Measuring the Impact of Urban Innovation Districts

Hegyi, Fatime Barbara
Zhu, Manran
Janosov, Milan

2021
# Table of contents

Executive summary................................................................................................................................. 3
Economical and societal impacts of innovation districts................................................................. 6
The policy context of innovation districts......................................................................................... 7
Identification of innovation districts............................................................................................... 9
Data exploration...................................................................................................................................... 14
Financial data cleansing.................................................................................................................... 14
Spatial clustering................................................................................................................................. 15
Overlaying category-level clusters.................................................................................................... 19
Cluster success ....................................................................................................................................... 20
Characteristics of innovation districts............................................................................................... 22
Cluster features and cluster success.................................................................................................. 31
Machine learning classifications........................................................................................................ 34
Summary............................................................................................................................................... 34
List of abbreviations............................................................................................................................ 40
List of figures......................................................................................................................................... 41
List of tables.......................................................................................................................................... 42
Abstract

Despite their significant impact on social and economic development, innovation districts are facing challenges due to inadequacy of policies in terms of horizontal and vertical coordination or due to the lack of integrative policy approach. Strategic and targeted policy support leads to the acceleration of the growth of innovation districts, impacting the development of cities in general. To reach the potential of innovation districts in benefiting their local communities and in enabling greater collaboration, in creating jobs, and in promoting regional competitiveness, it is important to facilitate the positive externalities created by innovation districts through targeted policies.

Hence the publication proposes a generic and algorithmic methodology to identify and measure the success of innovation districts. To achieve this, different sets of large-scale geospatial data have been combined with well-established machine learning methods and in-depth statistical analysis. As a result, a quantitative methodology is presented that can support the policy-making process in the identification of urban areas with a high concentration of innovation activities and with high potential for growth. First, this methodology allows the identification of such areas. Second, an evaluation framework is proposed that captures the success of these areas based on their economic performance. Third, these results are combined with descriptive statistical features to understand the main differentiators between successful and unsuccessful areas.

This exploratory research aims at providing a set of methods and findings that heavily build on recent advances on using large-scale datasets and data science to understand social problems, and in particular, the key driving indicators of deprivation and success of various entities, such as urban areas with high concentration of innovation activities.
Authors
Fatime Barbara Hegyi, Joint Research Centre, Seville
Manran Zhu, Central European University, Budapest
Milan Janosov, Datapolis, Budapest
Executive summary

Recognising the social and economic impact of innovation districts on urban development and on cities in general, this publication proposes a generic and algorithmic methodology to identify innovation districts within seven target cities (Amsterdam, Barcelona, Budapest, Copenhagen, Eindhoven, Paris, and Stockholm). Given the inadequacy of specific policies to support innovation districts, along with insufficient horizontal and vertical coordination of policies and lack of integrative policy approach (Galan-Muros and Hegyi, Forthcoming), the identification and measurement of the impact of innovation districts are crucial to be able to define policy instruments that can maximise their social and economic potential.

This exploratory research proposes a methodology for the identification of innovation districts, thus areas with high concentration of innovation activities and then proposes success metrics to describe the economic performance of innovation districts. Machine learning clustering and data exploration-based process is proposed to identify the minimum requirements of innovation districts. To quantify the success level and to characterize innovation districts, financial information of urban areas and a set of geospatial indicators are curated (e.g. demographics, received EU-funding, network-embeddedness). Finally, the previously identified innovation districts’ characteristics are contrasted to their computed success level.

Policy context. Given the fragmented and heterogenous policy approach towards innovation districts in Europe, such areas of dense innovation face difficulties or even fail to develop. There is a heterogeneous European policy environment as regards to supporting innovation districts. While not all European countries have policies specifically targeting the growth of such innovation dense areas, some European countries have committed to design and implement a set of targeted policies. Innovation districts lay in the intersection of a wide variety of policies, including urban development, regional development, economic development, or other sectoral policies. This requires the definition of the most appropriate governmental body that can efficiently implement targeted policies (Galan-Muros and Hegyi, forthcoming).

At European level, cluster policies have promoted a wide range of programmes, platforms and initiatives over the last decade (DG Grow, 2012). Current cluster policies contain elements that can be beneficial for innovation districts, while these policies are too focused on a specific industry or a set of related industries and rarely consider the role of knowledge providers or societal actors. Specific policies of innovation districts may be part of science-technology and innovation policies, which sometimes are not aligned with national or regional development strategies (Padilla-Pérez and Gaudin, 2014; Padilla-Pérez, 2013).

Again at European level, the development and implementation of research and innovation strategies for smart specialisation have been introduced as a legal precondition for using research and innovation funding (thematic objective 1) of the European Regional Development Fund. To minimise duplication and fragmentation of publicly funded activities, a methodology for developing these research and innovation strategies have been proposed that result in the identification of a limited set of clearly defined research and innovation priorities of regions and member states. Priority setting in this context requires the analysis of regional (local) potential for growth. As such potential, innovation districts in the design and implementation of such strategies should play a role. Recognising the position in contributing to long-term sustainable growth and to reducing disparities, the European Commission’s Cohesion policy has been supporting innovation districts, mostly targeted towards science and technology parks.

Highlighting the importance of adapting policies to the local context (Mukand and Rodrik, 2005), the methodology offered in this publication for identifying and measuring geographical concentration of innovation activities belonging to specific markets can allow policy experimentation and testing of new policy tools (Hegyi and Rakhmatullin, 2020). At the same time, efficient pilot projects can be combined with effective monitoring and evaluation mechanisms leading to adequate assessment of
pilot projects (EC2012, Hegyi and Prota, 2021). Hence, this publication offers indicators for the
collection of relevant data to measure the success of innovation districts, contributing to the
assessment of the impact of policy actions linked to geographies of innovation.

**Main motivation of the research.** Providing a methodology for the identification of innovation
districts and for assessing the growth and impact of these areas can provide input for governments
in supporting the emergence and development of urban areas with high potential for growth. The
support of such areas is important given their recognised contribution to social and economic
development, while specific policies to support them and synergies across policies lag behind. A lack
of understanding of the importance of innovation districts in creating employment, especially creating
high value-added jobs, limits the commitment of governments to direct public resources towards their
development. Therefore, this exploratory research proposes a quantitative way to identify
geographical concentration of innovative companies that are or that can potentially bring benefits to
a city or region to provide information for policy makers for adequate policy support. Furthermore,
this research analyses the main features that describe innovation districts based on company
(financial) data and European Union research and innovation funding data, while proposing a method
to differentiate potentially successful and unsuccessful innovation districts and identifying the main
differentiator factors.

**Results.** In this work, a spatial clustering-based methodology is proposed that can capture small
localized urban areas with a high concentration of innovation based on large-scale global company
data. This clustering methodology is combined with an evaluation framework to assess the level of
success of such areas based on economic indicators. Finally, a set of quantitative features describing
each innovation district is curated to feed a predictive model highlighting the main differentiating
factors between successful and unsuccessful innovation districts.

**Related and future JRC work.** The Joint Research Centre – together with other European
institutions – has been researching place-based innovation eco-systems to contribute to evidence-
based policy development and to local economic development overall. Several case studies have been
produced to better understand the key success factors and the nature of interconnectedness of actors
in areas with high concentration of innovation, including science and technology parks and innovation
districts. Rissola et al (2019) have focused on the Boston area and argue that innovation districts
may act as enablers for place-based innovation through the implementation of a variety of place-
grounded initiatives. In another publication, Rissola and Haberleithner (2020) provide a comparative
analysis aiming to contribute to policy development linked to urban innovation eco-systems through
five European case studies. Lund et al (2020) examine how to apply public-private partnerships, as
credible models for development and operation of public infrastructures to the development of –
among others – innovation districts and their future work aims to support the creation and operation
of innovation districts as engines of urban transformation and sustainable and inclusive growth.
Galan-Muros and Hegyi (Forthcoming) provide a review of the fragmented and heterogeneous policy
approach towards geographies of innovation, while point out the holistic and integrative approach
needed to successfully support the emergence and development of geographies of innovation as well
as the consideration of a wide range of policies in other areas that also affect these geographies.
Economical and societal impacts of innovation districts

The Urban and Territorial Development unit of the Joint Research Centre (European Commission) supports the urban and territorial articulation of European Union policy agenda for delivering science for policy support. The unit’s urban data platform\(^1\) stands as a key component of the Knowledge Centre for Territorial Policies\(^2\) by providing information on status and trends of cities and regions of the European Union. The platform aims to provide a vision for future dynamics and opportunities for cities and recommendations for the development of policies in specific areas linked to urban development and challenges.

Along the mission of the urban and territorial unit of the Joint Research Centre, this science for policy report aims at providing a set of methods and findings that heavily build on recent advances on using large-scale datasets and data science to understand social problems, and in particular, the key driving indicators of deprivation and success of various entities, such as urban entities and areas (Quercia and Saez, 2014, Fortunato, 2018, Barabasi, 2018, Hristova et al, 2018 and Janosov, 2020).

Literature suggests diverse definition for innovation districts. Katz and Wagner (2014) focus more on the physical attributes: “innovation districts are geographic areas where leading-edge anchor institutions and companies cluster and connect with start-ups, business incubators, and accelerators. They are also physically compact, transit-accessible, and technologically wired and offer mixed-use housing, office, and retail”. Burke and Grass highlight the potential of innovation districts in activating the dormant capabilities of a community (Burke and Grass, 2019). The European Institute of Innovation and Technology emphasises that innovation hubs "constitute the backbone of an innovation community and should have a strong management, enabling collaboration within the hub itself and with partners from other hubs.” Lastly, Mulas et al (2017, pp.35) define innovation districts as “A space that has community managers that integrate many of the other functions of creative community spaces. Innovation hubs’ main function is to coordinate all actors of the ecosystem and help manage the community of tech-innovators and entrepreneurs to grow sustainably”.

Galan-Muros and Hegyi (forthcoming) have explored the concept of innovation districts (in their publication referred to as geographies of innovation) through combining theoretical and practical approaches, resulting in a definition and classification of urban areas with high concentration of innovation. Their research proposes specific policy recommendations on how governments can better support the emergence and development of these urban areas “to enhance their performance and their contributions to greener, cleaner, more social and more developed cities and regions in Europe and beyond”.

The raison d’être behind researching the impact of innovation districts is the social and economic impact of innovation districts on urban futures. Compared to non-innovation neighbourhoods, geographies of innovation produce on average four times higher intensity of tangible innovations per employee as well as nine times higher density of job opportunities. Furthermore, for every innovation-focused job the geography creates, it generates four or five additional support jobs. There exists an inverse correlation between the concentration of innovation activities within a territory and a community’s unemployment level (Burke and Grass, 2019). As Burke and Grass argue:

“Innovation Districts generate a unique form of economic growth and vitality that delivers positive economic effects to the broader population. They activate the dormant capabilities of a community and generate exponential benefits for surrounding neighbourhoods and regions. They are drivers of economic growth and how to harness them for the sustainable benefit of their communities. Innovation Districts enable greater collaboration, create job

---

\(^1\) Please see the web page of the urban data platform: https://urban.jrc.ec.europa.eu/#/en

\(^2\) Please see the web page of the knowledge centre for territorial policies: https://urban.jrc.ec.europa.eu/#/en
opportunities, and promote regional competitiveness, complex dynamics of urban spaces, highlighting the positive externalities and benefits to broader society generated by Innovation Districts, enabling greater upward mobility than non-innovation districts” (Burke and Grass, 2019, pp. 13.).

Despite their tremendous impact on social and economic development, there is an inadequacy of specific policies supporting the development of innovation districts. To facilitate the policy development process, this publication proposes a generic and algorithmic methodology to identify innovation districts and then presents success metrics along which innovation districts can be measured. As Galan-Muros and Hegyi state, there is a general lack of understanding of the long-term benefits that innovation districts can provide, resulting in the lack of an integrative policy approach (Galan-Muros and Hegyi, forthcoming).

The next section provides an overview of the fragmented and heterogeneous policies approach to innovation districts in Europe, with the aim to highlight the importance of addressing specific challenges these urban areas are facing.

The policy context of innovation districts

Europe presents a rather heterogeneous policy approach for innovation districts. In some countries there are targeted policies aiming to support the emergence and development of innovation districts, indicating strong political commitment that focuses on strengthening innovation districts. These policies are implemented by diverse levels of governments within the framework of different policies. While there is a strong urban aspect of innovation districts, still there is a strong regional development component of policies directed towards specific organisations or groups of organisations that can be found in innovation districts, such as universities, enterprises, research institutes or clusters. Hence, the support of innovation districts lies in the intersection of various policies from educational policy, through science policy, urban development, innovation policy and economic development. Therefore, there exists a policy challenge in deciding on the most appropriate government body responsible for the design and implementation of an adequate policy (Galan-Muros and Hegyi, forthcoming).

A horizontal misalignment between policies within a country may result in overlapping programs and / or contradictory objectives that may result in the reduction of efficiency of public measures and thus in the reduction of potential impact of policies, such as cluster policies, STI policies, Smart Specialisation strategies or the European Commission Cohesion policy’s structural funds. At the same time, the vertical misalignment of policies (international, national, regional, and local policies) poses the challenge of uncoordinated aims and objectives.

Evidence of innovation districts show that the success of such high potential areas largely depends on the existence and the types of policies as well as the policy approach applied in a specific geography that can have a key influence on its emergence and development (Galan-Muros and Hegyi, forthcoming).

As Galan-Muros and Hegyi (forthcoming) point out, there is a need to design a set of evidence-based policies targeted to address the specific problems, needs and bottlenecks of the geographies of innovation in the area of implementation. Moreover, other research argues that innovation districts act as enablers of place-based innovation (Rissola et al, 2017).

The scope of the present publication serves to address these challenges, thus proposing a generic methodology that builds on large-scale global data and well-established data science techniques and allows to capture innovation districts, defined as small urban areas with a high concentration of innovative entities. Within the scope of the current explorative pilot project, the methodology has been tested on seven cities with different socio-economic characteristics, namely Amsterdam, Barcelona,
Budapest, Copenhagen, Eindhoven, Paris, and Stockholm. Thus, this exploratory project identifies and measures the progress and success of innovation districts within the seven target cities.

In the following sections, the basic statistical properties of the identified innovation districts are presented, followed by the presentation of the success metrics developed by the methodology to describe the economic performance of innovation districts. A machine learning clustering and data exploration-based process is proposed to identify the minimum requirements of innovation districts.
Identification of innovation districts

In this section, a generic and algorithmic methodology to identify innovation clusters is proposed. Then the methodology to identify innovation districts is presented across the target cities and their basic statistical properties are discussed. Following that, these findings are contrasted to information on well-known innovation hubs and universities. Next, success metrics have been developed to describe the economic performance of innovation districts based on financial data of companies within that area. Finally, a feature set is curated based on five types of measures:

a) quantities related to the company-profile of each cluster (based on Orbis database),
b) descriptors capturing the relation to EU funding (based on H2020 and ERDF),
c) variables capturing the local creative output as filed patents (based on Patstat),
d) parameters encoding the proximity of innovation-heavy entities, and
e) population levels as a basic demographic indicator. Finally, these features are related to the clusters’ success.

The next sections introduce the European wide database that have been included in the analysis and the data exploration process. The section will also provide an explanation for selection of specific financial indicators.

Data

In this section the Europe-wide datasets are described that have been used in the analysis, covering seven selected European cities: Amsterdam, Barcelona, Budapest, Copenhagen, Eindhoven, Paris, and Stockholm, spanning the period 2014-2020. For each city, in the spatial analysis, their local coordinate reference systems have been used, as follows: EPSG:28992 for Amsterdam, EPSG:2062 for Barcelona, EPSG:23700 for Budapest, EPSG:23032 for Copenhagen, EPSG:27572 for Paris, and EPSG:3006 for Stockholm.

The first two data sources describe two major sources of European Union funding: Horizon 2020 (H2020) and the European Regional Development Fund’s (ERDF) thematic objective 1, which both are financing research and innovation (Cordis dataset, 2020, Bachtroegler et al, 2020). Horizon 2020 is a major financial instrument of the European Union aimed at securing Europe’s global competitiveness by funding excellent science, industrial leadership, and projects tackling societal challenges. The program accounted for EUR 80 billion during the programming period of 2014-2020. While the European Regional Development Fund supports the development and structural adjustment of regional economies through strengthening economic and social cohesion in the European Union. During the programming period 2014-2020, it amounted to more than EUR 250 billion. This research focuses on the fund’s thematic objective 1 that supports research and innovation and accounted for over EUR 65 billion (including national co-financing).

The third database, PATSTAT (OECD, 2020) collects Europe-wide information about bibliographical and legal event patent data, providing patent intelligence and statistics. PATSTAT accounts for over 100 million patent documents.

Finally, the fourth database, Orbis (Bureau van Dijk, 2020) is a global company database comprising information about various aspects of companies and private entities. This database provides – among others - legal, ownership, financial, industry, and trade data of all companies within the selected cities. The Orbis database contains information on more than 400 million companies and entities across the globe, among which 40 million have detailed financial information. Due to its large amount of reliable data, it was used as the data source for characterizing firm-level behaviour in economic studies (Ribeiro et al, 2010 and Bloom at al, 2010). In this research, the focus was on companies in eleven
NACE industry sectors that can be linked to research and innovation activities in the seven target cities, which amounts to around 20 thousand companies and EUR 2.7 trillion in total assets.

**European Union research and innovation funding data**

Horizon 2020 is a financial instrument of the European Union aimed at securing Europe’s global competitiveness by funding excellent science, industrial leadership, and projects tackling societal challenges. Its objective is to facilitate the production of world-class science and the removal of barriers to innovation. The program accounted for nearly EUR 80 billion during the programming period of 2014-2020 (European Council, 2013).

The raw H2020 dataset corresponding to the target cities, spanning the seven years of the programming period of 2014-2020. The dataset contains 18,144 funding events distributed across 9,432 projects and 2,872 legal entities, with total funding of about EUR 7.8 billion. The majority of the EU funding rounds are also attached with corresponding NACE codes, which is the industry-standard classification system applied in the European Union, describing the thematic categories in a standardized terminology. Furthermore, 99.7% of the funded entities (either the entire entity or its particular funded department) are annotated by geographical locations. The dataset has been cleaned according to the followings:

First, all entities have been geocoded with their exact geographical coordinates based on their entity names by using Google Maps Geocoding API and then compared the geocoded locations of the funded entities to the administrative boundaries of their host cities to filter out the ones that fell outside of the city boundaries. This left the research with 12,569 entities, roughly about 70% of the total volume. The main reason behind this was to remove the ambiguity in city-categorization and NUTS areas within the H2020 dataset.

Those categories have been dropped that are unspecific and / or correspond to individual-dependent grants, in particular, that of the ‘Marie-Sklodowska-Curie Actions’ and ‘European Research Council (ERC)’ (e.g. postdoctoral research fellowships, research group starting grants), which affected another 30% of the remaining records.

Then those projects have been filtered out that have received phased supports, that did not include research or innovation activities. These projects stood up for 15% of the already cleaned data records.

Finally, all institutions have been dropped that were only participating as partner organizations, meaning that they have not benefited financial support, arriving at a cleaned set of 7,449 records, 3,993 projects, 1,435 entities, and EUR 3.2 billion of funds allocated, with 70% of the projects annotated by NACE topic categorization.

**European Regional Development Fund**

The European Regional Development Fund (ERDF) dataset, has altogether 140,884 records, containing 2,855 funding records located in NUTS3 code areas corresponding to the seven target cities. These funding rounds are split across beneficiary entities and cover funds of EUR 103.5 million in total. The funded projects’ topics are described by NACE categorization, however, unlike the H2020 dataset, the entities are not geolocated. Therefore, Google Maps Geocoding API14 has been used for geocoding the entities based on their names. After filtering out missing values, arriving at a set of 1,378 records with 545 entities and EUR 97.4 million distributed.

The Statistical Classification of Economic Activities in the European Community, commonly referred to as NACE15 (for the French term "nomenclature statistique des activités économiques dans la
Commu

nauté européenne”) is the industry-standard classification system applied in the European Union. In function NACE is similar to the SiC16 and NAICS17 system used in North America. NACE uses four hierarchical levels, with the first level identified by alphabet letters A to U (See Table 1 for the first level categories). In the research, industry sectors have been targeted that are related to EU funding and innovation. These sectors are A, F, P, K, Q, J, C, B, M, H, and E shown in Table 1.

Table 1 Number of companies involved in the analysis from Orbis per NACE code

<table>
<thead>
<tr>
<th>NACE</th>
<th>ECONOMIC AREA</th>
<th>AMSTERDAM</th>
<th>BUDAPEST</th>
<th>BARCELONA</th>
<th>EINDHOVEN</th>
<th>PARIS</th>
<th>STOCKHOLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Agriculture, forestry, fishing</td>
<td>0</td>
<td>17</td>
<td>66</td>
<td>6</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>B</td>
<td>Mining and Quarrying</td>
<td>22</td>
<td>2</td>
<td>11</td>
<td>5</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>C</td>
<td>Manufacturing</td>
<td>133</td>
<td>490</td>
<td>1494</td>
<td>65</td>
<td>81</td>
<td>560</td>
</tr>
<tr>
<td>E</td>
<td>Water Supply; Sewerage, Waste Management and Remediation Activities</td>
<td>7</td>
<td>25</td>
<td>49</td>
<td>3</td>
<td>6</td>
<td>21</td>
</tr>
<tr>
<td>F</td>
<td>Construction</td>
<td>86</td>
<td>303</td>
<td>1043</td>
<td>81</td>
<td>29</td>
<td>314</td>
</tr>
<tr>
<td>H</td>
<td>Transportation and Storage</td>
<td>71</td>
<td>314</td>
<td>499</td>
<td>59</td>
<td>22</td>
<td>156</td>
</tr>
<tr>
<td>J</td>
<td>Information and Communication</td>
<td>342</td>
<td>534</td>
<td>916</td>
<td>408</td>
<td>84</td>
<td>1169</td>
</tr>
<tr>
<td>K</td>
<td>Financial and Insurance Activities</td>
<td>424</td>
<td>99</td>
<td>183</td>
<td>113</td>
<td>42</td>
<td>578</td>
</tr>
<tr>
<td>M</td>
<td>Professional, Scientific and Technical Activities</td>
<td>681</td>
<td>740</td>
<td>1319</td>
<td>412</td>
<td>168</td>
<td>1576</td>
</tr>
<tr>
<td>P</td>
<td>Education</td>
<td>28</td>
<td>190</td>
<td>79</td>
<td>32</td>
<td>11</td>
<td>156</td>
</tr>
<tr>
<td>Q</td>
<td>Human Health and Social Work Activities</td>
<td>141</td>
<td>331</td>
<td>270</td>
<td>62</td>
<td>46</td>
<td>162</td>
</tr>
</tbody>
</table>

**PATSTAT database**

To measure innovation, PATSTAT database has been incorporated, which represents a large-scale database on published patents. The raw database, curated based on NUTS3 area codes, contains 79,238 records that describe 11,845 patents filed by 14,697 innovators and 26,682 applicants. While the database is not directly geolocated, it contains the patenting individuals’ and entities’ addresses, which has been geocoded and applied geographical filters according to the administrative boundaries of the target cities. These filtering steps resulted in a subset of 43,728 records on patents filed by 12,946 innovators and 16,976 applicants.

**Orbis database**

The Orbis database from Bureau van Dijk contains firm-level information of more than 400 million companies globally, among which 40 million have detailed financial information. The data used in this work is the 2019 snapshot of the database under the license of the Joint Research Centre (JRC)
at the European Commission. The Orbis database provides information on various aspects of companies, among which the data tables explained in the next paragraphs are particularly relevant to this research. Using these tables, 20,365 companies have been selected across the seven target cities (see Table 1), among which 11,321 have latitude and longitude information, while the rest have been geocoded.

The contact info table contains the companies’ name, address, latitude, longitude, city, and NUTS3 region. Companies in the seven target cities have been selected and the greater areas (NUTS3 region) of the cities to capture all relevant companies. For companies that don’t have latitude and longitude information, Google geocoding API has been used to find the company’s latitude and longitude from its address. As the last step, companies have been filtered out whose latitude and longitude fall outside the target cities’ administrative boundaries.

The legal info table contains the incorporation date, size, status, status date of the companies. The status of the company can be active, inactive, or unknown, and the status date is the date when the status value was recorded. To filter out companies that is not active during 2014-2019, those companies have been kept whose status is active, or whose status is inactive, but the status date is after 2014, meaning the company’s status changed to inactive after 2014. The distribution of the companies’ incorporation year is shown in Figure 1, with the probability defined as the fraction of the number of companies incorporated at a certain year over the total number of companies. As shown, the Orbis database’s records dated back to as early as 1950 and became more concentrated since 1980.

Figure 1 Company features overview – incorporation year distribution

The industry financial table contains data on the companies’ financial statements, such as balance sheet, income statement and cash flow statement. Since a detailed financial analysis of the companies’ health status is beyond the scope of this work, based on previous research (Kalemli-Ozcan, 2015 and CPB 2017), several financial indicators have been selected to represent the financial status of the companies based on the following principles:

1) The selected indicators should have low percentage of missing values in the ORBIS database (<5%),

2) The selected indicators should represent the most prominent aspects of the companies’ financial performance, such as assets and debts information on the balance sheet, profit and operation cost information on the income statement and cash flow statement,

3) The ORBIS database contains a summary of the companies’ financial data called the key financial table, which is less detailed than the industry financial table, but could serve as a reference for the selection financial indicators,
4) The selected indicators should be financial ratios in order to avoid the effect of the size of the companies.

Figure 2 shows all the financial features in the key financial table and their percentage of missing values for companies in Paris.

The financial ratio features in the key financial table can be grouped to the following categories:

- Current ratio (CURRENT_RATIO_X), capturing a company's ability to pay short-term debts,
- Solvency ratio (SOLVENCY_RAT.Asset_based_pct), capturing a company's ability to pay long-term debts,
- Return on assets (ROE, ROCE), indicating how profitable a company is relative to its total assets and
- Profit margin (PROFIT_MARGIN_PCT), measuring what percentage of sales has turned into profits.

In order to avoid the problem of high percentage of missing values for the features return on assets and profit margin, we replace them with ROA_USING_NET_INCOME_PCT and OPERATING.PL_EBIT in the industry financial table, which are alternative measures of the two indicators and have less missing values. As this exploratory research focuses on the possibility of identifying innovation districts based on financial data - while further possible research could include sustainability and circularity as indicators for measuring innovation-, the four financial ratios current_ratio_x, solvency_rat_asset_based_pct, roa_using_net_income_pct and operating_pl_ebit have been finally selected as the financial KPI of the companies.

Figure 2 Features in the Key Financial Table and their percentage of missing values

The last table used for the analysis is the so-called DMC table that contains information on the companies' current and previous directors. When directors move from one company to another, knowledge is exchanged, and bonds are created. A company that exchanges human resources with other companies is better positioned to learn from other companies' experience and benefit from the directors' personal connections and, therefore, more likely to succeed. To capture how successful a company is at this, a network of companies has been constructed where each node is a company, and a link exists between two companies if a person has worked as a director in both companies. The constructed company network contains 13,400 nodes and 23,322 edges. A company’s degree
(number of links a company has) in the network has been used as an indicator among company characteristics.

Data exploration

Figure 3 shows the basic comparative statistics of the main features from the EU funding datasets, which relate to the cardinality of funded entities, the diversity of funding categories, and the amounts of funding. Figure 3a highlights that Paris is dominated by the H2020 funding, while Paris and Copenhagen have received the lowest volume of ERDF funding, while in Stockholm and in Budapest the number of ERDF projects overcomes the number of H2020-funded projects. Figure 3b supports the earlier remarks on H2020 accounting for a larger volume of funds, which is also reflected in the number of participating entities (Figure 3c). Figure 3d highlights that the two most populated NACE categories, both for H2020 and ERDF projects, are category M (Professional, Scientific and Technical Activities), and P (Education).

Figure 3 Horizon 2020 and ERDF research and innovation funding statistics

Financial data cleansing

To measure the selected companies’ economic performance, their financial data has been used from the industry financials table. The main steps of the cleaning procedure for the financial data have been the following:
1. Drop financial records that have the consolidation date not equal U1 or U2³.

2. Drop financial records that have the feature: number of months not equal to 12. Some of the financial features, e.g., profit, are accumulated over time and therefore affected by the length of time on which it was calculated.

3. Drop financial records with missing values or negative values in the features total assets or shareholders find⁴. Negative total assets or shareholder’s equity are generally red flags for financial performance and shouldn’t appear in the records of active companies. They are likely to be misinformation and therefore are dropped.

4. Drop financial records that have number of employees smaller than 10, having chosen companies that are above a certain scale by having at least 10 employees, corresponding to significant innovation potential.

5. Drop financial records that have missing values in the feature closing date or have closing date before the year 2014⁵.

6. Keep the columns corresponding to the seven key financial indicators and drop the rest. The kept features are: total assets, shareholders fund, number of employees, current ratio x, solvency ratio (asset based), ROA (using net income), and operating profit (divided by total assets to get the operating margin). Values of the above ratios are aggregated (median) during the period 2014-2019 and then used to capture the company’s overall recent performance.

Spatial clustering

To capture innovation clusters, small and compact geographical areas with a significant company density, the methodology is built on the well-established spatial clustering technique called DBSCAN (Density-based spatial clustering of applications with noise) and its Python implementation (Ester et al, 1996 and Sklearn.cluster.dbscan, 2021). Clustering is conducted in each city for each NACE code separately and then a methodology has been introduced to overlay these category-specific clusters. The DBSCAN algorithm has been used providing two standard parameters given a distance metric, which for the sake of simplicity Euclidean was chosen. Considering the curvature of the Earth’s surface, while Haversine distance may be a more accurate metric than the Euclidean distance, at the characteristics scale of the chosen cities, this curvature can be considered negligible compared to the local geography’s inhomogeneities. These are usually referred to as ‘eps’ describing the maximum distance between two neighbouring points so that they can be considered to be in the same spatial cluster, and ‘min’ representing the number of points in an area for a point to be considered as a core point (a member of a specific cluster), including the point itself. While for the parameter ‘min’ has been introduced a limit of 5 (encoding a lower cut-off for a cluster size as 5 entities). Furthermore, the optimization of ‘eps’, the characteristic size of clusters, was considered as a hyperparameter. As illustrated by a schematic view on the city of Paris on Figure 4, clustering with low ‘eps’ results (small typical cluster size) in multiple irrelevantly small clusters (left), while large eps clustering only captures giant clusters spanning across administrative districts (right), while a carefully picked medium ‘eps’ can provide the right indications for the existence of relevant spatial clusters.

³ Consolidated statements (i.e., C1 and C2) reports the financial data of the headquarter together with all separate legal entities belonging to the group (Kalemli-Ozcan, 2015), and therefore are not comparable to unconsolidated statements. Statements with limited financial information (i.e., LF) are also dropped since they don’t provide enough information about the company.

⁴ These are the financial features that 1) appear both in the industry financial statement and the key financial statement 2) has the least amount of missing values.

⁵ Focus on the financial performance of the companies in the recent years, not the entire historical period.
This flexibility in choosing ‘eps’ allows the definition of a cost function ‘C’ that accounts for the company-density and prevents large inhomogeneities to occur in the cluster size:

\[ C = \text{number of clusters} \times \frac{\text{Total area of clusters}}{\text{Standard deviation of clusters' areas}} \]  

(1)

In addition to this cost function, the following selection criteria has been used for company clusters:

- Exclude giant clusters with more than 100 points of interests (POI),
- Exclude tiny clusters with less than 5 POIs (via DBSCAN parameter ‘min’),
- Use the concave hull of a cluster to measure its area,
- Both active and inactive companies were used to mark the boundaries of clusters.

Under these conditions, ‘C’ has been optimized by testing various values of ‘eps’ in the range from 100 meters to 1000 meters with a step size of 10 meters to find the best cluster ensemble for a given city and a given NACE\textsuperscript{6} category. As an example, the results of such an optimization process in terms of the cost function and the clustering layout for NACE category M (Professional, Scientific and Technical Activities) and the city of Paris are shown in Figure 5.

\[ \text{Figure 5 Cluster optimisation} \]

\textsuperscript{6} NACE: nomenclature statistique des activités économiques dans la Communauté européenne, Statistical Classification of Economic Activities in the European Community
The left panel shows the city of Paris (in the EPSG:27572 - Geodetic Parameter Dataset) coordinate reference system visualized on the two axis with the corresponding Orbis POIs and the identified clusters. The right panel shows the optimization of the C parameter as a function of eps.

Thus, clustering optimization for the city of Paris and the NACE category M showing both the clustered POIs and the concave hulls of the clusters by coloured areas, and the measured values of the cost function (1) smoothed by a moving window with a size of ±20m as a function of 'eps' at a resolution (step size) of 10 meters, showing a maximum value of C at 390 meters and with 20 clusters identified.

Based on this methodology, the total number of clusters identified in each city and category are summarized in Table 2.

<table>
<thead>
<tr>
<th>NACE</th>
<th>Amsterdam</th>
<th>Barcelona</th>
<th>Budapest</th>
<th>Copenhagen</th>
<th>Eindhoven</th>
<th>Paris</th>
<th>Stockholm</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c</td>
<td>11</td>
<td>18</td>
<td>53</td>
<td>1</td>
<td>6</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>e</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>f</td>
<td>5</td>
<td>12</td>
<td>29</td>
<td>6</td>
<td>1</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>h</td>
<td>5</td>
<td>10</td>
<td>18</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>j</td>
<td>13</td>
<td>25</td>
<td>10</td>
<td>15</td>
<td>4</td>
<td>23</td>
<td>19</td>
</tr>
<tr>
<td>k</td>
<td>9</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>m</td>
<td>31</td>
<td>3</td>
<td>12</td>
<td>19</td>
<td>8</td>
<td>20</td>
<td>29</td>
</tr>
<tr>
<td>p</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>q</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>11</td>
<td>7</td>
</tr>
</tbody>
</table>

Figure 6 presents the category-level clustering results for each NACE code in the case of Paris. Each map shows companies from the Orbis database and the identified clusters corresponding to that given NACE category with the same colour code. Each figure shows the number of clusters (#clusters) and the number of companies (#companies) within the NACE category after overlaying the category specific clusters.
Figure 6 Category-level clusters for Paris

Category indicates the NACE category
#clusters indicate number of clusters identified within the NACE category
#companies indicate number of companies within that NACE category after overlaying the category-specific clusters.
Overlaying category-level clusters

While the formerly introduced methodology identifies category-specific clusters, the main goal is to identify innovation hot-spots that show sparking innovation, and therefore cover multiple areas. To capture those areas, a four-step process has been applied, as shown on Figure 7a-d respectively. First, the category-level cluster polygons have been overlayed, followed by the city-polygon that has been split into a grid with a cell size of 25x25 meters.

The number of categories has been quantified via corresponding clusters. One can observe in each grid cell – which is an integer value ranging from zero up to nine, depending on the cities (for details see Table 2). Finally, a threshold of three on this grid-level value have been introduced expressing the degree of category-overlap and considering these highly overlapping areas as meta clusters of innovation. In other words, areas as meta innovation clusters have been kept that overlay with at least three different NACE-category clusters. The meta clusters found by this methodology are shown for all the seven target cities on Figure 8, while the number of meta clusters found in each city is shown in Table 3, showing that Copenhagen has the fewest – six total, while in the case of Budapest the methodology identified 21 potential innovation districts.
Figure 8 Meta-clusters

The figure shows the innovation clusters for each city, obtained by the overlaying methodology with an underlying grid having cell size of 25x25m.

Table 3 Number of meta clusters

<table>
<thead>
<tr>
<th>City</th>
<th>Amsterdam</th>
<th>Barcelona</th>
<th>Budapest</th>
<th>Copenhagen</th>
<th>Eindhoven</th>
<th>Paris</th>
<th>Stockholm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of meta clusters</td>
<td>12</td>
<td>13</td>
<td>21</td>
<td>6</td>
<td>7</td>
<td>14</td>
<td>10</td>
</tr>
</tbody>
</table>

The number of combined meta clusters found in each target city.

Cluster success

Based on the financial key performance indicators (KPIs) of individual companies (see sections on data and data exploration on the selection of KPIs), the following metric has been proposed to quantify the economic success of innovation districts in a binary fashion.
Four financial ratio indicators have been chosen from the Orbis Industry Financials data table as the KPIs to quantify the companies’ financial performance as described in the section on data and data exploration. They are: current ratio, solvency ratio, return on assets and operating margin. To measure the firm-level success, companies are compared to their own peers. If a company has a financial ratio value above the median value for all companies in the same city and sector, it is considered successful in that aspect, otherwise it is considered as unsuccessful. In this way, a binary success measure has been constructed for each of the four financial ratios, characterizing the financial performance of a company in four different aspects: liquidity, solvency, profitability, and operation efficiency. Since four binary measures are used to capture the performance of a company, the resolution problem of binary measures is to some extent avoided. An illustration of the final success measures for a company is shown in Table 4.

Table 4 Example of performance profile

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>success_liquidity</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>success_solvency</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>success_profitability</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>success_efficiency</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Example of the performance profile of company A and B that contains the four peer performance indicators.

Using the individual company’s success measures calculated above, a single binary success indicator is derived for an innovation district in the following manner. For company $a$, the values of its $i$th financial success indicators is denoted (rows 1-4 in the company profile in Table 4) by $s^a_i$, $i = 1, 2, 3, 4$. The $i$th financial success indicator $S^D_i$ for an innovation district (overlaid meta cluster) ‘D’ is then defined as the average over all the companies in the district:

$$S^D_i = \frac{1}{N} \sum_{c \in D} S^c_i$$

(2)

Where ‘N’ is the total number of companies in ‘D’. Assuming that the four financial ratios are equally important for quantifying the success of a district, we define the overall financial success indicator for ‘D’ as the average over all four indicators:

$$S^D = \frac{1}{4} \sum_{i=1}^{4} S^D_i$$

(3)

Finally, the binary financial success indicator $B^D$ for innovation district D is defined as:

$$B^D = \begin{cases} 1 & \text{if } S^D > M \\ 0 & \text{else} \end{cases}$$

(4)

Where ‘M’ is the 50 percentile of $S^D$ for all innovation districts across the respective cities. An overview of the economic success of the innovation districts measured this way is shown in Figure 9. It’s important to note that the success measure for different innovation districts is comparable only at the domestic level, since the binarization of the four financial KPIs were conducted for a certain country and a certain industry sector.
Characteristics of innovation districts

To understand what can potentially drive the success of innovation districts, several potential indicators have been identified that later have been tested against the cluster-success labels. The indicators are based on the i) aggregated descriptive statistics of the companies within a cluster, ii) the innovative output measured by patents filed from a cluster, iii) the information related to the EU funding received by companies within an area, iv) the proximity of innovation-related entities within a city, v) and an estimated local population as a simple demographic indicator. The result of these are presented with the city of Paris in the following sections.

Company-based information
As innovation clusters consist of diverse companies, their characteristics are therefore an important determinant of the overall prosperity of the clusters. Thus, in this section, different variables are overviewed, which include the volume and activity status of companies within an area, the profile diversity a particular cluster shows, and their interconnectedness to the overall Orbis company network, as described in more detail in the section on data and data exploration.

As regards to the number of companies per cluster, based on the activity status of the companies, the following three measures account for company-cardinality and activity:

- The total number of companies in a cluster that have ever been active
- The number of currently active companies within a cluster,
- The fraction of currently active companies as the ratio of the previous two measures.

The reason for differentiating between active and inactive companies is as follows. While clusters of active companies may certainly show the location of current innovation districts, it disregards any changes in temporal trends. Thus, while two innovation districts with a similar actual size could have substantially different histories. For instance, computing the fraction of active and inactive companies can assess the level of fluctuation in the number of entities, stability, and longevity.

Regarding the year of incorporation, to account for the temporal changes and characteristics, the average year of incorporation of companies within a cluster have been included. While to include NACE category diversity, each company is attached by a NACE category tag, which when aggregated to the level of clusters that will mean a set ‘N’ of NACE codes for each cluster. Based on this, NACE category diversity of cluster ‘i’ have been defined as the Shannon entropy of N, (Shannon and Weaver, 1949):

$$NACE\ diversity_i = -\sum_k N_{i,k} \cdot \log N_{i,k}$$  \hspace{1cm} (5)

Based on the company network, a binarized feature have been created that captures whether a certain cluster is connected to the network of Orbis actors or not by binarizing the average number of connections companies have within an innovation cluster. Figure 10 shows a selected set of features for the city of Paris, while table 5 shows the computed Orbis-network-feature table for Paris.
Figure 10 Orbis features for Paris

The colour-bars identify the values each feature takes in the different clusters.

Table 5 Orbis features for Paris

<table>
<thead>
<tr>
<th></th>
<th>year of incorporation</th>
<th>number of active companies</th>
<th>total number of companies</th>
<th>fraction of active companies</th>
<th>NACE diversity</th>
<th>Linked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meta-0</td>
<td>1997.8</td>
<td>43</td>
<td>47</td>
<td>0.914894</td>
<td>0.897447</td>
<td>1</td>
</tr>
<tr>
<td>Meta-1</td>
<td>1990.3</td>
<td>15</td>
<td>15</td>
<td>1.000000</td>
<td>0.811977</td>
<td>0</td>
</tr>
<tr>
<td>Meta-2</td>
<td>1997.8</td>
<td>208</td>
<td>217</td>
<td>0.958525</td>
<td>0.782908</td>
<td>1</td>
</tr>
<tr>
<td>Meta-3</td>
<td>1992.7</td>
<td>50</td>
<td>52</td>
<td>0.961538</td>
<td>0.841909</td>
<td>1</td>
</tr>
<tr>
<td>Meta-4</td>
<td>1996</td>
<td>40</td>
<td>45</td>
<td>0.888889</td>
<td>0.826570</td>
<td>1</td>
</tr>
<tr>
<td>Meta-5</td>
<td>1990</td>
<td>11</td>
<td>11</td>
<td>1.000000</td>
<td>0.546295</td>
<td>1</td>
</tr>
<tr>
<td>Meta-6</td>
<td>1995.0</td>
<td>301</td>
<td>336</td>
<td>0.895833</td>
<td>0.868072</td>
<td>1</td>
</tr>
<tr>
<td>Meta-7</td>
<td>2003.2</td>
<td>7</td>
<td>9</td>
<td>0.777778</td>
<td>0.809657</td>
<td>1</td>
</tr>
<tr>
<td>Meta-8</td>
<td>1994.7</td>
<td>32</td>
<td>32</td>
<td>1.000000</td>
<td>0.941434</td>
<td>1</td>
</tr>
<tr>
<td>Meta-9</td>
<td>1999.2</td>
<td>22</td>
<td>24</td>
<td>0.916667</td>
<td>0.871596</td>
<td>1</td>
</tr>
<tr>
<td>Meta-10</td>
<td>1992.9</td>
<td>64</td>
<td>65</td>
<td>0.984615</td>
<td>0.922704</td>
<td>0</td>
</tr>
<tr>
<td>Meta-11</td>
<td>1994.1</td>
<td>12</td>
<td>12</td>
<td>1.000000</td>
<td>0.864787</td>
<td>1</td>
</tr>
<tr>
<td>Meta-12</td>
<td>1994.7</td>
<td>33</td>
<td>33</td>
<td>1.000000</td>
<td>0.949894</td>
<td>0</td>
</tr>
<tr>
<td>Meta-13</td>
<td>1992.8</td>
<td>7</td>
<td>11</td>
<td>0.636364</td>
<td>0.878741</td>
<td>1</td>
</tr>
</tbody>
</table>
Patent information

To characterize the cluster-level properties of innovation strength, the metrics are based on the number of patents and the characteristics of the collaboration between the actors filing these patents. In relation to the number of patents, to quantify the volume of patent-level output of a cluster, the number of patents has been computed, that have been geolocated for each cluster. The co-patent-filing network of actors have been built, which resulted in a network of 1,377 nodes and 1,560 connections between them. The number of nodes that fell into the target cities’ innovation clusters was 210 for Paris, 110 for Amsterdam, only 12 for Budapest, 93 for Barcelona, 30 for Copenhagen, 558 for Eindhoven, and 389 for Stockholm. Finally, the connectedness of each cluster has been computed and attached by a binary value based on whether the entities within that cluster are connected to the PATSTAT network or not. Figure 11 shows a selected set of features for the city of Paris, while table 6 shows the computed PATSTAT-feature table for Paris.

Figure 11 PATSTAT features for Paris

![PATSTAT features for Paris](image)

The colour-bars identify the values each feature takes in the different clusters.

Table 6 PATSTAT features for Paris

<table>
<thead>
<tr>
<th>meta_id</th>
<th>PATSTAT_num_patents</th>
<th>PATSTAT_linked</th>
</tr>
</thead>
<tbody>
<tr>
<td>meta-0</td>
<td>26</td>
<td>1</td>
</tr>
<tr>
<td>meta-1</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>meta-2</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>meta-3</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>meta-4</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>meta-5</td>
<td>224</td>
<td>1</td>
</tr>
<tr>
<td>meta-6</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>meta-7</td>
<td>64</td>
<td>1</td>
</tr>
<tr>
<td>meta-8</td>
<td>73</td>
<td>1</td>
</tr>
<tr>
<td>meta-9</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>meta-10</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>meta-11</td>
<td>118</td>
<td>1</td>
</tr>
<tr>
<td>meta-12</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>meta-13</td>
<td>22</td>
<td>1</td>
</tr>
</tbody>
</table>
European Union funding

The analysis not only aimed to characterize the company-level details of the identified clusters but to relate them to the level of funding arriving at these clusters with the main goal of nurturing local innovation. This section captures this via the volume of funding being committed to the different clusters, and whether the clusters are linked to the Europe-wide H2020 collaboration network via joint projects. In addition, earlier research has shown that in various fields, from the film industry, art or science, that several network indicators, such as centrality measures and core-periphery structures can proxy the success of the network actors, e.g. by capturing entities in information-brokerage or influencer positions (Fraiberger et al, 2018, Burt, 2004, Juhasz et al, 2020, Janosov et al, 2020). Therefore, here the research attempts to replicate these results in the case of EU funding (ERDF TO1 and Horizon 2020) grants by introducing the collaboration networks of the funded entities and using them to compute simple network metrics.

The number of entities that received funding in either the H2020 or the ERDF categories have been attached to each cluster. Then for each cluster, the total amount of funding has been measured, and both the median and the average amount of funded projects (H2020 and ERDF). As Figure 3 also hints, in some cases (e.g. Paris), the amount of ERDF funding is low. While the ERDF projects did not show network characteristics and collaboration due to the nature of the funding, in the cleaned H2020 dataset a network of 1,018 nodes and 3,222 edges could have been built, presented on Figure 12. Within this network, 342 nodes can be found in innovation clusters in Paris, 122 in Amsterdam, 121 in Budapest, 174 in Barcelona, 94 in Copenhagen, 72 in Eindhoven, and 95 in Stockholm. Based on the information if funded entities within a cluster were connected to this network, each cluster is labelled as linked or not linked in H2020, presented in Table 5, where the nodes correspond to the funded entities, the size of the nodes is proportional to the total number of projects they had, the link between two nodes shows the number of shared projects between these two funded entities, and the node-colouring encodes the countries of the funded entities.

Figure 12 H2020 collaboration network across the analysed cities
Figure 13 shows a selected set of features for the city of Paris and Table 6 shows the computed EU funding feature table for Paris.

Figure 13 Funding features for Paris

<table>
<thead>
<tr>
<th>Table 7 EU funding features for Paris</th>
</tr>
</thead>
<tbody>
<tr>
<td>H2020 linked</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>Meta-0</td>
</tr>
<tr>
<td>Meta-1</td>
</tr>
<tr>
<td>Meta-2</td>
</tr>
<tr>
<td>Meta-3</td>
</tr>
<tr>
<td>Meta-4</td>
</tr>
<tr>
<td>Meta-5</td>
</tr>
<tr>
<td>Meta-6</td>
</tr>
<tr>
<td>Meta-7</td>
</tr>
<tr>
<td>Meta-8</td>
</tr>
<tr>
<td>Meta-9</td>
</tr>
<tr>
<td>Meta-10</td>
</tr>
<tr>
<td>Meta-11</td>
</tr>
<tr>
<td>Meta-12</td>
</tr>
<tr>
<td>Meta-13</td>
</tr>
</tbody>
</table>

Funding data expressed in EUR
Proximity of knowledge and innovation hubs

Besides funding, the geographical proximity of establishments linked to innovation may have a significant effect on the existence and success of a company cluster. As Galan and Hegyi argue in their endeavour of defining and classifying geographies of innovation, that most types of models of geographical concentration of innovation within cities are closely linked to universities (Galan-Muros and Hegyi, forthcoming). Hence, two types of POI-level entities have been integrated in the analysis: the location of universities and the location of official or well-known hubs of innovation in the target cities as follows.

The POI information of the target cities has been collected from OpenStreetMap using node categories amenity-university and amenity-college (Bennett, 2021). Based on the geographical location of these universities, shown on Figure 14, for each cluster, the distance between the clusters defined in this analysis and universities have been measured resulting in a proximity feature for the analysis.

To include official or well-known innovation centres in the proximity analysis, a list of entities for each city have been included in the analysis, based on experts’ evaluation:

- Amsterdam: Amsterdam Science Park, Cumulus Park, Amsterdam Zuidas Knowledge Quarter;
- Barcelona: 22@, Barcelona advance industry park (BCN ACTIVA), UPC Parc Scientific Barcelona, Parc de Recerca Biomèdica de Barcelona (PRBB);
- Budapest: Graphisoft Park, Infopark, PreziHQ;
- Copenhagen: Orestad Innovation City Copenhagen, Copenhagen Science City, Frederiksberg Knowledge / Science City, Copenhagen Health Innovation;
- Stockholm: Kista, Stockholm SciLifeLab, Karolinska Institute Innovations AB, KTHInnovation;
- Eindhoven: Brainport Eindhoven (as it is a broader area, its centroid has been added to the analysis), High Tech Campus Eindhoven.

These analyses are shown on the examples of Paris, Amsterdam and Barcelona and are visualized in Figure 14.
The figure a. shows the innovation hubs (marked by blue) and the university buildings (marked by red) across Paris, while b. measures the average distance between meta clusters and university buildings, innovation hubs, and the random baseline (grid cells).

Location of both university and innovation centres have been used to quantify their proximity to the clusters by attaching the distance to the nearest university and innovation hub to each cluster. In addition, the average shortest distance from every university and innovation centre location to the nearest company clusters have been compared and then contrasted this measure against the distance of every point within a city-level grid, which accounts for a base-line proximity value. As Figure 14b shows for the case of Paris, innovation hubs show about 45% closer proximity to the clusters than
the university buildings, which is still three times closer than, on average, any generic picked points within the city.

Table 8 Average proximity values

<table>
<thead>
<tr>
<th></th>
<th>H2020 linked</th>
<th>H2020 total funding</th>
<th>H2020 median funding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paris</td>
<td>517</td>
<td>288</td>
<td>1320</td>
</tr>
<tr>
<td>Amsterdam</td>
<td>537</td>
<td>792</td>
<td>2943</td>
</tr>
<tr>
<td>Budapest</td>
<td>1431</td>
<td>1141</td>
<td>3335</td>
</tr>
<tr>
<td>Barcelona</td>
<td>727</td>
<td>612</td>
<td>2096</td>
</tr>
<tr>
<td>Copenhagen</td>
<td>1886</td>
<td>1495</td>
<td>1974</td>
</tr>
<tr>
<td>Eindhoven</td>
<td>287</td>
<td>553</td>
<td>1438</td>
</tr>
<tr>
<td>Stockholm</td>
<td>1740</td>
<td>2469</td>
<td>5097</td>
</tr>
</tbody>
</table>

The average proximity of university buildings, innovation hubs, and generically picked grid cells uniformly scattered across the cities, to the innovation, expressed in meters.

**Basic demographics**

Finally, as a basic demographic indicator, the approximated residential population have been added within each cluster, based on the official demographics data from the cities’ OpenData portals. 481 neighbourhoods of Amsterdam, 81 admin-level-10 areas of Paris (more precisely, a 250x250m grid mapped into the admin areas), and 616 DeSO areas of Stockholm have been used (Amsterdam demographics (2021), Paris demographics (2021), Amsterdam demographics (2021). These population levels are shown on Figure 15.

Figure 15 Approximated resident population in target cities

The following section presents the derived 14 features (financial, spatial and proximity based and demographic features) for Paris, Amsterdam and Stockholm, while applying 13 types of features (missing demographics) for Barcelona, Budapest, Copenhagen and Eindhoven.
Cluster features and cluster success

Correlation analysis

First, more than a dozen of different features and financial performance-based success indicators have been derived in an explorative fashion across the studied datasets. Then to characterize the innovation clusters, these features have been related to each other by conducting a thorough correlation analysis. Success indicator introduced in the section on cluster success have been used to measure the economic success of the innovation clusters. The correlations between the cluster features and cluster success are shown in Figure 16.
To investigate which cluster features correlate most with the economic success of the clusters, the Pearson correlation has been used between the cluster features and the cluster success across all target cities, which results are shown in Table 9.
First, the relation between EU funding and the success of the innovation clusters has been assessed. For ERDF, there is an exclusively positive or zero correlation between the funding features and the clusters’ success across all seven target cities (zero correlation due to missing values). For H2020, the same trend can be observed, with positive or negligibly low correlation score in all target cities. For ERDF, the correlation is especially strong in the city of Paris, despite the comparatively fewer amount of ERDF funding Paris received (see Figure 3a). For H2020, the strongest positive effect happens in Budapest, which shows an interesting contrast to the fact that Budapest received relatively more ERDF funding than H2020.

In summary, the innovation clusters across the seven cities all correlate non-negatively with the EU funding they received, which hints that the clusters identified by the methodology do correspond to centres of innovation. In addition, it is possible to observe that the most successful innovation districts are not the ones that received the type of EU funding with the higher amount of total funding in a given city (e.g. H2020 in Paris and ERDF in Budapest), but rather the ones that received funding from the other type of fund with lower budget in the specific city.

The effect of the company features on the success of the clusters is more ambiguous. In Paris, Budapest, Barcelona, and Stockholm, the number of companies and the interconnectedness of the companies in an innovation cluster seem to play a positive role in the economic success of the cluster, but in other target cities, the effect can be zero or negative. However, across all seven target cities, the average incorporation year of the
cluster is always an indicator of success, meaning that the most successful districts tend to be the younger districts.

As for patents related features, the city of Copenhagen and Eindhoven stand out to be the ones whose innovation districts’ success correlate the most with the number of patents, which resonate with the fact that they both showcase a strong presence in technology (e.g. Eindhoven BrainPort). In the other target cities, the correlation between patents and the economic success of the clusters is not obvious. Looking at the spatial proximity of the innovation clusters to innovation concentrated regions such as the official innovation centres and universities, Eindhoven stands out again to be the city where the most successful innovation clusters tend to be the ones that are most close to official innovation centres and universities. Regarding the population feature, the correlation results show that the economic success of the innovation clusters doesn’t seem to depend on how populated the area is, except for Paris, Amsterdam, and Stockholm. In Paris and Stockholm, the population plays a positive role while in Amsterdam, the effect is negative.

**Machine learning classifications**

Finally, a machine learning classification experiment has been conducted where the complete feature table has been combined (despite demographics, which was only available for three of the target cities). In this classification feature table, each row corresponds to an innovation cluster from any given cities, described by the 13 indicators that have been derived, extended by the host city as a feature. Then each record has been labelled in a binary fashion as 0 or 1, corresponding to the binary success value they possess. Due to the limited number of cities and clusters, the initial 55 records were reduced to 49 after dropping missing values.

Training an XGBoost classifiers (John Lu, 2010 and Chen et al. 2016), the model has been set up with 200 estimators; a maximum depth of the individual decision trees of 5; a learning rate of 0.1; and a subsample size of 0.8. A grid-search has been conducted under these settings with - due to the small number of instances - with a three-fold cross-validation. Balanced samples have been used, thus picking the same number of 0 and 1 instances at random 20 times and obtained average accuracy values.

Resulting in one aspect show high level of clarity, while from other perspectives, leave many questions open for future research. In the first experiment, the clusters’ cities were used as features (after one-hot-encoding them). In this scenario, the prediction accuracy reached ~ 95% (above the 50% case of random sampling) in a solid way. However, dropping the city tag as a feature, and considered each cluster as an entity with 13 descriptors about company profiles, EU funding, and patent information – like they were not linked to any cities in space at all, the predictions only yielded a modest accuracy of ~ 60%. These findings suggest that the variation between the key drivers of innovation – within the limitations of the current datasets – is mostly driven by unique factors characteristics to the different cities. While this still leaves the possibility of tracing down the main drivers of within-city-success, the low cardinality of innovation districts expects new statistical methodologies and more detailed data sources to be incorporated to deeper understand this problem.

**Summary**
Previous research has identified several key challenges of innovation districts, such as clear quantitative definitions, inadequate and/or fragmented policies and funding schemes, and assessment processes biased towards intuitions and subjective factors instead of hard data, which cause such areas of high potential for innovation slow or fail to develop (Galan-Muros and Hegyi, forthcoming). Policy options to support innovation districts may range from dedicated and long-term resources, the application of comprehensive and integrated approaches, policy alignment across diverse levels of government and across different sectorial organisations.

In this research, first a quantitative, data-driven attempt was performed to design a methodological framework that identifies and evaluates innovation districts. To develop the methodology and test it on real-life data, several unique datasets on innovation and funding have been combined on a European scale. In addition, besides the methodological advances, this explorative work has already helped provide a deeper understanding of the geospatial characteristics of the success of innovation in seven target cities as well.

The core of the proposed methodology is built up as follows. First, a spatial clustering methodology has been used, building on well-established machine learning tools to identify innovation districts across cities. In the current project, density-based clustering have been restricted to the POIs present in Orbis, however, the methodology is applicable to any other set of company-related POI databases. In addition, a series of generic statistical analysis and network science steps have been applied to curate an extensive set of features of the currently used limited datasets characterizing some descriptors and the financial success of these innovation districts. Finally, a quantitative comparison building has been provided in correlation analysis and decision-tree-based classification to shed some light on the role of different cluster descriptors on cluster success.

As regards to machine learning classifications, results show, in one aspect, a high level of clarity, while from other perspectives, leave many questions open for future research. These findings suggest that the variation between the key drivers of innovation – within the limitations of the current datasets – is mostly driven by unique factors characteristics to the different cities. Thus, targeted policies are key for developing such areas with dense innovation activities, shaped to the needs and peculiarities of each city.

In this exploratory research, the focus has been on developing a methodology that allows the identification and measuring of innovation districts by combining several widely used statistical, machine learning, geospatial, and network science tools in a generic way and test them on a few primary data sources capturing urban innovation at a European scale. While the developed methodology performs in a promising way even on the limited sets of exploratory data bases, it also leaves multiple directions open for further research that fall outside of the scope of this paper. Other areas of future research include incorporating sustainability and circularity as variables for measuring innovation and extending the research to cities with diverse characteristics in terms of size or typology of predominant innovation, like agriculture-focused areas.
References


Bennett, J. (2010). OpenStreetMap, Packt Publishing Ltd.


List of abbreviations

DG GROW  Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs
DGSCAN  Density-based spatial clustering of applications with noise
EC  European Commission
ERC  European Research Council
ERDF  European Regional Development Fund
EU  European Union
H2020  Horizon 2020 funds
KPI  key performance indicators
NACE  European Industry-standard classification system
NUTS  Nomenclature of Territorial Units for Statistics
OECD  Organisation for Economic Co-operation and Development
POI  point of interest
ROA  return on assets
List of figures

FIGURE 1 COMPANY FEATURES OVERVIEW – INCORPORATION YEAR DISTRIBUTION 12
FIGURE 2 FEATURES IN THE KEY FINANCIAL TABLE AND THEIR PERCENTAGE OF MISSING VALUES 13
FIGURE 3 HORIZON 2020 AND ERDF RESEARCH AND INNOVATION FUNDING STATISTICS 14
FIGURE 4 DBSCAN SCHEMATIC VIEW POI (POINT OF INTEREST) CLUSTERING WITH DIFFERENT MAGNITUDES OF ‘EPS’ VALUES 16
FIGURE 5 CLUSTER OPTIMISATION 16
FIGURE 6 CATEGORY-LEVEL CLUSTERS FOR PARIS 18
FIGURE 7 CLUSTER OVERLAYING STEPS 19
FIGURE 8 META-CLUSTERS 20
FIGURE 9 ECONOMIC SUCCESS OF INNOVATION DISTRICTS 22
FIGURE 10 ORBIS FEATURES FOR PARIS 24
FIGURE 11 PATSTAT FEATURES FOR PARIS 25
FIGURE 12 H2020 COLLABORATION NETWORK ACROSS THE ANALYSED CITIES 26
FIGURE 13 FUNDING FEATURES FOR PARIS 27
FIGURE 14 PROXIMITY FEATURES 29
FIGURE 15 APPROXIMATED RESIDENT POPULATION IN TARGET CITIES 30
FIGURE 16 CORRELATION BETWEEN CLUSTER FEATURES AND CLUSTER SUCCESS 31
List of tables

TABLE 1 NUMBER OF COMPANIES INVOLVED IN THE ANALYSIS FROM ORBIS PER NACE CODE ........................................ 11
TABLE 2 NUMBER OF ORBIS POI CLUSTERS IDENTIFIED IN EACH CITY PER NACE CODE ........................................ 17
TABLE 3 NUMBER OF META CLUSTERS .................................................................................................................. 20
TABLE 4 EXAMPLE OF PERFORMANCE PROFILE ......................................................................................................... 21
TABLE 5 ORBIS FEATURES FOR PARIS ............................................................................................................................. 24
TABLE 6 PATSTAT FEATURES FOR PARIS ............................................................................................................................ 25
TABLE 7 EU FUNDING FEATURES FOR PARIS .................................................................................................................. 27
TABLE 8 AVERAGE PROXIMITY VALUES ............................................................................................................................ 30
TABLE 9 CORRELATION BETWEEN CLUSTER FEATURES AND CLUSTER SUCCESS FOR ALL CITIES .................................. 33
GETTING IN TOUCH WITH THE EU

In person
All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email
Europe Direct is a service that answers your questions about the European Union. You can contact this service:
- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696, or
- by electronic mail via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online
Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications
You can download or order free and priced EU publications from EU Bookshop at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).
The European Commission’s science and knowledge service
Joint Research Centre

JRC Mission
As the science and knowledge service of the European Commission, the Joint Research Centre’s mission is to support EU policies with independent evidence throughout the whole policy cycle.