Spatio-temporal localisation pattern of technological startups – RNN in predicting intra-urban startups' clusters

Maria Kubara

Faculty of Economic Sciences, University of Warsaw, Warsaw, Poland;
ORCID: 0000-0002-8768-8391

Correspondence details:
University of Warsaw, Faculty of Economic Sciences,
ul. Długa 44/50, 00-241 Warszawa, Poland,
maria.kubara@uw.edu.pl
+48 518 309 590

Abstract:

Little is known about the localisation decisions of the technological startups at the urban level. While the general trend of attracting innovative companies to the metropolitan areas is well known and thoroughly researched, much less is understood about the micro-geographical patterns appearing within the cities. Considering the growing number of papers mentioning that agglomerative externalities attenuate sharply with distance, such analysis of micro-scale localisation patterns is crucial for understanding if these effects are of importance for technological startups. Taking a sample of startups from the up-and-coming market in Central-East Europe in Warsaw, Poland their spatial organisation across the years will be tracked to investigate whether there is a defined pattern consistent with highly localised externalities operating within cities and how this pattern evolves over time. Additionally, the paper will show how recursive neural networks (RNN) may help in predicting the locations of technological startups' clusters. It will be presented how to include the spatial dimension in the model in a computationally effective way and how this augmentation improves the results, allowing the network to "understand" the spatial relations between neighbouring observations.

Keywords: startups, intra-urban, business localisation, agglomeration, RNN, spatial data science

JEL codes: C21, C33, C38, C45, R12

Highlights:

- Localisation patterns of technological startups at the intra-urban level show a high tendency for co-locating and clustering.
- While metropolitan areas on their own are appealing to innovative companies, the densest intra-urban locations are increasingly attractive for startups.
- Observed localisation patterns are thus consistent with highly localised agglomeration externalities operating at fine scales within cities.
- Machine learning and deep learning techniques prove to be highly effective in the micro-geographical analysis of localisation choices.
- Recurrent neural networks (RNN) can provide accurate predictions on spatio-temporal localisation trends of technological startups, predicting an intra-urban cluster localisation with 1km precision.

1. Introduction

Technological startups nowadays are locating more frequently in cities (Duvivier & Polèse, 2018). A growing literature on the geography of startup locations shows that the diversified urban environment becomes increasingly attractive for innovative businesses (Arauzo-Carod, 2021). Little is known however about the specifics of that process. In the literature, it is suspected that agglomerative economies are the most important drivers of this towards-city switch (Jang et al., 2017; van Oort & Atzema, 2004). Yet more recent papers suggest that the positive effects of agglomeration have strong attenuation with distance and their most significant influence is only experienced in fine geographical scales (Andersson et al., 2016; Ferretti et al., 2022; Jang et al., 2017; Rammer et al., 2019). This may indicate that companies that strive to utilize these effects need to co-localise, creating dense intra-urban clusters of business activity.

Identifying intra-urban startup localisation patterns is important from at least two perspectives. On the one hand, it allows to revisit the question of agglomerative economies' importance for innovative businesses, tracking the possible influence of such effects at the micro-scale, which is rarely seen in the literature. On the other, understanding the forces driving startups' localisation decisions can make a strong case for policymaking, showing where new startups tend to be created, which may be used as an indicator for city planning. To get a better understanding of these effects, in this paper, an attempt will be made to check whether we can see startup localisation patterns consistent with highly localized externalities operating at fine spatial scales within cities.

To verify that statement for technological startups, one needs to track their localisation decisions at a microgeographic scale, not larger than a city neighbourhood. Ideally, the startups' localisation decisions should be observed at their exact georeferenced addresses and compared over time with the newly created companies to track the stability of the spatial pattern – a goal that is not easily achieved using traditional spatial econometrics models. Such analysis however can be successfully conducted with the usage of modern machine learning methods on point data of startup location. Techniques such as DBSCAN or recursive neural networks can be effectively utilized to identify dense clusters of startup activity and to track the new business formation patterns over time. Moreover, the usage of deep learning tools like RNN enables making predictions regarding the spatial patterns of the business location and forecasting future hot spots of startup formation within cities. Such forecasts can be of high value for creating targeted startup support policies or land-use agendas for city planning.

The contribution of this paper is threefold. Firstly, it will add up to the literature on the attenuation of agglomeration externalities, verifying if the spatio-temporal localisation patterns of startups are consistent with small scale of impact of these effects. Secondly, a methodological contribution will be made, showing how modern machine learning techniques and deep learning models can be utilized in micro-geographical analyses, extending the possibility to quantify co-localisation tendencies, and to predict spatio-temporal trends. Finally, the empirical analysis will be conducted on the sample of technological startups in Warsaw, Poland, extending the knowledge about startup operations in up-and-coming markets of central-east Europe.

2. Motivation and related literature

Despite growing globalisation forces and declining cost of communication, today's knowledge-driven economy is continuingly benefiting from agglomeration externalities (de Groot et al., 2009). These effects are especially important for innovative industries, for which knowledge spillovers and information flows – forces hugely reliant on proximity – are the key to developing ideas and creating innovative products (Jang et al., 2017; Rammer et al., 2019).

The importance of agglomerative economies may be observed in the location decisions of technological startups. The literature reports, that there is a growing number of companies that decide to locate within urban spaces rather than suburban areas (e.g. Arauzo-Carod, 2021; Duvivier & Polèse, 2018). The switch in the location trends has been noticed by the popular press, popularising terms like "Silicon Alley" for the new technological cluster in Manhattan and "Silicon Roundabout" for a similar structure created in London (Duvivier & Polèse, 2018). Although, while the general towards-city tendency has been well recognised, still little is known about the intra-urban patterns of technological startup location.

The concentration of economic activity (both in industrial clusters and the metropolitan areas) brings positive effects to the entities localised nearby. These agglomeration externalities are intensively researched in the literature, investigating how they may impact the companies located within their range (e.g. Devereux et al., 2007; Jang et al., 2017; Neffke et al., 2010). Yet an increasing body of literature is showing that the spatial range of agglomeration externalities is smaller than was previously expected, forcing the switch in the analytical approaches from metropolitan to intra-urban level. Among the micro-foundations of urban agglomeration economies, as defined by Duranton & Puga (2004), the most limited range of influence seems to be connected to the learning effects. In the literature, it is suggested that the most effective impact of these effects may be restricted to areas as small as city neighbourhoods (Andersson et al., 2016; Ferretti et al., 2022), or even lesser 50-250m range (Rammer et al., 2019).

Learning effects are particularly important for technological companies. Stimulation of knowledge flows and information sharing between creative individuals are increasing productivity and raising the chances of creating an innovative product (Boschma, 2005; de Groot et al., 2009; Isaksen, 2004; Neffke et al., 2010). While it may seem that the rising popularity of remote communication techniques should diminish the role of physical proximity, especially for tech-oriented companies, the literature is proving otherwise. As it was shown by A. Isaksen (2004), the knowledge specific to technological companies is tacit and complex, and in order to effectively share it, face-to-face interactions are needed. This may push these companies to co-locate next to each other, creating clusters of innovative activity at the intra-urban level.

Despite the expected positive effect on productivity and innovativeness, localising within a cluster may be a hazardous decision for a startup. Dense co-location with other similar companies creates a push from both supply and demand side – making the companies compete for production factors, skilled employees, investors, and customers. What is more, in the urban space especially, a greater presence of companies in a specific area generates upward pressure on the office rental costs, creating an additional financial burden on the entering startups

(Huynh, 2014; Jennen & Brounen, 2009). Additionally, the same intensive knowledge exchange that can facilitate the innovative process, makes keeping the prototypes secret much more difficult, making startups prone to drainage of ideas (Kogler, 2015). When these negative forces prevail, the newly founded companies may be driven to exit the market prematurely. Thus, anticipating the negative consequences of concentration, some technological startups may decide to localise outside of the clusters, choosing the remote areas of the city.

Acknowledging the importance of the agglomeration externalities for innovation creation and their sharp attenuation with distance, I expect that technological startups will co-locate and create intra-urban clusters of innovative activity. If such a pattern can be observed at the micro-geographic level, it may be suspected that technological startups utilize the highly localized externalities operating at fine spatial scales within cities. At the same time, it may be the case, that a portion of newly founded companies chooses remote parts of the city, localising further away from the areas of business concentration. These conflicting trends of concentration and dispersion within a city space leave us with an open-ended question: which of those patterns will be identified and which will prove to be relatively stronger for technological startups? The answer to that issue will be the main goal of the empirical section of this paper.

3. Study design and dataset

3.1. Sample

Although there were many attempts at the unification of the "startup" definition (e.g. Nauman & Edison, 2010; Reisdorfer-Leite et al., 2020), in the literature, there is still no consensus about the description of this term. In their mapping study, Paternoster et al. (2014) have identified a few themes which dominate in the literature regarding startups in the technological area. According to the papers they screened, startups are usually defined as new companies, which lack resources, have a small low-experienced team of employees, develop mainly one innovative product and evolve rapidly in a very uncertain environment (Paternoster et al., 2014). However, depending on the focus of a certain paper, different factors were taken into consideration while deciding whether the company is a startup or not. Thus in this paper, the following working definition of a technological startup has been assumed – it is a newly founded business entity (operating for up to 5 years), whose main specialisation is in the domain of computer technology and software.

Specifically, the total sample consists of 11'100 business entities founded between 2010-2018 in Warsaw, Poland. Companies were selected concerning their specialisation statement. Companies that will be considered in this paper specialise in manufacturing of computers, electronic and optical products, manufacturing of the peripheral equipment for computers, publishing of computer games, computer programming, software-related consultancy and specialising in other information technology and computer service activities. Information about the companies was obtained from the ORBIS database and the address information was geo-coded with the usage of MapQuest API (MapQuest, 2018).

The choice of the city for this analysis was dictated by a few factors. Firstly, while there is plenty of research concerning innovative businesses in developed economies like the United States, United Kingdom or Germany (e.g. Banal-Estañol et al., 2019; Florida & Mellander, 2017; Geibel & Manickam, 2016; Pisoni & Onetti, 2018), little is known about the up-and-coming markets of Central-East Europe. Secondly, the Warsaw startup community is dynamically growing and attracting bigger sources of funding from both domestic and foreign capital. Following high anticipation, in 2021 the first "unicorn" was declared (Bełcik, 2021) – a startup whose value is assessed above 1 billion USD. The presence of "unicorns" is a strong sign for investors, that the market is highly attractive and profitable (Suwarni et al., 2020). Thirdly, Warsaw with a population of 1'794'200 people and 517.24 square kilometres meters of coverage (Urząd Statystyczny w Warszawie, 2021) is one of the European metropolises. Being a highly heterogeneous city, with areas of business concentration and less densely populated peripheries, it creates an interesting case study for the different location opportunities and consequences of such choices for newly founded businesses.

Following the call for analysis of entrepreneurial activity at a "very low level of aggregation", as suggested by Guzman & Stern (2016), the localisation choices of technological startups will be here investigated from the micro-geographic perspective. Individual localisation points will be considered, created by geocoding the business address with an accuracy of 0.1 meters[1]. The highest aggregation level that will be used is the 1km per 1km grid, which size is dictated by the accessibility of the population grid coming from the census (Portal Geostatystyczny, 2021).

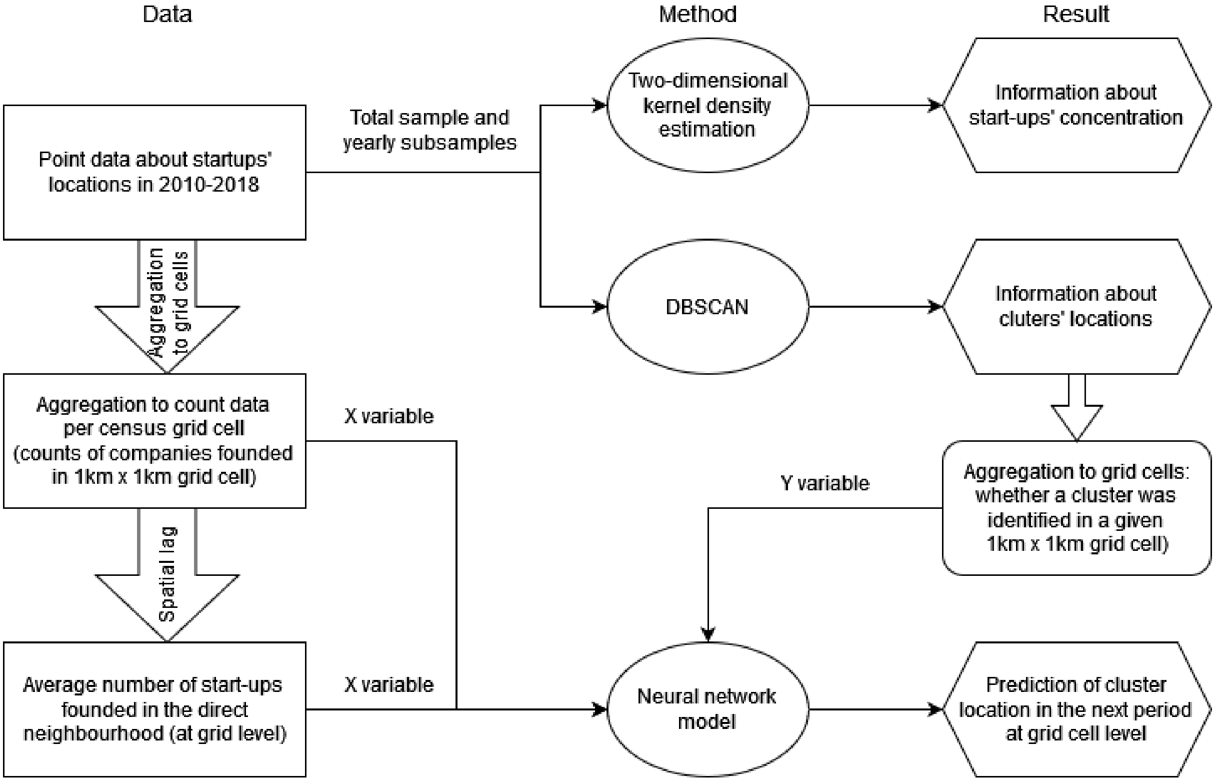
3.2. Methods

The empirical strategy of this paper is organised into three stages: recognition of startup concentration within the urban space, identification of clusters based on the density of startups, and prediction of future cluster localisations with 1km precision. The main objective, which is

[1] GPS coordinates with six decimal places allow for such precision. In our case we may additionally consider possible errors in the assignment of a particular address to a given geocoded location, however in the urban areas it is rarely the case (Davis & Fonseca, 2007; Goldberg & Wilson, 2007).

to verify if we can observe localisation patterns consistent with sharply attenuating agglomeration externalities, will be achieved with a mixture of machine learning and deep learning methods[2] (see Fig. 1).

Figure 1. Structure of the research methods, data formats and expected results


Data
Method
Result
Point data about startups' locations in 2010-2018
Total sample and yearly subsamples
Two-dimensional kernel density estimation
Information about start-ups' concentration
DBSCAN
Information about cluters' locations
Aggregation to grid cells
Aggregation to count data per census grid cell (counts of companies founded in 1km x 1km grid cell)
X variable
Aggregation to grid cells: whether a cluster was identified in a given 1km x 1km grid cell)
Y variable
Spatial lag
Average number of start-ups founded in the direct neighbourhood (at grid level)
X variable
Neural network model
Prediction of cluster location in the next period at grid cell level

Source: Own work in diagrams.net software.

3.2.1. First stage – concentration tracking with kernel density estimation

In the first stage, areas of concentration of startup activity within a city space will be recognised. This goal will be achieved by applying two-dimensional kernel density estimation with 25 bins on the geo-coded business locations (see Fig.1). That method allows one to recognise the concentration of points on a two-dimensional plane (Wand & Jones, 1994).

In this case, it means that it can be effectively utilized to identify hot spots of startup activity, pointing to the city areas which attract the most companies and create the most concentrated business clusters. Applied to the total sample, this method can show the general pattern of startup concentration in Warsaw, identifying the areas which are relatively the most

[2] Codes and datasets needed to reproduce this analysis can be found in Github repository https://github.com/mariakubara/RNN_startup_clusters

attractive for technological companies. Repeating the analysis on the yearly subsamples of newly created companies, we may spot any possible changes in the hot-spots localisations.

From the technical perspective, kernel density smoothing is a non-parametric estimation of a density function, where points are assigned relative importance depending on the number of points located around them (Wand & Jones, 1994). Results obtained for individual points are then smoothed to obtain an estimate of a density function (Wand & Jones, 1994). In the literature, this method was successfully applied in cases of hot-spot identification and recognising multimodality (see e.g. Hart & Zandbergen, 2014; Hu et al., 2018; Lin et al., 2011; Silverman, 1981; Wand & Jones, 1994).

3.2.2. Second stage – cluster identification with DBSCAN

While results from the previous stage can show the general tendencies in the intra-urban startup localisation pattern, especially from the concentration perspective, the next stage of the analysis will be focused on recognising and quantifying the clusters of startup activity (see Fig.1). In general, we may say that for a cluster to be created, a big enough number of companies need to locate closely next to each other. This working definition directly points to the fact that clusters depend on a density of a business location. Automatic identification of such density-based clusters can be done with the usage of the DBSCAN method, for which “big enough” and “closely” are just parameters to tune.

Specifically, DBSCAN is a density-based clustering method, which allows for the identification of clusters of arbitrary shapes (Ester et al., 1996). The algorithm tries to find a concentration of some minimum number of points (“big enough” parameter: *minPts*), which are located within a particular distance from each other (“closely” parameter: *eps*, interpreted as reachability distance radius). If such concentration is found, a cluster is recognised (Ester et al., 1996; Schubert et al., 2017).

In this case results from this method will provide information on how many clusters were created by technological startups, how big they were, where exactly they were located, and, last but not least, how many companies were not assigned to any groupings. These results will shed light on the co-location tendencies across startups, at the same time comparing the relative strength of concentration and dispersion trends in the urban space.

3.2.3. Third stage – recurrent neural networks

After recognising already existing trends and patterns of startup localisation, we may want to take a look into the future and try to predict areas with the highest startup attraction potential in the upcoming years at the intra-urban level. Such knowledge may be especially beneficial from the policymaking perspective and can have a significant impact on urban planning.

To achieve this goal a model will be build, which will be capable of predicting whether a certain city area will be a part of a startup cluster (see Fig.1). In this part aggregated data will be used, where the total city area will be divided into 1km per 1km grid cells[3]. The model will take information about the history of the attractiveness of a given area (taken as the 3-year sequence of the share of startups located in a given grid cell in consecutive periods) and predict if in the upcoming period a cluster will be formed there (exact cluster localisations are drawn from the DBSCAN results).

To find such a connection a model used here needs to be able to learn (fuzzy) patterns within the dataset and provide accurate predictions – a problem that can be effectively solved by neural networks. In this case, the recurrent neural network (RNN) will be the most appropriate, as it allows for processing the time-series and text data, uncovering the patterns hidden in data sequences. It is an augmentation of the standard neural network procedure, enabling the network to keep the "memory" of the previously processed inputs (Medsker & Jain, 1999). In recurrent neural networks, the same weights are recursively applied over a structured, sequence-sensitive input in a linear form – combining previous time steps and information drawn from with the next input (Medsker & Jain, 1999). RNNs were successfully utilized in numerous use cases, including short panel data (Fan et al., 2017; Gu et al., 2019; Zhang & Man, 1998).

Utilizing RNN in this paper's scenario will allow the model to learn the information stemming from the attractiveness sequence, remembering how the startup concentration evolved in time, and connect it to the probability of cluster formation. Adding a spatial lag as a second variable will allow the model to additionally consider the spatial relations within the sample. Results obtained in this stage will provide information about the stability of the concentration patterns and provide an accurate prediction of the future cluster localisations with 1km precision.

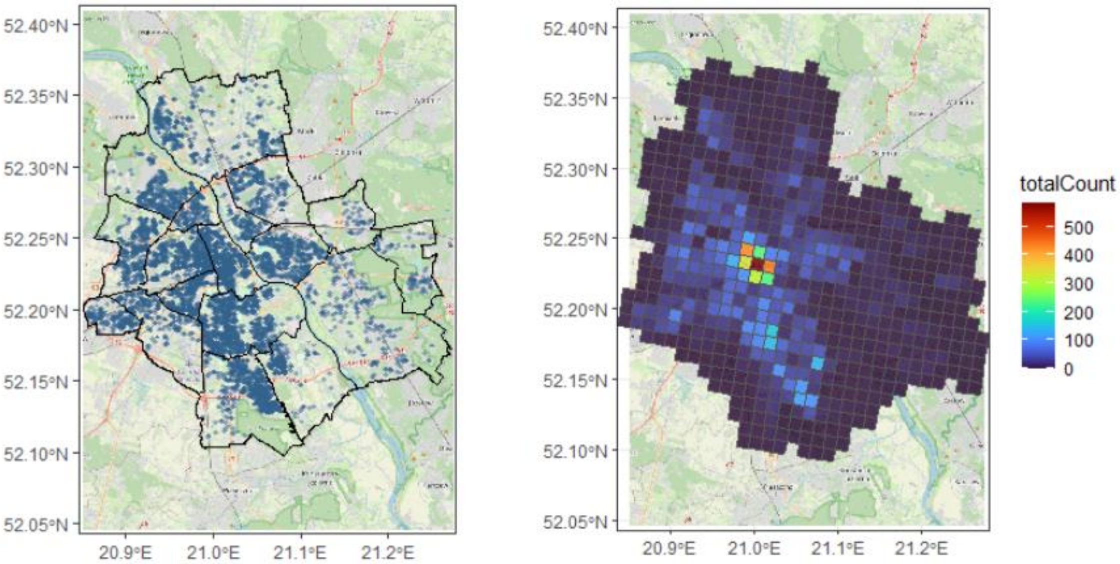
4. Results

4.1. Visual exploratory data analysis

[3] City division provided by the census grid stemming from INSPIRE project (Portal Geostatystyczny, 2021).

To build the reader's intuition about the situation we are considering, the results of the visual exploratory data analysis of the considered sample will be presented. Figure 2A presents the total sample of startups and the administrative borders of Warsaw with the background map published by the Open Street Map association (OpenStreetMap, 2022). The city is divided into 18 districts, which are diversified from the perspective of urban organisation and business concentration. This difference is well visualised by the raw presentation of the sample plotted on the city map (see Fig.2A). It can be seen that there are some preferred localisations that are commonly chosen for startup localisation (mainly the inner-city), while there are also some districts that are less preferred and probably disadvantaged economically, especially on the east and southern side. There is also a visible division of the city which is indicated by the Vistula river. One can spot that the western side of the city is more densely localised by innovative businesses, while on the easter side there are much fewer companies, which exist in smaller concentration. Looking at the grid representation of our sample in Fig.2B, it can be seen that the distribution of companies is uneven in the space. There is much more concentration of the startups in the city centre and the southern-west side of the city than in the remaining area.

Figure 2. Administrative borders of Warsaw and the distribution of the total sample of startups founded between 2010-2018



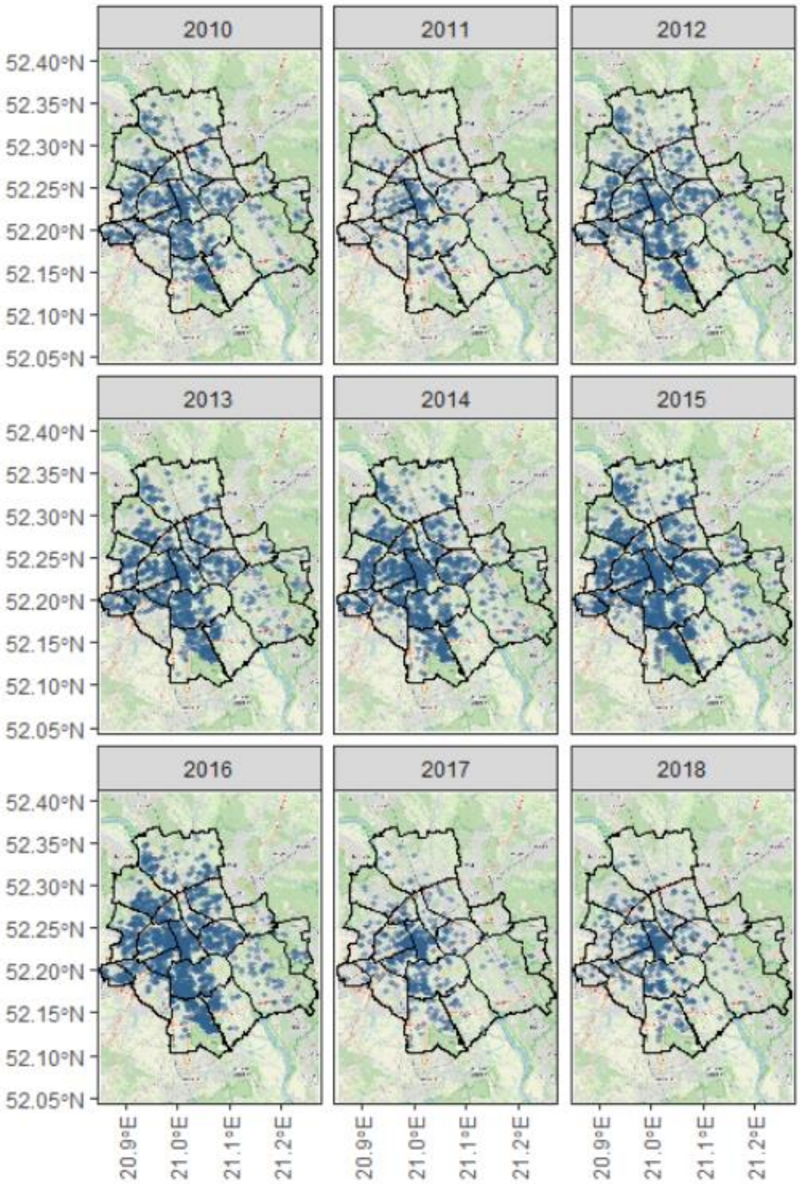
Details: Image A shows the administrative borders of Warsaw, including the division into districts, and the overall point distribution of the total sample (all startups established between 2010-2018 in Warsaw). The saturation of an individual point is set to 50%. The more points located in a certain area, the more intensive the overall colour will be. Image B shows the distribution of the companies of the total sample, recognised as counts of founded companies per given grid cell. Grid distribution is following the structure of the 1km x 1km census population grid from the INSPIRE project, which is also utilized in section 4.4.

Source: Own work in R, utilizing the OpenStreetMap background map for Warsaw, Poland.

One can also observe that there were some visible differences between the spatial patterns created by the startups in different years considered in our sample (see Fig.3). Except for the number of founded companies (which varied between 262 in 2011 and 2188 in 2016), there were also some changes in the spatial pattern created by the group as a whole. It can be seen that between 2012 and 2016 there was a growing tendency for the companies to concentrate, which seems to be changing in the last two years considered.

Due to the high number of business entities and low spectral dimension recognised by the "bare eye" such simplified analysis can be only used for some intuitive statements, which will be then verified by the usage of statistical methods.

Figure 3. Yearly subsamples of newly founded startups plotted on the map of Warsaw



Details: The saturation of an individual point is set to 50%. The more points located in a certain area, the more intensive the overall colour will be.

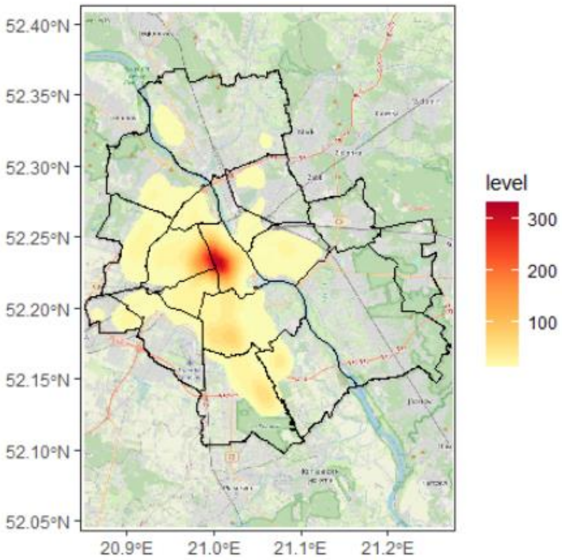
Source: Own work in R, utilizing the OpenStreetMap background map for Warsaw, Poland.

4.2. Results of the hot-spot analysis:

The first part of the formal analysis will consider the business concentration of the companies in general terms. It will be investigated if any significant hot spots were created and if such tendencies are visible in yearly samples or the whole sample only.

Figure 4 shows the result of kernel density estimation for the whole considered sample. It turns out that startups, in general, are localising densely on the left-hand side of the Vistula river. There is some concentration on the eastern side of the map, but it is much lesser than the one recognised on the western side. One very visible hot spot in the city centre can be observed, where business concentration is significantly higher than in other parts of the city. Though three other concentration areas in the southern part of the city can be observed as well. Those are less concentrated than the one in the centre, but there are indeed present. The localisation structure of startups in Warsaw is polycentric and not homogeneously distributed in the urban space.

Figure 4. Results of kernel density estimation for the whole sample of startups

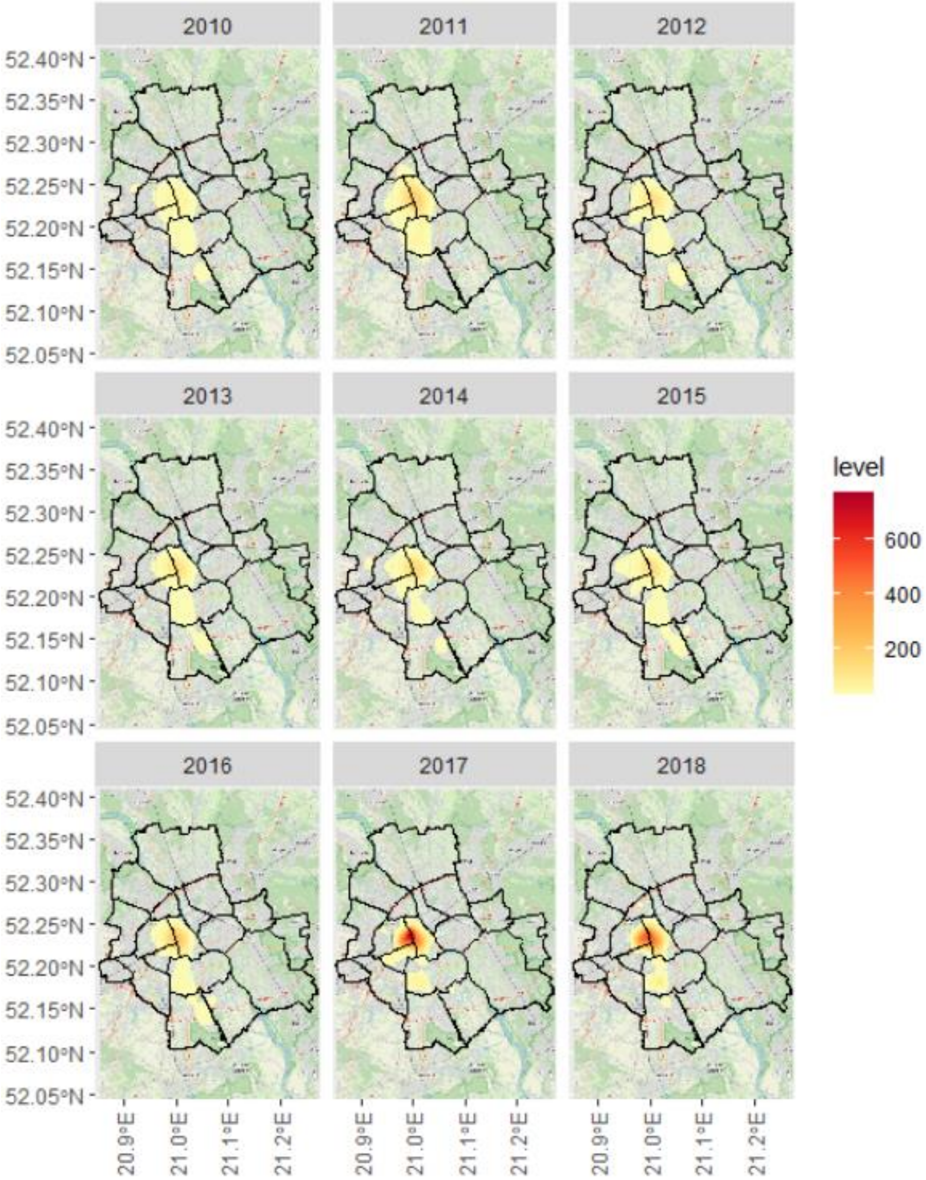


Source: Own work in R, utilizing the OpenStreetMap background map for Warsaw, Poland.

Following the yearly subsamples of newly created companies it can be observed that the concentration pattern was changing over time, separating the concentration effect (see Fig.5). By assessing the density of newly created businesses, it can be seen how the location tendencies evolved over time, showing only the most significant hot spots created by startups founded in a given year.

The most visible tendency observed in the yearly subsamples is the strengthening of the central hot spots. While in 2010 only some general concentration of startups on the left-hand side of the Vistula river can be observed, in the following years startups were more intensively choosing the hot spots rather than the overall western part of the city. Between 2016 and 2018 the importance of the central hot-spot grew, up to a point in which most of the startups' concentration was localised in the most central, prestigious city area.

Figure 5. Results of kernel density estimation for yearly subsamples of newly created startups



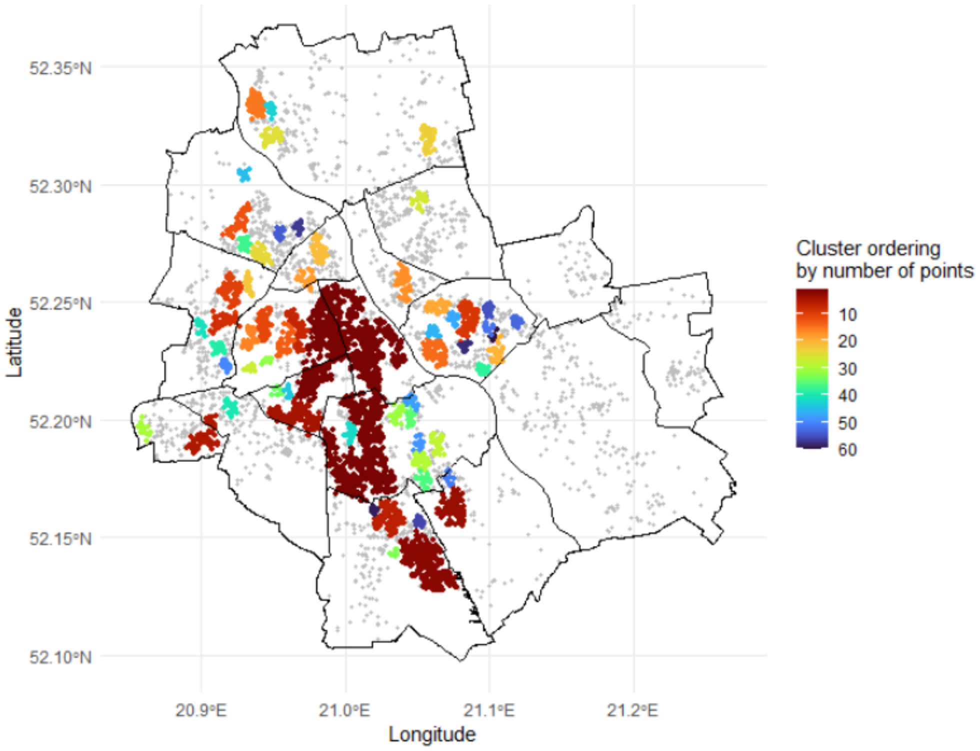
Source: Own work in R, utilizing the OpenStreetMap background map for Warsaw, Poland.

Results from kernel density estimation show that technological startups tend to co-locate. The strength of the concentration trend is changing over time, but the conclusion remains constant – the localisation pattern each year showed significant hot spots of startup concentration. In these hot spots, the companies are located densely, and their concentration is much higher than in the remaining parts of the city. These results suggest that the technological startups may follow the spatial patterns consistent with the highly localised externalities operating in cities. However, to assess the strength of the concentration trend, we need to compare the number of companies that co-locate and disperse in the city area. For this aim, the DBSCAN method will be utilized and its results will be discussed in the next section.

4.3. Results of DBSCAN

DBSCAN is a method that allows one to track even small groupings of companies, created within a target distance from each other. Such analysis can show how many clusters are created, what sizes they have and how many companies are localising outside of the business groupings. It allows tracking the information omitted by the previous method and verifying the relative strength of the concentration and dispersion trends.

Figure 6. Density-based clusters identified in the total sample



Details: DBSCAN parameters: eps=0.0035 (approx. 238m) and minPts=20. Clusters are ordered and coloured according to their size, e.g. cluster of order=1 is the biggest grouping with 4625 observations, cluster of order=60 is the smallest grouping with 9 observations. Grey points represent startups that were not allocated to any cluster.

Source: Own work in R.

At first, it will be checked how many clusters can be identified in the total sample of startups considered in the analysis. With the parameters eps=0.0035[4] and minPts=20, the algorithm recognised sixty clusters (see Fig.6). Their sizes varied between 4625 and 9 observations in each. The largest cluster (dark red points located centrally on Fig.6) was

[4] Eps in this paper is measured in degrees (as the longitude-latitude coordinates are). Around 52.25N, a 0.001-degree distance is equal to 68 meters (Morse, 2008). In this case, the distance for eps=0.0035 is approximately 238 meters, while eps=0.006 is equal to approximately 408 meters. Because there is no theory supporting the choice of appropriate values for eps and minPts (Lai et al., 2019), values for both parameters were initially defined using a knee plot and then tuned to the sample following the adaptive approach (Sawant, 2014).

localised in the city centre, on the western Vistula bank and was aligned with the hot spot identified by the previously discussed method. Despite the identification of that many groupings, there is also a large share of companies that locate outside clusters (see Fig. 6). In the total sample, the method has recognised 8288 of those firms which make for 74.7% of the analysed sample.

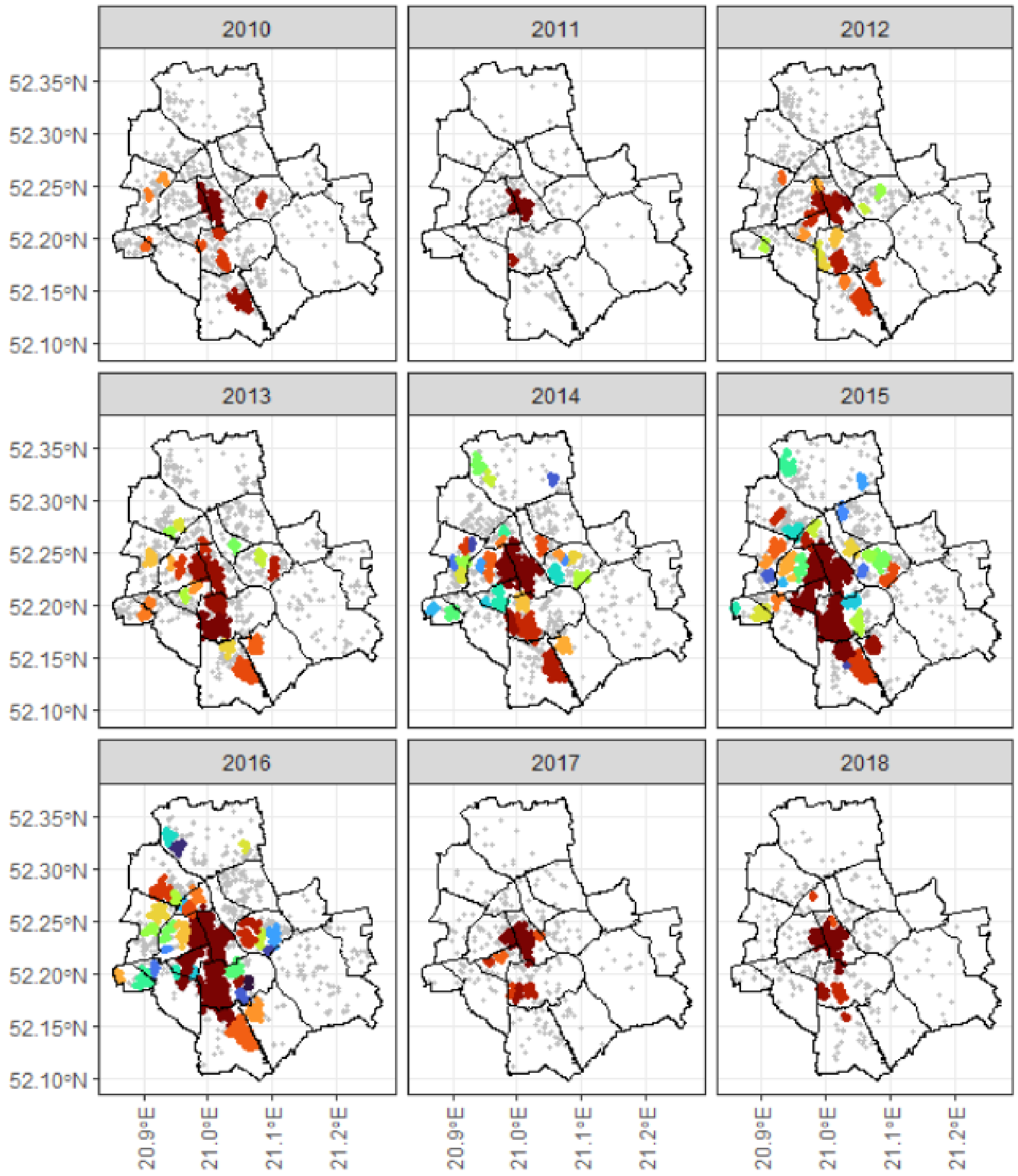
Table 1. DBSCAN results in detail

Sample	***Companies in the sample***	***Number of clusters***	***Companies in clusters***	***Companies outside clusters***	***Share of companies in clusters***
Total sample	11100	60	2812	8288	25.3%
2010	821	9	248	573	30.2%
2011	304	2	81	223	26.6%
2012	978	16	461	517	47.1%
2013	1360	17	767	613	56.4%
2014	1538	28	1002	536	65.1%
2015	2105	28	1615	490	76.7%
2016	2324	30	1831	493	78.8%
2017	832	7	540	292	64.9%
2018	838	6	534	304	63.7%

Details: Yearly subsamples include only technological startups which were founded in a given year, while the results of the total sample show clusters recognised for the aggregated group of all startups founded between 2010 and 2018.

Source: Own work.

Figure 7. Density-based clusters identified for yearly subsamples of startups



Details: Yearly subsamples include only technological startups which were founded in a given year, while the results of the total sample show clusters recognised for the aggregated group of all startups founded between 2010 and 2018. DBSCAN parameters: eps=0.006 (approx. 408m) and minPts=10. Grey points represent startups that were not allocated to any cluster.

Source: Own work in R.

The number and size of identified clusters differed throughout the years (Fig.7). In 2010 among 821 startups created only 30% were localised in nine clusters (see Tab.1). The remaining companies were spread in the urban space with a slight preference for the localisation in the western part of the city. In 2011 following the lower number of founded companies only two clusters were identified – one of which was located directly in the city centre. In the samples of

2012-2016, a growing number of groupings was recognised with increasing size and significance of the central cluster. The share of startups localised outside of any grouping was lowering – from 52.9% in 2012 to 21.2% in 2016 (see Tab.1). Though in 2017 and 2018 this pattern seemed to change. Fewer clusters were identified and most of the grouped companies were located in the city centre, concentrating in much smaller areas than before. At the same time, the share of dispersing companies has increased to the levels identified in 2014 (share of companies dispersing in the urban area was 35.1% in 2017 and 36.4% in 2018). However, while more companies decided to follow the dispersion trend, the concentration of the remaining businesses was stronger than ever. The clustering tendencies were much more localised, following the pattern consistent with the agglomeration externalities operating at fine spatial scales.

Some changes have been observed in the spatio-temporal localisation pattern of technological startups. Across the years the concentration trend has strengthened. At first this tendency was visible in the rising share of clustering companies, and then in the increasing density of the newly created groupings. While the share of clustering companies fluctuates, the concentration trend is relatively stronger than the dispersing tendency for technological startups. Clusters created by newly founded companies are dense and concentrated in similar localisations – following a polycentric pattern with hotspots located at the western side of the city. Following this insight, it can be stated that the majority of technological startups creates a spatio-temporal localisation pattern based on concentration and co-location. These two elements, binding companies together in small-scale distinct city areas are consistent with the highly localised agglomeration effects. Thus it can be suspected that small-scale agglomeration externalities may be important forces influencing the intra-urban location decisions of the majority of technological startups.

4.4. Recurrent neural networks in predicting the startup clusters

The knowledge about the trends in startup localisation patterns may be very beneficial. Being able to predict future hot spots could be of help for accurately planning city infrastructure, locating startup support facilities or proposing local programs promoting entrepreneurship in less preferred areas. While the theoretical roots of the intra-urban startup location decisions are yet not fully understood, it is possible to utilize deep learning methods such as neural networks to “learn” the patterns available in the data.

4.4.1. Model structure

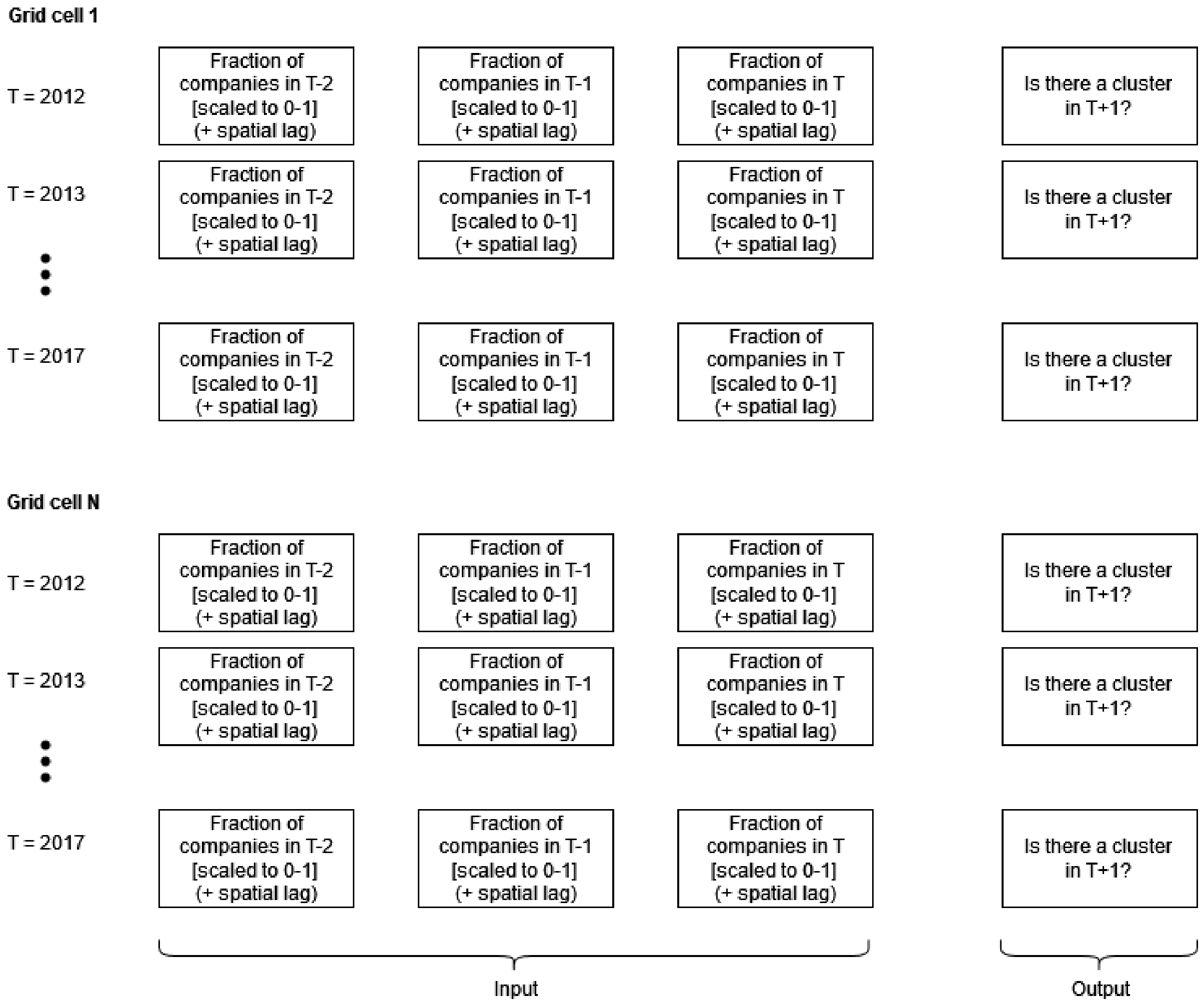
In this case, it is aimed to build a model which will be able to predict the startup's clusters localisations, as identified in the previous section with the usage of the DBSCAN method. Because the goal is to track a highly spatially autocorrelated process, it is crucial to incorporate the spatial dimension into the model. Common practice is to use convolutional neural networks for spatial data modelling. However, such an approach works best for long time series, following preferably not more than a few distinct data points. As our data is a short panel (601 grid cells observed for 9 years), such an approach was not available. Thus the decision has been made to use recurrent neural networks, which can follow short-term relationships between observations, and include the spatial dimension as an additional feature.

Specifically, the model will be fed with time-series data, containing the information about the fraction of companies localised in a particular grid cell in the past three years[5]. In the second specification, the model will be augmented with the additional spatial feature, which will store the information about the first-order spatial lag of the former variable (spatial average of the fraction of the companies founded in a given year from adjacent cells).

Input data for the model was created in the following way. Firstly, the calculation of the counts of companies in 1km x 1km grid cells that were founded in particular locations each year was made. Then the counts were transformed into fractions – each number assigned to a cell was divided by the yearly sum of founded companies. Then the data were rescaled to 0-1 to keep a consistent scale across the samples, which is a standard procedure for the neural network models (Beck, 2018). Additionally, the first-order spatial lag of those values was calculated for each year and each grid cell with the usage of a queen spatial weight matrix based on a contiguity criterion. Then the data was reorganised into 4-year sequences as presented in Fig.8. The three-year sequences of companies' fractions will be used as the input of the model, while the information about the presence of a cluster in the T+1 period will be treated as the output of the model.

Figure 8. Structure of input and output data

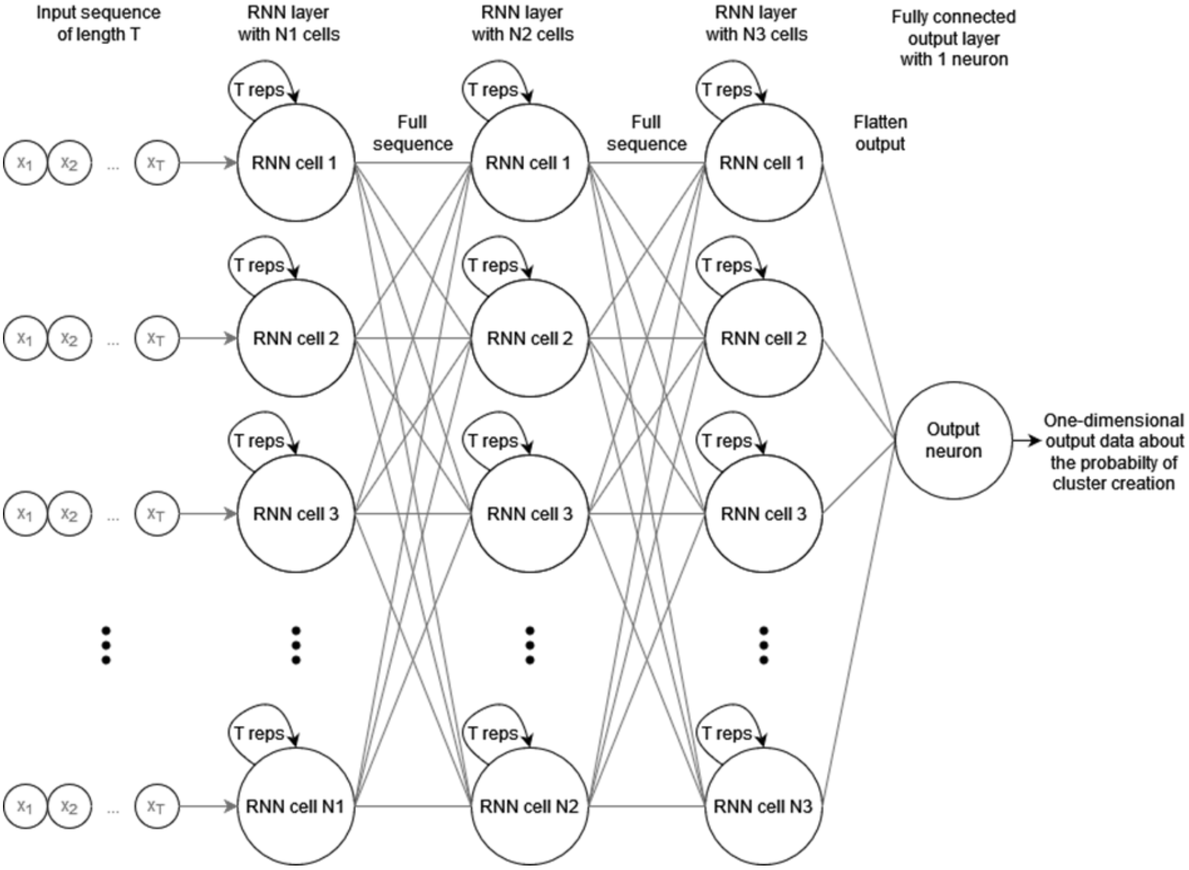
[5] The grid structure utilized for this task was already shown in Fig 1C.


Grid cell 1
T = 2012
Fraction of companies in T-2 [scaled to 0-1] (+ spatial lag)
Fraction of companies in T-1 [scaled to 0-1] (+ spatial lag)
Fraction of companies in T [scaled to 0-1] (+ spatial lag)
Is there a cluster in T+1?
T = 2013
Fraction of companies in T-2 [scaled to 0-1] (+ spatial lag)
Fraction of companies in T-1 [scaled to 0-1] (+ spatial lag)
Fraction of companies in T [scaled to 0-1] (+ spatial lag)
Is there a cluster in T+1?
T = 2017
Fraction of companies in T-2 [scaled to 0-1] (+ spatial lag)
Fraction of companies in T-1 [scaled to 0-1] (+ spatial lag)
Fraction of companies in T [scaled to 0-1] (+ spatial lag)
Is there a cluster in T+1?
Grid cell N
T = 2012
Fraction of companies in T-2 [scaled to 0-1] (+ spatial lag)
Fraction of companies in T-1 [scaled to 0-1] (+ spatial lag)
Fraction of companies in T [scaled to 0-1] (+ spatial lag)
Is there a cluster in T+1?
T = 2013
Fraction of companies in T-2 [scaled to 0-1] (+ spatial lag)
Fraction of companies in T-1 [scaled to 0-1] (+ spatial lag)
Fraction of companies in T [scaled to 0-1] (+ spatial lag)
Is there a cluster in T+1?
T = 2017
Fraction of companies in T-2 [scaled to 0-1] (+ spatial lag)
Fraction of companies in T-1 [scaled to 0-1] (+ spatial lag)
Fraction of companies in T [scaled to 0-1] (+ spatial lag)
Is there a cluster in T+1?
Input
Output

Source: Own work in diagrams.net software.

A four-layer neural network structure will be used, where three first layers are fully connected RNNs and the last one is a fully connected dense layer with one output. The general specification is shown in Fig. 9. RNN layers are fed with T-dimensional inputs, returning full sequences to their successors. The output is finally flattened in the last RNN layer to fit in the simpler neurons. After each RNN a random dropout ratio is used, which controls the fraction of weights learned in the previous training iteration (epoch) that will be removed (forgotten in the network) and tuned again in the next epoch (Gal & Ghahramani, 2016). This approach helps the model to constantly change its parameters, finding a solution that will be suitable for the whole phenomenon rather than just for the training dataset. Especially in the RNN networks, where there are many neurons with numerous repetitions, it is crucial to introduce measures such as random dropout ratio which help with mitigating the overfitting problem (Gal & Ghahramani, 2016), The learning ratio annealing parameter is additionally used to optimise the searching criteria for the best specification, as the model reaches a plateau. The first model's specification is using only one feature – which is the scaled fraction of companies in a given grid cell provided in the 3-year sequences. In the second specification, two features are used, including the spatial lag of the first variable. A specific number of neurons in each layer, dropout ratios, learning rate annealing parameter and an optimizer algorithm were tuned with the hyperparametric search.

Figure 9. The general structure of the neural network



Source: Own work in diagrams.net software.

Using the yearly data from 2010-2018 six sequences for each grid cell have been created (with T spanning from 2012 to 2017). The sequences for the T=2017 were set aside for the final testing of the model (with the T+1=2018 period for prediction). Having the remaining part of the sample a decision was made to create a 70%/15%/15% split of observations. Sequences were divided randomly into training, testing and validation samples, yet the split was made based on the grid ID to avoid data leakage. By doing so it has been ensured that if a given grid cell was chosen for testing, every sequence associated with it will be also included in the testing dataset. Finally, the final samples were created: the training dataset with 2100 observations (from 420 grid cells), the testing dataset with 455 observations (91 cells) and the validation dataset with 450 observations (90 cells).

For both specifications (one- or two-variable model) the optimal parameters were searched within the ranges shown in Table 2. Due to numerous possibilities and computational restrictions a hyperparametric search was done with a 5% sample of all combinations. Models were trained on 50 epochs with a batch of size 50. They were evaluated using the binary cross-entropy measure as a loss function, which is a metric based on the Kullback-Leibler information theory commonly utilized for optimising any binary classification problems (Ramos et al., 2018). In the simplest terms, it can be interpreted as a distance between the distribution of the true event and the estimated probability distribution for the empirical data – the lower the cross-entropy is, the better is the model at explaining given phenomenon. Additionally, the accuracy measure was followed for a better understanding of the explanatory power of a model. Choosing a model with the minimal loss value on a validation sample, the following sets of parameters were obtained: {N1=128, N2=64, N3=16, d1=0.2, d2=0.3, d3=0.3, lr=0.1, “rmsprop”} for the model with one variable, and {N1=32, N2=16, N3=16, d1=0.2, d2=0.2, d3=0.3, lr=0.05, “rmsprop”} for the model with two variables. Results of the final models are presented in Table 3.

Table 2. Parameters tested in the hyperparametric search

Parameters:	***Options:***
Nodes in layer one (N1)	{128, 64, 32}
Nodes in layer two (N2)	{64, 32, 16}
Nodes in layer three (N3)	{32, 16, 8}
Dropout ratios for RNN layers (d’s)	{0.2, 0.3, 0.4}
Learning rate annealing (lr)	{0.1, 0.05}
Optimizer	{“rmsprop”, “adam”}

Source: Own work.

Table 3. Results of the final models

Measure	***Model one-variable***	***Model two-variables***
Loss training	0.3351	0.2519
Loss validation	0.2976	0.2298
Loss test	0.2981	0.2358

Loss hold-out T=2017	0.1294	0.0965
Accuracy training	0.8690	0.8967
Accuracy validation	0.8711	0.8844
Accuracy test	0.8615	0.8857
Accuracy hold-out T=2017	0.9733	0.9750

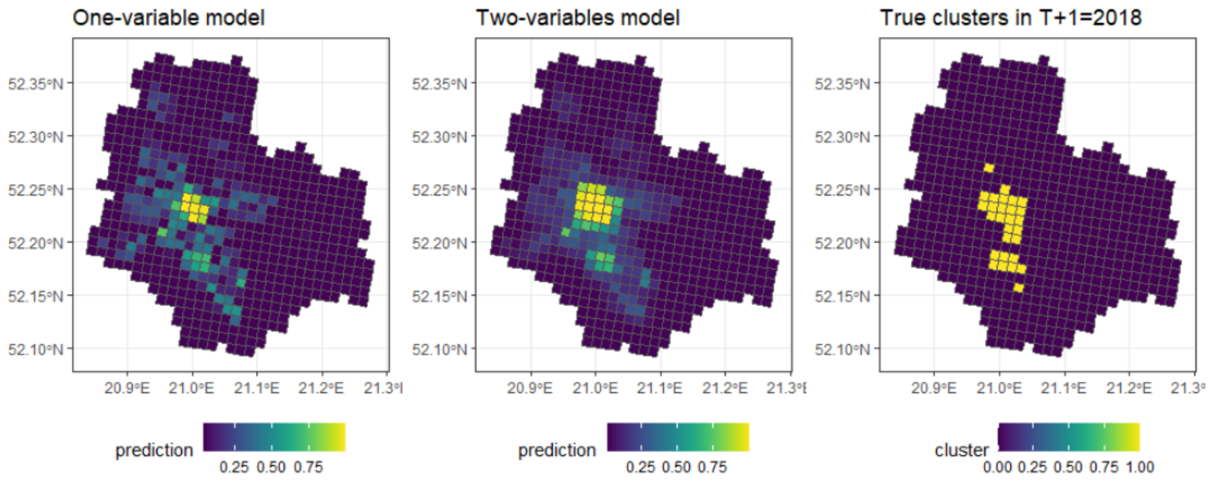
Source: Own work.

4.4.2. Results and predictions from RNN

With the usage of neural networks, two models were built which are successful in predicting the location of a future startup cluster (accuracy on the test data of 0.86 and 0.88 respectively). A model with two variables, which incorporates the spatial dimension of the data, has better results for the loss function (assessed with binary cross-entropy), as well as for the accuracy. Its results are also more stable across different samples – we get similar accuracy scores for training, validation and testing datasets. It is important to note that those better results are achieved with the much simpler specification of a model (considerably fewer neurons needed than in the case of the one-variable model). Augmenting the a-spatial RNN model with spatial features allows for getting more robust results for spatial panel data.

Considering the results for the hold-out sample of T=2017 we can see that the models perform much better than in the testing scenario (exceptionally low scores of the cross-entropy measure and the accuracy with values as high as 0.973 and 0.975). Such good results are probably because temporal patterns present in this sample at the grid cell level have already been learnt by the algorithm in the previous stages. In this case, only one time step has been added – relations from T=2017. Looking only at the numerical results (see Tab.3), it seems that this one-step-ahead forecast works extremely well for both specifications, with a slight advantage to the model with two variables (lower cross-entropy and a bit higher accuracy scores).

Figure 10. Cluster prediction probabilities and the presence of true clusters in T+1=2018 – models tested for the holdout sample of 2017 on Warsaw 1kmx1km grid



Source: Own work in R.

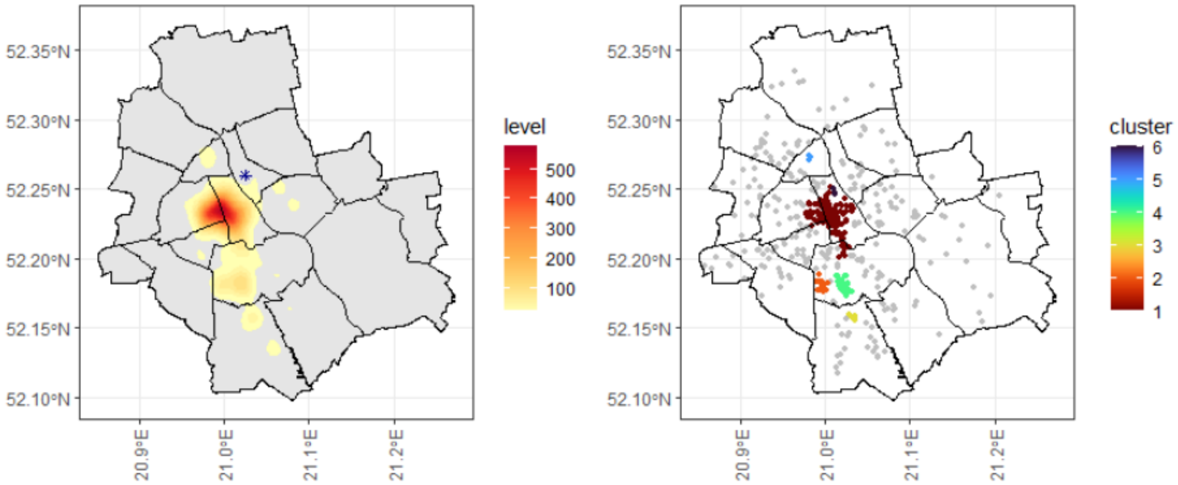
Visually considering the spatial distribution of predictions from both models we can easily observe that the two-variables model has much more stable results in terms of space (see Fig. 10). Adding the spatial lag as another feature to the model allowed it to navigate the spatial correlation between scores much better than the a-spatial model with one variable. The resulting probabilities are much smoother and more accurately correspond to the true cluster localisations in 2018. We can also see much less noise in the predictions (compared to the scattered pattern presented by the one-variable model). The model with two variables can account for the spatial dependencies in the sample and can produce much more robust results, identifying preferred city areas for the startup clusters creation.

Using the model with two variables we can predict that in 2018 we will have two main areas which will draw most of the newly founded startups (see Fig. 10B). There is a huge chance that a large cluster will be present in the city centre (bright yellow area), but we may expect some groupings in the southern direction from the city centre as well (Mokotów region – light green squares). We can expect that even though companies will probably locate in the other city areas, they will be rather distant from one another, and no other innovation cluster will be created (dark blue colours in most parts of the city). A larger number of startups is expected to be located on the left-hand side of the Vistula river (lighter colours in the western city area). However, the area in Praga Północ (east of the city centre) may be gaining more popularity (lighter blue squares right to the bright yellow hotspot). Probably, this notion can be strengthened by some targeted policies, which will increase the attractiveness of this urban area.

Those predictions can be easily verified. Following the empirical data from 2018 (Fig.11), we can see that most of the information drawn from the model prediction is true.

Indeed, in the 2018 sample, there was a large cluster created in the city centre, followed by another bigger grouping on the south from that cluster (Fig.11B). The rest of the companies were located loosely across the city space, with a preference for the left-hand side of the Vistula river (Fig.11B). While we can see an increasing number of companies located in the Praga Północ region (growth in concentration east of the city centre, city district marked with a dark blue asterisk in Fig.11A) there are still no startup clusters there (Fig.11B).

Figure 11. Spatial organisation of the startups founded in 2018



Details: Figure A shows the results of kernel density estimation for the 2018 sample. Figure B shows DBSCAN results for the 2018 sample with parameters: eps=0.006 (approx. 408m) and minPts=10. Grey points represent startups that were not allocated to any cluster. Dark blue asterisk is marking the Praga Północ region on Fig.11A.

Source: Own work in R.

The neural network model presented in this paper works very well for predicting the localisation of startups' clusters in Warsaw, Poland. It can also provide valuable information about the general tendencies regarding the popularity of certain city areas at a very detailed scale of 1km x 1km grid cell. Being fed with the current data, the model can be a very valuable tool for policy creation. However, those results are not limited to this one particular city. Following this successful model structure, we may easily recalibrate it to make predictions for another metropolis. As was shown in the paper, even short panel data can be used in this approach. Here, having the data about addresses of newly founded startups across nine years, was enough to build a model which predicts innovation clusters with above 95% accuracy.

Additionally to their methodological value, the results from this part are also contributing to the main thesis of the paper. The persistence of the concentration trend of the technological startups is allowing the model to get accurate results even with a considerably simple modelling structure. While across the years there were changes in the cluster sizes and their exact locations, it turns out that the 3-years sequence of area's attractivity suffices to decide whether in the upcoming year a new cluster will be founded there. Moreover, the effect of including the spatial feature is so successful thanks to the co-localising trend. When startups locate densely in one location (to utilize the small scale agglomeration externalities), the density of startups will probably be high in the direct neighbouring areas as well (as the externalities spread across the space). The co-localising tendency, which was discovered in the previous sections, is improving the stability of the model's results, showing that the process of cluster creation is highly dependent on the micro-geographical trends appearing in the neighbouring area. The accurate prediction of future hot spots of technological startup activity is possible because these companies are benefiting from the highly localised externalities, which require them to locate near the areas which were already proven attractive for startups in the previous years.

5. Conclusions

Technological startups are creating intriguing localisation patterns at the intra-urban level. Tracking their localisation choices at the micro-geographical scale allows one to see that the innovative business activity is not evenly distributed within the urban space. This paper shows that technological startups tend to co-locate and create dense clusters of business activity in the urban space. Such a pattern is consistent with the highly localised agglomeration externalities operating at fine scales within cities.

What was observed from the macro scale – the towards-city switch of the startup location – turns out to be just the starting point in describing the actual patterns of innovative business location as observed from the micro perspective. The intra-urban localisation patterns of technological startups suggest that different parts of the city are valued differently by the entrepreneurs. Companies are not just drawn to the metropolitan area, but rather to the dense business clusters localised within it. Size of those groupings and the density of business activity within them seem to be aligned with the agglomeration effects operating at small-scales within cities.

This paper is also contributing to the literature from a methodological perspective. It shows how machine learning tools can be utilized in spatial research, developing the field of spatial data science. The most significant input from this perspective is the example of how neural networks can be helpful in regional science, making it possible to predict spatial patterns in the data. Having only a few historical data recordings for given locations we can predict the future occurrences of a given spatial process with the usage of RNN. This type of model can be easily improved with the usage of a spatial feature, which holds the spatial lags of the primary input variable. Such a simple operation allows for the utilisation of a non-spatial model in the spatial context with low computational cost. The model shown in this paper can be easily utilized for future data points, helping the city authorities with targeting their future startup support agenda. Even though there is still much to be uncovered about the dynamics of the localisation process of those companies, we can build machine learning tools that can be helpful for decision support.

There is still a big gap in the literature regarding the impact of intra-urban localisation decisions on the lifecycle of technological startups. Following the survival rates dependent on the localisation and exploring the impact of cluster dynamics on the innovative activity are only the first ideas that may be followed in future research. The role of this paper was to shed the first light on the intra-urban organisation of technological startups and to tie it with the growing literature on the attenuation of agglomeration externalities. Patterns uncovered in this paper are opening the way for the new micro-geographical research on the startup location to come.

Literature:

Andersson, M., Klaesson, J., & Larsson, J. P. (2016). How Local are Spatial Density Externalities? Neighbourhood Effects in Agglomeration Economies. *Regional Studies*, *50*(6), 1082–1095. https://doi.org/10.1080/00343404.2014.968119

Arauzo-Carod, J. M. (2021). Location determinants of high-tech firms: an intra-urban approach. *Industry and Innovation*, *28*(10), 1225–1248. https://doi.org/10.1080/13662716.2021.1929868

Banal-Estañol, A., Macho-Stadler, I., Nieto-Postigo, J., & Pérez-Castrillo, D. (2019). *Early Individual Stakeholders, First Venture Capital Investment, and Exit in the UK Startup Ecosystem*. GSE, Graduate School of Economics.

Beck, M. W. (2018). NeuralNetTools: Visualization and Analysis Tools for Neural Networks. *Journal of Statistical Software*, *85*(11), 1. https://doi.org/10.18637/JSS.V085.I11

Bełcik, A. (2021, September 15). *Jest pierwszy polski jednorożec - Puls Biznesu - pb.pl*. Puls Biznesu. https://www.pb.pl/jest-pierwszy-polski-jednorozec-1127554

Boschma, R. (2005). Proximity and Innovation: A Critical Assessment. *Regional Studies*, *39*(1), 61–74.

Davis, C. A., & Fonseca, F. T. (2007). Assessing the certainty of locations produced by an address geocoding system. *GeoInformatica*, *11*(1), 103–129.

de Groot, H. L. F., Poot, J., & Smit, M. J. (2009). Agglomeration externalities, innovation and regional growth: Theoretical perspectives and meta-analysis. *Handbook of Regional Growth and Development Theories*, 256–281. https://doi.org/10.4337/9781848445987.00022

Devereux, M. P., Griffith, R., & Simpson, H. (2007). Firm location decisions, regional grants and agglomeration externalities. *Journal of Public Economics*, *91*(3–4), 413–435. https://doi.org/10.1016/J.JPUBECO.2006.12.002

Duranton, G., & Puga, D. (2004). Micro-foundations of urban agglomeration economies. In *Handbook of regional and urban economics* (pp. 2063–2117). Elsevier.

Duvivier, C., & Polèse, M. (2018). The great urban techno shift: Are central neighbourhoods the next silicon valleys? Evidence from three Canadian metropolitan areas. *Papers in Regional Science*, *97*(4), 1083–1111. https://doi.org/10.1111/PIRS.12284

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD. ACM Press*, *96*(34), 226–231.

Fan, J., Li, Q., Hou, J., Feng, X., Karimian, H., & Lin, S. (2017). A spatiotemporal prediction framework for air pollution based on deep RNN. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *4*(15). https://doi.org/10.5194/isprs-annals-IV-4-W2-15-2017

Ferretti, M., Guerini, M., Panetti, E., & Parmentola, A. (2022). The partner next door? The effect of micro-geographical proximity on intra-cluster inter-organizational relationships. *Technovation*, *111*, 102390. https://doi.org/10.1016/J.TECHNOVATION.2021.102390

Florida, R., & Mellander, C. (2017). Rise of the Startup City: The Changing Geography of the Venture Capital Financed Innovation. *California Management Review*, *59*(1), 14–38. https://doi.org/10.1177/0008125616683952

Gal, Y., & Ghahramani, Z. (2016). A theoretically grounded application of dropout in recurrent neural networks. *Advances in Neural Information Processing Systems*, 1027–1035.

Geibel, R. C., & Manickam, M. (2016). Comparison of selected startup ecosystems in Germany and in the USA. Explorative analysis of the startup environments. *GSTF Journal on Busienss Review (GBR)*, *4*(3). https://doi.org/10.5176/2010-4804_4.3.387

Goldberg, D. W., & Wilson, J. P. (2007). From text to geographic coordinates: the current state of geocoding. *URISA Journal*, *19*(1), 33–46.

Gu, Q., Lu, N., & Liu, L. (2019). A novel recurrent neural network algorithm with long short-term memory model for futures trading. *Journal of Intelligent & Fuzzy Systems*, *37*(4), 4477–4484. https://doi.org/10.3233/JIFS-179280

Guzman, J., & Stern, S. (2016). Nowcasting and Placecasting Entrepreneurial Quality and Performance. In *Measuring Entrepreneurial Businesses: Current Knowledge and Challenges* (pp. 63–109). University of Chicago Press.

Hart, T., & Zandbergen, P. (2014). Kernel density estimation and hotspot mapping: Examining the influence of interpolation method, grid cell size, and bandwidth on crime forecasting. *Policing*, *37*(2), 305–323.

Hu, Y., Wang, F., Guin, C., & Zhu, H. (2018). A spatio-temporal kernel density estimation framework for predictive crime hotspot mapping and evaluation. *Applied Geography*, *99*, 89–97.

Huynh, D. T. (2014). The effects of clustering on office rents: Empirical evidence from the rental office market in Ho Chi Minh City. *Theoretical and Empirical Researches in Urban Management*, *9*(1), 5–26.

Isaksen, A. (2004). Knowledge-based Clusters and Urban Location: The Clustering of Software Consultancy in Oslo. *Urban Studies*, *41*(5–6), 1157–1174. https://doi.org/10.1080/00420980410001675797

Jang, S., Kim, J., & von Zedtwitz, M. (2017). The importance of spatial agglomeration in product innovation: A microgeography perspective. *Journal of Business Research*, *78*, 143–154. https://doi.org/10.1016/J.JBUSRES.2017.05.017

Jennen, M. G. J., & Brounen, D. (2009). The Effect of Clustering on Office Rents: Evidence from the Amsterdam Market. *Real Estate Economics*, *37*(2), 185–208.

Kogler, D. F. (2015). Evolutionary Economic Geography – Theoretical and Empirical Progress. *Regional Studies*, *49*(5), 705–711.

Lai, W., Zhou, M., Hu, F., Bian, K., & Song, Q. (2019). A New DBSCAN Parameters Determination Method Based on Improved MVO. *IEEE Access*, *7*, 104085–104095. https://doi.org/10.1109/ACCESS.2019.2931334

Lin, Y. P., Chu, H. J., Wu, C. F., Chang, T. K., & Chen, C. Y. (2011). Hotspot analysis of spatial environmental pollutants using kernel density estimation and geostatistical techniques. *International Journal of Environmental Research and Public Health*, *8*(1), 75–88.

MapQuest. (2018). *Geocoding API - Overview | MapQuest API Documentation*. https://developer.mapquest.com/documentation/geocoding-api/

Medsker, L., & Jain, L. C. (1999). *Recurrent neural networks: design and applications*. CRC press.

Morse, S. P. (2008). *Computing Distances between Latitudes/Longitudes in One Step*. https://stevemorse.org/nearest/distance.php

Nauman, B. A., & Edison, H. (2010). Towards innovation measurement in software industry. *Unpublishing Masters Thesis. School of Computing at Blekinge Institute of Technology in Sweden*.

Neffke, F., Henning, M., Boschma, R., Lundquist, K. J., & Olander, L. O. (2010). The Dynamics of

Agglomeration Externalities along the Life Cycle of Industries. *Regional Studies*, *45*(1), 49–65. https://doi.org/10.1080/00343401003596307

OpenStreetMap. (2022). *OpenStreetMap*. https://www.openstreetmap.org/about

Paternoster, N., Giardino, C., Unterkalmsteiner, M., Gorschek, T., & Abrahamsson, P. (2014). Software development in startup companies: A systematic mapping study. *Information and Software Technology*, *56*(10), 1200–1218. https://doi.org/10.1016/J.INFSOF.2014.04.014

Pisoni, A., & Onetti, A. (2018). When startups exit: comparing strategies in Europe and the USA. *Journal of Business Strategy*.

Portal Geostatystyczny. (2021). *INSPIRE - Portal Geostatystyczny*. https://geo.stat.gov.pl/inspire

Rammer, C., Kinne, J., & Blind, K. (2019). Knowledge proximity and firm innovation: A microgeographic analysis for Berlin: *Urban Studies*, *57*(5), 996–1014. https://doi.org/10.1177/0042098018820241

Ramos, D., Franco-Pedroso, J., Lozano-Diez, A., & Gonzalez-Rodriguez, J. (2018). Deconstructing Cross-Entropy for Probabilistic Binary Classifiers. *Entropy 2018, Vol. 20, Page 208*, *20*(3), 208. https://doi.org/10.3390/E20030208

Reisdorfer-Leite, B., Marcos de Oliveira, M., Rudek, M., Szejka, A. L., & Canciglieri Junior, O. (2020). Startup Definition Proposal Using Product Lifecycle Management. In F. Nyffenegger, J. Ríos, L. Rivest, & A. Bouras (Eds.), *Product Lifecycle Management Enabling Smart X. PLM 2020. IFIP Advances in Information and Communication Technology* (Vol. 594, pp. 426–435). Springer, Cham. https://doi.org/10.1007/978-3-030-62807-9_34

Sawant, K. (2014). Adaptive methods for determining DBSCAN parameters. *Journal of Innovative &, Engineering*, *1*(4), 329–334.

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems*, *42*(3).

Silverman, B. W. (1981). Using Kernel Density Estimates to Investigate Multimodality. *Journal of the Royal Statistical Society: Series B (Methodological)*, *43*(1), 97–99.

Suwarni, R. N., Fahlevi, M., & Abdi, M. N. (2020). Startup valuation by venture capitalists: An empirical study Indonesia firms. *International Journal of Control and Automation*, *13*(2), 785–796.

Urząd Statystyczny w Warszawie. (2021). *Urząd Statystyczny w Warszawie / Dane o województwie / Stolica województwa / Ludnosc*. Ludność. https://warszawa.stat.gov.pl/dane-o-wojewodztwie/stolica-wojewodztwa/ludnosc/

van Oort, F. G., & Atzema, O. A. L. C. (2004). On the conceptualization of agglomeration economies: The case of new firm formation in the Dutch ICT sector. *The Annals of Regional Science 2004 38:2*, *38*(2), 263–290. https://doi.org/10.1007/S00168-004-0195-8

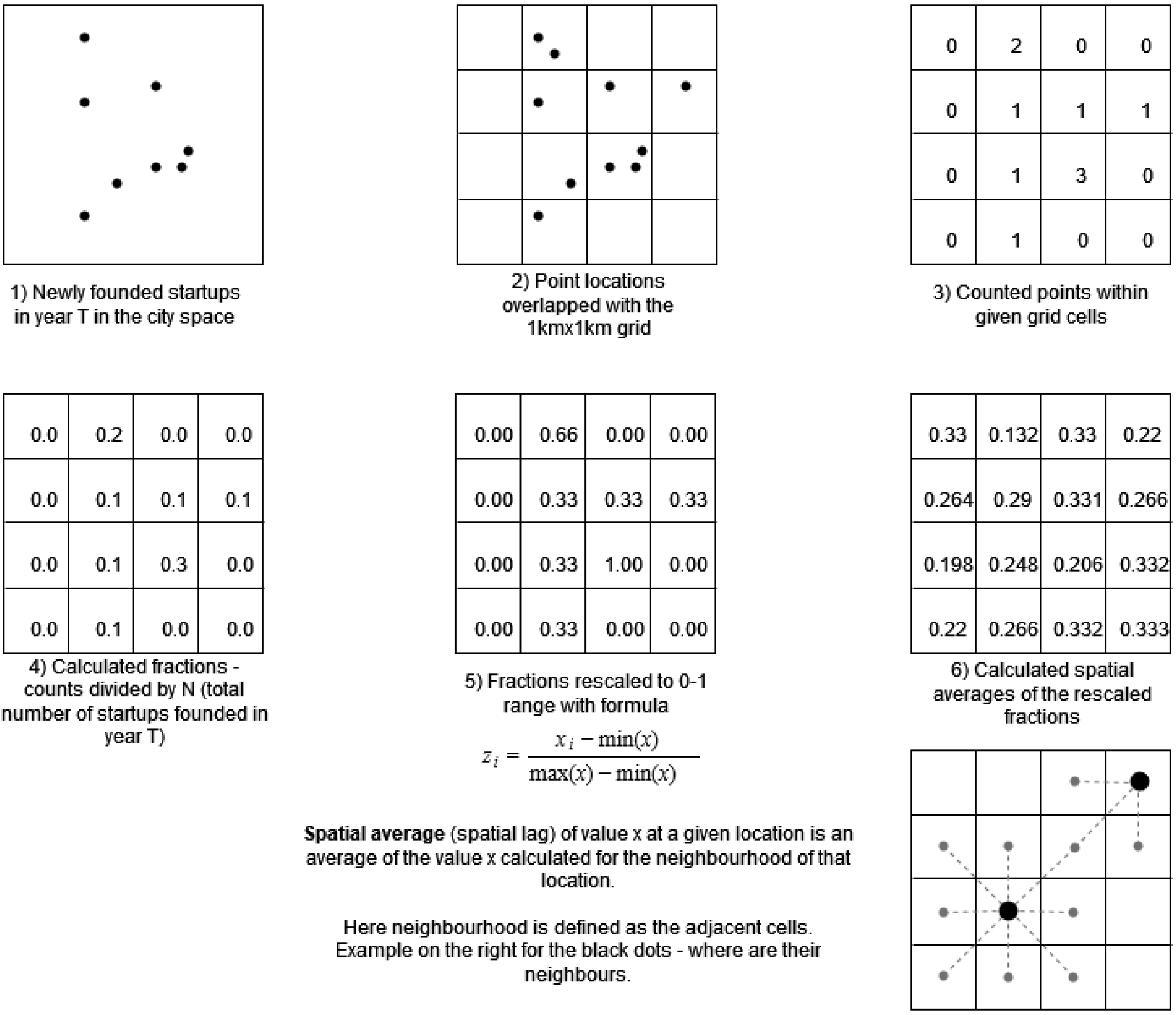
Wand, M. P., & Jones, M. C. (1994). *Kernel smoothing*. CRC Press.

Zhang, J., & Man, K. F. (1998). Time series prediction using RNN in multi-dimension embedding phase space. *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, *2*, 1868–1873. https://doi.org/10.1109/ICSMC.1998.728168

Appendix

Input data preparation procedure, utilized for the model in part 4.4, is presented in the Appendix Fig.1. The steps are here visualized for a simplified example of squared “city space” for a given year T, where 10 startups were founded (N = 10). The point pattern of startups founded in a year T (step 1) is overlapped with the 1km x 1km census grid (step 2). Then the counts of companies within a given grid cell are counted (step 3). In the next step, each count is divided by N, which is the number of startups founded in a given year T (step 4). Fractions calculated in the previous step are rescaled to the 0-1 range to keep the consistent scale of the input, which is required by the neural networks (step 5). Rescaling is done for each yearly sample separately. Then, for the second model structure, the spatial averages of the rescaled fractions are calculated (step 6). The neighbourhood is here defined as the full set of adjacent cells (following the structure of the queen’s contiguity matrix).

Appendix Figure 1. Visualization of the input data preparation for the neural network model (part 4.4)



1) Newly founded startups in year T in the city space

2) Point locations overlapped with the 1kmx1km grid

0	2	0	0
0	1	1	1
0	1	3	0
0	1	0	0

3) Counted points within given grid cells

0.0	0.2	0.0	0.0
0.0	0.1	0.1	0.1
0.0	0.1	0.3	0.0
0.0	0.1	0.0	0.0

4) Calculated fractions - counts divided by N (total number of startups founded in year T)

0.00	0.66	0.00	0.00
0.00	0.33	0.33	0.33
0.00	0.33	1.00	0.00
0.00	0.33	0.00	0.00

5) Fractions rescaled to 0-1 range with formula

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

0.33	0.132	0.33	0.22
0.264	0.29	0.331	0.266
0.198	0.248	0.206	0.332
0.22	0.266	0.332	0.333

6) Calculated spatial averages of the rescaled fractions

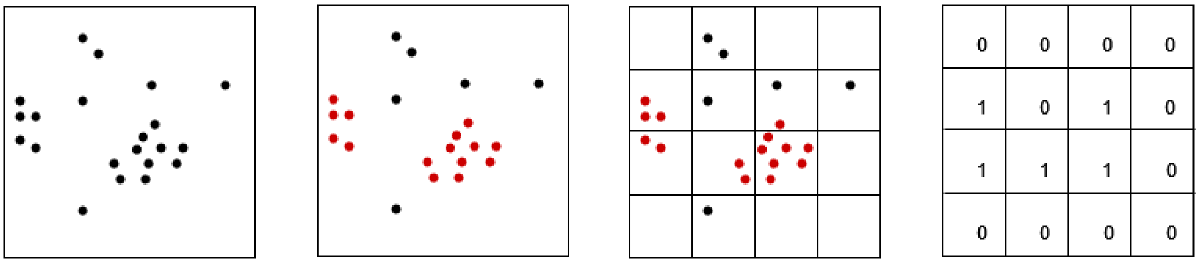
Spatial average (spatial lag) of value x at a given location is an average of the value x calculated for the neighbourhood of that location.

Here neighbourhood is defined as the adjacent cells. Example on the right for the black dots - where are their neighbours.

Source: Own work in diagrams.net software.

The way of preparing data for the output variable of the neural network model (labels of whether in a given cell a cluster is present or not) is presented in Appendix Fig.2. Here the results from part 4.3 are utilized: on the point data of startups founded in a year T (step 1) DBSCAN algorithm is run (step 2). Then the results from DBSCAN are overlapped with the 1km x 1km grid (step 3). In the last step, a binary variable is created, based on the DBSCAN results. If among the points which belong to a given grid cell, at least one was recognised as a part of a density-based cluster, the variable will take a value of 1. Otherwise, the binary variable will take a value of 0.

Appendix Figure 2. Visualization of the output data preparation for the neural network model (part 4.4)



0	0	0	0
1	0	1	0
1	1	1	0
0	0	0	0

1) Newly founded startups in year T in the city space

2) DBSCAN identification of clusters (red dots)

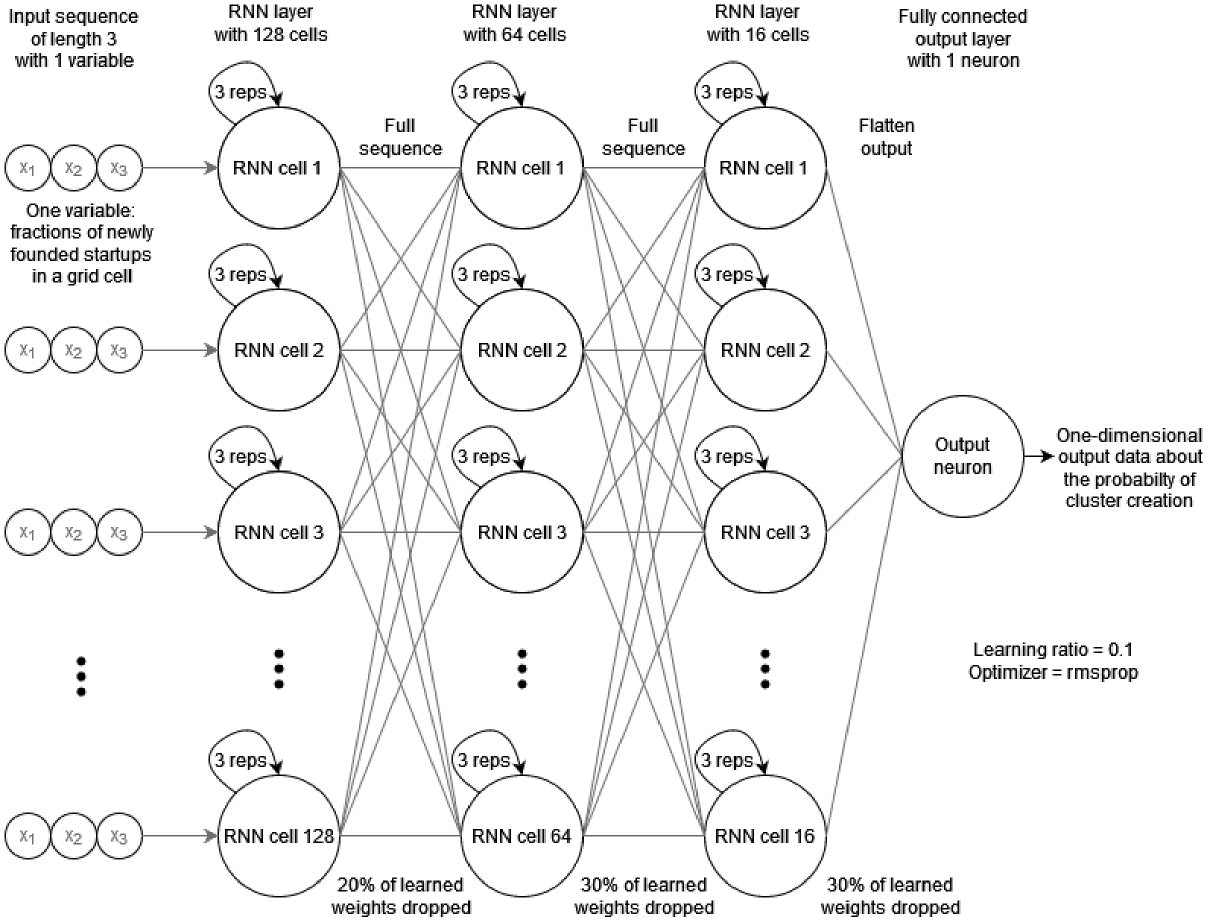
3) Point (and cluster) locations overlapped with the 1kmx1km grid

4) Cluster presence in a grid cell transformed into a binary variable - 1 for identified cluster, 0 otherwise

Source: Own work in diagrams.net software.

The detailed structure of the neural network models from part 4.4 is presented in Appendix Fig.3 and Appendix Fig.4. The number of neurons in each layer, dropout ratios, learning ratios and optimisers were chosen based on the hyperparametric search. To get the best-tuned specification a general model structure was run with different combinations of parameters. After each run, the loss value (binary cross-entropy) was calculated on the validation sample and saved in the output file. Having a full set of results from different models, the specification with the best score (minimum binary cross-entropy) was chosen.

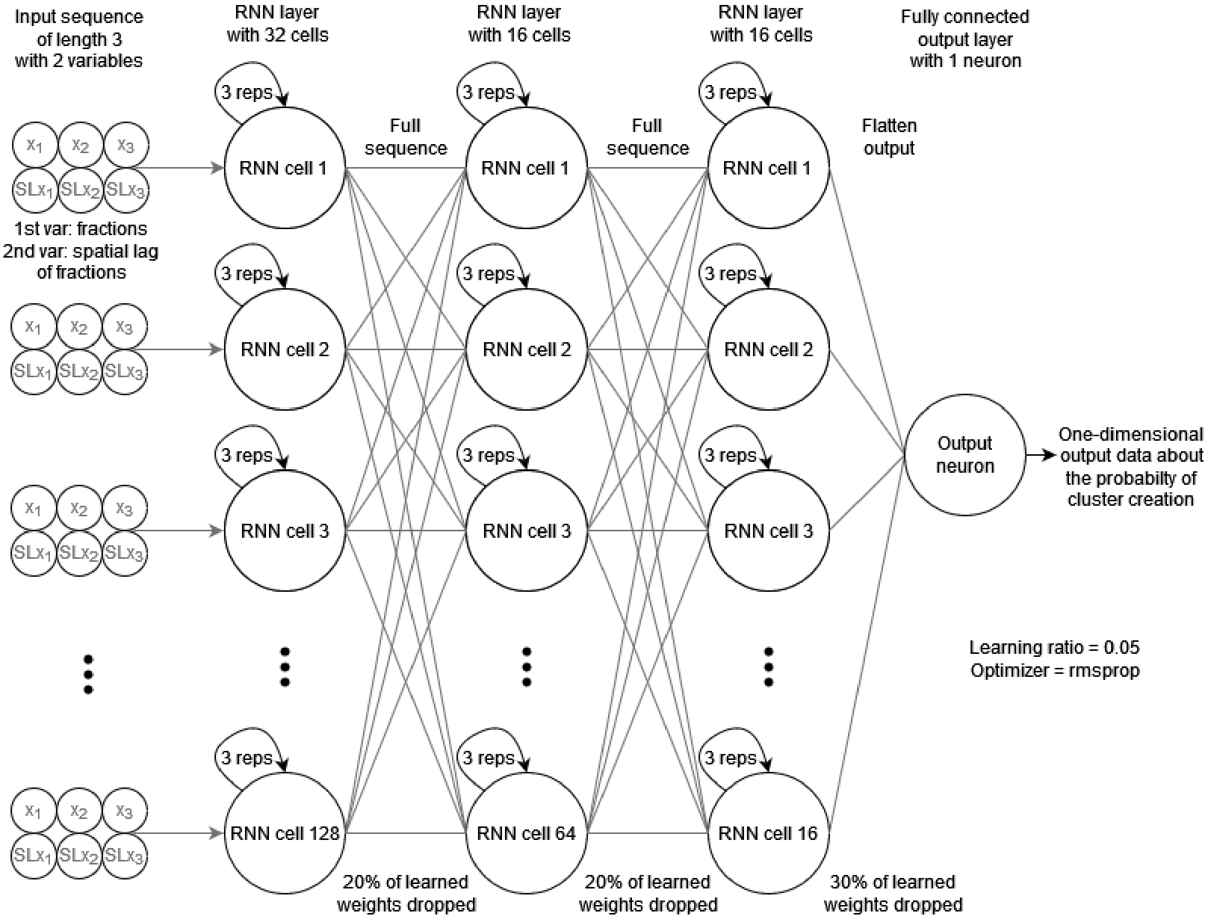
Appendix Figure 3. The final structure of the first model specification (part 4.4 – model with one variable)



Input sequence of length 3 with 1 variable
RNN layer with 128 cells
RNN layer with 64 cells
RNN layer with 16 cells
Fully connected output layer with 1 neuron
3 reps
Full sequence
Flatten output
x_1 x_2 x_3
One variable: fractions of newly founded startups in a grid cell
RNN cell 1
RNN cell 2
RNN cell 3
RNN cell 128
RNN cell 64
RNN cell 16
Output neuron
One-dimensional output data about the probabilty of cluster creation
Learning ratio = 0.1
Optimizer = rmsprop
20% of learned weights dropped
30% of learned weights dropped
30% of learned weights dropped

Source: Own work in diagrams.net software.

Appendix Figure 4. The final structure of the second model specification (part 4.4 – model with two variables)



Source: Own work in diagrams.net software.