BibTeX:

Using Data from Location Based Social Networks for Urban Activity Clustering

Roberto Rösler and Thomas Liebig

Abstract Understanding the spatial and temporal aspects of activities in urban regions is one of the key challenges for the emerging fields of urban computing and emergency management as it provides indispensable insights on the quality of services in urban environments and helps to describe the socio-dynamics of urban districts. This work presents a novel approach to obtain this highly valuable knowledge. We hereby propose a segmentation of a city into clusters based on activity profiles using data from a Location Based Social Network (LBSN). In our approach, a segment is represented by different locations sharing the same temporal distribution of check-ins. We reveal how to describe the topic of the determined segments by modelling the difference to the overall temporal distribution of check-ins of the region. Furthermore, a technique from multidimensional scaling is adopted to compute a classification of all segments and visualize the results. The proposed method was successfully applied to Foursquare data recorded from May to October 2012 in the region of Cologne (Germany) and returns clear patterns separating areas known for different activities like nightlife or daily work. Finally, we discuss different aspects related to the use of data from LBSNs.

1 Introduction

Residents do not use urban space homogeneously. Whereas some areas consist of residential quarters, others represent nightlife or industrial districts. Thus, the "usage pattern" of a city centre differs from an industrial region or a trendy neighbourhood; the shops in the city centre are visited during regular opening times whereas the bars attract people especially in the evening hours.

The identification of places with similar usage is an interesting topic for authorities, urban analysts and residents, as it provides valuable insights. For example it can be used for assessing the quality of services in urban environments and helps to describe and understand the socio-dynamics of these areas or to support town

Roberto Rösler
Fraunhofer IAIS, Schloss Birlinghoven, 53757 Sankt Augustin, Germany
e-mail: roberto.roesler@iais.fraunhofer.de

Thomas Liebig
Department of Computer Science LS8, TU Dortmund University, 44221 Dortmund, Germany
e-mail: thomas.liebig@tu-dortmund.de

development planning. Furthermore it can be used to explore the genuine use-related urban structure on an up-to-date and microgeographic level, which could be contrasted with official planning data (may be out dated, on a much higher level or incomplete) and hence reveal important information about planning deviation. Another application could be, to improve official databases. In this context it has also a second function, as the underlying empirical, open and fine-grained data capturing the socio-dynamics might provide some unbiased knowledge, which is generally unavailable to individuals without local knowledge and access to special data sets. Furthermore, the identification of similar regions is crucial for evaluating (scoring or ranking) the performance of places under an economic view e.g. shopping facilities or nightlife areas. Shop planners would benefit from this information, when making a selection of possible sites for daytime dependent business. In the field of disaster management, local activity profiles are of high value for planning preventive actions by responsible agencies because the profiles provide knowledge of typical spatio-temporal activities in different districts. Hence, it enables the planning authorities to facilitate effective and forward-looking action plans and gives them the opportunity for an optimized resource management.

This paper addresses the question of how to identify spatial regions with similar temporal activities using widely available up-to-date data about the interaction of people and places. Our approach utilizes location based social network data as input for a spectral clustering.

The data contains so-called check-ins for spatial locations combined with person identifiers and a feature type (i.e. bar, restaurant, etc.) for the location (referred as venues). The temporal aggregation of the check-in frequencies per hour results in a vector per location containing 24 integers. These vectors and therefore the associated spatial objects are used for clustering. This is in contrast to [1], where the spatial situations (i.e. the presence or flow aggregates among all locations at a particular time-stamp) are subject to clustering. The method we propose utilizes spectral clustering and thus provides the capability to find arbitrarily shaped clusters without posing any constraints. The intuitive colouring of the resulting clusters helps to understand the activity profiles of the considered regions. To achieve this, we apply Sammon`s projection (compare Section 3).

The proposed algorithm has been successfully applied to Foursquare check-ins recorded from May 2012 to October 2012. The novel contributions of this paper are threefold: (1) The clustering based on constructed activity profiles provides a deeper understanding of the spatio-temporal structure of a city. (2) It is shown, how a new combination and extension of existing approaches allows a more natural way to cope with typical real-world clustering problems like the estimation of parameter values, the choice of a useful similarity metric or assumptions about the cluster type. (3) Moreover, to our best knowledge, this is the first approach using LBSN data for microgeographic modelling outside the “densely monitored regions” (foremost in the US), which are used by most of the studies analysing LBSNs. The hereby-studied dataset is sparser. With this dataset, we still achieved reasonable results, however, a systematic analysis of the impact and requirements on sampling density are not subject of this paper.

A systematic literature survey and outlook on future extensions completes this work.

The remainder of the paper proceeds as follows. Section 2 highlights other related approaches for analysing or using data from location based social networks. In Section 3, we introduce our novel approach to model local spatio-temporal activities. Afterwards, in Section 4, we conduct our experiments with a subsample of the Foursquare data. We close with a discussion and an outlook on future work in Section 5.

2 Related Work

Area of Application

The field of Urban Computing analyses how modern ICT infer and integrates with urban life and is one of the emerging research fields. There is also an obvious interest in the relation between urban dynamics and data gathered from various sources depicting human activity to improve and fasten the understanding of social processes and interactions [17]. One approach is about using cell phone data to model time-dependent behaviour of a city by clustering and, similar to our approach, interpreting the resulting patterns [24]. However, obtaining cell phone data is much more complicated (access, size, preparation and legal aspects) and often restricted to certain purposes (not to mention the public concern about the violation of privacy by analysing mobile-device data[2]). In contrast to that, parts of the data from LBSNs are public to everyone and feature a considerably less complicated structure.

The same holds when it comes to data from surveys. The research in [14] shows an example where data from a large-scale survey is used to model the urban spatio-temporal structure in the Chicago metropolitan region. Even though the data is publicly available and forms a good representation of the total population, it only covers one region. In addition, the question arises if the individual design makes surveys from different regions comparable to each other. Nevertheless, the fusion of data from various sources like cell phone data, surveys and LBSNs seems to offer a major advantage in understanding urban life.

Data from the new field of LBSNs stimulates different researchers to analyse urban life. In [2] the aim is to model the spatio-temporal characteristics of urban land use based on information from Foursquare whereas [22] uses the feature type of venues to identify user communities and urban neighbourhoods. Similar to that, the 'Livehoods Project' [7] takes a more natural approach to characterize and distinguish different social areas. For that, the authors used check-in data from Foursquare giving them a more realistic picture than using official municipal organiza-

[2] http://www.telecompaper.com/news/german-govt-to-limit-telefonica-plans-to-sell-customer-data--905518 (last visited: 14.11.2012)

tional units. Hence, in contrast to our approach the 'Livehoods Project' is more focused on the delimitation of social than functional areas.

Besides the understanding of urban life also in emergency management, data from location-based social networks and microblogging services seem to form a valuable source to get some early information of potential crisis events. Different approaches therefore use data from Twitter [8,22,27] or the combined data from different LBSNs to detect disasters [4] and present information to officials and emergency personnel.

General Aspects Using Data from LBSNs

LBSNs form a completely new phenomenon for the research community and therefore require some basic understanding of the motivation why, where and how people share information about their location and mobility behaviour and how they deal with the aspect of privacy [18]. Other researchers analyse characteristics of human mobility through data from social networks like in [21]. For example, [5] explore some general aspects of movement patterns, returning probability and economic and geographic constraints using a dataset of 22 million check-ins worldwide. They also analyse the textual content given by short messages or announcements from the check-ins to identify significant terms and sentiments about the locations visited by the users. The authors in [6] analysed users movement in relation to their social relationships represented by the structure of their social network. Their findings show that the social relationship explains a significant amount of human movement even though most of it is explained already through periodic behaviour. With the possibility to explain historical movements, it seems logical that other researchers also look at the task of making predictions of future movements on the basis of the user`s history and the structure of the social network [6,10].

Spatial Topic Modelling

One important field that brings together the data from Location Based Services and Urban Computing is topic modelling. Here different authors analyse the existence of local geographic topics which essentially are locations connected by user movements and share some common theme – like 'sports' or 'business trip' [19,16]. While Foursquare data seems to be dominating most of the research papers mentioned here, the authors in [29] explore Whrrl[3], a different source which is not active anymore. They discover temporal patterns related to venues and their categories. They show that even if the feature type (e.g. college, restaurant) is not consistently used across all venues, the distribution of check-ins in time reveals distinguishable patterns between categories. Besides using data from LBSN other researchers show how to extract topics, make location prediction (this is an important step because still a lot of user generated content comes without coordinates) [12] and use identified sentiments to improve services for tourists [26], all with data from micro-blogging services like Twitter.

[3] http://en.wikipedia.org/wiki/Whrrl (last visited: 14.11.2012)

Privacy
The last discussed aspect here (but not less important) is the dimension of privacy for both, users and analysts, when dealing with data or topics related to LBSNs. A Characterization of non-private information and status messages of users and venues which are available from Foursquare (e.g. tips, mayorship status) is provided in [23]. The authors furthermore estimate the home city of a person using only publicly accessible information. In [15] a new framework for preserving residential privacy for users from Foursquare is proposed.

3 Modelling local spatio-temporal activities

Location based social networks (LBSNs) allow people to share location based information (e.g. position, time, location description, etc.) with other users. In return they get incentives from the LBSN provider for being an active user in the community or they benefits from local shops for visiting them. While the user 'checks in' at a place (called venue) his location will be shown on his mobile phone and also be depicted to his friends. In addition, it is possible to rate venues and attach notes, pictures or other information to check-ins. Use cases for this new type of service cover a broad spectrum from exploration over recommendation to location-based gaming [3].

Throughout this paper we use data collected from one of the largest LBSNs, called Foursquare, with a community of more than 25 million users worldwide who produce over one million check-ins per day (October 2012[4]). This rich data source gives us the opportunity to analyse the spatio-temporal interaction between individuals and places and to project the results onto a crisp classification of local clusters, which are described by their activity profiles.

Before explaining our methodology, we hereby state our notation: V is a set of n_V Foursquare venues i, U is a set of n_U Foursquare users u, C is set of check-ins where each check-in c_{vu}^{t} consists of a venue v, a user u and a timestamp t. For two venues we can calculate the geographic distance $d(i,j)$ for all $i,j \in V$ using the given coordinates from i and j.

According to [1] we address the uncertainties in spatio-temporal data by aggregation. Thus, in a first step aggregates are estimated for every venue i containing the count of all check-ins c_i per hour to get a vector of the hourly distribution of check-ins v_i, where the t^{th} Element (written as v_i^t) represents the number of check-ins at the venue i at hour t with $t = 0 \dots 23$ - we call this the *activity profile* of a venue. The next preparation step is data filtering. Thus, we remove all venues that have less than two check-ins or two unique users.

[4] https://foursquare.com/about/ (last visited: 14.11.2012)

Affinity Measure

For clustering the venues according to their activity profile, we define a similarity measure in the following way: Interpreting the profile of a venue as a short time series, we calculate a distance between two venues based on a comparison of the shapes of their profiles; see [28]. The idea is, that if v_i and v_j are similar, the change from t to $t+m$ should be similar for both venues for all possible values of t and m. For a chosen t and m this means a comparison of the shift from v_i^t to v_i^{t+m} with the shift from v_j^t to v_j^{t+m}. The possible values of the shift q among t and $t+m$ are qualitative, therefore $q(v_i^t, v_i^{t+m})$ gets the label 'increase' if $v_i^t < v_i^{t+m}$; 'decrease' if $v_i^t > v_i^{t+m}$ and 'no-change' if both are the same. The similarity between two venues v_i and v_j with respect to a shift from t to $t+m$ is denoted as as $sim(q(v_i^t, v_i^{t+m}), q(v_j^t, v_j^{t+m}))$ and is defined in the following Table 1.

$sim(q_1, q_2)$		q_1		
		increase	no-change	decrease
q_2	increase	1	0.5	0
	no-change	0.5	1	0.5
	decrease	0	0.5	1

Table 1 Similarity of the shifts $\boldsymbol{q_1}$ and $\boldsymbol{q_2}$ between the venues $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$.

At last we define an affinity measure between two venues as

$$Aff(v_i, v_j) = \frac{\sum_{t<t'} sim(q(v_i^t, v_i^{t'}), q(v_j^t, v_j^{t'}))}{NC}$$

where NC is the number of all possible comparisons (253 in our case).

Clustering Algorithm

For our analysis, we apply the spectral clustering as in [20] because it finds arbitrarily shaped clusters and does not pose any constraints on them (in contrast to the k-means, for example, which assumes cluster to be convex). For this, we follow the preparations given in [7] to create first the $n_V \times n_V$ affinity matrix $A = (a_{i,j})_{i,j=1,\dots,n_V}$ where:

$$a_{i,j} = \begin{cases} Aff(v_i, v_j) + \alpha & \text{if } j \in N_m(v_i) \text{ or } i \in N_m(v_j) \\ 0 & otherwise \end{cases}.$$

Here $N_m(v_i)$ refers to the m nearest venues to v_i with respect to their distance d. The small constant α assures that every venue is connected to its neighbours (α was set to 0.01). The affinity matrix A together with the number of desired partitions k is given as input to the spectral clustering algorithm described in [20]. The result is a partition of all venues into k disjoint clusters $C_1, \dots, C_k$ where every

cluster C_i could be mapped on a subgraph $G(A_i)$ of graph $G(A)$ where A_i is a set of the corresponding vertices to C_i. As in [7], we apply a post-processing step to get spatially contiguous partitions. Here we replace every subgraph $G(A_i)$ with a set of clusters produced by splitting it into its connected components.

Evidence Accumulation Clustering (EAC)

There exist many different clustering algorithms each having its advantages like simplicity, a small number of parameters, the ability to identify arbitrary shaped clusters or the capability to handle large data sets [13]. The difficulty is that every clustering algorithm and even any set of parameters will produce a somewhat different solution. This makes it hard to decide, which result should be kept. In our case, there is no prior knowledge about the number of clusters in the region under study. Consequently, it is not easy to estimate the right value for the parameter k. An approach to overcome this problem is called evidence accumulation clustering (EAC) and was proposed in [9]. The notion behind this method is to build clusters with different algorithms and parameterizations and then to aggregate all solutions into one final partition using every single partition as a voting if instances should be placed together. If two venues will be placed together in most solutions, it is reasonable to assign them to the same cluster in the final partition. In this context, this method could also be understood as a tool to enhance the validity of the resulting partition by reducing the impact resulting from a single non-optimal clustering (like not well-separated clusters).

We adopt the idea of evidence accumulation clustering, to combine different runs of the spectral clustering, where k is sampled from the interval $\mathrm{k_{min}}$ to $\mathrm{k_{max}}$ in every single run. The algorithm is as follows:

Input:	R – number of runs; k_{min} and k_{max} – interval for the possible values of k n_V – the number of venues
Output:	Final partition P^{final}

1. Initialize the $n_V \times n_V$ association matrix $S = (s_{i,j})_{i,j=1,\dots,n_V}$, which will contain the "votes" of every iteration, with zero
2. Do R times:
   2.1. Randomly select k from the interval k_{min} to k_{max}
   2.2. Run the spectral clustering with the chosen k and the precomputed affinity matrix A to produce a partition P
   2.3. Update the association matrix S according to the following rule:
   for every pair of venues (i, j) in the same cluster in P set $s_{i,j} = s_{i,j} + \frac{1}{R}$
3. Transform the association matrix S into a distance matrix S' by $1 - S$
4. Extract the final partition P^{final} with complete linkage clustering using the distance matrix S'. On the resulting dendrogram the final cutpoint is obtained by the highest 'cluster lifetime' which is the longest 'gap' between two successive merges.

Algorithm 1 Slightly modified version of the EAC Algorithm based on [9].

4 Experiment

4.1 Data collection

The data used here contains Foursquare check-ins from May 2012 to October 2012. We collected all public check-ins found by searching the Twitter Timeline[5] for Foursquare tags by restricting them to the German language (for this we used the search options from Twitter). Then the information about all check-ins was gathered by querying the Foursquare API[6]. Afterwards we verified the country attribute of the check-ins and the extracted coordinates and discarded all check-ins not belonging to venues in Germany. In addition, venues were enriched with information using the Foursquare API again. Essentially, this information contains the hierarchy of the feature type used in Foursquare (e.g. Arts & Entertainment, Travel & Transport). Finally, the check-in ID, the user ID, the venue ID, the venue name, the check-in date and time, the main feature type of the venue and the coordinates for every check-in are known.

Our analysis focuses on the city of Cologne as an example for a major German city with more than 1 million inhabitants. After removing a few inconsistent items, the final dataset for Cologne consists of 11,890 check-ins from 2,093 users over more than 1,008 venues. The size of our dataset is considerably smaller than in other approaches (for example [7] and [19]) but most of them focus on major regions in the US with an disproportionate number of active users compared to the majority of the regions in Germany. So the question is raised if such micro-geographic approaches based on public data from LBSNs are also applicable in urban regions in Germany.

Figure 1 and Figure 2 depict some general characteristics of the data. In Figure 1 the hourly check-in frequency averaged over all days is plotted; clearly showing three peaks: commuter traffic in the morning, lunch time and the evening rush hour. This graphic could also be considered as the Foursquare perception of the activity profile of Cologne. Figure 2 shows the relative frequency for all main feature types where the check-ins are dominated by the categories 'Travel & Transport' followed by 'Food' and 'Professional & Other Places' (work).

[5] https://dev.twitter.com/docs/using-search (last visited: 14.11.2012)

[6] https://developer.foursquare.com/ (last visited: 14.11.2012)

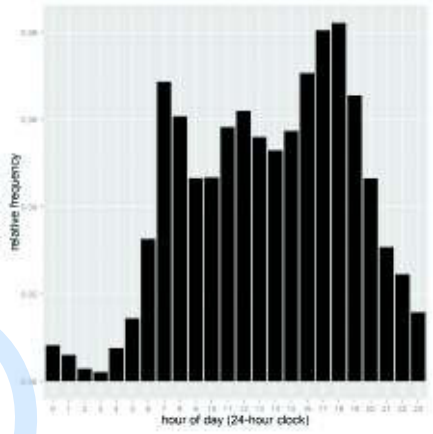

Fig. 1 Histogram of the check- ins per hour.

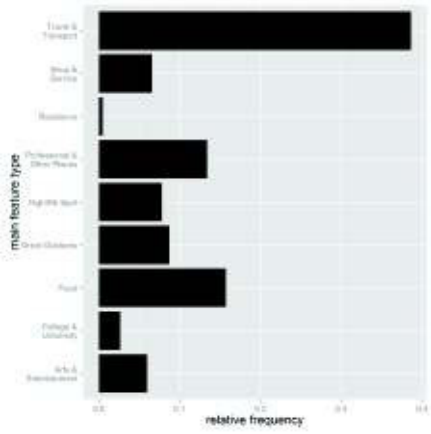

Fig. 2 Histogram of the check-ins per main feature type.

4.2 Analysis & Results

Computing the Local Clusters

For carrying out our analysis, we used the following parameters: To build our affinity matrix we choose $m = 5$ (five nearest neighbours) which gave us mostly contiguous clusters. The EAC was carried out with $R = 2{,}000$. The proposed value ensured convergence for our analysis. We set $k_{min} = 5$ and $k_{max} = 60$ and thus accounted for both, small and big clusters, targeting the inherent uncertainty about the correct value for k.

Our first step after data preparation was to build up our affinity matrix A according to the given parameters. Secondly we ran the proposed EAC algorithm to compute the association matrix S. Figure 3 shows the results. Due to the maximal cluster lifetime of 0.057 we decided to take 31 as our value for the number of regions (this assured that the venues to be placed in the same final cluster must have been placed together in at least 82% of all runs).

General Cluster Structure

The final clustering is displayed in Figure 4. In favour of a clear overview, we focus the shown map extent to the inner city of Cologne (every cluster is represented by a different colour and the convex hull). It is clearly visible that the city centre consists of some small clusters while the surrounding districts are represented by large partitions. This is plausible since often city centres are functionally much more heterogeneous, while suburban areas are more homogenous (like large residential or industrial areas). While the clustering seems reasonable, we now want to

show how each of the computed partitions describe some special topic depending on the activity profile.

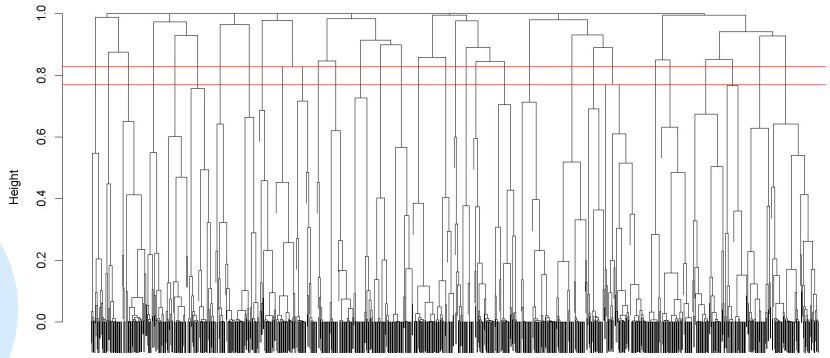

Fig. 3 Determining the number of clusters in the final partition through exploration of the resulting dendrogram (the maximal cluster lifetime of 0.057 is between the two red lines – see Algorithm 1 for the definition).

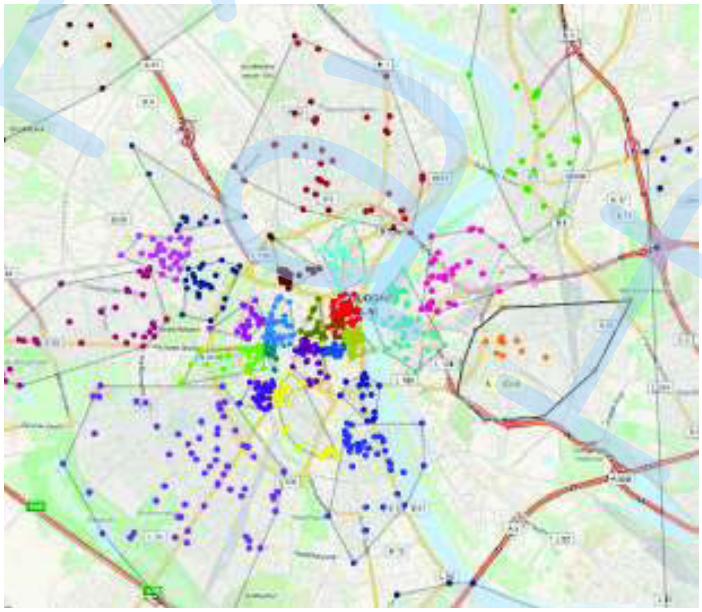

Fig. 4 Results from the clustering of Foursquare check-ins in Cologne city centre and surroundings.

Activity profiles

Subsequently, we compute the activity profile of a cluster, which is defined as the sum over all activity profiles of the corresponding venues. The next step is to sub-

tract it from the overall activity profile after transforming all absolute frequencies into relative frequencies. This yields the graphic shown in Figure 5 in which the difference to the regional activity profile is displayed for every cluster. If a bar has a value near zero, the activity in this cluster (which means check-ins) is very similar to the activity at a regional level with respect to that time slot.

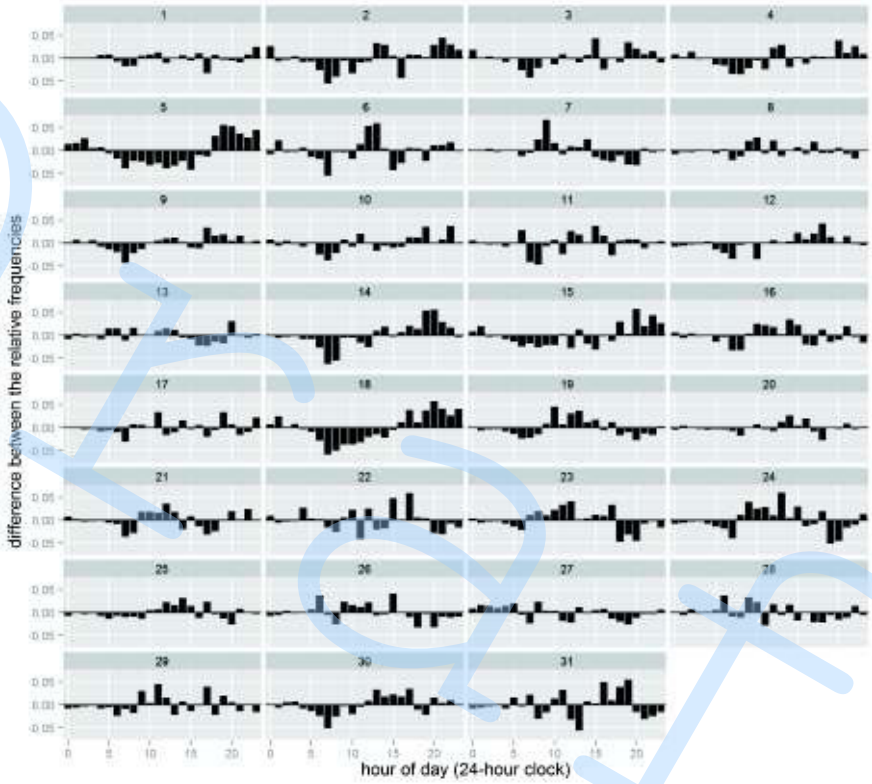

Fig. 5 Per cluster difference in the activity profiles between cluster and regional level – the ordinate axis displays the difference between the activity in the cluster and at the regional level for every hour.

There are some typical profiles like the one displayed for cluster 5, 15 and 18 which show the main hotspots of the Cologne's nightlife (Township 'Ehrenfeld', 'Belgisches Viertel' and 'Zülpicher Strasse'). While the number of check-ins is significantly less during the day, there is a lot more activity in the evening and at night. It is also typical that the last visitors of bars and clubs go home long time after midnight. It is interesting to see that cluster 2 (building the southern part of the region named 'Belgisches Viertel') also seems to have an attractive nightlife but contrary to the other nightlife hotspots, there are many individuals nearby around

lunchtime. On the other hand, there are also clusters showing completely different aspects of daily life. For example, cluster 1 is putting up an area around Cologne main station. It matches almost perfectly the overall activity profile (so the differences are near zero) because all of the mentioned peaks from Figure 1 are in some kind related to aspects of public transport. For example, the clusters with IDs 19, 23, 24 and 30 are typical instances for work-related areas. There the main activity takes place during typical working hours between ten and fifteen o'clock.

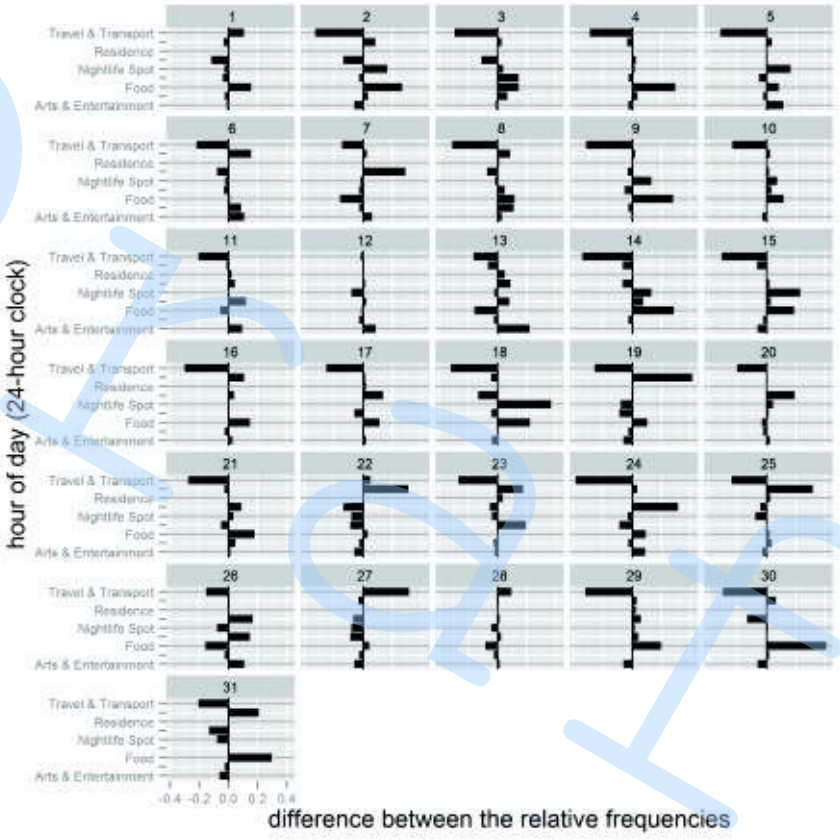

Fig. 6 Individual deviation of the distribution of the check-in category for every cluster from the average distribution of the check-in categories.

To demonstrate the plausibility of our findings, we create a different plot (Figure 6) showing the difference between the relative frequency of the main feature types between cluster and regional level. In favour of readability, we left out the label of every second feature type on the y-axis, which are the same like the corresponding one in Figure 2. Here it is conspicuous that the clusters forming the nightlife also consist of an outstanding number of venues categorized as nightlife

(compared to the average). For the cluster representing the area around Cologne main station venues from the categories ‘Food’ and ‘Travel & Transport’ are dominating. While the second category seems to be instantly intuitive, also the first one is plausible because most of the small shops at the main station sell food and beverages. The description for the class of the work-related clusters is somewhat more heterogeneous – those venues often have the main feature type ‘Shop & Service’, ‘Great Outdoors’ (botanic garden and places to rest along the river Rhine), ‘Professional & Other Places’ and ‘Food’ – most places restricted to opening hours which explain the typical shape of the activity profile.

Classifying activity profiles

So far, we can create a description for every constructed cluster. Still an overall description of the entire region is missing, which relates all clusters according to their activity profile. This could also be considered as a clustering of the constructed partitions where partitions featuring the same profile should be kept together. To get an intuitive and visualizable solution we therefore choose Sammon`s mapping [25], a method from multidimensional scaling which does a projection from the 24-dimensional space (the profile) into a space of lower dimensionality (we apply here a two-dimensional colour plane). Here the similarity between two clusters with respect to their profiles is expressed through closeness. A good way to visualize this in terms of our solution is to choose the colouring according to the mapping, so that similar clusters get similar colours. For the discovered nightlife hotspots we depict this colouring in Figure 7. The small graphic in the upper right corner shows a part of the results from the Sammon’s projection focussing on the identified nightlife areas 5, 15 and 18 (coloured in green). As expected, they are all arranged close together. The adjacent clusters 4 and 14 also seem to have a significant number of venues belonging to nightlife. However, looking at the activity profile indicates that closing time is earlier and the nightlife aspect is less pronounced.

In summary the general classification can be described by the following rules of thumb: while green indicates an active nightlife (or activity in the evening), blue clusters are more often characterized by ‘daylight activities’ and red stands for partitions not differing too much from the average regional profile shown in Figure 1.

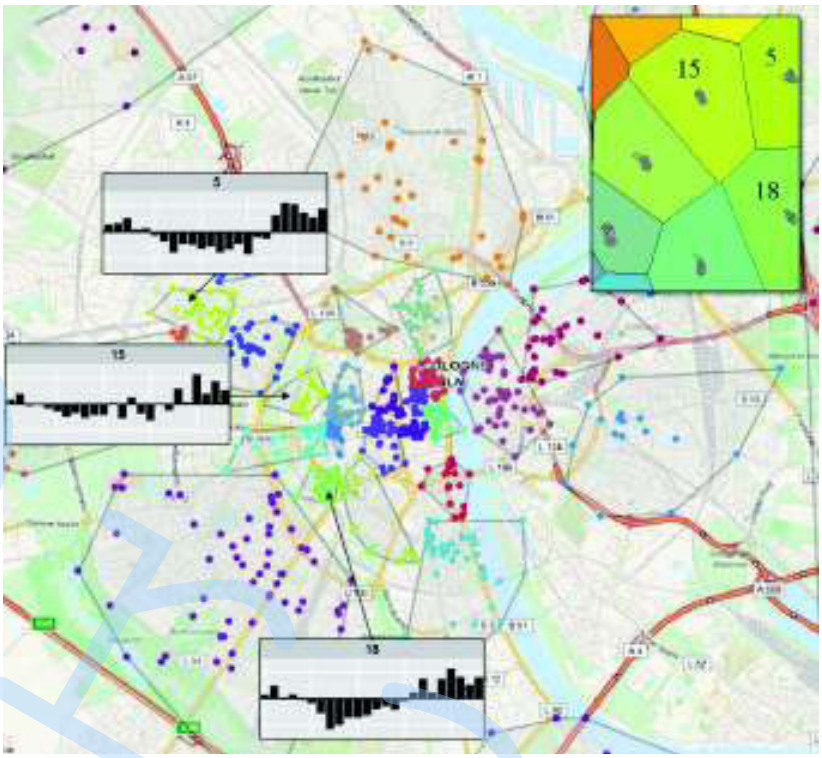

Fig. 7 Results from the Sammon's mapping for the inner city of Cologne.

5 Discussion/Future Work

In this work, we presented a novel approach to identify, describe and classify areas according to the temporal distribution of individual human activities. Therefore, we utilized a new source of data obtained from LBSNs. A major advantage of using this type of information is its general availability. It consists of events (called check-ins) triggered by users and bundle spatial (the location), temporal (the time of the event) and personal (the user ID) information. An important difference to the usage of coordinates is, that the spatial information here is connected with a physical location (a venue) and attached with a human readable description. Thereby people visiting the same physical location (e.g. a restaurant or a shopping area) could be matched on the same coordinates.

We used this information to build an activity profile for every venue in the area around Cologne. The profile is based on a similarity measure, which is particularly suitable for short time series.

We then applied spectral clustering to obtain partitions that contain venues with similar activity profiles. To overcome the problems arising from the "instability"

of clustering methods we used evidence accumulation clustering to deduce the parameters needed directly from the data (in this case the number of clusters).

Then our method proposed for classifying the partitions is straightforward. We computed the difference between the activity profile of the partition and the activity profile of the regional level for every cluster and visualized all differences. Especially partitions with the emphasis on nightlife and workplace could easily be identified. Furthermore, we classified every cluster using a technique from the field of multidimensional scaling to obtain an intuitive visualization. Therein the similarity between clusters is expressed through similar colours.

Based on this we showed how to explore regions of similar activity and how to characterize them by colour. Because of the promising results, we think that the overall approach could be a starting point for a better understanding of urban dynamics.

There are three main findings resulting from this work: (1) The construction of activity profiles for every location allows a clustering based on the temporal distribution of the venues. It thereby features a description of the spatio-temporal structure of a city. (2) We show how a new combination of existing approaches allows the creation of local and contiguous clusters without suffering from problems like the uncertainty about the right parameter values or assumptions about the cluster type. (3) To the best of our knowledge, this is the first approach using LBSN data for microgeographic modelling in the largest country in Europe. We conclude that while the size of the obtained data is still small compared to studies in the US, it is nevertheless possible to feature a much better understanding of social processes and interactions in urban regions. The outlook is even better because there is an on-going increase in the use of mobile devices and LBSNs.

In terms of future work, we intend to focus on three main aspects from which we expect the most benefits:

The first one is the integration of different sources of data coming from LBSNs and microblogging services, which will be a challenging task. Especially matching venues from the different ‘ecosystems‘ is clearly non-trivial. This is because there is no standardization concerning the names, feature types or localization. On the other hand, it will directly increase the sample size and thus likely improve results. Additional data from microblogging services like Twitter could fill gaps especially in regions where the usage of LBSNs is not sufficient and vice versa.

The second direction for future work is the exploration of general limitations by using Volunteered Geographic Information (VGI) [11] for urban analysis. For example, it is well known that this data does not provide a representative sample from the whole population. This can be important when interpreting the results from our method (and of course the work from others when using VGI data). Mentioned in the part of related work, a possibility to overcome this situation could be an approach that uses different sources of mobility data. The fusion of data from cell phones, surveys and VGI will provide valuable information for analysing activity patterns and possibly enhances the transformation from urban re-

gions to ‘smart cities’. In this case the public discussion about privacy should be taken into account as well.

The third and last direction will focus on the algorithmic parts, particularly the extension of the classification. For example, the method could support the interpretability by providing an intuitive and automatic description for every cluster and interactive tools to let the user dive into the results. In addition, the computation should be made ‘big-data-ready’ to cope with the massive amount of data produced in LBSNs and microblogging services.

References

1. Andrienko, N., Andrienko, G., Stange, H., Liebig, T., Hecker, D.: Visual Analytics for Understanding Spatial Situations from Episodic Movement Data. KI - Künstliche Intelligenz, Springer (2012) 241–251
2. Aubrecht, C., Ungar, J., Freire, S.: Exploring the potential of volunteered geographic information for modeling spatio-temporal characteristics of urban population. In: Proceedings of the 7th International Conference on Virtual Cities and Territorie. 7VCT ’11, Lisbon (2011) 57–60
3. Bawa-Cavia, A.: Sensing the urban: Using location-based social network data in urban analysis. In: The First Workshop on Pervasive Urban Applications. PURBA ’11, San Francisco (2011)
4. Chen, L.J., Li, C.W., Huang, Y.T., Shih, C.S.: A Rapid Method for Detecting Geographically Disconnected Areas after Disasters. In: IEEE International Conference on Technologies for Homeland Security. HST ’11, Greater Boston (2011) 501–506
5. Cheng, Z., Caverlee, J., Lee, K., Sui, D.Z.: Exploring Millions of Footprints in Location Sharing Services. In: The Social Mobile Web. ICWSM ’11, Barcelona (2011)
6. Cho, E., Myers, S.A., Leskovec, J.: Friendship and Mobility: User Movement in Location-Based Social Networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD ’11, New York, NY, USA, ACM (2011) 1082–1090
7. Cranshaw, J., Schwartz, R., Hong, J.I. and Sadeh, N.: The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. To appear in The 6th International AAAI Conference on Weblogs and Social Media. Dublin, Ireland (2012).
8. De Longueville, B., Smith, R.S., Luraschi, G.: ”Omg, from here, i can see the flames!”: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In: Proceedings of the 2009 International Workshop on Location Based Social Networks. LBSN ’09, New York, NY, USA, ACM (2009) 73–80
9. Fred, A.L.N., Jain, A.K.: Evidence Accumulation Clustering based on the K-Means Algorithm. In: Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition, London, UK, Springer (2002) 442–451
10. Gao, H., Tang, J., Liu, H.: Exploring Social-Historical Ties on Location-Based Social Networks. In Breslin, J.G., Ellison, N.B., Shanahan, J.G., Tufekci, Z., eds.: ICWSM, The AAAI Press (2012)
11. Goodchild, M.F.: Citizens as sensors: the world of volunteered geography. GeoJournal 69(4), Springer (2007) 211–221
12. Hong, L., Ahmed, A., Gurumurthy, S., Smola, A.J., Tsioutsiouliklis, K.: Discovering Geographical Topics in the Twitter Stream. In: Proceedings of the 21st international conference on World Wide Web. WWW ’12, New York, NY, USA, ACM (2012) 769–778

13. Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: A Review. ACM Comput. Surv. 31(3) (1999) 264–323
14. Jiang, S., Ferreira, Jr., J., Gonzalez, M.C.: Discovering Urban Spatial-Temporal Structure from Human Activity Patterns. In: Proceedings of the ACM SIGKDD International Workshop on Urban Computing. UrbComp ’12, New York, NY, USA, ACM (2012) 95–102
15. Jin, L., Long, X., Joshi, J.B.: Towards Understanding Residential Privacy by Analyzing Users' Activities in Foursquare. In: Proceedings of the 2012 ACM Workshop on Building analysis datasets and gathering experience returns for security. BADGERS ’12, New York, NY, USA, ACM (2012) 25–32
16. Joseph, K., Tan, C.H., Carley, K.M.: Beyond “Local”, “Categories” and “Friends”: Clustering foursquare Users with Latent “Topics”. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing. UbiComp ’12, New York, NY, USA, ACM (2012) 919–926
17. Kindberg, T., Chalmers, M., Paulos, E.: Guest editors’ introduction: Urban computing. Pervasive Computing, IEEE 6(3) (2007) 18–20
18. Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J., Zimmerman, J.: I’m the Mayor of My House: Examining Why People Use Foursquare - a Social-Driven Location Sharing Application. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI ’11, New York, NY, USA, ACM (2011) 2409–2418
19. Long, X., Jin, L., Joshi, J.: Exploring Trajectory-Driven Local Geographic Topics in Foursquare. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing. UbiComp ’12, New York, NY, USA, ACM (2012) 927–934
20. Ng, A.Y., Jordan, M.I., Weiss, Y.: On Spectral Clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems, MIT Press (2001) 849–856
21. Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: An Empirical Study of Geographic User Activity Patterns in Foursquare. In: In Proceedings of the 5th Int'l AAAI Conference on Weblogs and Social Media. ICWSM ’11, Barcelona (2011) 570–573
22. Noulas, A., Scellato, S., Mascolo, C., Pontil, M.: Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-Based Social Networks. In: The Social Mobile Web. ICWSM ’11, Barcelona (2011)
23. Pontes, T., Vasconcelos, M., Almeida, J., Kumaraguru, P., Almeida, V.: We Know Where You Live: Privacy Characterization of Foursquare Behavior. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing. UbiComp ’12, New York, NY, USA, ACM (2012) 898–905
24. Reades, J., Calabrese, F., Sevtsuk, A., Ratti, C.: Cellular census: Explorations in urban data collection. Pervasive Computing, IEEE 6(3) (2007) 30–38
25. Sammon, J.W.: A Nonlinear Mapping for Data Structure Analysis. IEEE Trans. Comput. 18(5) (May 1969) 401–409
26. Shimada, K., Inoue, S., Maeda, H., Endo, T.: Analyzing Tourism Information on Twitter for a Local City. In: First ACIS International Symposium on Software and Network Engineering. SSNE ‘11. (2011) 61–66
27. Thom, D. Bosch, H. Koch, S. Worner, M. and Ertl, T.: Spatiotemporal Anomaly Detection through Visual Analysis of Geolocated Twitter Messages. In: Proceedings of the Pacific Visualization Symposium. PacificVis’12, IEEE Press (2012) 41–48
28. Todorovski, L., Cestnik, B., Kline, M., Lavrac, N., Dzeroski, S.: Qualitative Clustering of Short Time-series: A Case Study of Firms Reputation Data, Helsinki University Printing House (2002) 141–149
29. Ye, M., Janowicz, K., Mülligann, C., Lee, W.C.: What you are is When you are: The Temporal Dimension of Feature Types in Location-based Social Networks. In: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. GIS ’11, New York, NY, USA, ACM (2011) 102–111