Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2017) 000–000

Procedia
Computer Science

www.elsevier.com/locate/procedia

Areas of Habitation in the City: Spatial Analysis of St.Petersburg Based on Social Media Data

Artem Konyukhov[a], Aleksandra Nenko[a], Marina Petrova[a]

[a]*ITMO University, 199034, 14 Birzhevaya line., St Petersburg, Russia*

Abstract

In this paper, we present a study on areas of habitation in St.Petersburg, Russia, which are considered as parts of the city used by dwellers on an everyday basis for performing routine practices and different from the official administrative division. Research methodology is based on defining clusters with social and spatial proximity using geo-referenced data generated by citizens in social media during their everyday life and is a refinement of the "Livehoods Project" approach (2012). The dataset for St.Petersburg is retrieved from VKontakte social network and comprises 6128 venues with 763079 check-ins collected for December 2017 - February 2018 time period. The number of clusters received is 155, they differ from the actual administrative division of St.Petersburg. Inter-cluster and intra-cluster similarity indexes reflecting congeniality of users are calculated, functional load for clusters is defined based on Google Places data. Interpretation of a sample of areas of habitation is given from a point of view of environmental features.

Peer-review under responsibility of the scientific committee of the 6th International Young Scientist conference in HPC and Simulation.

Keywords: areas of habitation; social media; location-based data; mixed methods; urban planning; livehoods

1. Introduction

Starting from Chicago school one of the continuous queries of urban researchers is detecting and defining urban communities and milieus where they concentrate and develop certain lifestyles. People tend to form spatial communities or "areas of habitation" where human activity concentrates [1]. Administrative division of the city does not always reflect such "organic" clusters of urban life. However mapping them allows developing urban polycentricity, to check infrastructural sufficiency and to support the vibrancy of urban life.

Areas of habitation are considered here as parts of the city which are used by dwellers on an everyday basis for performing routine practices. Well developed areas of habitation coincide with the high level of quality of life in the city [2]. In these areas, residents can satisfy their daily needs with minimal time and financial costs. They also form communities of practice united by a similar style of life connected with shared urban venues. In this paper, we present a study on the areas of habitation in St.Petersburg, Russia, based on the analysis of user-generated check-in data retrieved from a local social network. Research questions of the paper are formulated as follows: what are the areas of habitation in St.Petersburg? Do they coincide with the municipal division of the city? What are the functional characteristics of these areas? What are the environmental factors which form these areas?

Peer-review under responsibility of the scientific committee of the 6th International Young Scientist conference in HPC and Simulation.

2. Methodology

2.1. Methodological Approach and Dataset

Our methodological approach is based on the idea that areas of habitation can be detected based on geo-referenced data from social media generated by residents of the city during their everyday life [3]. This data reveals proximity ties (spatial cohesiveness) between urban places used by one user. The data source used is the most popular social network in Russian speaking countries VKontakte (VK), data was retrieved via API. The dataset contains around 80000 venues parsed for 2010-2018 years. After geocoding and deleting places located outside St. Petersburg the dataset shortened to 6128 places. For every place check-ins were collected, a total of 763079 from 128406 users for December 2017 - February 2018 time period.

2.2. Initial computational algorithm

The computational algorithm which was initially applied is based on the "Livehoods Project" conducted by Cranshaw et. al. 2012 [4]. The authors have proposed a spectral clustering model on a city-scale based on a dataset of 42787 check-ins of 3840 users at 5349 venues collected in Pittsburgh, PA, US from a location-based online social network. The received clusters were called "livehoods" . To define clusters the researchers have introduced the social proximity index which accounts for the distance between the neighboring venues and also for the similarity of the users who have check-ined there. The index is calculated as a cosine similarity of the vectors representing the venues.

Suppose that V is a set of n_V VK venues and for each $i, j \in V$ we compute Euclidean distance $d(i, j)$. To compute geographical distance between the venues we transform lat/lng coordinates to UTMZone36V projection. We consider the set U of n_U VK users and the set C of these user's geolocated messages (check-ins) in the venues which make up the set V. Each venue $v \in V$ is represented as a "bag of check-ins" to v. Let u^{th} component of the vector c_V be the count of users' check-ins in v. For this matrix we compute a social similarity index $s(i, j)$ between each pair of venues $i, j \in V$.

$$s(i, j) = \frac{c_i \cdot c_j}{||c_i|| \cdot ||c_j||} \tag{1}$$

Based on the social similarity index values we compile an affinity matrix $A = (a_{i,j})_{i,j=1,\ldots,n_V}$. For a venue v the $N_m(v)$ is the m closest venues according to the $d(v, \cdot)$

$$a(i, j) = \begin{cases} s(i, j) + \alpha & \text{if } j \in N_m(i) \text{ or } i \in N_m(j) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where α is a small constant that blocks venues from having no connections to any others.

The received graph $G(A)$ (see Fig.1a) is subjected to spectral clustering algorithm presented in the "Livehoods project" .

```
Data: V, A, G(A), $k_{min}$, $k_{max}$, $\tau$
Result: Clusters $A_i, ..., Ak$
Compute the normalized Laplacian matrix $L_{norm}$.
Let $\lambda_1 \leq ... \leq \lambda_{k_{max}}$ be the $k_{max}$ smallest eigenvalues of $L_{norm}$.Set $k = arg\max_{i=k_{min},...,k_{max}-1} \Delta_i$ where $\Delta_i = \lambda_{i+1} - \lambda_i$.
Find the $k$ smallest eigenvectors $e_1, ..., e_k$ of $L_{norm}$.
Let E be an $n_V \times k$ matrix with $e_i$ as columns.
Let the $y_1, ..., y_{n_V}$ be the rows of $E$ and cluster them into $C_1, ..., C_k$ with k-means. This induces a clustering on
  $A_1, ..., A_k$ by $A_i = \{j|y_j \in C_i\}$.
For each $A_i$, let $G(A_i)$ be the subgraph of $G(A)$ induced by vertices $A_i$. Split $G(A_i)$ into connected components.
  Add each component as a newcluster, removing $G(A_i)$.
Let $b$ be the area of bounding box containing coordinates in $V$ , and $b_i$ be the area of the box containing $A_i$. If
  $b_i/b$ ¿ $\tau$, delete cluster $A_i$, and redistribute each $v \in A_i$ to the closest $A_j$ under single linkage distance $d(v, A_j)$.
```

Algorithm 1: “Levihoods project” spectral clustering alghorithm

For this paper $m = 10$, $\alpha = 0.001$, $k_{min} = 30$, $k_{max} = 45$, and $\tau = 0.4$.

2.3. Validation of the algorithm

Parallel to calculation we have conducted a sociological survey to validate the clusters received with the algorithm. The survey ran for a month from mid-March till mid-April 2018. The respondents were people who live or work in a particular cluster. The methodology used was semi-structured interviews, 39 interviews were collected. During the interviews respondents were asked to tell about their everyday life practices and venues they visit. They showed their everyday routes and places on a map and commented on regularity of visits. The interviewees also commented on the emotional perception of the area, places they like and dislike, attractors and barriers. The interviews have shown that the areas of habitation mapped by respondents do not correspond or correspond only partly to the ones defined by clustering algorithm. Sociological validation has hinted a number of factors the algorithm did not account for, such as: (a) barriers created by the built environment (industrial zones, motorways, railways); (b) natural barriers (Neva river, channels); (c) pedestrian accessibility.

Another proof that the initial algorithm is not working is the results of k-nearest neighbor algorithm application to similarity graph construction. Resulted graph $G(A)$ has only 493 ties with weight more than α out of 54798 ties. To test the results we have clustered the unweighted 10-nearest neighbour graph and have received almost similar results. This means that social similarity of venues was not grasped by the k-nearest neighbor algorithm for our dataset.

The map of clusters for St.Petersburg received with the original “Livehoods project” clusterization algorithm can be assessed online at spblivehoods.github.io.

2.4. Refinement of the algorithm

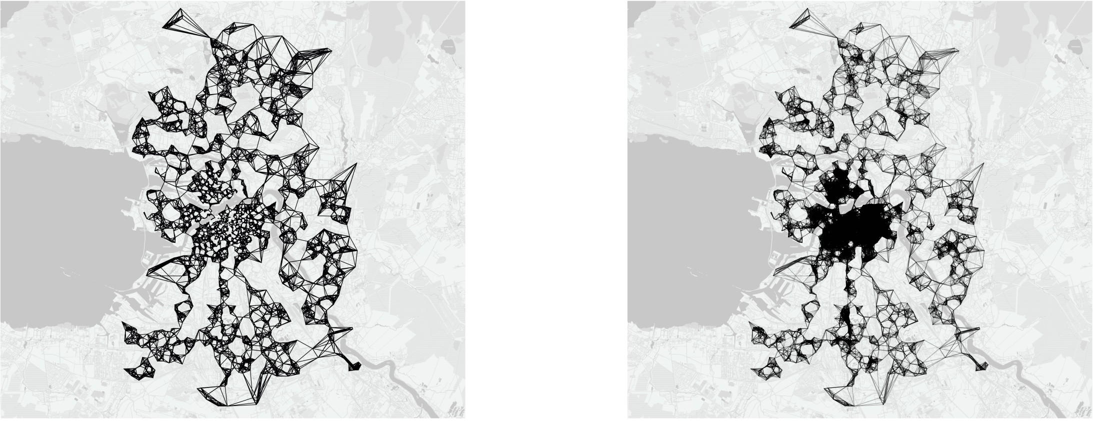
To receive more coherent results we refine the clusterization algorithm with an account for pedestrian accessibility. The social proximity index is calculated for a pair of neighboring objects, which are located at a 5 minutes distance from each other, which in spatial terms is defined as 500 m radius around each venue.

We compute a new affinity matrix $E = (e_{i,j})_{i,j=1,...,n_V}$ $n_V \times n_v$. For a given venue v we let $N(u, \varepsilon)$ be the set of venues which distance from u is smaller than ε.

$$e(i, j) = \begin{cases} s(i, j) + \alpha & \text{if } j \in N(i, \varepsilon) \text{ or } i \in N(j, \varepsilon) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

New similarity graph $G(E)$ was constructed by connecting each venue node by undirected edge with venues in ε meters radius. The weight of the edge between venues i, j is set according to $s(i, j)$.

The walkable distance between the venues in non-central areas of the city is often more than 500 m, so the received graph is unconnected, while spectral algorithm requires a graph consisting of one connected component. To receive the connected graph we define an affinity matrix $C = UNION(A, E)$. The resulting graph $G(C)$ (see Fig.1b) contains more than 10060 edges with weight greater than α.



(a) G(A)

(b) G(C)

Fig. 1: Representation of similarity graphs on St.Petersburg map.

Finally we have used the spectral clustering algorithm(Alg. 1) without the last two postprocessing steps.

Algorithm parameters: $\varepsilon = 500$, $m = 10$, $\alpha = 0.001$, $k_{min} = 30$, $k_{max} = 250$.

The upgrade of the algorithm has given a graph with 15496 ties with weight more than α out of 54798 ties, the number of clusters is 155. The results of the new clusterization can be assessed online in an interactive format at spblivehoods.github.io. Median check-in count is 28, the minimal number of check-ins in one place is 10, maximal is 3103.

2.5. *Cluster comparison*

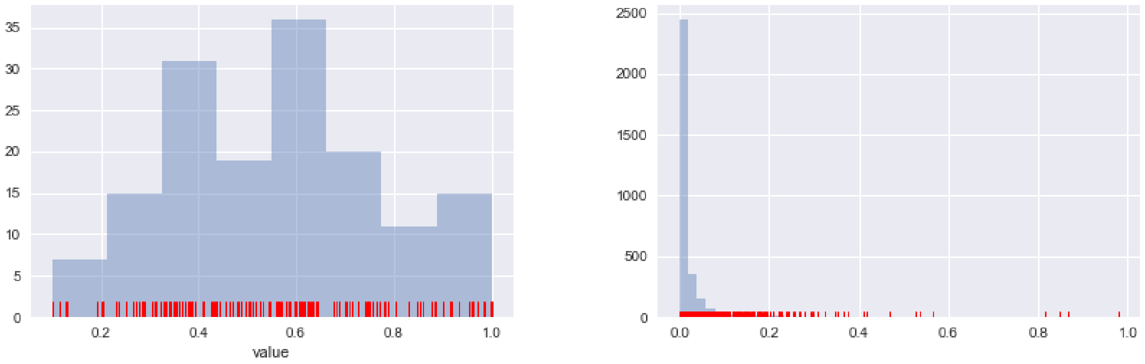
To compare clusters we compute two indexes - intra-cluster similarity index and inter-cluster similarity index.

Intra-cluster similarity index $S(G(A_n))$ indicates to what extent the venues in the cluster are similar in terms of the users who check-in in them.

$$S(G(A_n)) = \frac{1}{n} \sum_{i \in E(G(A_n)} d_G(A_n)(i), \text{ where } d_G(A_n)(i) = \frac{1}{n} \sum_{j \in E(G(A_n))} s(i, j) \tag{4}$$

Inter-cluster similarity index defines to what extent the clusters are similar to each other in terms of users who check-in in the venues included into the cluster, i.e. the measure is indicating existence of a strong connection between the clusters or that their users might form one community of practice. The inter-cluster similarity index is computed here the same way as the “similarity index” in the “Livehoods Project” [4, p. 3]: each cluster A_i is represented as a n_U dimensional vector c_{A_i}, where u^{th} component is a number of check-ins user u had to any venue in A_i. Then the cosine similarity between all pairs of clusters $s(A_i, A_j) = \frac{c_{A_i} \cdot c_{A_j}}{||c_{A_i}|| \cdot ||c_{A_j}||}$ is calculated.

Figure 2 shows the distribution of the venues by the indexes of intra- and inter-cluster similarity. The most frequent value of intra-cluster similarity is 0.6, which is quite high and means that users quite often check-in the same venues of the cluster (Fig.2a). However, there are few clusters which are similar to each other (Fig.2b).



(a) Intra-cluster similarity index

(b) Non-zero inter-cluster similarity index

Fig. 2: Distribution of intra- and inter-cluster similarity index.

3. Research results and discussion

The borders of the clusters received do not coincide with the municipal division of St.Petersburg: the central areas intersect several municipal districts, the non-central areas are much smaller than one municipal district (see Fig.3). The possible interpretation is that central areas are much more developed in terms of public life which forms clusters of its own, while non-central public life is underdeveloped and shrinks to the zones closer to the subway stations.

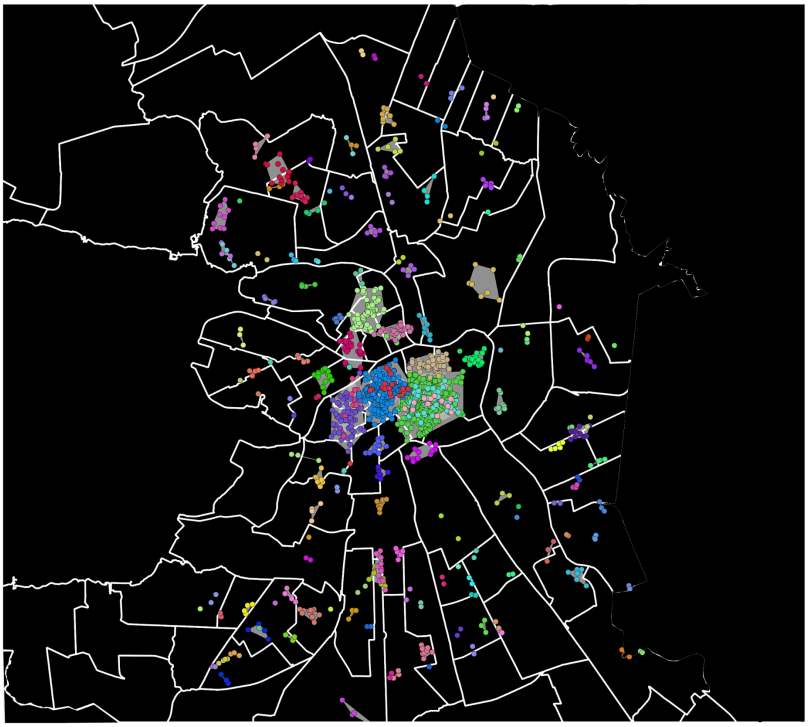


Fig. 3: Cluster borders VS. borders of St.Petersburg municipal areas.

We have compared a sample of 7 clusters located in the city center and non-central areas with the areas of habitation extracted during the sociological survey (see Fig.4). After the algorithm refinement, there is still no full coincidence of these areas: the squares of the areas of habitation are bigger than the defined clusters, especially in the non-central locations. This could be explained by the nature of the data used: check-ins reflect mostly public spaces, such as cafes, bars or museums, and less frequently everyday life places, such as supermarkets or drugstores which were detected by the survey.

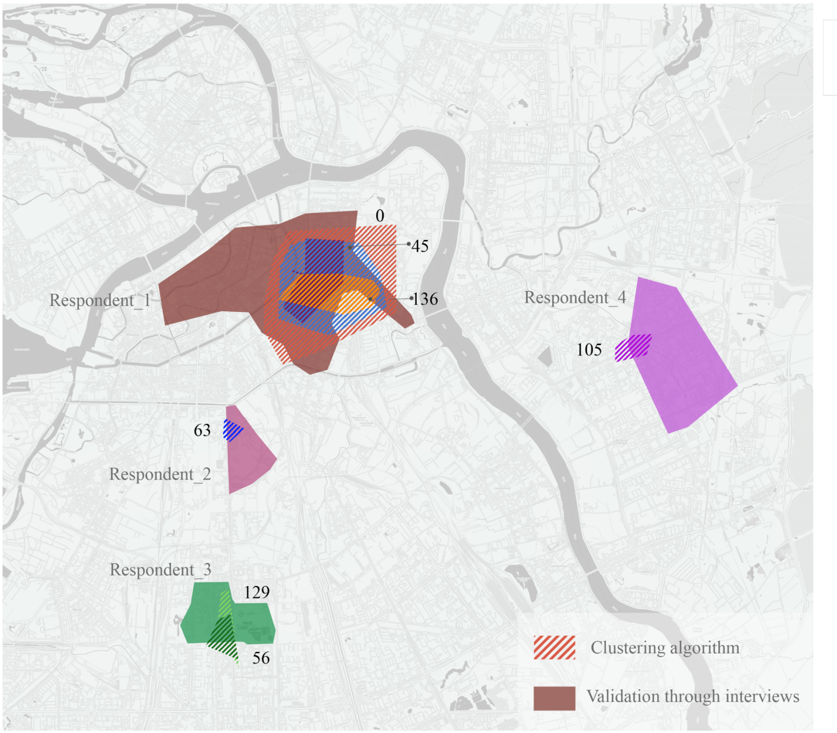


Fig. 4: Borders of the clusters and borders of the areas of habitation defined during the sociological survey.

We have analyzed the character of the sampled clusters based on the functional layout of the venues which they consist from and in connection to the city area they are located in. The functions of the venues were defined based on data from Google and were collected via Google Places API text search, which allows receiving the venue title by geolocation. For a more detailed account, we have chosen 3 clusters representing different areas of the city with a various typology of the built environment, namely a central city area, a former industrial territory, and a sleeping quarter.

The central St.Petersburg area is represented by clusters “0”, “45” and “136”, which intersect or are “nested” within each other (see Fig.5). “0” cluster is the largest one in the city center and intersects the borders of seven municipal districts. It is formed by Suvorovsky avenue and Kirochnaya street in the North-East, Aleksandra Nevskogo Square in the East, Obvodny channel embankment and Bagrationovskaya square in the South, and Gostiny Dvor subway station in the West. The cluster is centered on Vosstaniya Square (which is also a subway station). The cluster includes 625 venues, the least popular venue has 10 check-ins, the most popular one - 826 check-ins. The main functions here are food (157 venues), bars (66 venues), shopping (56 venues), art (46 venues) and points of interest (33 venues), entertainment and nightlife (53 venues) and others. This distribution indicates a concentration of leisure, tourism and cultural practices in the central area. Intra-cluster similarity index is quite low - 0.3 (as compared to the most frequent 0.6), the users of the cluster are slightly similar to the users of the nearby central clusters “45” and “136” (inter-cluster similarity is 0.2 and 0.1 respectively). Distribution of functions in the latter is similar to “0” cluster.

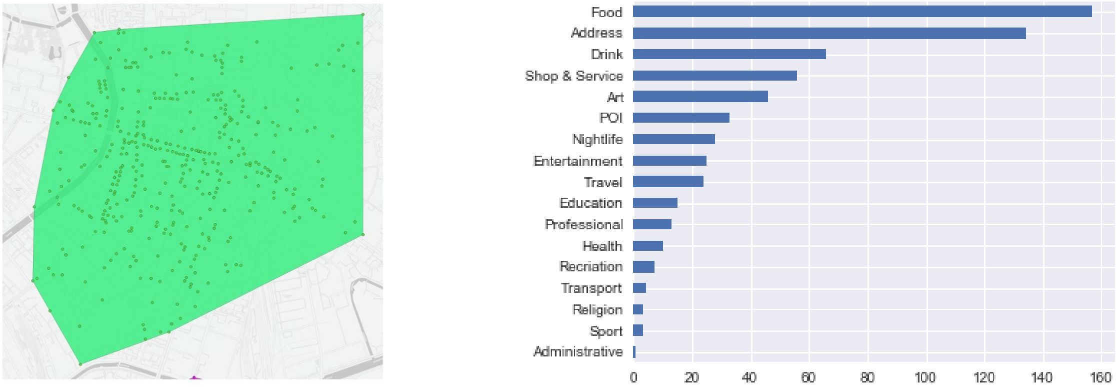


Fig. 5: “0” city center cluster borders and functional load.

St. Petersburg urban transformation is characterized by the redevelopment of the vast industrial territories which form the “grey belt” around the city center. In the late 90s with the collapse of the Soviet Union, the majority of the factories has stopped running. Nowadays the former industrial territories undergo the process of lengthy redevelopment mostly as residential areas with a small share of public functions. A typical cluster in this area “63” is located along Moskovsky avenue and is bordered by Frunzenskaya subway station in the North and by the intersection of Kiyevskaya street and Moskovsky avenue in the South. Compared to the central clusters, “63” has a very limited number of venues - 7, the least popular venue has 13 check-ins, the most popular one - 144 check-ins. Functional load of the venues is food, transport (subway station) and residential function (address). Intra-cluster similarity index is higher than in the city center - 0.51, the inter-cluster similarity with other clusters is low.

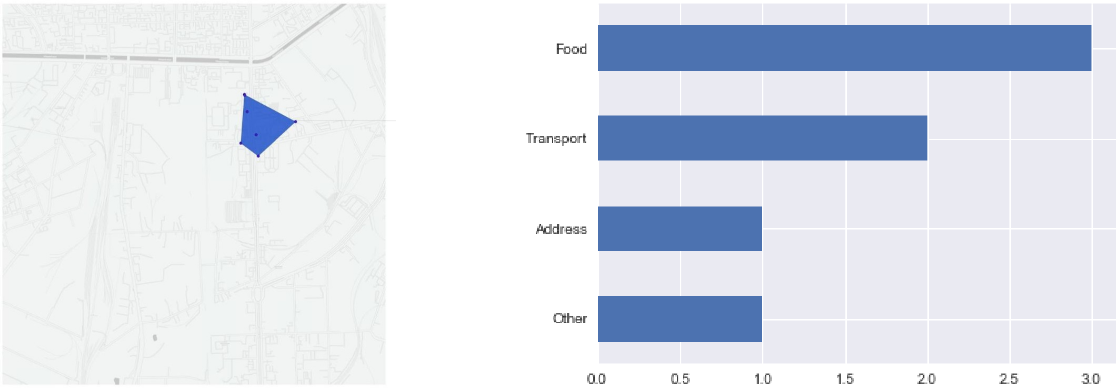


Fig. 6: “63” industrial area cluster borders and functional load.

St. Petersburg is a rapidly growing city with new residential complexes appearing at the city periphery as well as in the historical but remote areas. “139” cluster is located next to Ozerki subway station, where multi-storied residential buildings are being actively built. “139” has a bigger number of venues than the “63” - 12, the least popular venue has 10 check-ins, the most popular one - 83 check-ins. Activity here is also centered on “food” function represented by cafes with a small average check. However, there is a bar and a shop, unlike in “63” cluster. The intra-cluster similarity is similar to “63” cluster - 0.51, but there is a higher inter-cluster similarity - 0.03 with the cluster number “137” next to Sportivnaya subway station.

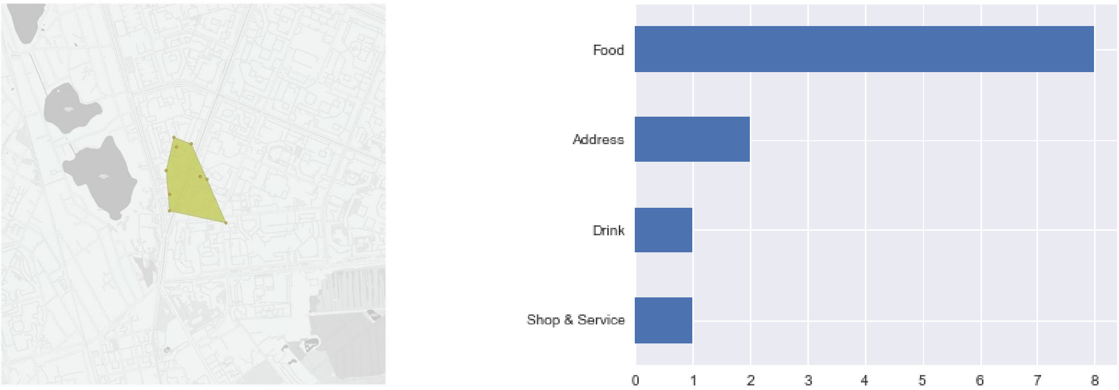


Fig. 7: “139” sleeping quarter cluster borders and functional load.

Clusters in the city center are bigger than the ones in the non-central areas. This can be explained by the fact that the city center has a bigger number of venues which are visited and check-ined by people. They also overlap with each other demonstrating different “lifestyles” formed by different social groups visiting central venues. Analysis of the functional load of the central clusters shows that they have a more distinguished leisure character. The non-central clusters are smaller and more dispersed in urban space, they tend to form closer to the subway stations. Their functional load shows a more residential character. The utmost function which is present in all of the clusters is food - cafes, restaurants, bistros form the core of the public life in each area of habitation of St.Petersburg. This might be interpreted both as a feature of the Northern city with priority on indoor leisure as well as the consequence of underdeveloped infrastructure for more variable public activities.

4. Discussion

The original clustering algorithm of the “Livehoods project” does not work for St.Petersburg database while it is not validated by sociological survey and does not account for pedestrian accessibility and specific natural barriers.

We propose to account for natural and built environment barriers which form the borders of the cluster and pedestrian accessibility within the cluster. In this paper, we incorporate the measure of pedestrian accessibility defined as a 500 m radius from the venue. Our main methodological contribution is a refinement of affinity matrix between venues that more effectively blends spatial affinity and social affinity. In further research, we will recalculate pedestrian accessibility through isochrones, which will provide an accurate account on the actual time needed to walk a specific distance in a given graph of pedestrian routes. Additionally, we will check for the natural and built environment barriers if the pedestrian routes around the venues will intersect the border of the barrier.

The clusters of areas of habitation built with the refined algorithm are showing more alignment with the results of the sociological survey, however not full correspondence. The received clusters have to undergo further sociological validation.

Interpretation of areas of habitation should be done carefully with an account for the nature of data. The data from social media represents demonstrative behavior when people are willing to signify the place they are in, which are mostly public spaces (cafes, bars, museums) and less everyday life places (supermarkets, drugstores).

Further analysis would be undertaken to test the hypothesis that clusterization gives specific economic effects, such as leveling up of the average check, and social effects, such as the formation of networks of places for social groups with a particular lifestyle.

The algorithm can be applied for the analysis and planning of polycentricity in St.Petersburg. It can help to work out strategies for areas development in terms of specific functions and coherent lifestyles, for example, bar culture area, educational milieu, public and cultural area within a sleeping quarter.

5. Acknowledgement

This paper has been written based on data collected within the framework of scientific and research work of students and PhD students №617039 “Methods to study behavioral patterns based on subjective perception of urban environment”.

References

[1] Hanson, J., Hillier, B. (1987) “The Architecture of Community: Some New Proposals on the Social Consequences of Architectural and Planning Decisions.” Arch. 8 Cwnport/Arch. Behavior 3 (3): 251-273.

[2] Westerink, J., Haase, D., Bauer, A., Ravetz, J., Jarrige, F., Aalbers, C. (2012) “Dealing with Sustainability Trade-Offs of the Compact City in Peri-Urban Planning Across European City Regions.” European Planning Studies journal 21 (4): 473-497.

[3] Shelton, T., Poorthuis, A., Zookb, M. (2015) “Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information.” Landscape and Urban Planning 142: 198-2;

[4] Cranshaw, J., Schwartz, R., Hong, J.I. Sadeh, N. (2012) “The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City.” AAAI Publications, Sixth International AAAI Conference on Weblogs and Social Media.