. Two peaks are evident, the first in 2014 and the second in 2018, corresponding to the years when the majority of calls containing the most frequent words were implemented. It is noteworthy that from 2019 onwards, the lines began to decline, showing a trend towards flattening in the years 2020-2022, during the Covid-19 pandemic.

In this context, it is interesting to examine Figure 3, which focuses on the years 2020-2022 and illustrates how, although calls related to civil services continue to predominate, attention has shifted towards projects related to employment and occupation, as well as others addressing contributions and incentives.

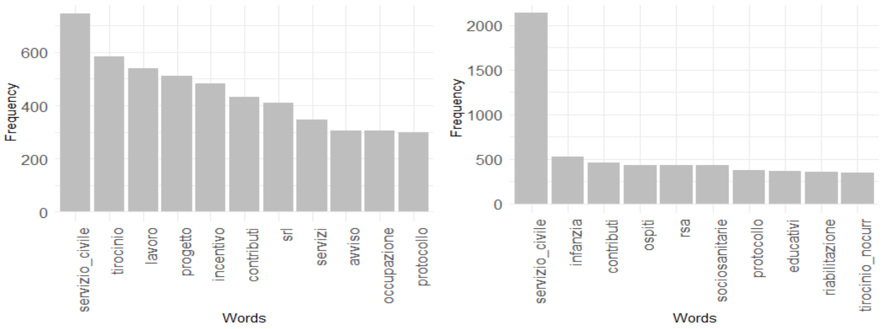

Figure 3. Frequency of the top 10 words from 2020 to 2022 using the Term Document Matrix (left plot) and the Term Frequency-Inverse Document Frequency (right plot).

As mentioned earlier, there exists an economic and social gap among Italian regions, confirmed by the number of initiated projects. In the North-Central area, there were 44 255 projects (86.82%), while in the South, there were only 6 716 projects (13.18%).

Figures 4 and 5 present the results of the repeated analysis, dividing the calls into the macro areas 'North-Central' and 'South'.

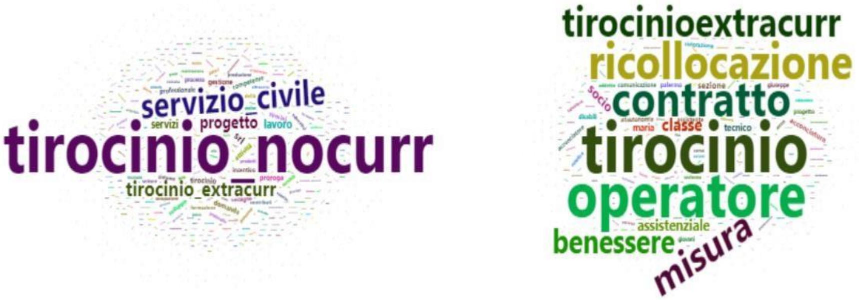

Figure 4. Comparison of the most frequent words in projects with the Term Document Matrix between the "Central_North" macroarea (left wordcloud) and the "Southern" macroarea (right wordcloud).

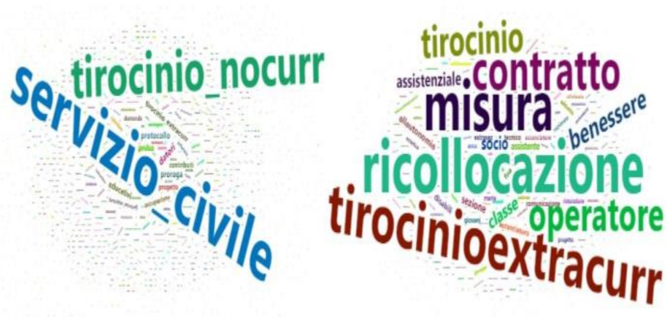

Figure 5. Comparison of the most frequent words in projects with the Term Frequency-Inverse Document Frequency between the "Central_North" macroarea (left wordcloud) and the "Southern" macroarea (right wordcloud).

While the most common words in the North-Central region coherently reflect what is illustrated in Figure 1, new key terms emerge in the South, such as "relocation". "contract", "well-being" and "welfare". These highlight the ongoing challenge related to occupational crisis situations in the region. Indeed, one of the objectives of the calls is to mitigate the effects of business difficulties in the involved areas and promote the preservation of employment levels.

In Figure 6, the trend from 2011 to 2022 of the frequency of the most common words in the calls has been compared, dividing them by macro area.

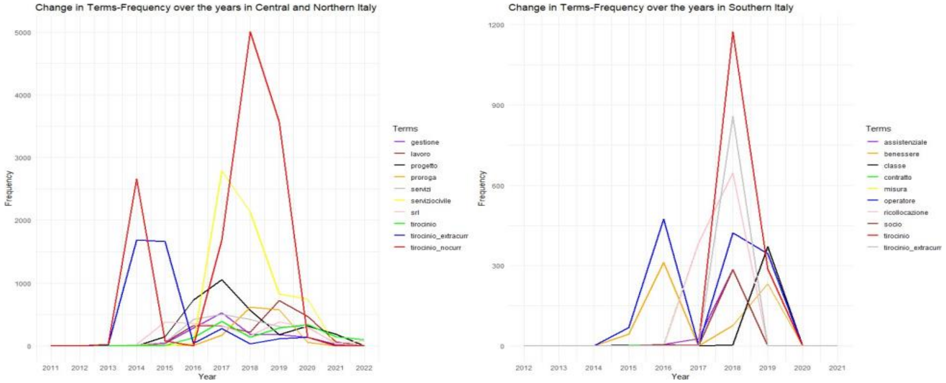

Figure 6. Comparison of change in frequency of the first 10 words over time from 2011 to 2022 with the Term Document Matrix between the "Central_North" macroarea (left plot) and the "Southern" macroarea (right plot).

In the North, the projects containing the most frequent words began in 2013, reaching a peak in 2014, followed by a period of contraction in 2016. In the South, they started later, in 2014, with a peak in 2016 and a sharp decline in 2017. Both macro areas experienced another peak in 2018. However, while in the North-Central region, a certain level of project initiation has been maintained even after 2020, even if in a smaller size; in the South, there has been a total interruption of new projects. What happens from 2020 to 2022? While the left graph in Figure 7 confirms the previous observations about the Central- Northern region, the right graph displays the most frequent words in projects for Southern Italy. It becomes evident that the funding is primarily directed towards contributions to individuals for acquiring services, including training, with a particular focus on specific roles such as "esthetic treatment technician." The specific objective is to reduce the rate of early educational failure and scholastic and training dispersion.

The remaining words are related to projects launched for specific municipalities in the Sicily region, including "Catania," "Acireale," "Agrigento," and "Caltagirone," which are about social inclusion and the fight against poverty. These initiatives aim to increase, consolidate, and qualify care services and infrastructure.

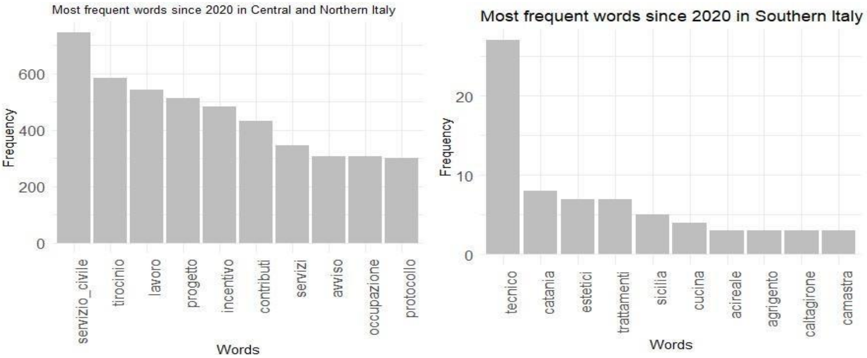

Figure 7. Frequency of the first 10 words from 2020 to 2022 with the Term Document Matrix between the "Central North" macro area (left plot) and the "Southern" macro area (right plot).

In returning to the comprehensive analysis of the calls, it is valuable to identify the words that exhibit a correlation with higher funding amounts. Let's focus on Figure 8, which considers not only the frequency of words but also the associated funding values.

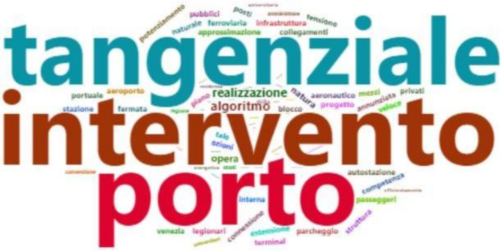

Figure 8. Words in the tenders weighted by the value of European funding

Considering that cohesion policies aim to reduce territorial disparities, it is not surprising that among the projects with a more significant economic impact are those focused on the development of infrastructure, such as roads, ports, and railway networks. This helps us understand that, although many projects focus on education and training (civil service, internships, etc.), the most substantial funding and truly impactful projects are directed towards the transport and mobility sector.

4. Conclusions

This study investigates what the projects implementing cohesion policies describe, through text mining analysis on administrative documents produced by local authorities.

Results shows how a huge part of funded projects focus on civil service and internships, hence they raise questions concerning economic policy and long-term development. While such initiatives can be effective in providing temporary support during periods of declining employment, it is critical to recognize that they primarily address the immediate consequences of unemployment rather than facing the structural causes that contribute to the declining employment itself.

Furthermore, the inappropriate use of civil service projects and internships may reflect a lack of strategic investment in innovation, education and professional training, which are crucial pillars for stimulating sustainable and inclusive economic growth the long term. Without adequate attention to these sectors there is a risk that national economic growth will remain limited, and society will continue to rely mostly on temporary measures to address employment challenges.

Additionally, it is important to consider the role of employment quality in the economy. While internships can offer work experience or a first entry into the job market for youths, if they are not accompanied by opportunities for professional and salary growth, they may not be able to provide a solid foundation for long-term work gratification and stability and for overall economic prosperity.

Therefore, although it is crucial to provide immediate support through civil service and internships, it is equally important to take a long-term perspective and invest resources and efforts in creating quality job opportunities, innovation, and professional training. Only through a balanced approach that addresses both immediate and long-term needs will it be possible to promote sustainable and inclusive economic growth.

References

Agenzia per la Coesione Territoriale (2018). Temi Cpt; Rapporto sui tempi di attuazione delle Opere Pubbliche. Dipartimento per lo Sviluppo e la Coesione Economica (2014). I tempi di attuazione e di spesa delle Opere Pubbliche.

Agrello, Pietro (2019). La politica di coesione: l’esperienza italiana. Rivista italiana di public management; Vol 2, n.1 (147-166). Marshall, A. (1890). “Principles of economics”. Macmillan, London.

Arbolino, Roberta & Boffardi, Raffaele (2017). The Impact of Institutional Quality and Efficient Cohesion Investments on Economic Growth Evidence from Italian Regions. Sustainability 9, no. 8: 1432. https://doi.org/10.3390/su9081432.

Choudhary, A. K., Oluikpe, P. I., Harding, J. A., & Carrillo, P. M. (2009). The needs and benefits of Text Mining applications on Post-Project Reviews. Computers in Industry, 60(9), 728-740.

European Commission (2022). The 8th Cohesion Report. Molle, Willem (2007). "European Cohesion Policy".

Feinerer, I., & Hornik, K. (2018). tm: Text Mining Package. R package version 0.7-6. URL: https://CRAN. R-project. org/package= tm. Feinerer, I., Hornik, K., & Meyer, D. (2008). Infrastruttura di text mining in R. Journal of statistical software, 25, 1-54.

Fratesi, Ugo & Wishlade, Fiona G. (2017). The impact of European Cohesion Policy in different contexts. Regional Studies, 51:6, 817-821, DOI: 10.1080/00343404.2017.1326673.

Gagliardi, Luisa & Percoco, Marco (2016). The impact of European Cohesion Policy in urban and rural regions. Regional Studies; DOI: 10.1080/00343404.2016.1179384.

Garbero, A., Carneiro, B., & Resce, G. (2021). Harnessing the power of machine learning analytics to understand food systems dynamics across development projects. Technological Forecasting and Social Change, 172, 121012.

Iammarino, S., Rodrıguez-Pose, A., and Storper, M. (2019). Regional inequality in Europe: evidence, theory and policy implications. Journal of economic geography, 19(2):273–298.

Jesús Crespo Cuaresma, Gernot Doppelhofer & Martin Feldkircher (2014). The Determinants of Economic Growth in European Regions, Regional Studies, 48:1, 44-67, DOI: 10.1080/00343404.2012.678824.

Liberati, P., Resce, G., & Tosi, F. (2022). The probability of multidimensional poverty: A new approach and an empirical application to EU-SILC data. Review of Income and Wealth.

OpenCoesione. (2024). Projects with Extended Path. Retrieved February 9, 2023 from https://opencoesione.gov.it/it/opendata/#!progetti_section

Petrakos, G., Kallioras, D., & Anagnostou, A. (2011). Regional convergence and growth in Europe: understanding patterns and determinants. European Urban and Regional Studies, 18(4), 375-391. DOI: 10.1177/0969776411407809.

Resce, G., & Maynard, D. (2018). What matters most to people around the world? Retrieving Better Life Index priorities on Twitter. Technological Forecasting and Social Change, 137, 61-75.

Silge, J., & Robinson, D. (2017). Text mining with R: A tidy approach. " O'Reilly Media, Inc.".