Towards Approaches and Techniques for Analysing WiFi Location Data

John McAuley  
*Technological University Dublin*, john.mcauley@tudublin.ie

Clement Roux  
*Trinity College Dublin, Ireland*, clement.roux@tcd.ie

James Little  
*ThinkSmarter*, james.little@thinksmarter.com

Follow this and additional works at: [https://arrow.tudublin.ie/scschcomcon](https://arrow.tudublin.ie/scschcomcon)

**Recommended Citation**


This Conference Paper is brought to you for free and open access by the School of Computing at ARROW@TU Dublin. It has been accepted for inclusion in Conference papers by an authorized administrator of ARROW@TU Dublin. For more information, please contact yvonne.desmond@tudublin.ie, arrow.admin@tudublin.ie, brian.widdis@tudublin.ie.

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License.
Towards Approaches and Techniques for analysing WiFi location data

Clement Roux, John McAuley, James Little
Trinity College Dublin, Dublin Institute of Technology, Ireland

Abstract. Cities have an increasingly dense network of connected access points that enable individuals and groups to access the web on a range of different devices. Although, in most cases, the data collected from these networks is anonymised, when aggregated, it provides a rich illustration of how cities, public places and urban spaces are used. However, making use of this data can prove challenging as rarely, due to anonymization and data protection measures, is the data labelled and often the accuracy of the location information is questionable. Given these constraints, we have investigated a set of methods that can be applied to anonymous WiFi-location-based data. Through two specific use-cases, we show successful examples of each method, describing insights and intuitions along with providing an explanation of their application.

1 Introduction

The objective of our work is twofold. Firstly, we investigate a range of techniques that can be applied to anonymous WiFi data and second, we seek to better understand how these techniques can support a range of different use cases. To do this, we are working closely with a team of WiFi engineers and WiFi installation experts that develop WiFi networks for large indoor and outdoor installations. For example, the team work closely with a conference centre in Barcelona to establish how best to utilise WiFi networks during their events while at the same time the team are working alongside Dun Laoghaire Library in Dublin to see how WiFi can better support their patrons. In each situation, the datasets have specific constraints that can only be alleviated through better network configurations or through the deanonymisation of data.

Collected WiFi signals result in spatiotemporal data that combines geographic coordinates with a timestamp and a mac address of a specific device. Often this data is gathered through a process of passive probing in which the access point connects momentarily to a device, identifies the device and its locations but then drops the connection. This is a continuous process performed by a large majority of WiFi access points. Our interest lies in maximising the potential of this data so that it can support a range of different applications such as urban planning, crowd management, public security, transportation optimisation and advertisement.

In the rest of this paper, we focus on two specific use cases in which we use WiFi data to address two specific yet common problems. The first use-case
investigates how WiFi data can be used for group identification. More specific, we collected data from the FA cup final that took place this year between Arsenal and Chelsea and sought to identify the individual groups of supporters based on their spatiotemporal behaviour. The analysis relies on the premise that two distinct groups of supporters participated in the event and attempts to identify each group using a novel approach to binary classification. The second use case illustrates a location accuracy analysis performed on a dataset gathered during a conference at the Fira conference centre in Barcelona.

The rest of the paper is organised as follows. First, we consider related work. As discussed in the introduction, our approach to acknowledge the constraints of many WiFi-generated datasets and to investigate ways to address these constraints. Our focus, as a result, is on papers that address methods and techniques. Next, we present the first case study, describing the data, methodology and results. Then, we present the second case study that looks at accuracy of these types of dataset. We conclude the paper with a discussion and future work.

2 Related Work

Indoor and outdoor WiFi probe requests along with their accuracy indicators such as RSSI are increasingly used for location-based analytics and knowledge discovery as they become cheaper and more reliable [20]. However, raw WiFi data is generally irregular and noisy, even more when configured outdoor which is probably why research has mainly focused on indoor configurations and analysis [30]. WiFi data is often studied in combination with other smartphone sensors inputs such as the GPS, Bluetooth, acoustic information [25], accelerometer and magnetic field [6] to provide more precise localisation information and a wider set of features. However, studies of large and public open space crowds often have restricted access to the devices information. The standard WiFi signals often remain the simplest and cheapest solution, from which derived applications are numerous, such as queuing time prediction [26] or zone occupancy tracking [14, 28].

Closer to the first use cases main focus, classification of shoppers state (walking vs standing, fast vs slowly, inside vs outside) based on indoor WiFi signals is performed in [34]. Contrary to us, the authors benefit from good levels of accuracy due to the relatively small indoor experimental setup being a retail shop. They also have access to the ground truth and labels to evaluate their classification model.

The crowd sensing, flocks detections and group identification use different methodologies to perform pool devices together such as clustering algorithms like DensityJoin-Cluster [12], graph theory using spatial connectivity features [21] or unsupervised algorithms like SVM [23]. Also, [24] use a rule-based learning model to classify devices within roles in a hospital (employee, visitor, patient). The rules are defined by local experts for labelling 5% of the dataset. The remaining dataset is labelled with the Bayesian Net learner supervised learning. The result
is then evaluated against national statistics. This method is hardly adaptable to our context which is highly depending on a few hours event. Slightly different and original, [22] extract the spatiotemporal transition features to predict places attributes. They assume, knowing where devices come from and where they go next holds information about the current places attributes. Even though successful, this methodology is not applicable to our dataset because of the restricted number of zones available compared to theirs.

The literature finally shows little example of group identification and crowd classification limited to unlabelled WiFi signals collected from outdoor large open spaces. The methodology we suggest is a mixture of existing techniques. To some extent, the work from [21] shows some similarities with ours because they focus on spatial connectivity features. However, they rather study motion and flows for getting insights about movements of students and professors working on several campuses. The adjacency matrix [9, 10] and community detection based on shared properties (shared devices and shared grid-cells) are also of interest to make classification. A similar approach based on trajectory similarities is studied in [18]. Those heuristics are visual and intuitive approaches for spatiotemporal data exploration, group identification and online recommendation.

Location analytics, group identification and accuracy analysis often use data visualisation as a supporting tool for analysing mobility behaviours, group structures and signal propagation, to mention just a few. More specifically, the graphs node-link model is a way to represent places, relationships and movements [2, 7, 11, 15]. In regard to our second use case, heat maps, grid maps and level plots are generally used for comparison and evaluation purposes in model and accuracy analyses such as in [8, 19, 20, 27].

Location accuracy evaluation metrics exist when there is a ground truth dataset against which similarity and distance metrics can be computed, such as the Euclidean distance between an estimated point and its ground truth or the Frchet distance between an estimated trajectory and its ground truth [4, 20].

3 Use Case 1: Group identification of large stadium events

In this section, we consider the first use case, the identification of groups of football supporters in large crowds. Although this is a specific use case, and supported by the fact we have two delineated groups within the stadium, the ability to identify and segment groups has a range of use cases beyond the current application.

3.1 Outline of use case: Classification of large crowds

Increasingly, local, corporate and civic entities are installing WiFi and other access points (such as Bluetooth) in public places. These points receive and persist probe requests from WiFi enabled devices within their proximity. A connected network of these devices captures the behaviour of individuals, groups and
crowds as they traverse the environment in and around each device. Analysing this data can provide urban planners and local enterprises with a number of useful insights that can be help to create more successful urban spaces [1], advance security measures [14], support the management of crowds [16], suggest online recommendations based on offline activity [18] and provide marketing and retail opportunities for local shops [33]. The identification of groups is a use case common to each scenario. Whether segmenting customers for a particular marketing campaign or identifying potential troublemakers at a sports event, the ability to classify groups within crowds has many potential applications.

Many of the existing solutions are developed on data in which no clearly delineated and apriori groups exist nurses versus patients for example but their behaviour is not particularly different nor is their identification as one group or another particularly straightforward. In contrast, we focused on a dataset with two clearly delineated groups—supporters of Chelsea and Arsenal at the Wembley Cup final 2017. We collected data from several WiFi access points before, during and after the cup final and used mobility patterns, groups locations and spatial behaviours to target specific supporter profiles. The advantage of our approach is that we had two clearly delineated groups and sought to develop a technique that would help to verify this delineation with binary classification.

Our particular use case had a practical application—the configuration, definition and continued improvement of WiFi subzones for outdoor events. The configuration of a WiFi zone can significantly impact the effectiveness of WiFi signals and accuracy of location-based WiFi analytics. Events that rely on WiFi require configuration that can handle the volume of users while also enabling better accuracy and connection certainty. In this particular use case, the event was organised so that each group of supporters (from each of the two teams) were provided with their own Fan Zone or recreational space outside of the stadium. Arsenals fan zone was in the Arena square and Chelseas fan zone was in the Events pad. Also, the supporters arriving from Olympic way (the main thoroughfare that brings supporters to the stadium) and going to the stadiums main entrance were separated so that each side of the road was dedicated to one team (see figure 1). We sought to use this information to perform different binary classifications based on the devices probe requests information that could be identified that day.

### 3.2 Data Description

The dataset was collected during the FA Cup Final, a major English football event between the premiership teams Chelsea and Arsenal. The data collection took place on the 27th of May 2017 using several Cisco Meraki access points located outside the Wembley stadium. The data was collected over a time span of 24 hours.

Each record in the dataset is the result of a triangulation process applied on multiple probe requests initiated by a single device. This results in a single data point generated by a single device and located on a map at a point in time. Each data point therefore contains one anonymised MAC address which we assume
to be related to one device, one time stamp, one set of geographical coordinates (latitude and longitude), one geographical zone and one confidence factor (CF). The latter is a distance representing half the side of a square of which the centre is the data point itself. This square represents the points uncertainty surface which shows where the device is located within a 95% confidence interval. Therefore, the bigger the CF the less accurate the estimation of the points location. These estimations were problematic with the Wembley dataset because the area is large and the network configuration required many APs. Even though WiFi-based indoor positioning techniques have been addressed previously, and have been found to benefit from good accuracy levels [32], the outdoor equivalent is not as accurate or addressed in the literature. Instead, common outdoor positioning techniques are performed using GPS or fingerprinting technologies, which have been shown to provide better location accuracy [13, 29]. This is one of the reasons why we decided to address the challenge of visually classifying supporters based on a large area outdoor WiFi-based positioning system. Even though the accuracy is not good, we assume it is possible to find strong relationships between devices or locations to reveal groups of supporters.

The data zones defined by Cisco and their respective locations are shown in figure 2.a. The orange points represent Arsenals fan zone and the yellow points represent Chelseas fan zone. As can be seen, the Olympic Way zone is not separated into two subzones nor groups.

![Fig. 1. Map showing Olympic way going to the main entrance and a picture of the Chelsea (blue) and Arsenal (red) supporters heading to the stadium before the FA cup final match, the 27th of May 2017. Credit: Guy Bell/Alamy Live News](image-url)
3.3 Methodology

Without labelled data, we focused on experimenting with a set of unsupervised techniques and to use a combination of analysis and visualisation to support our verification. Because the network is located in a public space and because of the inaccuracy that can occur to configuring networks for large open spaces, we avoid using the data points location estimations directly and instead focus on associated devices two or more devices collocated (same place, same time) and their relationship with either fan zone. Associated devices can be defined by the number of shared locations of two or more devices. The larger the number of shared locations, the greater the similarity between devices. Similarly, the association between two zones can be defined as a number of shared devices - the greater the number of shared devices, the more similar the zones. In both definitions, we assume that the supporters of the same team tend to spend more time in similar locations than supporters from opposing team. This is the case in English soccer matches because fans are generally segregated into different sections of the stadium. A group of supporters can be considered as a group of devices that share the same locations at the same time throughout the event. By the same reasoning, a group of nearby subzones sharing the same devices throughout the event will likely form a bigger zone dedicated to a team. This approach yields two adjacency matrices, the first or grid-cell matrix pools sublocations together (groups of grid-cells) based on the number of shared devices, whereas the second or devices-based matrix pools devices together based on shared locations.

Clustering the grid-cell matrix: This method consists in finding groups of grid-cells based on the proportion of shared devices between each pair of grid-cells. In other words, we cluster grid-cells which have in common a large

Fig. 2. (A) Raw data points split by their original zones. (B) Four groups of grid-cells resulting from the hierarchical clustering of the transformed grid-cells adjacency matrix between 11h31 and 16h59, taking into account the most 25% accurate devices’ location and the 20% strongest links between pairs of grid-cells.
proportion of collocated devices for a particular time span. The binary matrix (grid-cells X devices) is multiplied against its transposed to generate the square (grid-cells X grid-cells) adjacency matrix representing the number of shared distinct devices between each pair of grid-cells. Each cell of the square matrix is then divided by the total number of distinct devices that were identified at least once in either one of both grid-cells locations. This converts the absolute-based matrix into a proportion-based matrix where each proportion represents the percentage of distinct shared devices against the total number of distinct devices identified between the corresponding pair of grid-cells. We assume the bigger the proportion, the closer the pair of grid-cells and vice-versa. Two distinct directions are then tried. Firstly, the proportion matrix is converted into a distance matrix by subtracting the opposite of each proportion to one. Wards hierarchical clustering is then applied on the distance-like matrix [31] and the tree is cut into the desired number of groups (3). Secondly, the fast-greedy modularity optimization algorithm is executed on the proportion adjacency matrix. The latter interprets dense sub-graphs as communities. In this method, the algorithm converges to communities of which number is not provided as an input. The more relationships between a group of grid-cells, the more likely to be interpreted as a community by the algorithm [5].

Clustering the device matrix: This method consists in finding groups of devices based on the proportion of shared locations between each pair of devices. Unlike the previous approach, the device matrix contains the number of distinct shared locations between each pair of devices. By the same reasoning

![Fig. 3. Groups resulting from the fast-greedy modularity optimization algorithm applied on the adjacency graphs of shared devices between pairs of grid-cells (20 x 20 grid). The colours are not consistent in time. Only 75% of the most accurate data points (Confidence Factor first 0.75 quantile) from 8:00 to 22:00 with a minimum number of devices per grid-cell equal to its median.](image-url)
than previously the adjacency matrix is divided by the square matrix of weights containing the unique number of locations between each pair of devices. The resulting matrix therefore represents the percentage of distinct shared locations against the total number of distinct locations where the corresponding pair of devices were seen. We assume the greater the number of shared locations between two devices, the greater the likelihood that both devices belong to the supporters of the same team. Like previously, we apply the Wards hierarchical clustering on the distance-matrix with a tree cup parameter set to 4 in order to get four groups. Unlike the groups of grid-cells, all the devices cannot be directly mapped on a 2D plane as it would suffer from occlusion. Instead, the grid-cells are mapped and coloured as a superposition of the represented groups colours where groups proportion is encoded as the colour transparency. That way, zones where one group is highly represented would turn into the corresponding groups colour.

3.4 Results

Each approach reveals multiple clusters that reflects the topological configuration of the event. As illustrated in figures 2.b, 3 and 4 both fan zones, the Olympic way (vertically positioned) and Engineers way (horizontally positioned) are clearly visible. Four distinct zones are evident in figure 2.b, Chelseas and Arsenals fan zones appear respectively in green and blue. In addition, three vertical zones (light blue, red and grey) appear on Olympic way illustrating different groups separated by Police on Olympic way. The light blue and red areas represent Arsenal supporters and Chelsea supporters respectively. The grey zone is

![Fig. 4](image-url) Four groups of devices resulting from the hierarchical clustering of the transformed devices adjacency matrix between 11h00 and 16h59, taking into account the most 25% accurate devices’ location and the 20% strongest links between pairs of devices. The groups proportions are encoded with colours and transparency. The size of the grid-cells represents the number of distinct devices per location.
most likely noise. Figure 3 shows the event as a set small multiples with each image representing a specific time span between 8:00 am to 21:59 pm. Each time span has its own colour signature expressing dense and high levels of shared devices between grid-cells. The same colour between two small multiples does not necessarily represent the same group. Instead the colours are used to differentiate the groups of similar grid-cells. As can be seen, both fan zones are always defined with different colours and the Olympic Way is separated vertically in two parts from 10:27 to 21:59. This is coherent with Figure 3 and this supports our assumption that clustering the adjacency matrix can reveal subzones matching the environments properties. Similarly, the figure 4 shows the fan zones. It is interesting to note that only a small number of devices are located in between both fan zones.

3.5 Discussion

As evident from the visualisations, our approach yields some success. We can clearly identify groups that reflect the order and composition of crowds at the event. Even though there are limitations in terms of accuracy with WiFi-based location data, it was possible to reveal strong relationships between supporters and geographical areas using this technique. The method is flexible and includes several parameters such as the number of groups and the size of the grid-cells that can be used to address a range of different scenarios. In our case, we knew two major groups existed in the data Arsenal and Chelsea supporters - but there are also are subgroups whose spatial behaviour might be different.

4 Use Case 2: Accuracy analysis on large indoor spaces (supporting WiFi configuration)

In this section, we describe the second use case in which we analyse the accuracy of location-based estimations using RSSI values. As with the previous use case, there is a range of the applications of this approach although we focus on specifically on network configuration.

4.1 Outline of use case

The accuracy of geographic location is variable with many existing WiFi solutions. The accuracy of a position is based on a triangulation or estimation of devices location based on several access points and generally expressed as a measure of uncertainty in regard to that estimation. There are multiple parameters that can impact this accuracy such as the WiFi spatial configuration, the number of access points, whether the connection area is outside or inside a building and the chosen types of access points (in our case the provider is Cisco Meraki material and services) to name a few. Comparative methods evaluating accuracy have been studied for unipersonal data points and trajectories using the ground truth [20, 32]. However, there has been little work on developing and evaluating
techniques based on aggregated data whereby the location is determined based on an aggregation from multiples devices.

The locations estimation accuracy is generally based on the raw RSSI values provided by the original probe requests and defined as an average error distance when the dataset is aggregated [17]. In this use case, we specifically have access to a confidence factor (CF) directly deriving from the triangulation between multiple probe requests RSSI values.

Generally, zones locations that exist within a specific geographic area are used to support location based analytics. Often the topographical configuration of the building as a room, a floor, a street portion, a fan zone, a road intersection and so on are used to define the zones. Although this makes sense in a lot of different situations, it often hinders the accuracy of the location and is not the optimal way to configure the network. Zone analysis can evaluate the configuration and possibly suggest better configuration of the network.

Without ground truth information, accuracy cannot be evaluated with high level of precision. This use case is an illustration of how to perform a preliminary zones accuracy analysis based on a distance indicator, which in our case is the confidence factor. We suggest two heuristic approaches, one visual and one analytical. First, a level plot is used to map the estimated subzones accuracies and better understand its variability across one zone. Second, a zones accuracy metric is defined as a numerical aggregation of a zones ability to contain its estimated points. Both tools are useful to report and compare the accuracy of the points location estimations, as well as to support the zones configuration process and its related location analytics efficiency.

4.2 Data Description

The dataset used for conducting the accuracy analysis was generated during the ESC congress 2017 which took place in the Barcelona Fira conference centre. As can be seen in figure 5, the centre is a multi-room building of which we focus on the two biggest zones called hall 2 and hall 3. The dataset consists of 48 hours of

![Fig. 5. The red data points are located inside Hall 2 and Hall 3 of the Barcelona Fira Gran Via conference centre. The black points are located outside the Hall 2 and the Hall 3.]
data collected from the 26th to the 27th of August 2017. Each record contains the estimation of the devices location (latitude, longitude), a confidence factor, the locations zone label and a time stamp. All the information except the time stamp are used in the following accuracy analysis.

4.3 Methodology

Two distinct approaches are used for conducting the accuracy analysis. First, data visualisation is used as a tool for plotting the levels of accuracy confidence on a map. Second, the topological properties of each data point are used to compute an aggregated statistic expressing the ability of a zone to reflect reality.

Data visualisation and more specifically level plots can be used as a tool for analysing subzones accuracy. Heatmaps using bilinear interpolation is performed as a mean for comparison and evaluation in [27]. Similarly we perform bilinear interpolation in R using the akima package [3] to map the confidence factor on a 2D plane. Mapping and interpolating the points confidence factor is a way to visualise precisely the levels of uncertainty in the locations estimation. Parts of the room might show high levels of confidence factor which can be interpreted as places where RSSI signals are weak and vice versa.

Hall 2 and Hall 3 zones are randomly sampled down to approximately 10 thousand records each. Both samples are used separately to compute the corresponding level plots. The figure 6 shows the Hall 2 and the Hall 3 confidence factors level plots. The greener the subzone the more accurate the locations estimation and the more likely the points estimation to be located there, the redder the subzone the less accurate the locations estimation and the less likely the points estimation to be located there. Both level plots show similar levels of accuracy. Most of both halls show levels of confidence factor between 0 to 50 meters where accuracy is the best in the room. However, both halls border subzones show more inaccuracy with levels of confidence factor between 100 to 150 meters. The estimated devices locations in those areas cannot be considered reliable as it is. Finally, small areas of both zones show high levels of inaccuracy with levels of confidence factor between 250 to 350 meters.

The zones accuracy indicator is a probabilistic measure of a zones ability to match each points inner and outer states between the points location estimation and its unknown ground truth. To reflect reality, a zones estimated inner-points should maximise their probability of being real inner-points, and the estimated zones outer-points should maximise their probability of being real outer-points. Based on this observation, we assume that the accuracy of a zone is closely related to its inner and outer points average probability to be located inside the zone. More precisely, the bigger the zones inner-points average probability to be located inside the zone and the smaller the zones outer-points average probability to be located inside the zone, the better the zones accuracy.

\[
\max(\text{ZoneAccuracy}) = \max(\max(P(\text{Inside Zone}|\text{ZoneInnerPoint})/\max(P(\text{Inside Zone}|\text{ZoneOuterPoint}))
\]
We define uncertainty square as the surface where an estimated point is located within a 95% level of confidence. Two points uncertainty squares are illustrated in figure 7. Following this definition, we assume that the probability of a point to be located inside a zone can be computed as the proportion of the uncertainty squares surface which is located inside the zone. As illustrated in figure 7, 60% of the inner-points uncertainty square resulting from its confidence factor is located inside the zone. This can be interpreted as a 60% likelihood of being a real inner-point and 40% likelihood of being a real outer-point. Similarly, 35% of the outer-points uncertainty square resulting from its confidence factor is located inside the zone. This can be interpreted as a 35% likelihood of being a real outer-point and 65% chance to correctly be defined as an estimated inner-point.

**Fig. 6.** Level plot of the Fira Barcelona Gran Via - Hall 2 and Hall 3 - over 48 hours of conference

**Fig. 7.** Example of an inner-point and an outer-point. The estimated inner-point has 60% chance to correctly be defined as an estimated inner-point and 40% to wrongly be defined as an estimated outer-point. The estimated outer-point has 65% chance to correctly be defined as an estimated outer-point and 35% chance to wrongly be defined as an estimated inner-point.
a real inner-point (instead of an estimated outer-point) and 65\% likelihood of being a real outer-point.

Finally, we assume a zones aggregated accuracy indicator can be defined as the difference between both inner and outer points average probabilities to be located inside the zone. The more accurate a zone, the bigger the average probability of its inner-points to be located inside the zone and the smaller the average probability of its outer-points to be located inside the zone.

\[
\text{zoneAccuracy} = P(\text{InsideZone}|\text{ZoneInnerPoint}) - P(\text{InsideZone}|\text{ZoneOuterPoint})
\]  

(2)

Because inner and outer points are computed separately, data points from all zones are needed. Firstly, the data points identified in different floors are filtered out and then random sampling is performed on the dataset so that approximately 20 thousand data points are used in the analysis.

After every points uncertainty squares are created, the proportion of each squares surface located inside the selected zone is computed. As previously explained, the resulting proportion is interpreted differently whether it is an inner or an outer point. Indeed, the bigger the inner points surface being located inside the zone and the lower the outer points surface being located inside the zone, the more likely the estimation to be true. The average uncertainty squares proportions of inner points and outer points located inside the zone are computed separately. As can be seen in Figure 9 and Figure 10 a large proportion of inner-points squares are fully located inside the zone and a large proportion of outer-points squares are not located inside the zone.

Fig. 8. Both distributions of inner and outer points squares’ proportions located respectively inside the Hall 2 and the Hall 3.
Finally, the aggregated zones accuracy indicator is computed as the difference between the inner and the outer points averages. The bigger the zones accuracy indicator, the more likely the estimated points locations to be truly inside the zone, and vice versa.

4.4 Results

Both halls confidence factors level maps (Figure 7) allow to compare both zones together as well as each subzone accuracies. It provides a good indication about places where the estimated location is uncertain and compromised. Moreover, both halls aggregated accuracy indicator was computed based on each records confidence factor. We can say, based on this metric, that Hall3 (zone accuracy = 82.45) outperforms Hall2 (zone accuracy = 78.87) on its ability to match each points inner and outer states between the points location estimation and its unknown ground truth.

5 Conclusions and future works:

This project is an attempt to develop methodologies to approach and analyse spatial data and more specifically crowd sensing, group identification and metrics for computing estimated location accuracy. It goes towards an extended evaluation-oriented study and started with two use cases mainly relying on our intuition and detail analyses as data scientists and visualisation engineers when facing such problems. It meets application-oriented requirements and it can be valuable to practitioners manipulating large and uncertain datasets. We believe those type of analyses will become part of the smart cities developments as the latter will generate bigger, cheaper and probably noisier datasets based on WiFi signals which now are ubiquitous and globalised.

Our next research project will measure the preciseness of our clustering method using a labelled dataset. This will allow to evaluate quantitatively the ability of the adjacency matrix method to identify groups of supporters in large open-spaces outdoor football events. Moreover, we plan to provide some guidance on how to fine-tune the clustering algorithms to optimise crowd sensing and increase the likelihood of discovering groups in the crowd.

References


