A Footfall Density Estimation Method for Location-Based Targeting Advertising using H3 Geospatial Indexing System

Name: Eleftherios Sergios

Student No: 14045828

Supervisor: Dr Ana Basiri

BENVGSC6: MSc Smart Cities and Urban Analytics

Institution: UCL

Date: 30 August 2019

Word Count: 10,144

This dissertation is submitted in part requirement for the MSc in the Centre for Advanced Spatial Analysis, Bartlett Faculty of the Build Environment, UCL.

Abstract

In this dissertation a novel workflow estimating footfall density in urban areas is designed based on the H3 Geospatial Indexing System. H3 is a hexagonal grid with multiple levels of granularity that covers the entire surface of the Earth. Its grid cells are predefined unique hexagons whose individual characteristics are pre-calculated e.g. hierarchical indexes, areal unit areas, distances from each other etc. Assigning H3 hexagons to geographic data defines geospatial indexing and results in the creation of a global geodatabase that allows for large-scale fast and efficient geospatial computations. Two location data sets of Global Positioning System (GPS) logs contributed by mobile users in June 2019 and Points of Interest (POI) are used. Both datasets contain point data within the metropolitan area of Greater London. The geographic coordinate pairs of these points are assigned to unique H3 grid cells as a result of the geospatial indexing discretization process. Supervised random forest regression is applied to both datasets to forecast the footfall density across the H3 grid. Implementing this machine learning technique results in achieving an estimation accuracy of 88% with the average absolute error of 2 GPS events when predicting footfall density on an average day of the week. Our results prove that H3 grid cell aggregated trajectory data, in combination with POI, can be used to estimate and potentially predict footfall density more efficiently than other methods using different grids. One could argue that this study is an important step towards a more comprehensive treatment of strategies aimed at model optimization of trajectory mining, analysis and forecasting.

I, Eleftherios Sergios, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis. It is 10,144 words in length.

Signed: 

30 August 2019

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

Ad-Tech	Advertising Technology
DGGS	Discrete Global Grid System
EU	European Union
GIS	Geographic Information Systems
GLA	Greater London Authority
GPS	Global Positioning System
LBA	Location-Based Advertising
LBS	Location-Based Services
MAE	Mean Absolute Error
MAEP	Modifiable Areal Unit Problem
MBR	Minimum Bounding Rectangle
ML	Machine Learning
POI	Point of Interest/Points of Interest
R&D	Research and Development
RoI	Return of Investment

Introduction

Today, the massive usage of Global Positioning System (GPS) enabled devices brings a great opportunity to have large volumes of location data to extract useful and insightful information about the mobility of people. These can be used by different fields and disciplines, be it retail, academia, private research and development (R&D), as well as online targeting advertising. For advertisers and more specifically for the specialized field of Advertising Technology (Ad-Tech), location data is used to analyse users' movement traces and their spatio-temporal patterns that can tailor information to be included in the advertisements to be delivered at the right time, in the right place and to the right person. In this way, the mobile users can be informed about more relevant products and services, when and where they are thought to be interested in.

Despite the great potential of utilizing location data, little logic and investigation has been implemented around targeted spatial and location based advertising (van 't Riet et al., 2016). This is due to the complexity and size of user location data, which limits the advertising companies, hence most advertisements still follow the traditional practices (Hühn et al., 2017). As there is room for improvement in the mobile location targeting, this dissertation aims at proposing and implementing a workflow that could help practitioners understand existing and future footfall density within a city i.e. how people move around and where they will be in different times of the day and week, in relation to the nearby POI.

To visualise and explore footfall density consistently and dynamically on a global scale, a framework of regular tessellation is required for geospatial

indexing of our data. Such frameworks are called Discrete Global Grid Systems (DGGS) and are used for data representation based on their geocoded grid cells. Among grid regular shapes that tessellate, although squares and triangles have been preferred in the past, lately hexagons have gained the interest of researchers and professionals in the field of Geographic Information Systems (GIS) (Uher et al., 2019). That is mostly due to hexagon's equal uniform distance between neighbouring hexagons and higher representational efficiency as it achieves better circle approximation and less shape distortion than the other regular geometric shapes (Sahr, 2019). Therefore, the H3 hexagonal grid provides a more approriate platform for analysing dynamic phenomena of trajectories.

The contribution in methodology within this research, allows to shed light upon tailored Location Based Advertising (LBA) strategies and help Ad-Tech industry unlock audience opportunities. After examining three different types of geospatial grid indexing systems, Geohash (Balkić et al., 2012) , Google S2 (Ekawati and Suprihadi, 2018) and Uber's H3 Hexagonal Hierarchical Spatial Index (Brodsky, 2018), we choose the latter as the most suitable one for its minimum quantization error, maximum geometric cell shape preservation and uniform distance between adjacent grid cells.

Research Goal

Although prior literature has explored the potential of using movement data for evaluating the success of delivering advertising campaigns (Zubcsek, 2017), the methodologies for forecasting prospective footfall around POI has not been thoroughly examined (Rosenkrans and Myers, 2018).

The three main objectives of this research are (a) the discretization of location data, POI and trajectories within the urban contexts, (b) predicting the footfall density using GPS mobile data event counts within each areal unit (hexagon) of the H3 Geospatial Indexing System, and finally (c) forecasting the changes in the densities over time and predicting the footfall change rates across the city. These objectives address the research hypothesis of this dissertation which investigates the suitability of the H3 geospatial indexing system as a framework for analysing and forecasting point densities on a city scale, as well as the contribution of the POI to density estimation.

Since previous research around methods of implementing and evaluating footfall density predictions on discrete global grids is limited, the main purpose of this research is to build a model for analysing POI and accurately predicting their nearby footfall densities using the hexagonal geospatial indexing system of H3 with its discrete areal unit structure and global functionalities.

Data

In order to identify these movement patterns and forecast footfall densities within the H3 grid, two main datasets are used, GPS events and POI.

GPS Events

This study conducts experiments on a GPS dataset of mobile user events provided by the advertising technology company Blis (“Blis | Mobile Location and Behavioural Advertising Solutions,” n.d.), over one month, i.e. 1st to 30th April 2019. The dataset geographically covers a rectangle bounding the administrative areas of Greater London accessible on GADM (“GADM,” n.d.). The Minimum Bounding Rectangle (MBR) with Latitude, Longitude coordinates of top left corner at 51.721924, -0.725098 and bottom right corner at 51.207745, 0.435333 is shown in figure 1.

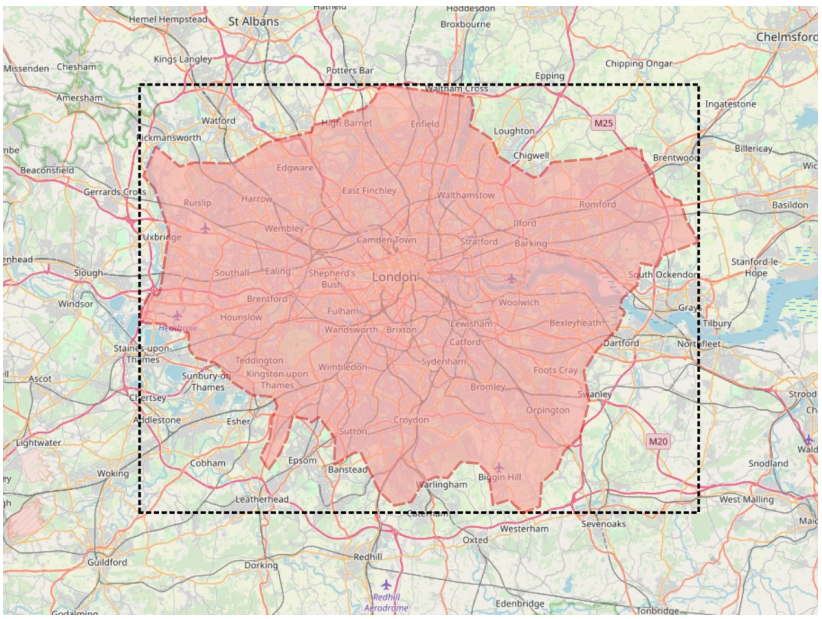

Figure 1: The rectangle around the boundary of Greater London.

POI

The second main dataset used is the POI which represents the physical world and regarding advertising, is highly beneficial to investigate their distribution across the city. These are the places that one would expect prospective customers to be visiting in high frequency i.e. high POI density could potentially lead to high footfall density.

Our POI dataset was downloaded from Digimap (EDINA, 2019) and contains 357,081 points within a MRB of top left Easting, Northing at 454800.0, 220799.999 and bottom right at 605200.0, 145199.999 in British National Grid projection. Since the study focuses on London, the points outside of the administrative boundary of the metropolitan area of London are discarded along with the ones that belong to categories and subcategories not relevant to the objective we address in this project.

The POI classification scheme used in this project is designed by Ordnance Survey (“POI,” n.d.) and has 9 main categories:

- accommodation, eating and drinking
- commercial services
- attractions
- sport and entertainment
- education and health
- public infrastructure
- manufacturing and production
- retail
- transport

The final number of POI is 121,978 and a sample of them can be seen in table 1.

Name	group name	lat	lon
Tennis Courts	Sport and entertainment	51.494961	-0.280982
Superdrug Pharmacy	Retail	51.376985	-0.100763
Perfect Pizza Ltd	Accommodation, eating and drinking	51.520803	0.018691
Sayah Trading Services Ltd	Retail	51.580659	-0.083146
Smooth & Simple	Accommodation, eating and drinking	51.522476	-0.071536

Table 1: Sample of POI data from Digimap (EDINA, 2019).

Having the number of POI from Digimap relatively higher than other open data sources such as OpenStreetMap (“OpenStreetMap,” n.d.) whose equivalent POI number for this project was calculated at only 26,377, reassure us to use our POI as a valuable dataset to predict footfall density within H3 cells.

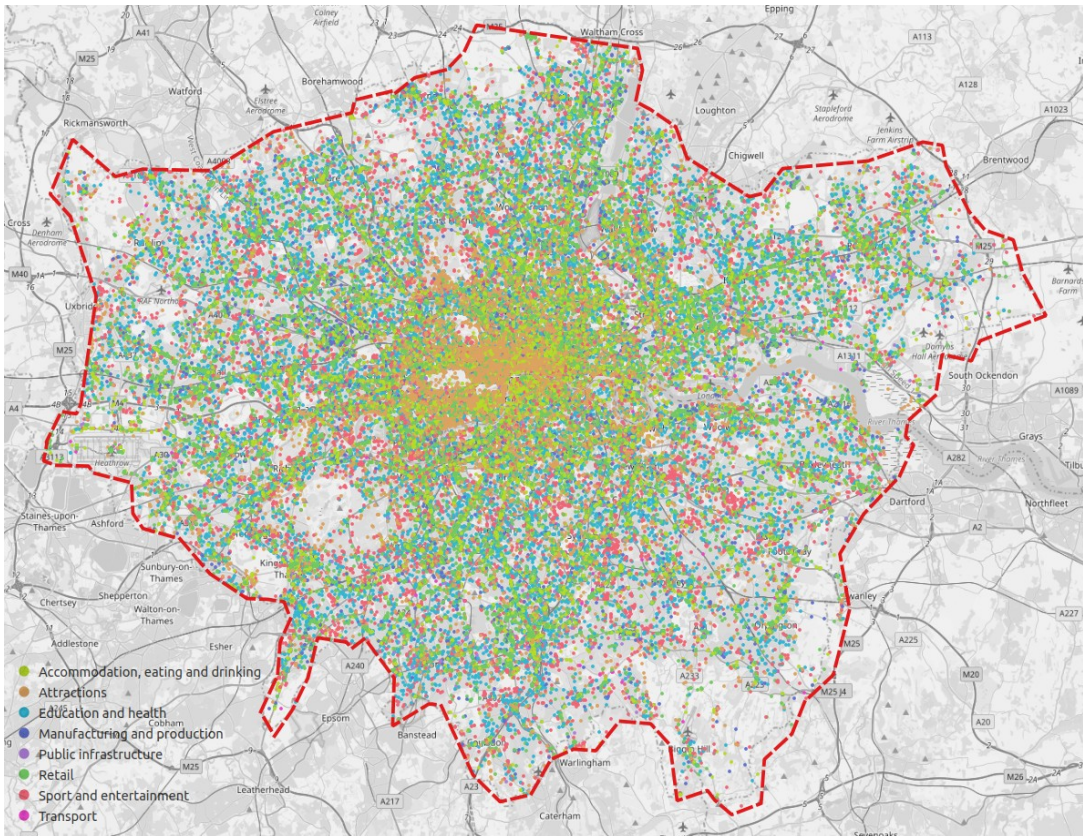

Figure 2: All 121,978 points of interest within the administrative boundary of Greater London in red dashed line. The POI category of “commercial services” was not included in this studv.

Methodology

The steps in this study are described as follows:

- In order to study and understand the principles and logic applied to similar methodologies, we first present a few projects related to analysing and predicting discretized point densities in grid converted urban areas.
- In the next step, we elaborate on the field of study of this project which is location-based targeting advertising. Important questions are being addressed such as, what is it about geo-targeting and makes it so special? How is it being used today, and why would footfall estimation method be useful to this field?
- Then, we describe the nature of trajectories as type of data and its characteristics, shedding light on the unique insights we can generate related to people movement and potential technical problems associated to data cleaning methods and quality.
- In the next section we discuss about geospatial indexing as a process of converting spatial data into discrete forms and compare three different discrete global grids, Geohash, Google, S2 and H3. We examine the available literature, online sources, technical documentation and user comments from web forums to understand how they work, how they can be used and both their advantages and limitations. Later, we evaluate their performance by testing them on three different cities, London, San Francisco and Sydney.
- In the following step we apply the H3 grid on our location data and discretize the set of POI and GPS events. Each point of both datasets

gets assigned to a unique H3 cell, allowing us to calculate densities for both POI and GPS events.

- At the final part of our project, we present our preferred machine learning (ML) technique, that of random forest and discuss the reasons behind choosing this specific one for addressing the research goal. The part of explaining the model building process follows, along with the results.
- In the conclusion, we share our thoughts on the findings and suggest future work and method development.

The technical workflow using H3 can be seen in the diagram below:

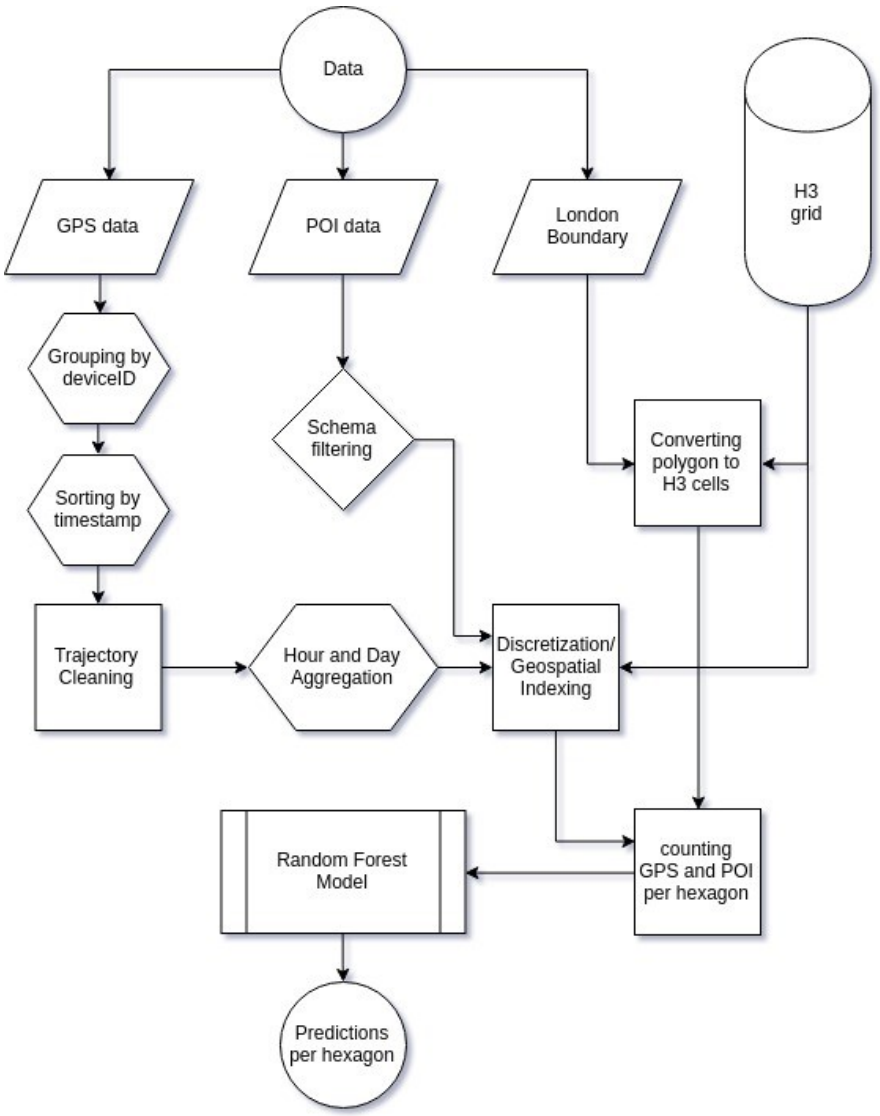

Figure 3: Technical workflow diagram using H3.

Related Work

Regarding previous research literature, some examples are briefly mentioned related to point density prediction methods using grids.

In a project from Harvard Data Science (“NYC Taxi Data Prediction,” n.d.), a team of researchers attempted to predict taxi pickup density across New York City with taxi trip data from January 2013 to June 2015 using Geohash as the grid to discretize the pickups, and random forest as the machine learning method to achieve prediction from day to day and hour to hour. The principles are quite similar with our research since there are trajectory movement data, that of taxi trips which are being bucketed into unique grid cells, that of Geohash and later forecasted via random forest. The same exact approach another group of students from California State University, Sacramento followed to tackle a similar objective to the previous team in Harvard which is predicting taxi demand across New York City (“Taxi Demand Prediction System,” n.d.). There is a great number of projects and various methodologies related to taxi supply-demand forecasting proposed, most likely due to the availability of open taxi trip data in New York City from FOIL (“Committee on Open Government,” n.d.). Using sample of the same datasets, Davis et al. (2018) attempted to address taxi demand forecasting by comparing the performance of two tessellation systems, that of Geohash and Voronoi.

Regarding footfall destiny, a different team used mobile cellular tower data to approach count estimation by designing two custom grids, a Voronoi and a hexagonal, of which each hexagonal cell covered an area of radius 150 m, a

typical size in tessellation systems for cellular networks (Ng et al., 2017). Working with mobile user data, another group of researchers explores the potential of crime concentration prediction within London by utilizing the telecommunications company Telefonica’s footfall data and grid system, designed to support its product, 'Smartsteps' (Bogomolov et al., 2014). Having data aggregated in Smartsteps’ grid cells, crime prediction is treated as a binary classification problem and addressed via using random forest as the main machine learning technique.

From examining these projects, we can clearly see the value of trajectory data discretization through the utilisation of a tessellation grid system in studying various methods of density forecasting within an urban context.

Location-Based Targeting

With mobile devices such as smartphones and smart wearables being used extensively on a worldwide scale, the mobile marketing industry has exploded with location targeted advertising spending to reach over $32 billion in USA alone by 2021 (BIA/Kelsey, 2017). Companies have realised that with so much mobile data they are able to segment audience and send users mobile ads when and where they are more receptive, in order to increase their earnings. People’s behavior and psychological state can vary with respect to the place they are located, e.g. a place can make a person more open to new ideas and information (Dholakia and Dholakia, 2004).

It is worth highlighting that not only the use of mobile devices has increased over the years but also users’ receptivity to impressions, suggestions etc. (Fong et al., 2015). It has been proved that people are more receptive to information when that information is related to their physical context, and hence more likely to engage with advertisements regarding places close to their current location as it could generate interest in the product or service due to users’ proximity to it (Fong et al., 2015)

According to Bauer and Strauss (2016), LBA has been used for quite a long time since the very first road billboards, which were occupied with location-based content such as driving instructions of how to get to a shop nearby or the distance from the closest McDonald’s restaurant on the motorway. LBA is the 'marketer-controlled information specially tailored for the place where users access an advertising medium' (Bruner and Kumar, 2007, p.3).

The movement data of every day shoppers and prospective consumers is very important and useful to the advertising industry as it reveals retail patterns and it helps in new site selection. Counting the number of individuals within an area and within a period of time is called '*footfall density*' (Sundararaj, 2017), and it renders the foundation of retail competition among different companies trying to convert as many shoppers as possible into their customers. Prior literature has shown that connecting with people not only at the right moment but also at the right place increases advertisement response because individuals appreciate more the opportunities that can be exploited almost instantly (Prelec and Loewenstein, 1991). Therefore, it comes naturally that professionals in the field are willing to exploit all possible benefits of mobile technologies including especially that one of targeting users both by location and time (Ghose et al., 2012).

Despite the importance of audience geobehavior, nowadays, most of the mobile marketing campaigns are based on empirical knowledge shared among professionals. In terms of LBA, usually the goal is to target potential consumers around a physical store to inform them about a product and convert them into the store. But which place out of all possible areas in the vicinity of a store is the most suitable one to target when there is no indication of upcoming audience population? As Maniaka (2019), an advertising operations manager from location-based advertising company Blis explains:

> *for an ad campaign the site selection around a POI is implemented based on personal intuition and experience. Sometimes it works but not always because no one is really aware of how the footfall around a shop will be spatially distributed. If we knew we would target precisely*.

In this regard, predicting people's locations could help practitioners analyse retail catchment areas, and study the competitive effects of the distance and the attractiveness to increase the revenue through increasing the probability of delivering successfully advertisements. The concept is basically simple and focuses on the distribution of visitors at proximal distances to a POI. If only there was an indication of the approximate areas of high footfall density around a POI, then these areas would be precisely targeted, where interest in the respective amenity information would be higher due to user's close location (Luo et al., 2014). Therefore, we could say that the number of prospective shoppers that would visit the advertised amenity depends mostly on footfall density around that specific amenity. Approximating that footfall density would encourage attraction which then would be followed most likely by effective conversion (Fang et al., n.d.).

Applications related to spatio-temporal targeting studies are related to contextual marketing theory which looks at the impact of emerging location technologies on consumer behaviour and highlights that nowadays context is more valuable than the content (Kenny and F Marshall, 2000). In practice, by leveraging on footfall data to identify potential target audiences, companies can increase conversion rates and improve the Return on Investment (RoI) of their advertising campaigns. As a result, product and service information can get delivered successfully to the right recipients and all parties get benefit from it. Companies will be able to multiply their earnings and consumers receive appropriate instructions about where to go and what to purchase.

Trajectories

The footfall data required in this study is a product of human mobility data, generated from smartphone devices and represented in the form of trajectories. The advancement of mobile technologies has made possible the extraction of massive amount of spatio-temporal data and hence the location-acquisition of trajectories in scale. As depicted in figure 4 and defined by Zheng (2015, p.29),

> *a spatial trajectory is a trace generated by a moving object in geographical spaces, usually represented by a series of chronologically ordered points, for example, $p^1 \rightarrow p^2 \rightarrow \cdots \rightarrow p^n$, where each point consists of a geospatial coordinate set and a timestamp such as p = (x, y, t).*

Our GPS points follow the common type of trajectory data, being that of device ID, latitude, longitude and timestamp. Grouping the points by device ID and sorting by timestamp generates a sequence of GPS points which forms ultimately the trajectories of mobile users.

Figure 4: The segments of a single trajectory with its nodes in green.

In table 2 the four main columns of GPS raw data are shown with timestamp format following the Unix Time i.e. counting the seconds after the 00:00:00 Thursday, 1 January 1970 (“Unix time,” 2019)

deviceId	latitude	longitude	timestamp
[illegible]	51.630932	-0.058942	1555628400000
[illegible]	51.496001	-0.167009	1556305200000
[illegible]	51.506063	0.280654	1555531200000
[illegible]	51.515471	-0.162692	1556319600000
[illegible]	51.516989	-0.140714	1555812000000
[illegible]	51.507906	-0.124209	1556136000000
[illegible]	51.547948	-0.167446	1555941600000
[illegible]	51.390600	-0.091900	1555070400000
[illegible]	51.580014	-0.202522	1555110000000
[illegible]	51.583000	-0.020600	1555830000000

Table 2: Some columns of a sample of raw GPS data points with blurred deviceIds for privacy reasons.

As shown in figure 5, the difficulty level of extracting meaningful insights from trajectories increases dramatically when the number of these traces increases.

Figure 5: Randomly picked 10,000 trajectories in Southern England.

In terms of preserving privacy, researchers need to be very cautious about what kind of information they disclose when analysing trajectory data. Despite signal quality, movement traces can reveal different kinds of sensitive information about an individual's personal life in a very detailed manner (Marc Langheinrich author, 2019) e.g. where one goes for shopping, which gym he/she attends and most importantly the location of someone's residence. These cases are related to either *'identity linkage'* or *'attribute linkage'*. The former focuses on matching a person to his/her immediate trajectory because it can be unique and characteristic, whereas the latter on identifying a person via inferring his/hers identity based on the trajectory (Chen et al., 2013). Especially when it comes to publishing such data, like in this study, some attributes that may accompany movement data should not be shown, like the deviceId in table 2, as a result of data privacy and protection concerns ("2018 reform of EU data protection rules," n.d.).

The sources to capture people's trajectory data are diverse. However the main one and most popular is the smartphones (Timašjov, 2016). Most of the current mobile devices are GPS-enabled allowing the users' coordinate locations to be captured (Roy and Pebesma, 2017). Recording trajectories can be either active, when users themselves report their own locations, for example via check-ins in a social location-based network like Facebook, or passive via the smartphone's enabled background location logging (Zheng, 2015).

This research uses GPS data stamped, due to several reasons, including being free-to-use and global availability as there are GPS satellites around the planet

(Paul E. Ceruzzi author, 2018) providing relatively good level of accuracy i.e. ranging from 3.4 m to 24.4 m in the City of London (Adjrad and Groves, 2017). Determination of the exact GPS location is measured by the reception timings from the navigation satellites to the receiver antennas (“GNSS Frequently Asked Questions - GPS,” n.d.).

While GPS is free-to-use, globally available and a privacy preserving system, the quality of the GPS signal depends on the physical context where it is received. For example in an urban landscape, populated by buildings and other city structures, the signal can be blocked, reflected, or attenuated (Zheng, 2015). Also, the accuracy of the positioning solution can vary depending on the position of the satellites broadcasting the signals, generating invalid signal triangulations as seen in figure 6.

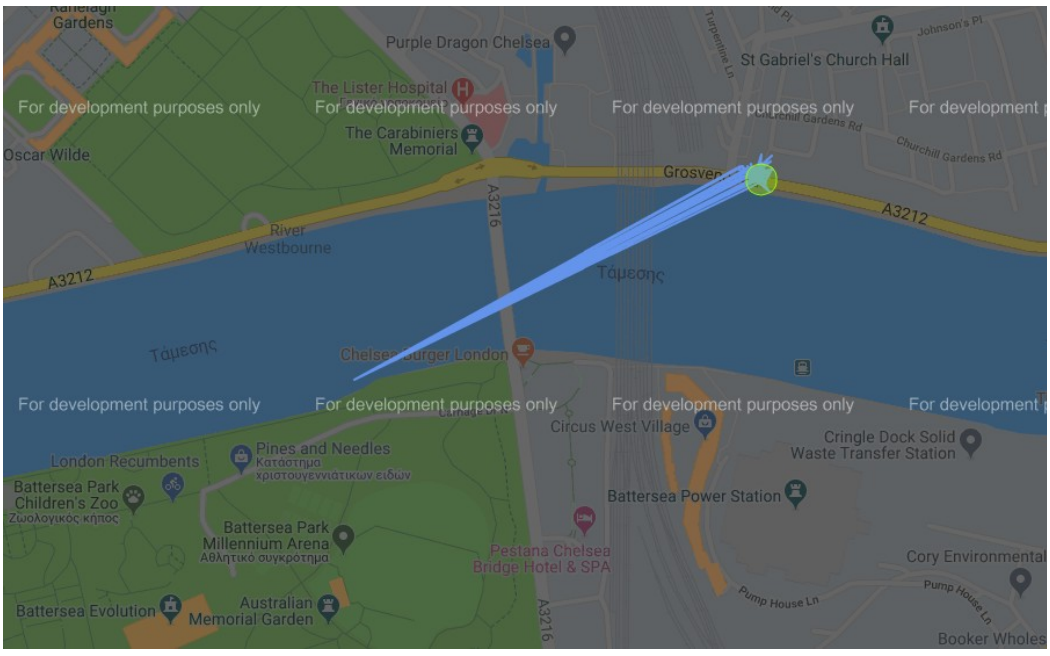

Figure 6: A trajectory's noise point that falls on the water.

Trajectory Cleaning

In our study, there are in total 8,589,632 GPS points of 1,071,401 unique device IDs. That means that the dataset contains trajectories of almost 1 million people. However, this set of trajectory data needs to be carefully evaluated, processed and cleaned. The first stage is filtering data by setting some thresholds including the travel speed of pedestrians (calculable between the subsequent GPS points) that can remove outliers if it exceeds the maximum speed of 20km/h (Basiri et al., 2016). These thresholds are used either to filter out noise points or approximate their real position. In the second stage multiple common assessment criteria are applied on the remaining trajectory data, as seen below.

- Point Lacking of Precision: latitude, longitude values with less than 3 decimal points are rejected due to inaccuracy, as they can not cover distances smaller than 110 m which is too big for the purpose of this study (“Decimal degrees,” 2019).
- Point on Water: latitude, longitude does not intersect with any country and the point is located on water.
- Point on either Prime Meridian or Equator or Both: all longitude values are 0.00 falling on the Greenwich line (Prime Meridian) or all latitude values are 0.00 falling on the equator line or both coordinates are 0.00. This absolute zero phenomenon is perceived as invalid signalling.
- Empty device ID: when a GPS point does not contain a unique device ID string and thus is not used to form a trajectory as it can not be grouped by device ID, as explained earlier.

Finally, from the remaining cleaned GPS points the ones within the study area of Greater London are selected, resulting in 7,649,570 points of 979,987 unique device IDs. Due to the dataset's size only a sample of these trajectories is visualised in figure 7.

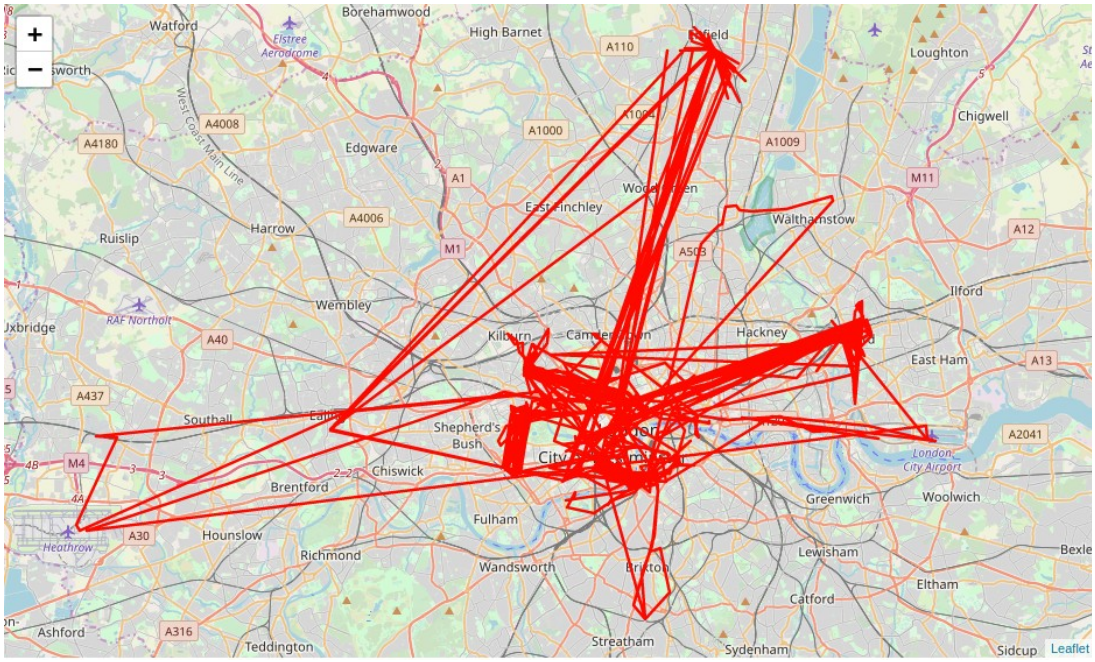

Figure 7: Sample of the cleaned trajectories. Due to data privacy we cannot zoom in further to illustrate the cleaned GPS trajectory data.

Geospatial Indexing

Integration of spatial data has become one of the most challenging tasks in geo-computational analysis (Peterson, 2017). Traditionally, location data along with their spatial metadata features are stored and used for different GIS purposes (Mahdavi-Amiri et al., 2016) or geo-analytical pipelines (Longley, 2005). However, due to the high volume (Leszczynski and Crampton, 2016), complexity and heterogeneity of this geospatial data (Goodchild and Haining, 2003), it has been difficult to put in place analytical workflows reasonably maintainable and scalable and as a result, companies are not able to easily leverage their location data (Zhao et al., 2016). Filtering geographic observations through a process of discretization could potentially help researchers to aggregate them into areal units and more efficiently run analysis (Cheng and Adepeju, 2014).

These areal units could be arbitrary boundaries, drawn for policy making purposes, census tract, service provision and many other administrative purposes. Such boundaries could partially address the geospatial challenges, but still when running data analysis on a city level the urban contexts may vary significantly. The administrative area polygon boundaries of a city, although practical for multiple reasons, can vary in many ways, e.g. the area or shape. This change, which is potentially a change over time, may result in several further challenges (Longley and Tobón, 2004), including lacking of a common ground for spatial data integration. Even if there was a city in the world whose boundaries were homogeneously designed, it would be hard to compare its analysis results with another city as their boundaries would probably be different

with no comparable shapes and sizes, adding in that way a global scaling complexity. For example, looking at the boroughs of London and New York City (see figure 8), it is clear how their boundaries are different in number, size, area, and hence not allowing for comparing data assigned to them.

Figure 8: The boroughs of London on the left and boroughs of New York City on the right are completely different in shape and size.

When it comes to challenges on a neighbourhood scale, Uber's engineer Joseph Gilley (Uber Engineering, 2018a) explains the consequences and problems of arbitrary boundaries. In high volume events like holidays, concerts or sport matches the demand for rides increases dramatically during such occasions, and for Uber, the supply of available cars for dispatching must meet this demand. To achieve that, Uber introduced geofenced surge pricing by identifying areas of localised demand so drivers are incentivised to visit these areas to pick up customers. Although, for entire city regions it was helpful, it could not serve as a solution for highly localized demand within neighbourhoods. Due to these arbitrary geofenced boundaries, there could be a surge cliff with high customer search in one area while on the other side of the

boundary line low search (see figure 9). This phenomenon is caused by the so called '*boundary effect*'.

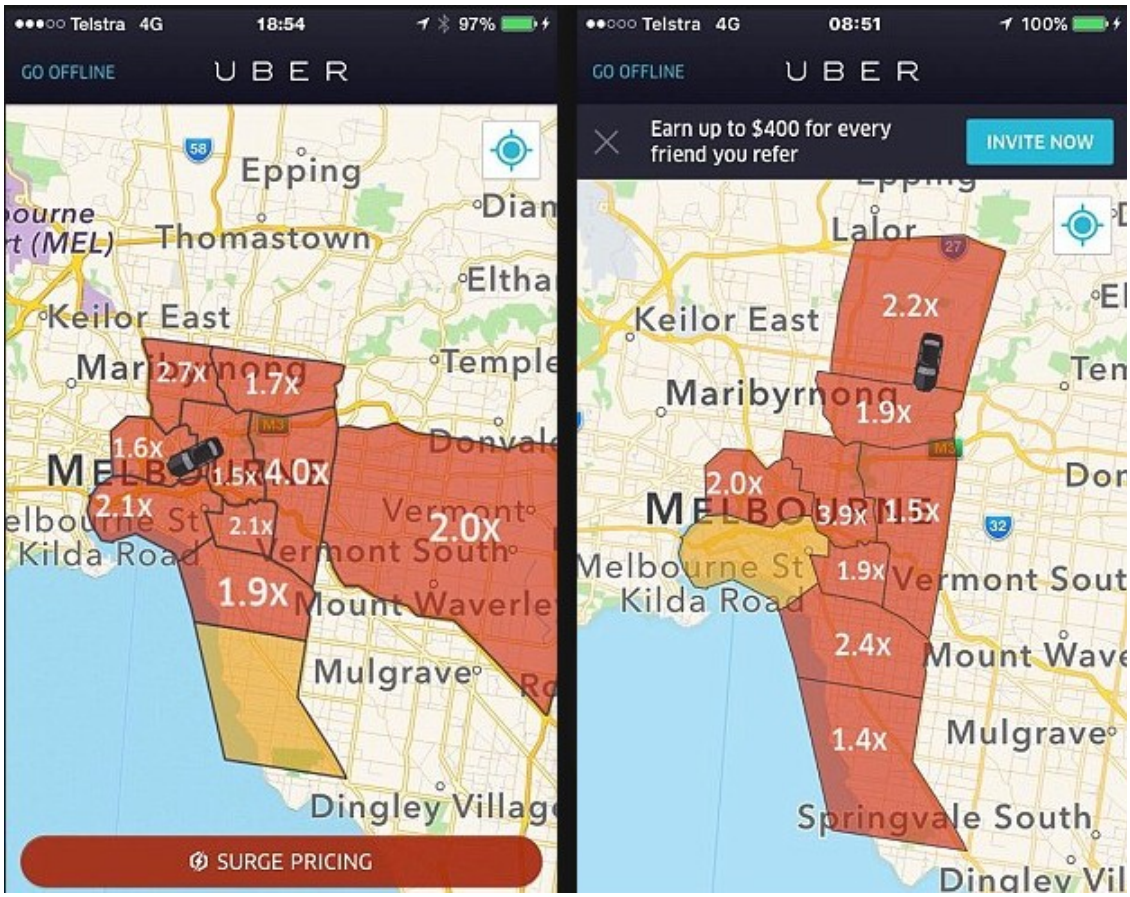

Figure 9: Surge pricing images from Uber platform before H3 implementation showing arbitrarily geofenced areas that caused the boundary effect ("Uber's surge pricing secrets revealed as customers brace for price hikes over Christmas | Daily Mail Online," 2015).

Boundary Effect

By drawing or using predefined boundaries around areas of interest, the geographic entities that can get affected are points. Therefore analysing spatial point patterns within such regions is likely to cause the problem of boundary effect which is examined through the *'edge effect'* and *'shape effect'* (Fotheringham and Rogerson, 1993). The edge effect is related to breaking the

dependences with what takes place just outside the drawn boundary (Griffith, 1985), without necessarily implying that data within the area of study cannot cross the border lines (Fortney et al., 2000) or get affected by factors outside these lines (Van Meter et al., 2010).

Regarding the shape effect, it derives from the arbitrary artificial shape of a boundary which can also affect statistical calculations i.e. it is more likely to create denser similar point clusters within an area when its unit is more elongated (Fotheringham and Rogerson, 1993). Especially in cases of connecting points where movement is implied, like trajectories or flow patterns, Isard (1960) states that even if boundaries share similar size, the difference in their shape can create bias, thus making them incomparable. In his research, migration rates between Ohio and Tennessee are not suitable for comparison because despite their equal sizes, the latter's boundary is much more elongated allowing for more short movements crossing the state (see figure 10).

Figure 10: Ohio and Tennessee areas are similar but the latter's elongated shape can lead to different analysis.

Another spatial problem that is related to the boundary effect, is the modifiable areal unit problem (MAUP) (Rogerson, 2006). When it comes to spatial density studies, the MAUP is a statistical bias effect that derives from the fact that a given dataset's density is strongly bounded to the study area it is placed into and for a different study area the same dataset could result into totally different density measures ("The Modifiable Areal Unit Problem and GIS," 2018). When the study areas are arbitrarily defined and modified by the researcher, the results of data aggregation would be an artefact of those areas and as a result the spatial analysis could not describe realistically the underlying geographic conditions (Viegas et al., 2009). An example to illustrate MAUP can be seen in figure 11 where the same dataset has different interpretations due to different boundaries.

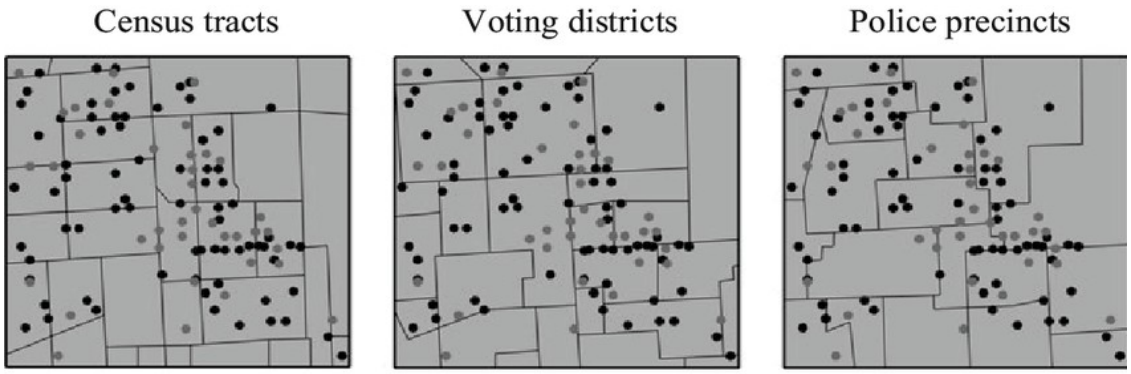

Figure 11: Different areas (names above images) lead to different data aggregation of the same data set, and hence different analysis results (Ramasubramanian and Albrecht, 2018).

To address these problems, DGGS could potentially be used as a different model for storage (Raposo et al., 2019), representation and analysis of spatial big data (Li, 2013), due to their 'regular hierarchical partitions of subsets of the Earth's plane' (Sahr et al., 2003). In one of the studies, Viegas et al. (2009)

attempted to address the MAUP that was caused from the aggregation of spatial transportation data into different census zoning systems in Lisbon Metropolitan Area by developing various size and scale grids with similar principles of a DGGS and transforming the project into a grid-based analysis, as shown in figure 13.

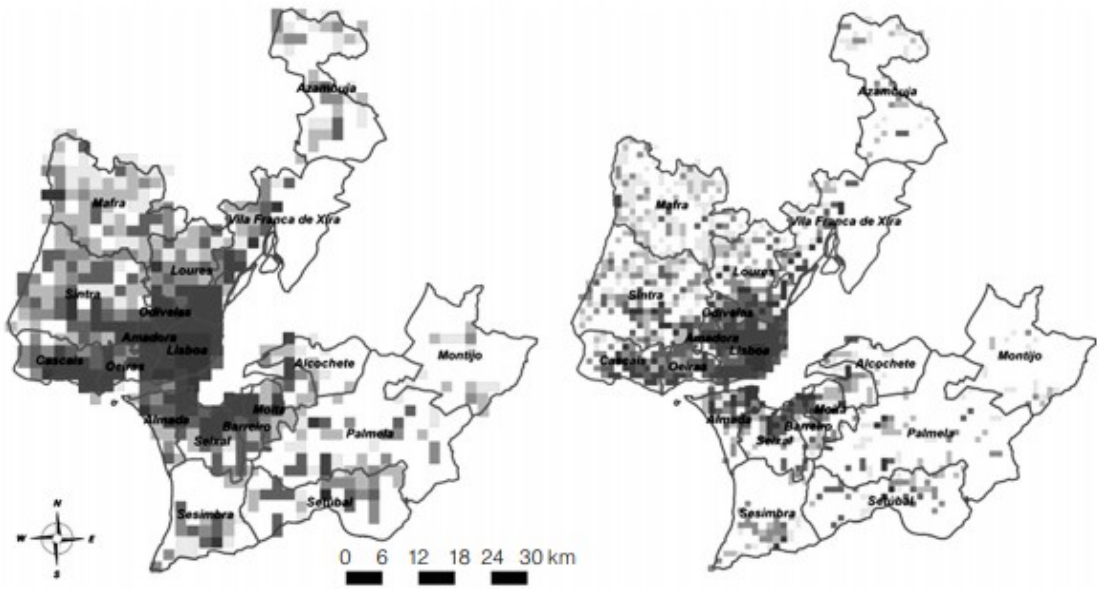

Figure 12: Spatial trip distributions of Lisbon bucketed in grid cells (Viegas et al., 2009).

Discrete Global Grid Systems

A DGGS is a group of areal units or cells that cover the entire Earth's surface without any overlaps or gaps ("Information & Examples – STCL," n.d.). Since the main focus of a DGGS is the planet's surface, topologically they approximate it as a sphere/ spheroid or geoid depending on each system's logic (Sahr et al., 2003). Systems like DGGS can form tessellations of hierarchical order following increasingly finer resolution grids (Sahr, 2019), while their cells can take the form of various kinds of regular shapes including rectangles, triangles, hexagons as shown in figure 13 (Peterson, 2017).

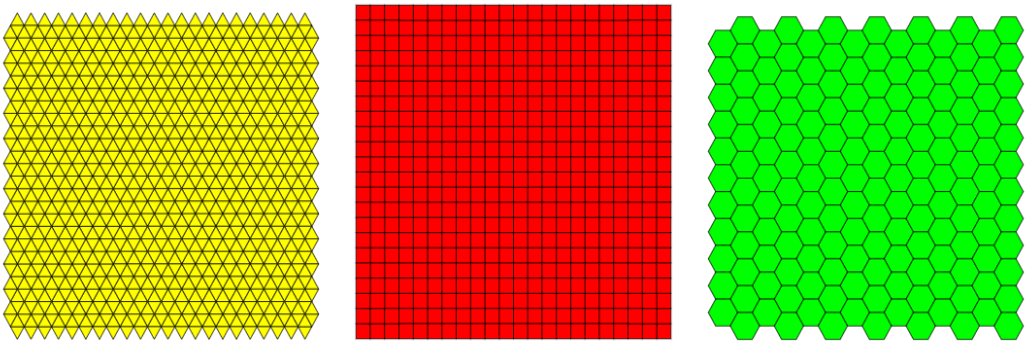

Figure 13: Tessellations for triangles, squares, hexagons ("Spatial Modelling Tidbits: Honeycomb or Fishnets? – Towards Data Science," n.d.).

The interesting feature of such grids is that each cell has been assigned with a unique ID and a point which allows researchers to run spatial indexing tasks while aggregating their features within that cell by using either area computations (geometric intersections) or approximations to the associated centroid of each cell (Sahr et al., 2003). This property is important for two main reasons. The first one is the lack of bias regarding the spatial patterns and the

second is that utilizing the point representation of the cells can avoid computationally expensive tasks like spatial areal interpolation which deals with re-aggregating values from one set of polygons to another intersecting smaller or bigger size polygon set (Lloyd, 2014).

All the above are the reasons this research uses DGGS as a geospatial indexing tool in order to develop efficient algorithms of spatial data integration based upon a global grid which is used as a uniformed spatial database and allows for bucketing events into homogeneously identifiable cells. Although these grids are not able to align perfectly to streets or local areas in urban contexts, they can be used as a tool to represent and approximate neighbourhoods as a group of grid cells, available for advanced analysis.

Historically, these types of systems had been initially created for cartographic and navigational projects. Because of their discrete structural nature, today they are also used for spatial data storage, minimising complexity in the geographic computational tasks (Purss et al., 2016). This research aims at exploring in-depth the global geospatial grids of GeoHash, Google S2 and Uber H3 while comparing their usage as geospatial storage systems of GPS footfall and POI data for location information analysis. Comparing them helps to understand their spatial and structural characteristics as well as the advantages and limitations of each from the perspective of their (a) suitability as a means for spatial data representation, (b) analysis complexity, (c) prediction, and (d) accuracy.

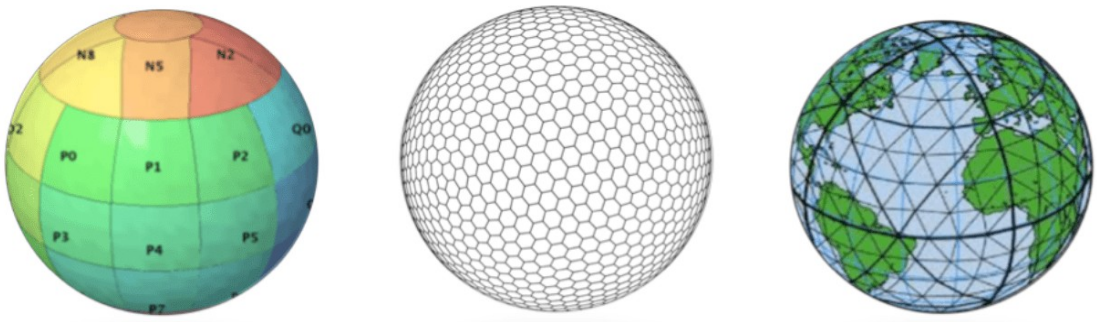

Figure 14: Examples of DGGS ("Discrete Global Grid System (DGGS) - A new reference system - Geoawesomeness," 2014).

Through the experimental process of this research three different spatial grid encoding Python libraries are used:

- polygon_geohasher (Bonsanto, 2019)
- H3 or 'Uber's Hexagonal hierarchical geospatial indexing system' (Hexagonal hierarchical geospatial indexing system, 2019)
- s2geometry or 'S2 Geometry Library' (Computational geometry and spatial indexing on the sphere, 2019).

All three of them share a discretization functionality, which is transforming two dimensional coordinate pairs of latitude and longitude into comparable hash strings. Each library represents one grid for converting boundaries into sets of native grid cells.

Boundaries are picked from three different cities across the world, London, San Francisco and Sydney. These cities vary in latitude and longitude in order to check if a location's position in terms of its geographical coordinate values affects at any way the grid's cell area and shape leading to any type of distortion. At the end of this part, one of the three grids will be chosen as the

most appropriate grid in terms of its practical applicability to address the research's objective.

Geohash

Geohash algorithm was created by Gustavo Niemeyer in 2008 as a geospatial information encoding method with the main purpose of converting geographical coordinate pairs (latitude and longitude) into unique string formats. This allows integration in web URLs ("Geohash," 2019). The algorithm covers every possible position on the Earth, ensuring that any point can be targeted with some level of certainty (Huang et al., 2018).

In practice, the Geohash algorithm works as a geofencing technology that divides the globe's geographic areas recursively into bounding boxes (or rectangles) up to the point when the necessary resolution is achieved (Fathy et al., 2017). Its bounding box levels are 12 and their coverage dimensions range approximately from 5000 km to 3.7cm ("Geohash grid Aggregation | Elasticsearch Reference [7.1] | Elastic," n.d.).

By splitting the world into multiple regions, we could say that geohashing as a technique follows a predefined hierarchical grid of spatial data streams. The very first level of such global division, may start with four major different quadrants. Latitude and longitude play the role of Y and X axes on a geographic Cartesian system (PubNub, 2014.) and when further detail is necessary, these quadrants can be split even more (see figure 15), up to the point that they can

reach hyperlocal data scale for various purposes e.g. finding target taxi pickup locations (Fathy et al., 2017).

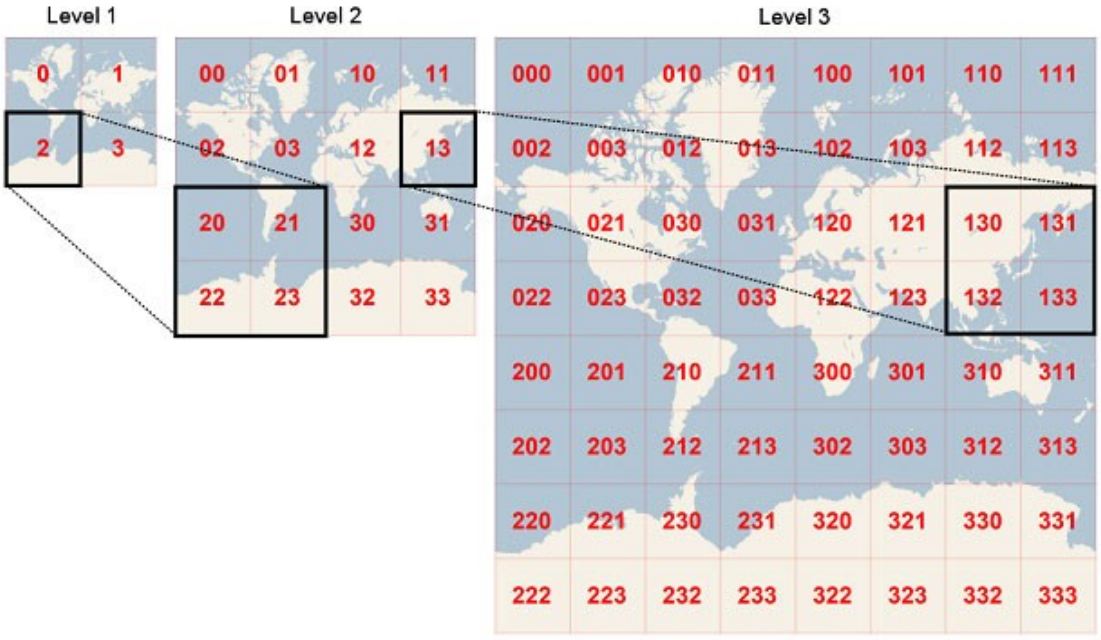

Figure 15: Geohash subdivision ("Geohashing Chat by User Proximity Tutorial | PubNub," 2014).

As an example of converting geographic coordinates, in the case of Trafalgar square in London where the latitude and longitude are 51.50820508 and -0.12806894 respectively, Geohash locates it within the bounding box of 'gcpvj0f' of resolution size 7 (see figure 16). Geohash has been used in many location projects academic and commercial and multiple algorithmic packages have been built on it. However, it is a planar geometry library that does not take into account the spheroid nature of Earth.

Figure 16: Trafalgar square into Geohash size 7 (“Geohash encoding/decoding,” n.d.).

Google S2

S2 library was firstly written by computer scientist Eric Veach in 2011, developed by Google and unlike traditional grids, Earth’s surface is not perceived as a two dimensional flat plane (Ekawati and Suprihadi, 2018). The main difference with Geohash is that S2’s grid takes into account the shape of Earth as a spheroid and introduces a three dimensional sphere that approximates the geometry of the planet as a sphere (“Announcing the S2 Library,” 2017). That is important as it tries to approximate a global GIS database suitable for low distortion geospatial indexing.

The spherical geometric grid of S2 is a revolution in DGGS field, as it is not dependent on planar map projections but on spherical ones which minimize the topological distortion to maximum of 0.56% (“Overview,” n.d.). The main idea

behind S2's characteristic projection is that the spherical representation of Earth is transformed into a hierarchy of cells as seen in figure 17, each of them consisting of four sub-cells ("S2 Cells," n.d.). This cell structure can project the planet onto six cube sides (see figure 18 and 19) and one fractal continuous *'space-filling'* curve (see figure 20). This can be done, based on Hilbert curve, which runs through all the cells across the entire spherical surface passing near every possible point on it and preserving locality ("Earth Cube," n.d.).

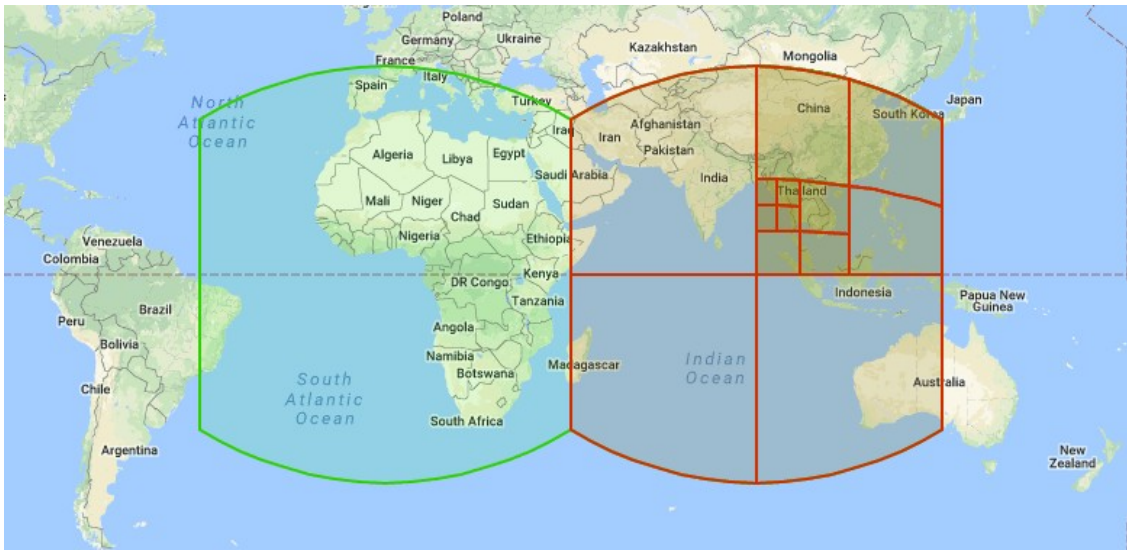

Figure 17: Two of the six main face cells of S2, one of which has been subdivided ("S2 Cells," n.d.).

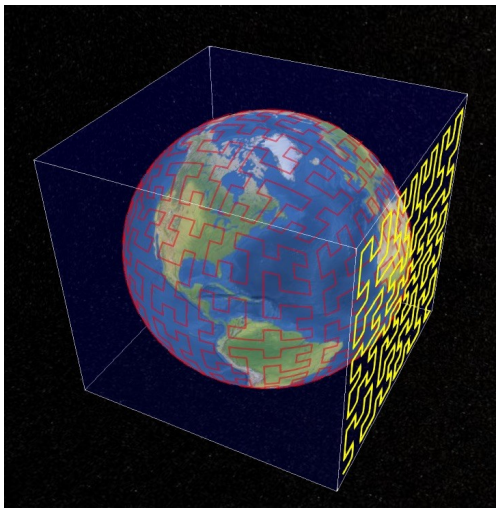

Figure 18: The cube Earth projection of S2 (Imgur, 2016).

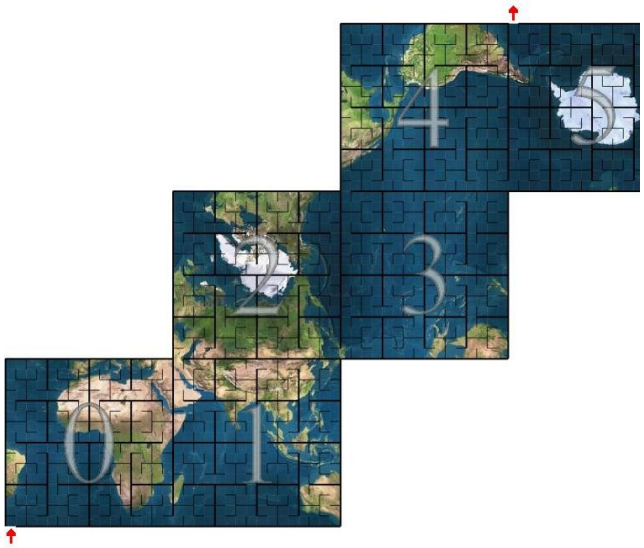

Figure 19: The top-level six face cells of S2 unwrapped (“Earth Cube,” n.d.).

The S2 cells begin with the *'face cells'* which are the biggest and first level of subdivision while the smallest ones are called '*leaf cells*'. Their levels range from 30 to 0 covering areas from very little as approximately 0.7 cm^2 to 85,000,000 km^2 (“S2 Cells,” n.d.). No matter the size, they support efficient storing as each one is uniquely encoded on an uint64 and their unique numbering system is used to maximize locality (Maurya, 2018).

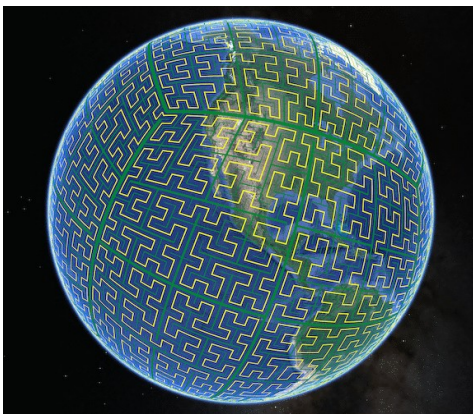

Figure 20: The S2 space-filling curve after five levels of subdivision (“S2 Cells,” n.d.).

One of the most interesting tools of S2 is the S2 *'RegionCoverer'* which can approximate an arbitrary region or a point as a set of cells ("Geometry on the Sphere: Google's S2 Library - Google Drive," n.d.) and an example of it is shown in figure 21. This makes S2 suitable for running geospatial searching and indexing operations (Maurya, 2018).

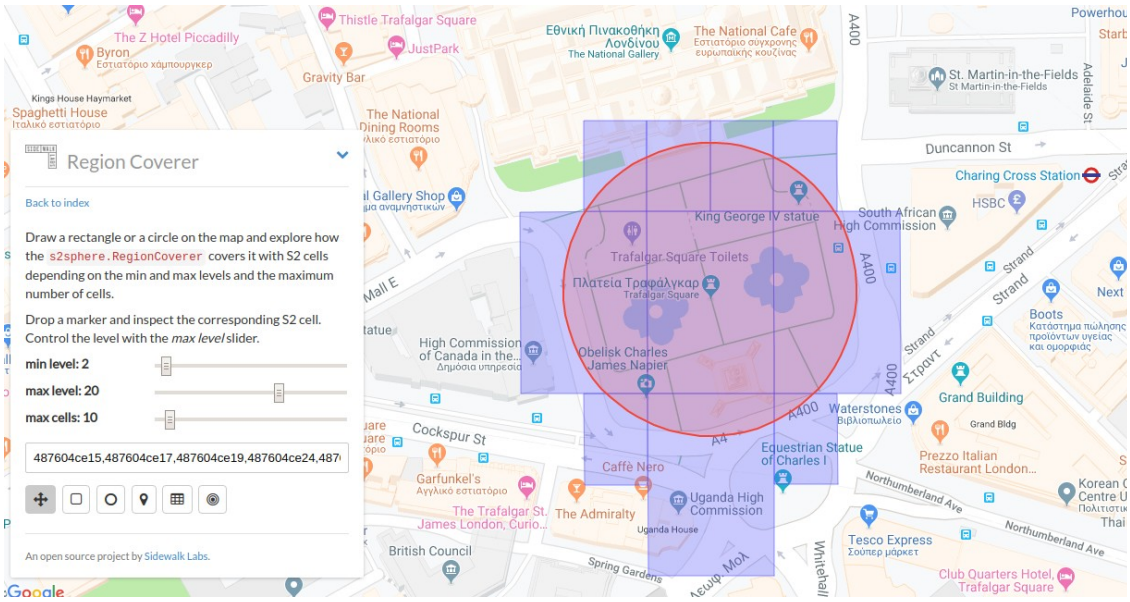

Figure 21: Conversion of circle over Trafalgar square into Google S2 cells using RegionCoverer ("Region Coverer," n.d.).

Unfortunately, at the time of this project this library is quite recent without enriched documentation and is currently undergoing development. Therefore, there is little literature and academic experimentation around it. On top of that the coding capabilities of the library, though extensive, are hard to use and not as intuitive as libraries for Geohash and H3.

Uber H3

H3 Geospatial Indexing System was designed by Uber (Sahr, 2019) as a DGGS for visualisation, exploration and analysis of spatial data on a city-wide scale and was open sourced on Github in 2018. It follows mainly a hexagonal tiling with hierarchical indexes on a sphere (Wang et al., 2019) as shown in figure 22, which, like Google's S2, is a geometric approximation of the Earth ("H3," n.d.). H3 has been used by Uber's logistics platform (see figure 23) to model the supply and demand for surge pricing (Uher et al., 2019).

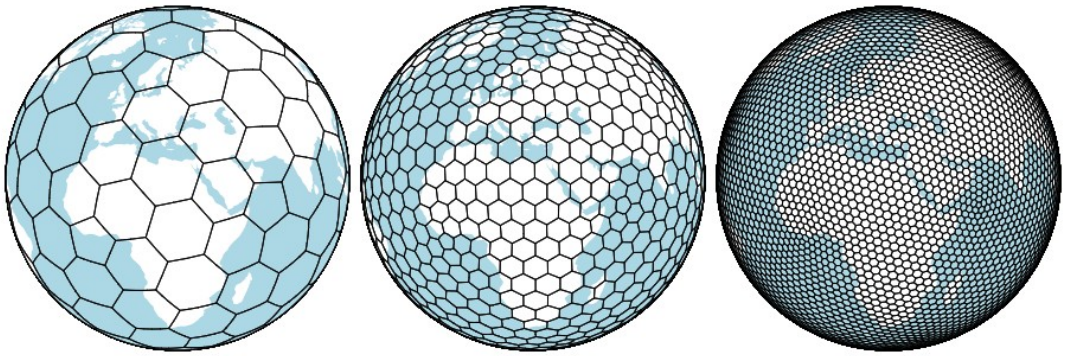

Figure 22: Multiple levels of granularity for the H3 global grid ("Exploring H3 and tessellations on the sphere," 2018).

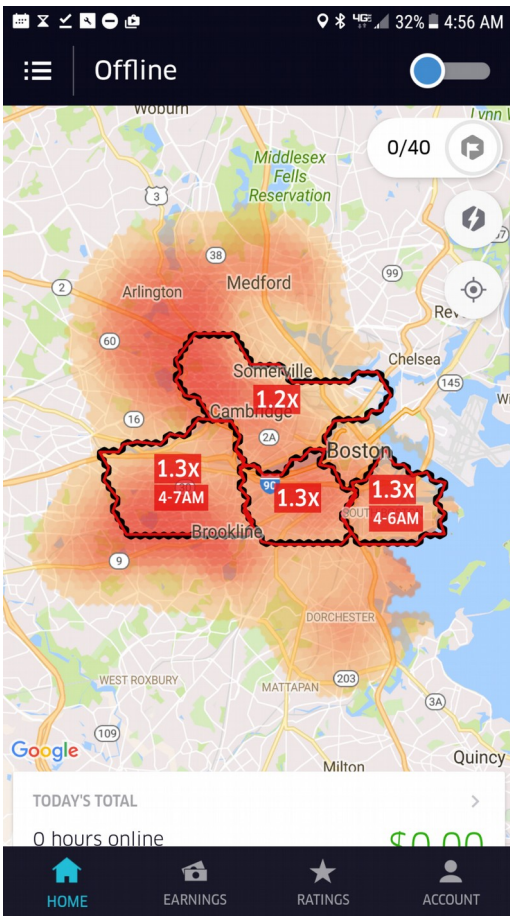

Figure 23: Image from Uber platform visualising morning surge areas based on H3 grid ("Early morning surge," 2017).

Although, possibly more developed than S2, based on the developers activity in both libraries' repositories on Github.com (H3: https://github.com/uber/H3-py, S2: https://github.com/google/s2geometry) still today, literature and academic research is limited around H3. Therefore this research looks mostly at Uber's online sources for information.

Like Google's geoindexing library, H3 also takes into consideration the third dimension of the Earth and avoids planar projections as Geohash follows. However, there is a major difference between S2's and H3's type of projection.

While S2 uses the cube or hexahedron to project the sphere, in the case of H3 the grid is developed on the faces of an icosahedron and its cells are then projected on the surface of Earth as shown in figure 24 (“H3,” n.d.). To do that, H3 utilizes the Dymaxion projection of Earth (see figure 25).

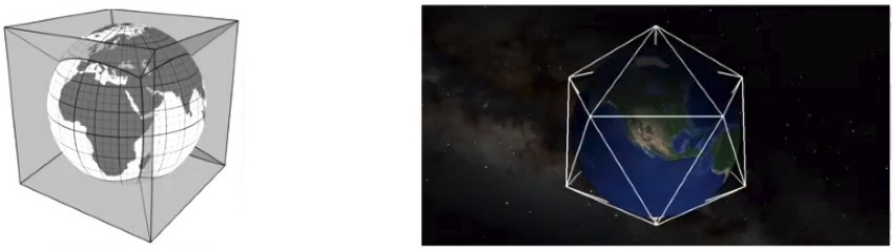

Figure 24: Projection systems for Google S2 cube on the left and Uber H3 icosahedron on the right (Uber Engineering, 2018a).

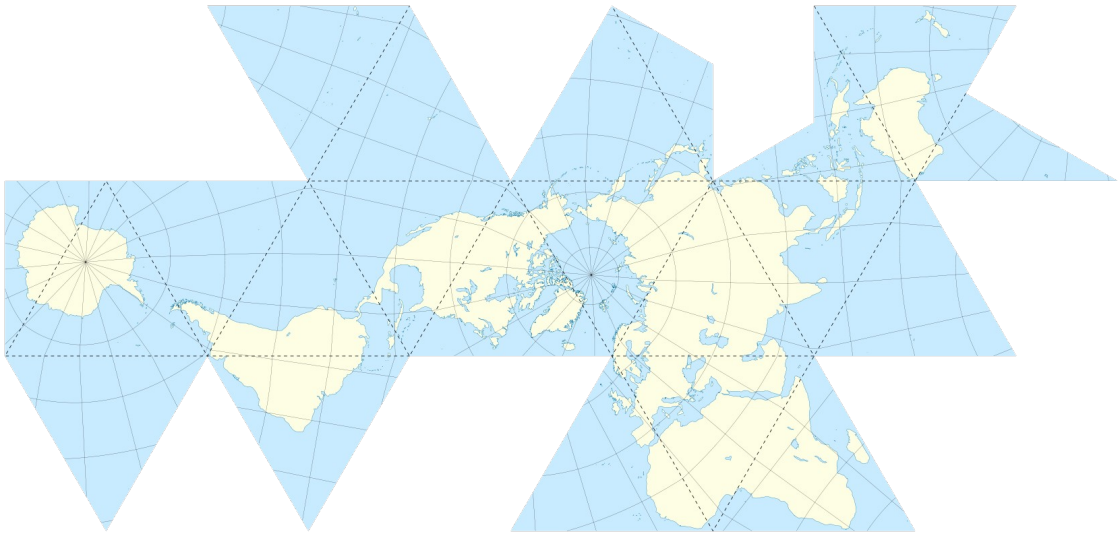

Figure 25: Dymaxion map (“MondayMap,” 2016).

One of the advantages of such projection mechanism is the minimisation of distortion which is less than S2’s. The reason behind the smaller distortion, could be that a bigger projection angle may result in larger distortion levels. H3 grid reduces its distortion by reducing the maximum projection angle to the side of an icosahedron as shown in figure 26.

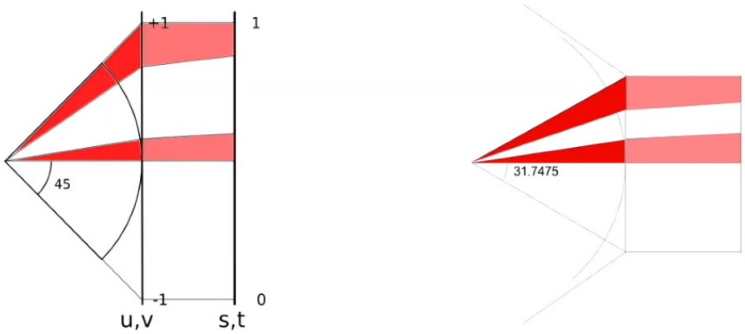

Figure 26: Projection angles for Google S2 on the left and Uber H3 on the right (Uber Engineering, 2018a).

According to the documentation of H3, tiling the entire spherical representation of the Earth with only just hexagons is not feasible, and for that reason in each resolution 12 pentagons have been introduced on the 12 icosahedron vertices of the grid to complete the system ("H3," n.d.). These pentagons fall mostly in the ocean and hence not affecting the efficiency of H3's implementation. The resolution levels available in H3 are 15, with level 0 as the base cell level of 110 hexagons and 12 pentagons while level 15 being the finest one. Their coverage ranges from $1m^2$ to 4,250,546 km^2 ("H3," n.d.).

A unique tool provided in H3 library, that logically derives from the geometric characteristic of hexagons to approximate circles, is the *'kRing'* function ("H3," n.d.). Its main purpose is querying cells within a grid distance k from a cell and it is basically useful for radius around a point analytics as shown in figure 27.

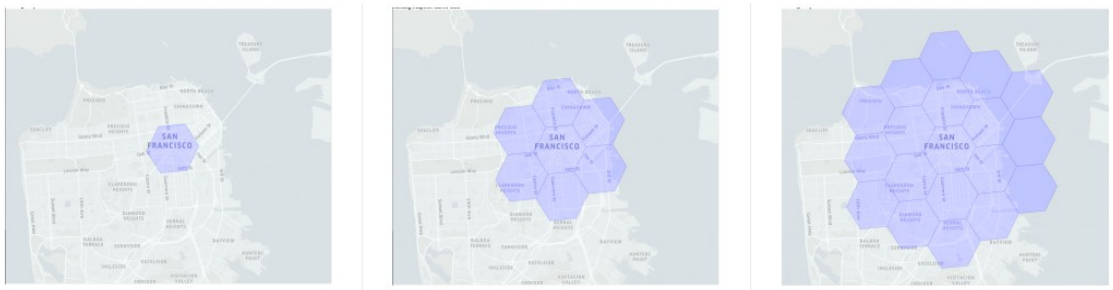

Figure 27: Neighbouring hexagons using kRing function (Brodsky, 2018).

Grid Comparison

Depending on the use case and the extent of acceptable trade-offs per project, each grid has its advantages and limitations.

Cell Division

Starting with the square grids, Geohash and S2, are much more accurate and precise when it comes to hierarchical division of cells due to square's geometric nature. Increasing the spatial resolution of a fishnet grid can be achieved by simply dividing the cell into four sub-cells. The square division is perfect, with no error. In the case of hexagons, divisions are not exact at different scales and finer resolution cells can only approximately be contained within a parent cell. Hexagons can not be subdivided perfectly, and therefore squares are preferred for hierarchical indexing as seen in figure 28.

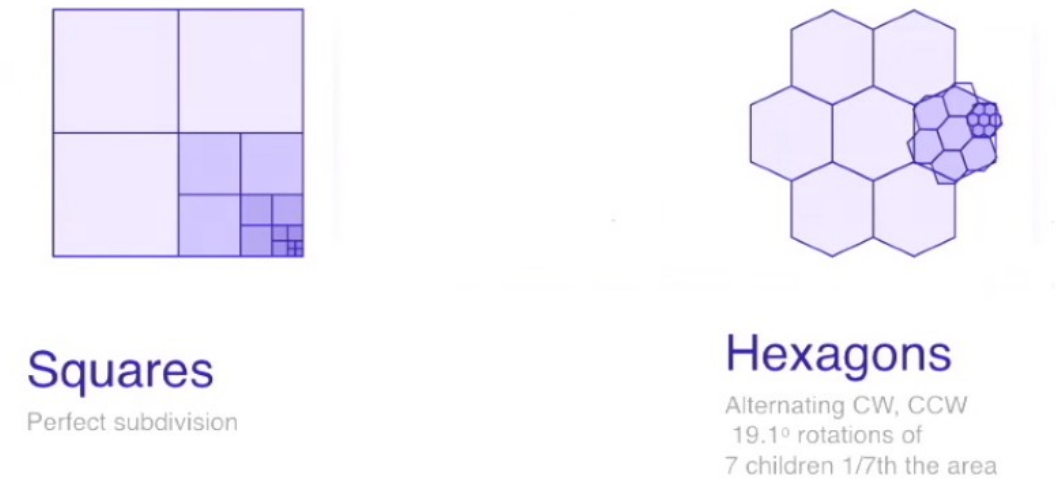

Figure 28: Perfect subdivision for squares whereas for hexagons the subdivision is approximated (Uber Engineering, 2018b).

At this point, its worth mentioning that H3 has a similar function to Google S2's RegionCoverer which is called '*Polyfill'* and allows converting arbitrary polygon boundaries into hexagons as visualized in figure 29. However, due to hexagons' geometric properties not allowing for perfect subdivision the results allow for areas not covered by any H3 cell as seen in the figure below on the right.

Figure 29: Bedford boundary converted into H3 hexagons and compacted cells on the right resulting in areas not covered.

Cell Distortion

All three grids encounter shape distortion errors in their cell sizes due to their attempt to flatten the Earth's surface. None projection system can perfectly do that, however some have less distortion than others. Geohash has the most distortion as it considers Earth's surface as a flat 2-d plane and Google's S2, though much less distortion than Geohash, yet is more than H3 because of the six sides of the cube and its projection angle (Uber Engineering, 2018b). Both in S2 and H3 shape distortion changes according to the location's distance away from the centre of the projected plane. Due to S2's projection's greater angle its cell shape can get more distorted, and therefore when shape gets distorted, the surface changes, and then its geometry and area changes as well (Uber Engineering, 2018b).

As it can be seen in the figure 30, the S2's cell shape in San Francisco is elongated and does not look like its equivalent in Amsterdam, which is geometrically closer to a square. This phenomenon of elongation can cause a boundary effect as described in previous sections and therefore, S2 cells are not potentially the most appropriate for point density analysis. On the other hand, the H3 cells in both cities maintain approximately their hexagonal shape.

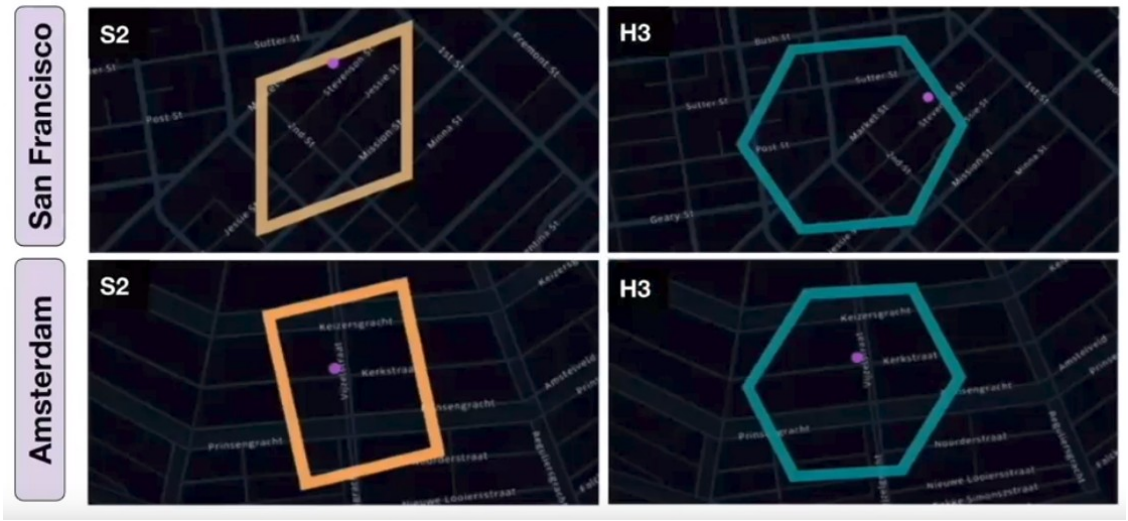

Figure 30: Uber's office in San Francisco and Amsterdam in purple dots. Left column: Google S2 cell shapes gets distorted, Right column: H3 cell shapes remain almost the same.

Distance to Nearest Neighbouring Cells

Since footfall density is the main focus of this research, and by footfall we refer to trajectories and movement data, one of the areas to investigate is neighbour traversal. Geometrically speaking, hexagons are the only cell shape to have just one distance between a cell's centre point and its neighbours', compared to two distances for squares (across edges and diagonals) and three distances for triangles (across edges, two edges and diagonals) as seen in figure 31 ("Red Blob Games," 2019). In H3, distance between centroids is the same in all six

directions with adjacent hexagons, which helps including more neighbours in analysis as opposed to square grids (Birch et al., 2007). This geometric property of hexagons becomes a valuable quality for this research and allows for flexible connectivity analysis of movement data across a consistent hexagonal tilling.

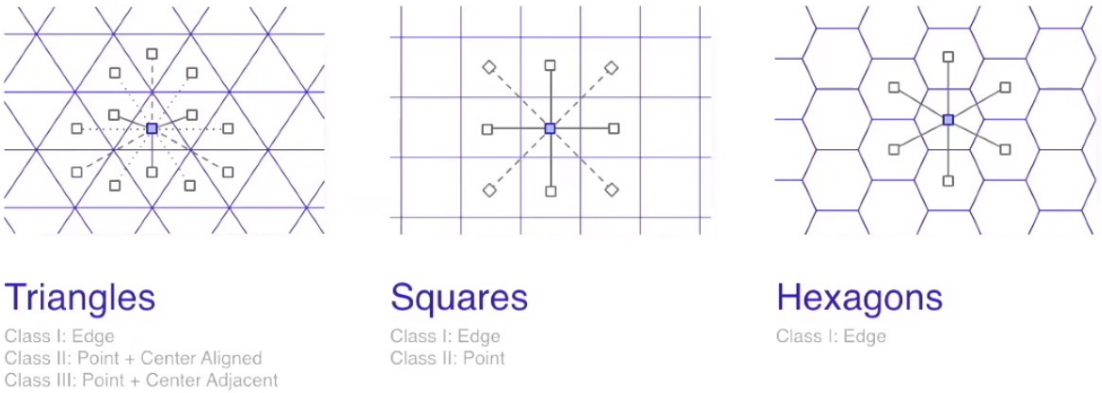

Figure 31: Different tessellation systems and their distances from neighbouring cells (Uber Engineering, 2018a).

Grid Analysis and Selection

Following the research, all three geospatial indexing grid systems are put into test in order to examine their suitability regarding the project's objective. As mentioned above, three cities are studied, London, San Francisco and Sydney with their boundary centroid coordinate pairs in latitude, longitude (51.497, -0.102), (37.755, -122.44), (-33.889, 151.202) respectively.

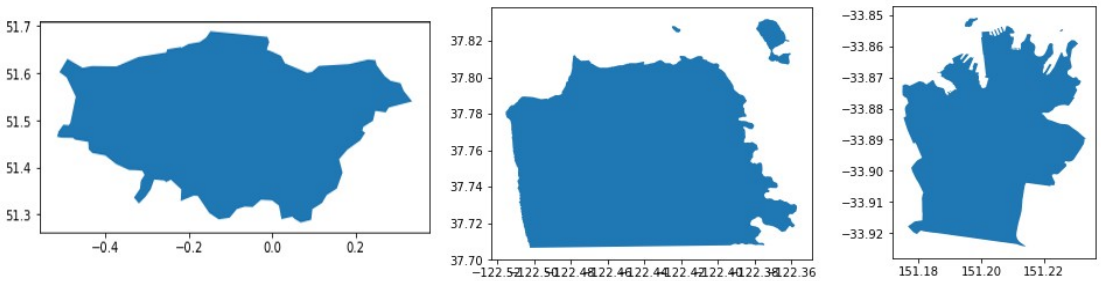

Figure 32: From left to right: London, San Francisco, Sydney.

Since not all libraries allow for fast and memory efficient polygon to grid cells transformation, the bounding boxes of all cities were taken for examination, instead of their actual point-dense boundaries which could cause computationally expensive conversions. The resolution size for each system was selected so that the resulting cell population distributions and cell areas are similar for each city and hence comparable. The following table 3 illustrates these points.

Grid	Cell Size	Area (sqkm)
Geohash	7	0.023
S2	16	0.015
H3	10	0.019

Grid: Cell Size	London	San Francisco	Sydney
Geohash: 7	184,437	10,672	2,322
S2: 16	180,331	12,050	2,406
H3: 10	198,414	12,417	2,441

number of cells per grid

Table 3: Top: cell area per grid, Bottom: cell population per grid and city.

Overlaying the grid cells of each city on the map helps us understand the nature of each system and highlights the points discussed in previous sections. In the following figures we visualise first the Geohash grids, followed by S2 and lastly by H3.

Figure 33: Geohash grid resolution 7 at 1:10,000 scale (left: London, middle: San Fransisco, right: Sydney).

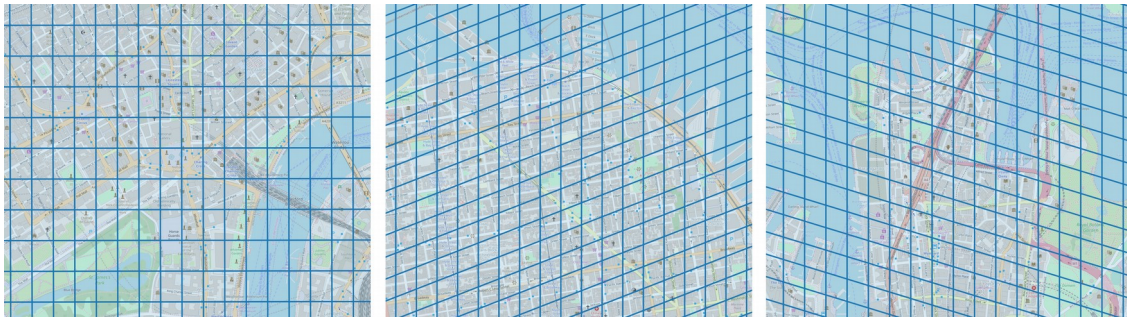

Figure 34: Google S2 grid resolution 16 at 1:10,000 scale (left: London, middle: San Fransisco, right: Sydney).

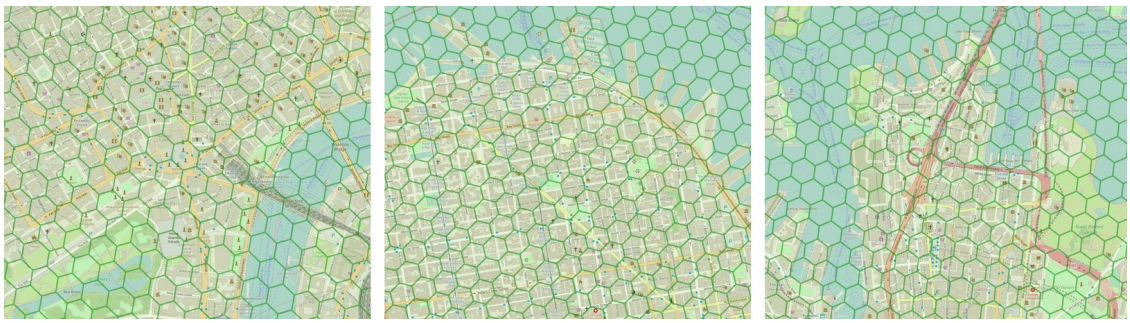

Figure 35: Uber H3 grid resolution 10 at 1:10,000 scale (left: London, middle: San Fransisco, right: Sydney).

From the figures above we could say that the S2 cell shapes are possibly different from one city to the other, while Geohash cells are quite similar between San Francisco and Sydney whereas London's cells are stretched. Regarding H3 cells, the shape is maintained across all cities while the size in London is a bit bigger. To delve deeper into more details about the grids of the three cities, we run analysis on cell areas, perimeter length and cell centroid distances to minimum and maximum point to measure distortion.

Another way, possibly more appropriate to approximate distortion would be to calculate the ratio of each cell's area to its minimum enclosed circumscribed circle's area. The greater the number the less the distortion. However, due to

that process being computationally expensive and time consuming the distances to closest and furthest points are used instead, which are also good indicators for geometric distortion. The results of these metrics are shown in figure 36.

```
GEOHASH                      S2                           H3

London:                      London:                      London:
geom_area      14578.248892  geom_area      16498.061455  geom_area      15635.755522
geom_length      496.406314  geom_length      532.004585  geom_length      465.681279
max_dist          90.067185  max_dist         109.704037  max_dist          79.538279
min_dist          47.707218  min_dist          61.663881  min_dist          65.527142
dtype: float64               dtype: float64               dtype: float64

San Francisco:               San Francisco:               San Francisco:
geom_area      18442.759162  geom_area      15057.486233  geom_area      13500.208831
geom_length      546.841050  geom_length      498.616633  geom_length      433.303653
max_dist          97.305440  max_dist          89.617838  max_dist          75.338752
min_dist          60.498022  min_dist          51.359740  min_dist          59.731000
dtype: float64               dtype: float64               dtype: float64

Sydney:                      Sydney:                      Sydney:
geom_area      19350.551881  geom_area      19121.401232  geom_area      18097.770534
geom_length      558.719617  geom_length      565.544093  geom_length      500.770814
max_dist          99.172746  max_dist         113.618035  max_dist          83.582831
min_dist          63.516904  min_dist          67.400421  min_dist          72.174995
```

Figure 36: Average values for area, length of perimeter and distances to the closest and furthest point as proxy of distortion.

At this stage calculating the mean values is more useful since it takes outliers into consideration as opposed to median. Focusing first on areas, we notice that all grids encounter value differences across three cities, something that researchers should bear in mind when comparing different cities and be cautious how they use cells across the globe. Despite these differences, on could argue that there is a good approximation in similar cell areas, as it is well known that perfect cell size equality in such grids is difficult to achieve (Sahr et al., 2003). The lengths are also relative similar while the distances from the cell centroid to minimum and maximum cell points show that H3 maintain reasonably similar values, proving that H3 could be potentially more resilient to cell shape distortion.

Grid Selection

Except for the grids analysed above there are also other geospatial indexing solutions like OSM tiling, ISEA, Geodesic DGGS etc. (Mocnik, 2019). There is a belief, though, based on professional and academic experience that these specific three grids, are the ones mostly used in the field of location based targeting and the ones that also provide many functionalities applicable to most of commercial and open source programming tools. Especially the square grids which are among the most popular and can accommodate for typical data structures (Sahr et al., 2003). However, for the purpose of this research, H3 hexagons are chosen as the fundamental geospatial grid unit and H3 Uber's hexagonal geospatial indexing library in Python programming language is used to implement it. It should be highlighted that Uber's hexagonal grid is not proposed as the one-size-fits-all system. It is not used to solve every possible geospatial problem but only to help us address the project's objective.

Running measurements and analytics on movement data is the main focus of this research, and not just searching. If geospatial querying was the focus of this research then one of the square grids would be selected that supports perfect subdivision. Therefore, we acknowledge that H3 is not a grid that gets perfectly divided, however the resolution we need for footfall density estimation does not require analysing multiple resolutions at the same time. Also, since we deal with movement data we need to have a grid that keeps equal distances across cells, and as far as density measures are concerned we need the cell shapes to be as consistent as possible on a global scale. Lastly, the area is something not being used for this research as there will be no comparison

between footfall densities across different cities. However, it would be interesting to identify cities that share similar cell areas and develop this study further. It is feasible, based on the projection of H3, that these urban areas share similar distances to the centre of the side of the icosahedron projection of H3 and thus easily discoverable.

This research focuses on London as its main area of interest and by using H3 the city's geographic polygon boundary is converted into hexagonal cells which are used as the spatial unit of study. Converting a polygon like London's boundary into H3 cells is illustrated in figure 37.

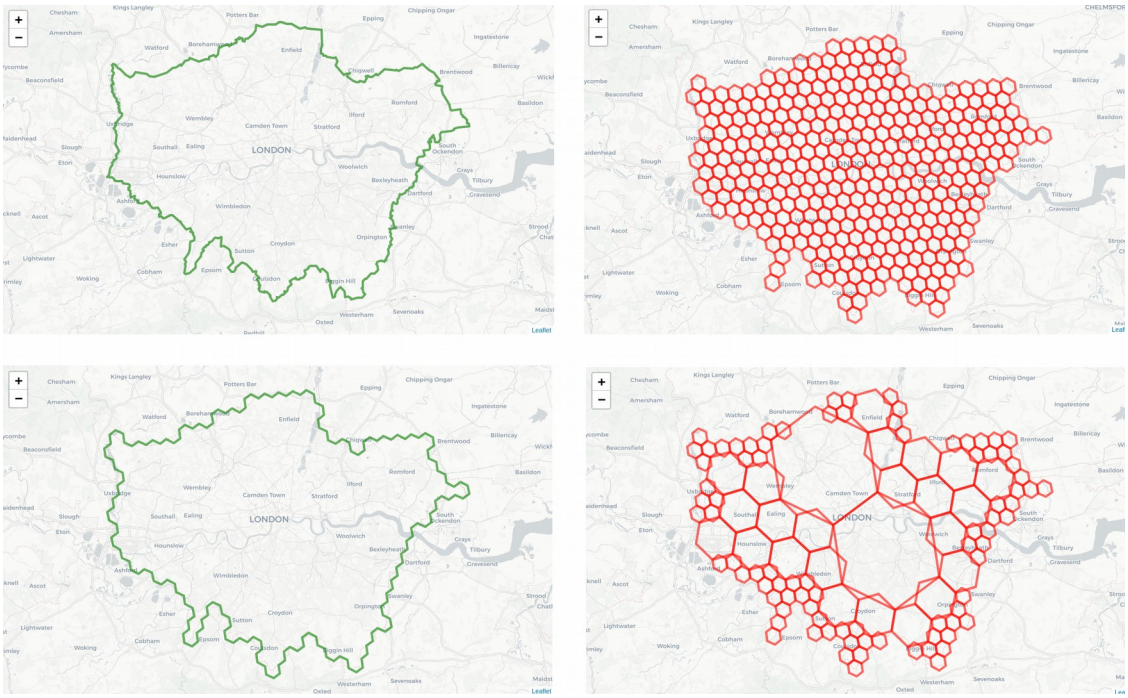

Figure 37: London boundaries converted into H3 cells of resolution 7 (upper left: London boundaries, upper right: H3 hexagons res7, low left: outline of upper righ, low right: compactisation of upper right.

After investigating the different resolution sizes of the hexagonal cells of H3 the most suitable resolution for this research it terms of reaching ideal spatial coverage and minimizing the cost of computation would be size 9 of average

hexagon area 0.105 km^2 (“H3,” n.d.). Spatially an H3 cell of resolution 9 can be seen in figure 38.

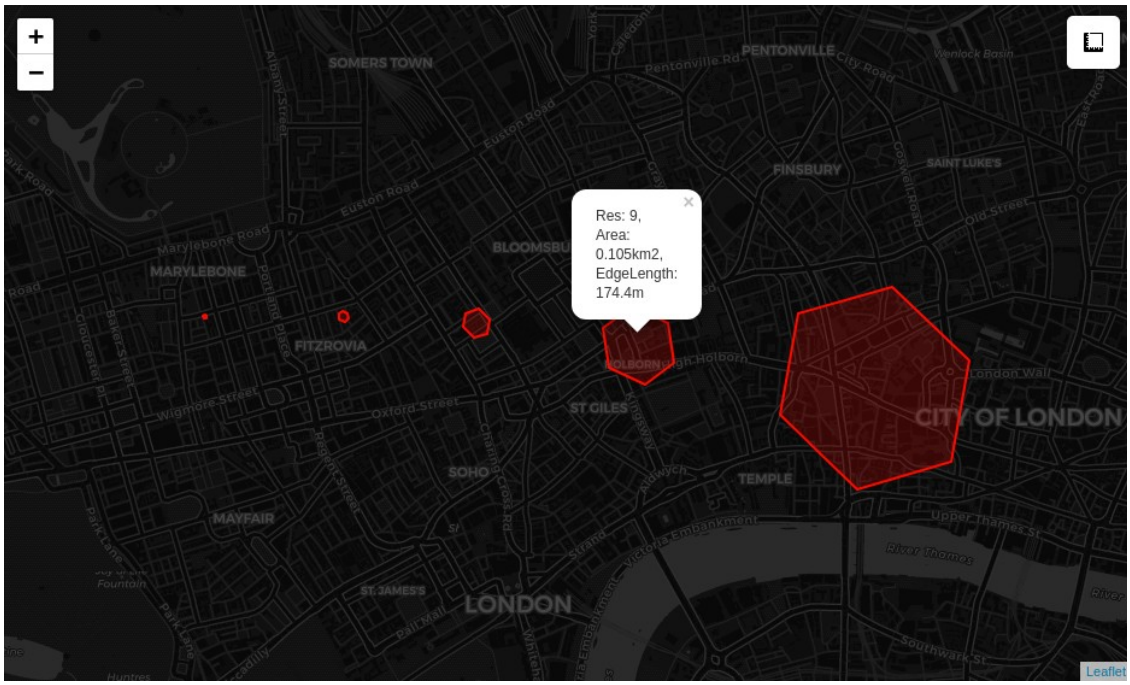

Figure 38: H3 cells of different resolutions in London. Sizes from left to right: 12, 11, 10, 9 (selected for this research) and 8.

Transforming London’s boundary into hexagonal cells of resolution size 9, results in 17,047 cells covering the administrative area of Greater London.

Location Data Discretization

The literature shows that partitioning the surface of the Earth into identifiable cells on a grid could potentially contribute to analysing big spatial data. In this section, the POI and trajectory GPS data are assigned to H3 cells via the process of geospatial indexing or discretization. Matching coordinate pairs in two dimensional space to strings in one dimensional discrete space is what defines discretization (Werner, 2015).

Our geospatial indexing strategy for the POI includes the following two steps:

- First, for each data point, we get its coordinate pair of latitude and longitude and map it to an H3 cell of resolution 9, as shown for some POI in table 4.

Name	group name	lat	lon	hex9
Tennis Courts	Sport and entertainment	51.494961	-0.280982	89194adb56fffff
Superdrug Pharmacy	Retail	51.376985	-0.100763	89194ac2663ffff
Perfect Pizza Ltd	Accommodation, eating and drinking	51.520803	0.018691	89194ad2413ffff
Sayah Trading Services Ltd	Retail	51.580659	-0.083146	89195da6caffffff
Smooth & Simple	Accommodation, eating and drinking	51.522476	-0.071536	89194ad3467ffff

Table 4: The POI data of table1 with an extra column of the unique H3 cell ID resolution 9.

- Second, we count how many times each discrete hex9 cell ID appears in the dataset which accounts for the POI density (see table 5).

hex9	total_cnt
89195da49cfffff	397
89194ad32d7ffff	346
89194ada44bffff	344
89195da4987ffff	339
89195da4913ffff	299
89194ad324bffff	252
89195da4903ffff	251
89195da49b3ffff	244
89195da4907ffff	244
89195da498fffff	240

Table 5: Sample of H3 cell Ids within London and GPS counts within each individual one.

Visualizing in figure 39 each cell's POI density, highlights clearly the concentration of amenities across London. From the map it is noticeable where the high streets of London are located and the areas of strong retail clustering. This supports the idea that a geospatial indexing grid can be used to represent neighbourhoods, streets, attractors, geographic areas in general.

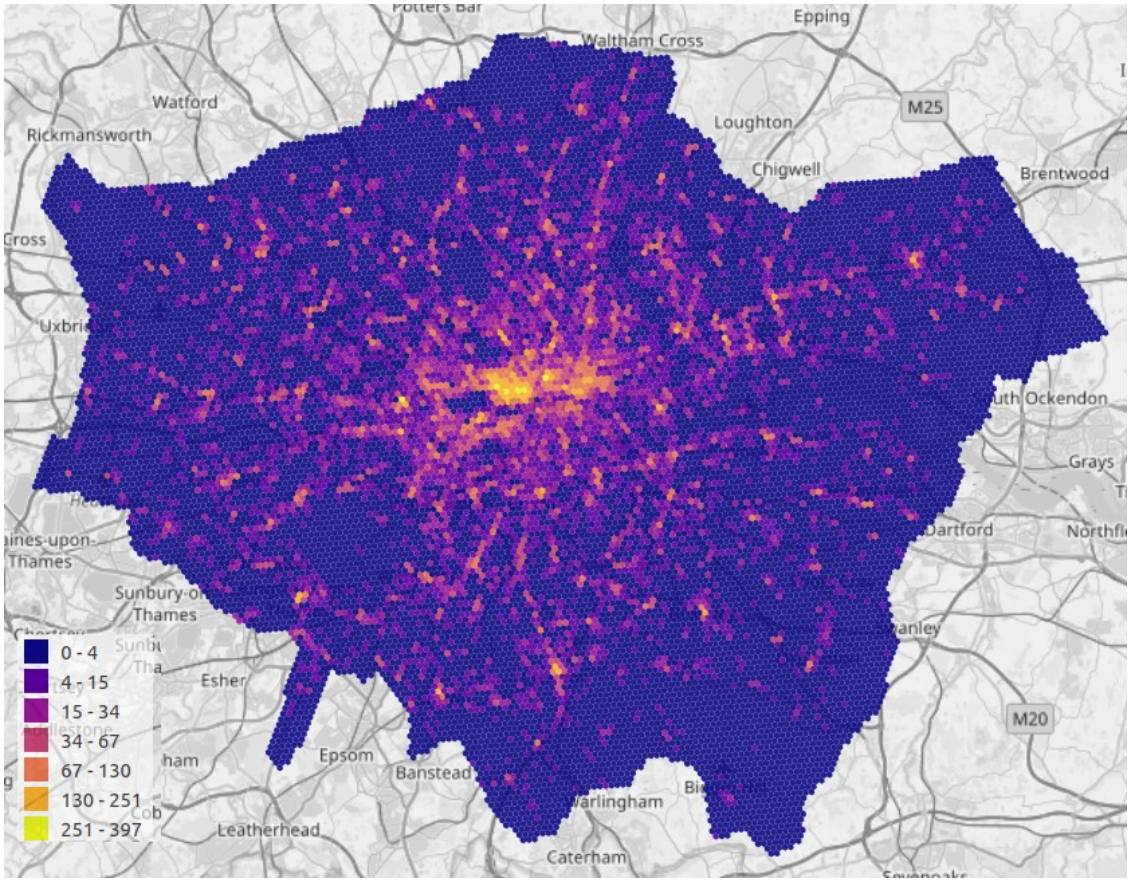

Figure 39: POI counts per H3 hexagonal cell Id.

Our geospatial indexing strategy for the trajectories follows the same first step as for the POI. However, in the second step we need to address the dimension of time:

- First, for each GPS point, its latitude and longitude is mapped to an H3 cell of resolution 9, whose hash string IDs look like the ones in table 6 under the column hex9.

timestamp	latitude	longitude	hex9
1555750800000	51.509730	-0.137311	89195da4917ffff
1556161200000	51.539327	-0.142434	89195da4c37ffff
1556434800000	51.560443	0.069717	89194e68ccfffff
1555455600000	51.565587	-0.138795	89195da6a03ffff
1555927200000	51.557739	-0.216276	89195da436bffff

Table 6: Sample of GPS points and the mapped unique H3 cell Ids.

- Second, we count how many times each discrete hex9 cell ID appears in the dataset per hour of each day. Assigning the events into hourly time buckets allows to aggregate for every hour the total amount of unique GPS events in each hexagon and hence map trajectories to the H3 grid. That introduces time granularity so we can then be able to predict footfall density for an hour of a day, and not just overall throughout a day. In figure 40, the chronological sequence of the maps illustrates how footfall density changes throughout a typical Saturday.

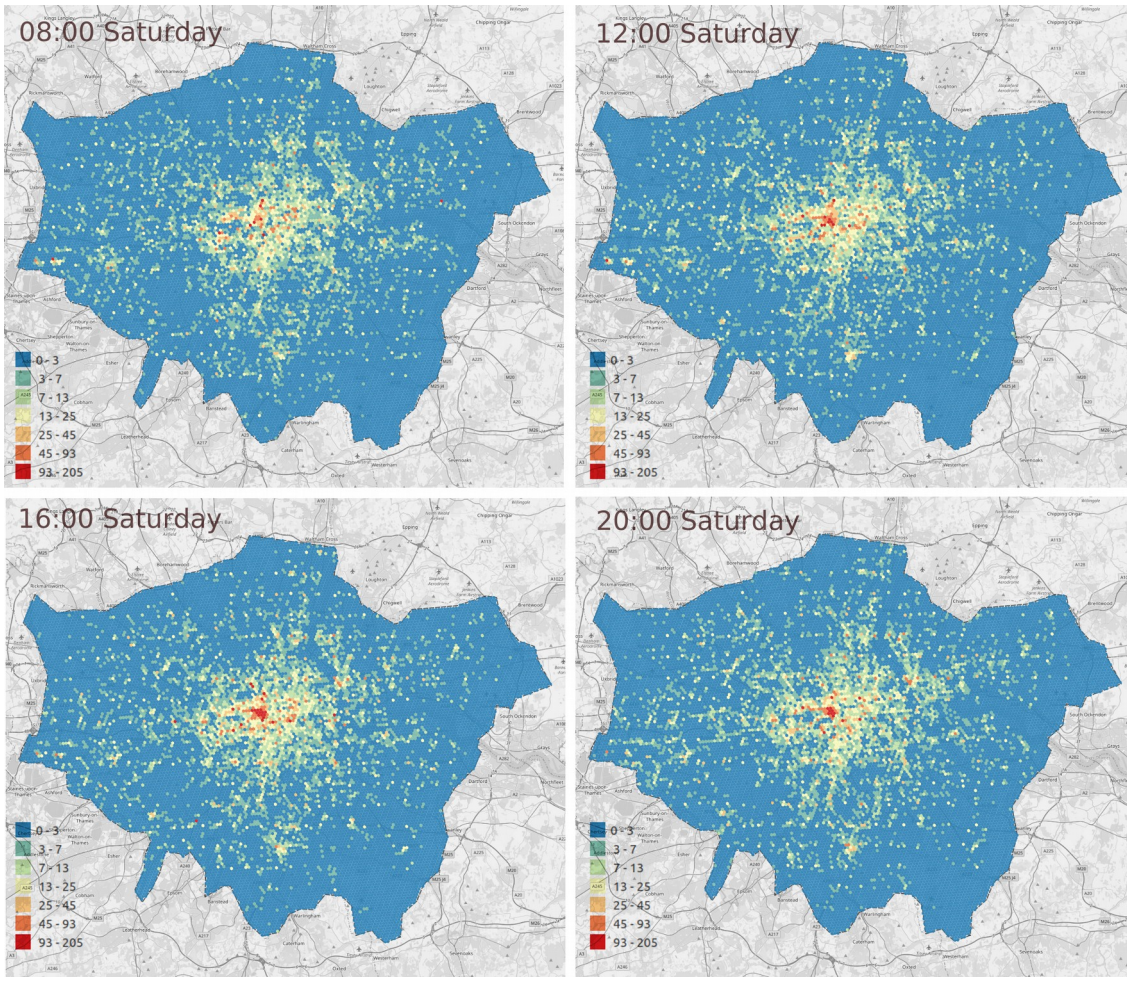

Figure 40: Total GPS counts on a Saturday in April 2019 per H3 hexagonal cell Id in four different times.

- Finally, the centroids of the cells that our POI and GPS events are assigned to are calculated (see table 7) and used in the machine learning method that follows.

hex9	hex9_lat	hex9_lon
89194e696b7ffff	51.622278	-0.055962
89195dac4b7ffff	51.588124	-0.391525
89194e6d557ffff	51.487046	0.117202
89195da49b7ffff	51.507842	-0.125709
89194ad16dbffff	51.500886	-0.161343

Table 7: Sample of the cell Ids in London with their centroid coordinates.

Random Forest

The machine learning method used in this study is the supervised learning algorithm of Random Forest. This algorithm, first proposed by Tin Kam Ho of Bell Labs in 1995 (Ho, 1995), is an ensemble of multiple predictors, be it regressors or classifiers, coming from several decision trees. This fact makes it one of the most efficient machine learning algorithms (Géron, 2019). While decision trees are useful, when employing only one, it will most likely not yield the best possible prediction. Therefore, running several ones on different subsets of the original training dataset and averaging their results reduces the high variance of a single estimator thus maintaining low bias, preventing overfitting and improving prediction accuracy (Breiman, 1996). The final prediction of a random forest is ultimately the average value of all the decision trees (Breiman, 2001).

In this study random forest is chosen as the most suitable technique for big data analysis, since it allows for efficient processing of large-scale datasets (Bao et al., 2018). Except for the data size, random forest handles binary, categorical and numerical features without the need of heavy data pre-processing, rescaling or transformation (Géron, 2017). In terms of computation time, this algorithm allows parallelisation of the decision trees, meaning that they can be distributed to multiple cores to run, and hence speeding up the process. This means that in case we would like to apply the study on multiple cities, random forest allows for the project to easily scale up. Also, since we deal with GPS data, extreme values are likely to still be included in the dataset, even after aggressive data cleaning. In this case, decision trees, which build the random

forest, are robust to outliers because they are capable of isolating them in small areas of feature space (Bao et al., 2018). Lastly, regarding variable selection the algorithm usually avoids using the variables that are non-important in explaining the target variable (Hastie et al., 2009).

Model

In this study the model of random forest is used to predict footfall density for an average day of the week and hour of the day, within all H3 hexagonal grid cells distributed across Greater London. Since the data includes the targeted scores, supervised learning is the appropriate machine learning technique and due to the numerical nature of these scores, regression for estimating continuous values is preferred. In terms of implementation, Python is the programming language of choice and the random forest regression model from skicit-learn will be imported (“scikit-learn”, n.d.), as one of the most popular and widely used machine learning libraries in that language (Garreta, 2017).

Feature Selection

The main features used to train the random forest model are:

- **hex9_lat, hex9_lon**: the coordinates of the H3 cell centroids as numerical variables, instead of the string representation of the cells, which are treated as multiple categorical one-hot encoded variables. During one-hot encoding process, one column is created for each value within the categorical variable. For each column, a row gets a 1 if the row contained that column’s value and a 0 if it did not. If cell IDs were to be used, we would end up with 17,047 columns and that would probably cause computational blockers regarding fitting in memory. Therefore, H3 cell centroid coordinates are used instead.

• **time_num, day_num**: binned time and day of the week, which transforms the two categorical variables into numerical ones, encoded within a range of 0 to 1, which is the floating centre of the bin.

• **time_cos, time_sin, day_cos, day_sin**: cyclical feature engineering of time and day, which derives from mapping these two variables onto a circle. The calculation of the x,y component is achieved by using the sin and cos trigonometric functions. The cyclical representation of time and day is useful because it brings the lowest value next to the largest which in other case numerically it would not stand e.g. 23:59 next to 00:00.

• **poi_cnt**: POI density per cell, i.e. the population of amenities within each hexagon.

• **gps_cnt**: footfall density, which is the number of GPS events per cell per hour per day and the variable or label this model will attempt to predict.

It is important to keep the most relevant features for the objective of this project, as it helps the model generalize better and avoid overfitting. In terms of feature scaling we do not need to scale the values of the different variables as random forest are scale invariant (Li and Martin, 2017). A sample of the resulted data can be seen in table 8 with all the columns of the random forest model.

hex9_lat	hex9_lon	time_num	time_cos	time_sin	day_num	day_cos	day_sin	poi_cnt	gps_cnt
51.590766	-0.223533	0.427083	-0.896873	0.442289	0.489583	-0.997859	0.065403	24	2
51.511073	-0.226968	0.760417	0.065403	-0.997859	0.394345	-0.787627	0.616153	35	7
51.488318	-0.146542	0.218750	0.195090	0.980785	0.031250	0.980785	0.195090	5	10
51.474807	-0.208560	0.260417	-0.065403	0.997859	0.465774	-0.976966	0.213396	31	6
51.601963	-0.095339	0.968750	0.980785	-0.195090	0.281250	-0.195090	0.980785	2	8

Table 8: Sample of data with all the independent features and the dependent at the end gps_cnt.

Evaluation

To evaluate the performance of the random forest model, the data is randomly divided into training and testing sets, following a 80% and 20% split, typical in machine learning workflows (Géron, 2019). As in most of the ML models, we train the model on a training set so it 'learns' how to predict the target variable and then we check the model's prediction error on the testing set, by keeping the target hidden and comparing actual and predicted values.

Regression Performance Metrics

The performance metric used is the mean absolute error (MAE) which measures the average absolute differences between the true and predicted values. The lower the MAE, the better the model predicts. However, we need to assign a baseline threshold and try to keep the model's MAE below that. The targeted variable of gps_cnt follows a distribution with mean 5.3 GPS events, maximum value 785 and standard deviation 10.4 as seen in figure 41.

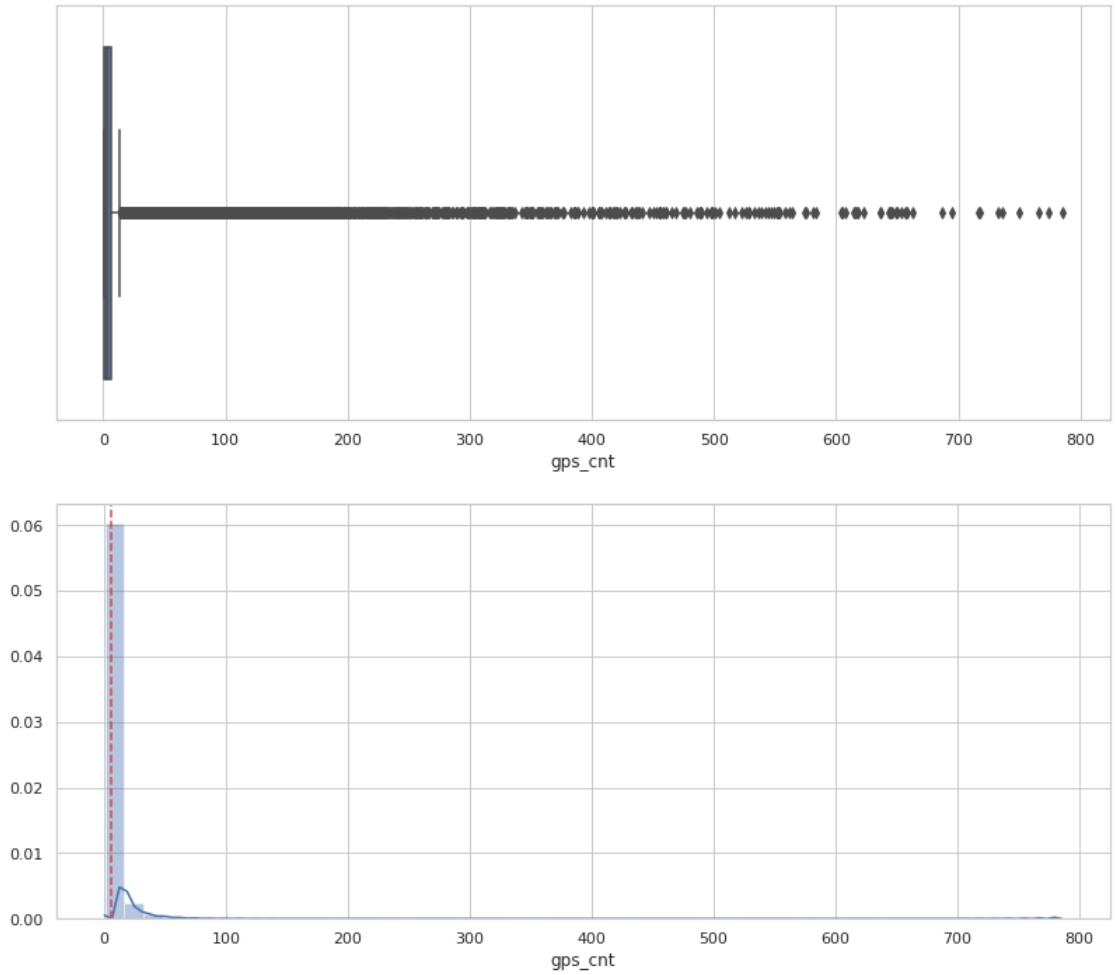

Figure 41: Top: Box plot of gps_cnt with extreme values on the right, Bottom: Distribution of gps_cnt values with mean value on red dashed line.

In such distribution where the standard deviation is so smaller than the maximum value, one could say that its mean has been affected by extreme values (or ‘outliers’ in other problem definitions). Therefore, Interquartile Range (IQR) filtering is applied to remove the high values (Upton and Cook, 1996) and generalize our baseline threshold. Although the IQR is the distance between the upper and lower quartiles, i.e. $IQR = Q_3 - Q_1$, when detecting outliers the IQR filtering ranges from the first quartile minus 1.5 times the IQR to the third quartile plus 1.5 times IQR. Any data outside this range is considered to be an outlier and hence removed (“Identifying outliers with the 1.5xIQR rule,” n.d.).

The resulted dataset is described in figure 42 where there are still some extreme values on the right of the whisker plot on top.

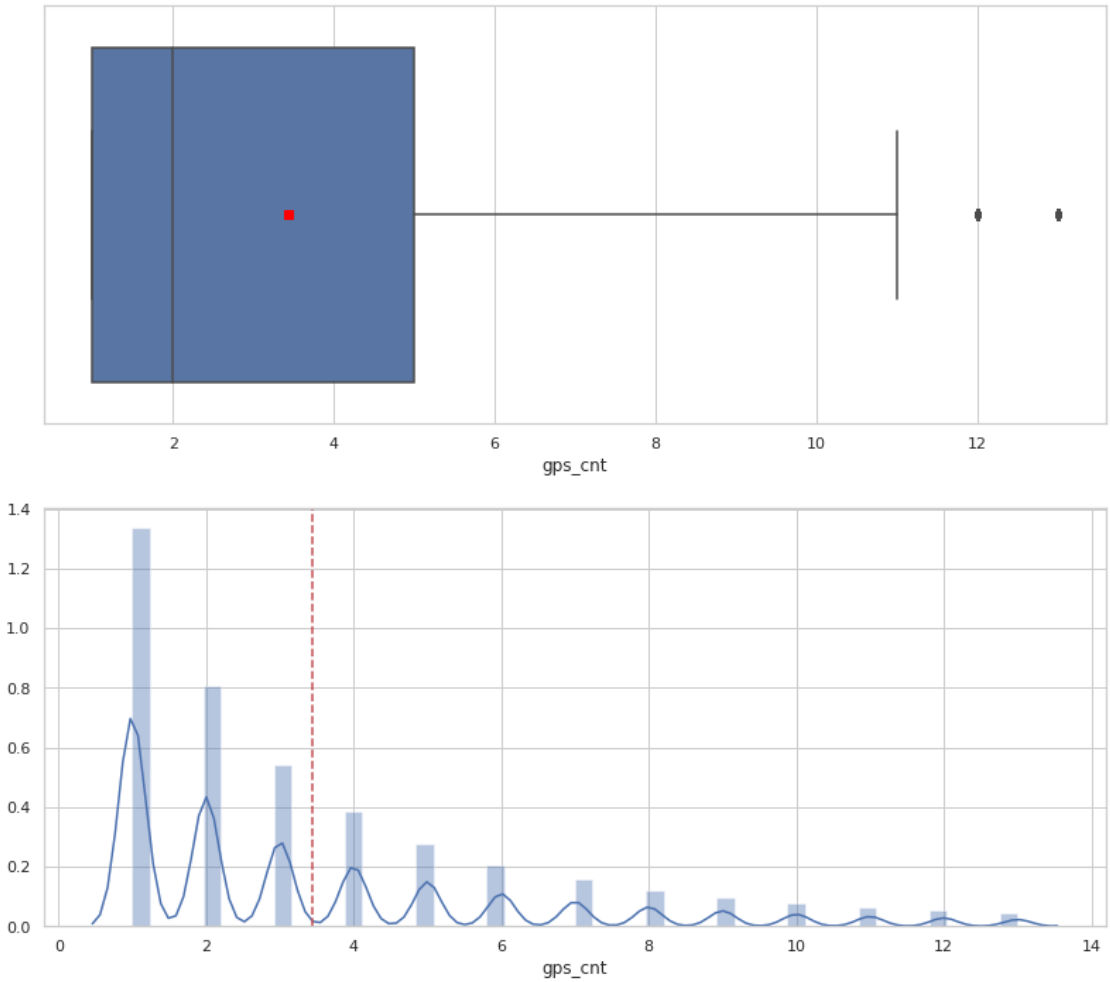

Figure 42: top: 1.5 IQR box plot of gps_cnt with less extreme values on the right than with no IQR (red box mean), Bottom: 1.5*IQR distribution of gps_cnt values with mean value on red dashed line.*

However, we set stricter rules and therefore we define for filtering range only the IQR to get lower mean value. After the elimination of the extreme values, we end up with a gps_cnt distribution of mean 2.46 and standard deviation 1.5 as shown below.

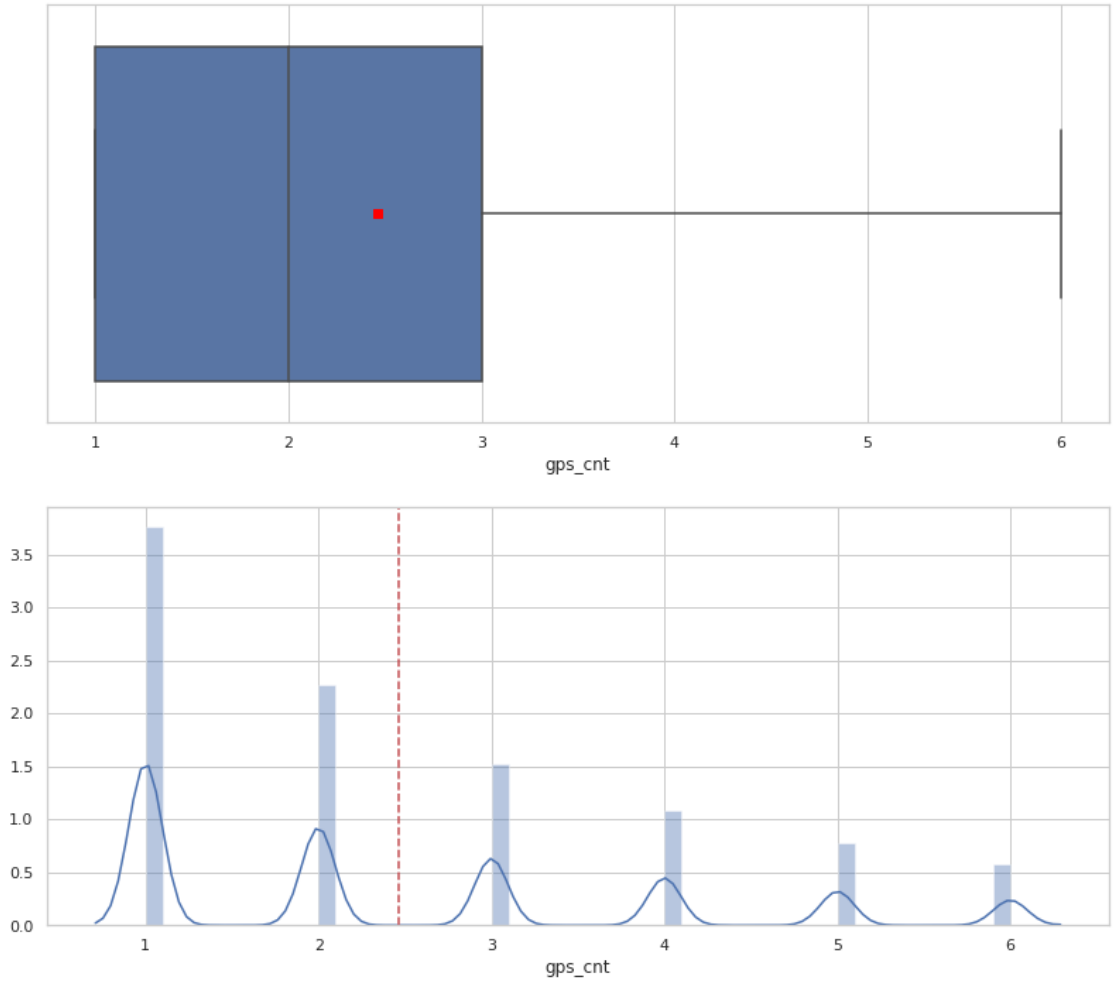

Figure 43: Top: 1 IQR box plot of gps_cnt with no extreme values on the right (red box mean), Bottom: 1*IQR distribution of gps_cnt values with mean value on red dashed line.*

A summary of the IQR filtering iterations is shown in the table 9. In the middle, the 1*IQR mean value dictates the baseline threshold of our model and our goal to achieve MAE below the value of 2.46 GPS events per hexagon.

No IQR		1*IQR		1.5*IQR	
count	1439544.000000	count	1136092.000000	count	1331389.000000
mean	5.313884	mean	2.460417	mean	3.441172
std	10.476902	std	1.544571	std	2.857279
min	1.000000	min	1.000000	min	1.000000
25%	1.000000	25%	1.000000	25%	1.000000
50%	3.000000	50%	2.000000	50%	2.000000
75%	6.000000	75%	3.000000	75%	5.000000
max	785.000000	max	6.000000	max	13.000000

Table 9: IQR filtering iterations of gps_cnt.

The R-Squared is also used as a performance metric to understand how the independent variables explain the variance in our model (Guanga, 2019).

Results

Our random forest model performs very well with an R-squared on the test data of 0.8829, which means that the model explains over 88% of the variation in the footfall density distribution. Also, the resulted MAE is 1.93 GPS events which is indeed lower than the threshold of 2.46 that we set up as baseline.

	real_cnt	pred_cnt
163531	5	2.31
261786	4	4.88
115358	9	4.10
66854	56	52.81
245868	4	4.19
284068	4	2.33
19705	14	10.93
182165	6	5.80
115497	13	8.59
24763	1	2.01

Table 10: Ten randomly picked real values and their predicted ones.

Feature Importance

In order to quantifiably measure the contribution of each feature in to the prediction model, we extract the variable importance values from the random forest model. This can help us to identify the variables with the most predictive power meaning that their values have a significant impact on the targeted variable. Therefore, regarding the features with high importance to the random forest estimation model, clearly POI density holds the highest contribution, followed closely by location (longitude, latitude) and then less by day and time as seen in table 11 and figure 44.

```
Variable: poi_cnt   Importance: 0.27
Variable: hex9_lon  Importance: 0.25
Variable: hex9_lat  Importance: 0.18
Variable: day_num   Importance: 0.09
Variable: time_num  Importance: 0.06
Variable: time_cos  Importance: 0.06
Variable: time_sin  Importance: 0.03
Variable: day_cos   Importance: 0.03
Variable: day_sin   Importance: 0.03
```

Table 11: Variable importance values.

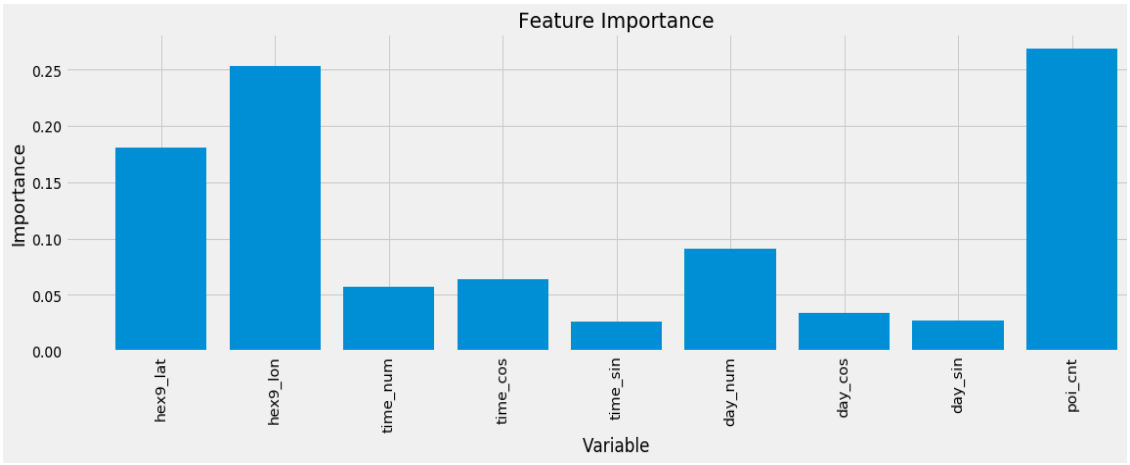

Figure 44: Importance across independent features.

Demonstration

The figure below shows the area around University College London, centred at latitude: 51.52335, longitude: -0.13380 with a radius of 1.5 km, converted into H3 hexagonal cells.

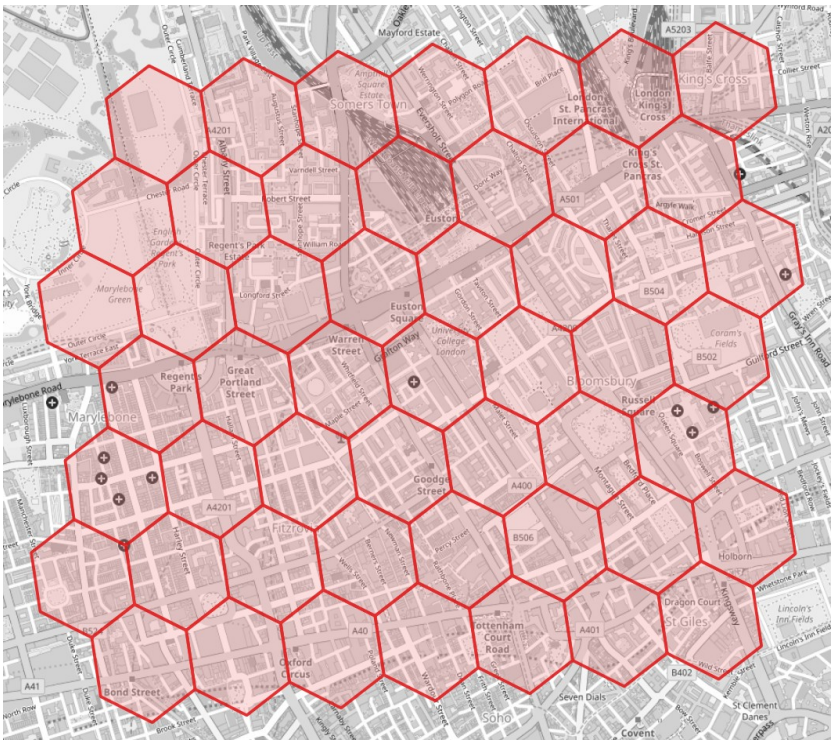

Figure 45: The area around the UCL campus in central London converted in H3 cells.

In an attempt to assess the results of our model, we visualize on figure 46 the predicted number of footfall density on a given Monday at 13:00 using the actual number of GPS events on the left and the random forest regressor model prediction on the right.

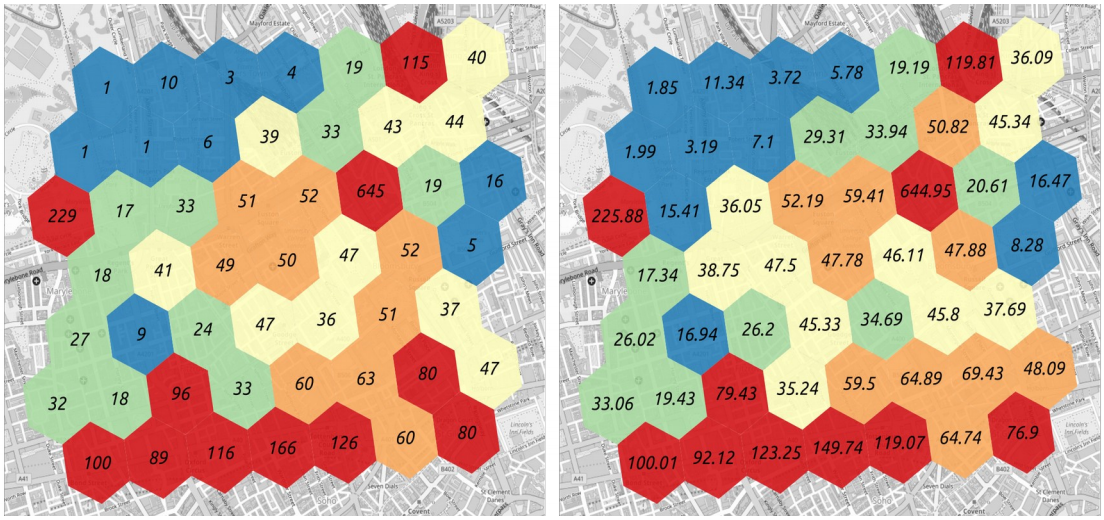

Figure 46: The area around the UCL campus in central London with actual values on the right and predicted ones on the left.

From this example it is clear that the real and predicted values are close to each other as seen in table 12, proving that our model performs well and hence being able to estimate footfall densities with high levels of accuracy.

	hex9	pred_cnt	actual_cnt ▲
1	89195da4d...	644.950000...	645
2	89195da48...	225.879999...	229
3	89195da49...	149.740000...	166
4	89195da49...	119.069999...	126
5	89195da49...	123.250000...	116
6	89195da4d...	119.810000...	115
7	89195da49...	100.010000...	100
8	89195da49...	79.4300000...	96
9	89195da49...	92.1200000...	89
10	89194ad32...	76.9000000...	80

Table 12: Top 10 hexagons in terms of footfall density around UCL in central London.

Conclusion

This research proposes and implements a novel method to estimate footfall density based on GPS trajectories and POI data geospatially indexed to the H3 hexagonal grid. Using H3's discritization functionality of geospatial data and geometric features we managed to develop a scalable indexing model to estimate footfall density within cities. Comparing to related methodologies using different grids, H3 reduces shape error in area representation when converting a boundary into hexagons and supports more efficiently movement data due to hexagon's uniform distances from neighbouring cells. Therefore, H3 could be more appropriate for analysing, exploring and predicting footfall densities. Using our model we were able to achieve 88% accuracy with a MAE of less than 2 GPS events per grid cell, which is much less than the dataset's mean of 5.3 GPS events per hexagon.

The results of this study show that using a geospatial indexing system like H3 could potentially be a key step towards efficient location-based targeting and footfall density forecasting. Our estimation model empowers the decision makers in the advertising industry to optimally design and provide location based marketing services by identifying not only upcoming future opportunities around POI but also retail patterns related to different types of products and services.

This research shows a strong relationship between POI and movement data since the former contributes the most amongst all the model features to increasing the predictive power of the random forest model. Having a better

understanding of such relationship would certainly help to design a city that can provide better service to the citizens. In addition to the potential improvements of this model, the proposed estimation method could also contribute to the field of trajectory data mining, introducing a different way of analysing movement patterns via the use of a geospatial grids.

While the use of H3 helps us have a better estimation, prediction, and generally a better understanding of the relationship between POI and footfall densities, further study to examine the proposed workflow in different areas and measure the importance of other factors, including cultural norms, census, demographic and socio-economic data would yield interesting findings. Also the use of finer data granularity (both spatially and temporally) would potentially result in different (perhaps higher) level of prediction accuracy.

In addition, it is expected that the higher frequency in the GPS sampling rate the higher the precision for the predictions is achieved. This also can be viewed as smaller errors, as trajectories would be more granular and detailed due to more movement nodes. One of the areas for further study is temporal patterns. A wider time window of trajectory data would allow to recognise different temporal patterns, e.g. bank holidays, sport events, extreme weather conditions etc. that could potentially change the footfall values.

This study proves that forecasting trajectory patterns representing people’s location histories within a city and extracting people’s most frequently visited locations from raw data can provide valuable information about human mobility

patterns and potentially such tools could contribute not only to location targeting advertising but also urban planning and management, vehicle tracking, monitoring, crowd monitoring, transportation planning, emergency management etc. (Luo et al., 2018).

Regarding future technical development, the conversion of the discrete hexagonal coordinates (hex9_lat, hex9_lon) to axial coordinates in the feature selection stage of the random forest model could be tested. During Uber Open Summit 2018, engineer Chong Sun (Uber Engineering, 2018c) highlights that better predictions are achieved when transforming the H3 hexagonal grid into axial one. In another growing market, that of the gaming industry, hexagonal grids are popular for character movement simulation, and conversion to axial coordinates has also proved to lead to more efficient estimations (“Introduction to Axial Coordinates for Hexagonal Tile-Based Games,” n.d.). However, additional research would appear to be necessary as this conversion hides potential problems of unused space as seen in the figure below (“Red Blob Games,” n.d.).

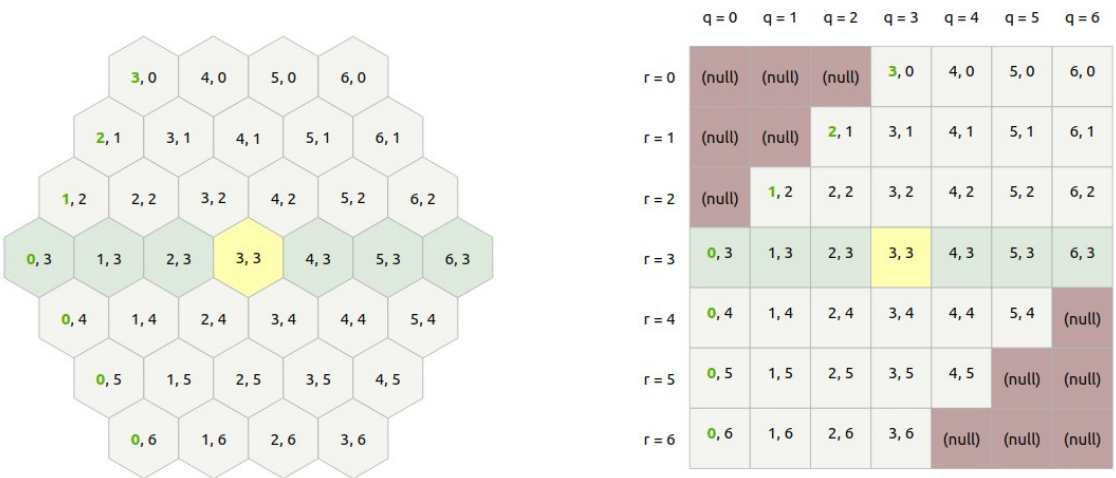

Figure 47: After the conversion null space appears on the top left and bottom right sides (“Red Blob Games,” n.d.).

A GitHub repository dedicated to this research can be found at the following link github.com/tetekos/footfall-density-uber-H3. This repository describes the coding implementation of the research's methodology in the programming language of Python. Any contribution to the repository would be appreciated, since the robustness of the proposed model and the findings of this study could unlock interesting research opportunities in trajectory mining and prediction. We hope that there will be further future research on this study to contribute to the optimization of location based targeting and footfall density estimation based on geospatial indexing systems.

Bibliography

2018 reform of EU data protection rules [WWW Document], n.d. . European Commission - European Commission. URL https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en (accessed 6.29.19).

Adjrad, M., Groves, P., 2017. 3D-mapping-aided GNSS exploiting Galileo for better accuracy in dense urban environments, in: 2017 European Navigation Conference (ENC). Presented at the 2017 European Navigation Conference (ENC), pp. 108–118. https://doi.org/10.1109/EURONAV.2017.7954199

Announcing the S2 Library: Geometry on the Sphere, 2017 . Google Open Source Blog. URL https://opensource.googleblog.com/2017/12/announcing-s2-library-geometry-on-sphere.html (accessed 6.20.19).

Balkić, Z., Šoštarić, D. & Horvat, G., 2012. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7327, pp.290–298.

Banerjee, S. (Sy), Roy Dholakia, R., 2012. Location□based mobile advertisements and gender targeting. Jnl of Res in Interact Mrkting 6, 198–214. https://doi.org/10.1108/17505931211274679

Banerjee, S. and Dholakia, R.R., 2008, “Does location based advertising work?”, International

Bao, J., Liu, P., Qin, X., Zhou, H., 2018. Understanding the effects of trip patterns on spatially aggregated crashes with large-scale taxi GPS data. Accident Analysis & Prevention 120, 281–294. https://doi.org/10.1016/j.aap.2018.08.014

Basiri, A., Jackson, M., Amirian, P., Pourabdollah, A., Sester, M., Winstanley, A., Moore, T., Zhang, L., 2016. Quality assessment of OpenStreetMap data using trajectory mining. Geo-spatial Information Science 19, 56–68. https://doi.org/10.1080/10095020.2016.1151213

Bauer, C., Strauss, C., 2016. Location-based advertising on mobile devices: A literature review and analysis. Manag Rev Q 66, 159–194. https://doi.org/10.1007/s11301-015-0118-z

Birch, C.P.D., Oom, S.P., Beecham, J.A., 2007. Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. Ecological Modelling 206, 347–359. https://doi.org/10.1016/j.ecolmodel.2007.03.041

Blis | Mobile Location and Behavioural Advertising Solutions [WWW Document], n.d. . Blis. URL https://blis.com/ (accessed 7.6.19).

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., Pentland, A., 2014. Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data. arXiv:1409.2983 [physics].

Bonsanto, A., 2019. Converts a polygon into a set of geohashes with arbitrary precision.: Bonsanto/polygon-geohasher.

Breiman, L., 1996. Technical note: Some properties of splitting criteria. Machine Learning, 24(1), pp.41–47.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

Brodsky, I., 2018. H3: Uber's Hexagonal Hierarchical Spatial Index [WWW Document]. Uber Engineering Blog. URL https://eng.uber.com/H3/ (accessed 7.6.19).

Bruner, G.C., Kumar, A., 2007. Attitude toward Location-based Advertising. Journal of Interactive Advertising 7, 3–15. https://doi.org/10.1080/15252019.2007.10722127

Chen, R., Fung, B.C.M., Mohammed, N., Desai, B.C., Wang, K., 2013. Privacy-preserving trajectory data publishing by local suppression. Information Sciences 231, 83–97. https://doi.org/10.1016/j.ins.2011.07.035

Computational geometry and spatial indexing on the sphere: google/s2geometry, 2019. . Google.

Davis, N., Raina, G., Jagannathan, K., 2018. Taxi Demand Forecasting: A HEDGE-Based Tessellation Strategy for Improved Accuracy. IEEE Transactions on Intelligent Transportation Systems 19, 3686–3697. https://doi.org/10.1109/TITS.2018.2860925

Decimal degrees, 2019. . Wikipedia. (accessed 6.29.19).

Dholakia, R.R., Dholakia, N., 2004. Mobility and markets: emerging outlines of m-commerce. Journal of Business Research 57, 1391–1396. https://doi.org/10.1016/S0148-2963(02)00427-7

Discrete Global Grid System (DGGS) - A new reference system - Geoawesomeness [WWW Document], 2017. URL https://geoawesomeness.com/discrete-global-grid-system-dggs-new-reference-system/ (accessed 7.6.19).

Early morning surge [WWW Document], 2017 . Uber Drivers Forum. URL https://uberpeople.net/threads/early-morning-surge.161590/ (accessed 7.6.19).

Earth Cube [WWW Document], n.d. . S2Geometry. URL http://s2geometry.io/resources/earthcube.html (accessed 6.20.19).

Ekawati, R., Suprihadi, U., 2018. Analysis of S2 (Spherical) Geometry Library Algorithm for GIS Geocoding Engineering. TELKOMNIKA 16, 334. https://doi.org/10.12928/telkomnika.v15i4.6985

Exploring H3 and tessellations on the sphere [WWW Document], 2018. URL https://observablehq.com/@tmcw/H3 (accessed 7.6.19).

Fang, Z., Luo, X., Keith, M.E., n.d. How Effective Is Location-Targeted Mobile Advertising? 5.

Fathy, Y., Barnaghi, P., Tafazolli, R., 2017. Distributed spatial indexing for the Internet of Things data management.

Fong, N.M., Fang, Z., Luo, X., 2015. Geo-Conquesting: Competitive Locational Targeting of Mobile Promotions. Journal of Marketing Research 52, 726–735. https://doi.org/10.1509/jmr.14.0229

Fortney, J., Rost, K., Warren, J., 2000. Comparing Alternative Methods of Measuring Geographic Access to Health Services. Health Services & Outcomes Research Methodology 1, 173–184. https://doi.org/10.1023/A:1012545106828

Fotheringham, A., Rogerson, P., 1993. GIS and Spatial Analytical Problems. International Journal of Geographical Information Systems 7, 3–19. https://doi.org/10.1080/02693799308901936

GADM [WWW Document], n.d. URL https://gadm.org/ (accessed 6.29.19).

Garreta, R. et al., 2017. Scikit-learn : machine learning simplified.

Geohash encoding/decoding [WWW Document], n.d. URL https://www.movable-type.co.uk/scripts/geohash.html (accessed 7.6.19).

GeoHash grid Aggregation | Elasticsearch Reference [7.1] | Elastic [WWW Document], n.d. URL https://www.elastic.co/guide/en/elasticsearch/reference/current/search-aggregations-bucket-geohashgrid-aggregation.html (accessed 6.18.19).

Geohash, 2019. . Wikipedia.

Geohashing Chat by User Proximity Tutorial | PubNub [WWW Document], 2014. URL https://www.pubnub.com/blog/2014-05-07-geohashing-chat-by-proximity/ (accessed 7.6.19).

Geometry on the Sphere: Google's S2 Library - Google Drive [WWW Document], n.d. URL https://docs.google.com/presentation/d/1Hl4KapfAENAOf4gv-pSngKwvS_jwNVHRPZTTDzXXn6Q/view#slide=id.i130 (accessed 6.20.19).

Géron, A., 2019. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition / Géron, Aurélien. 2nd ed.,

Ghose, A., Goldfarb, A., Han, S., 2012. How Is the Mobile Internet Different? Search Costs and Local Activities. Information Systems Research 24, 613–631. https://doi.org/10.1287/isre.1120.0453

GNSS Frequently Asked Questions - GPS [WWW Document], n.d. URL https://www.faa.gov/about/office_org/headquarters_offices/ato/service_units/techops/navservices/gnss/faq/gps/#1 (accessed 6.29.19).

Goodchild, M., Haining, R., 2003. GIS and spatial data analysis: Converging perspectives. Papers in Regional Science 83, 363–385. https://doi.org/10.1007/s10110-003-0190-y

GPS.gov: Performance Standards & Specifications [WWW Document], n.d. URL https://www.gps.gov/technical/ps/ (accessed 6.29.19).

Griffith, D.A., 1985. An Evaluation of Correction Techniques for Boundary Effects in Spatial Statistical Analysis: Contemporary Methods. Geographical Analysis 17, 81–88. https://doi.org/10.1111/j.1538-4632.1985.tb00828.x

Guanga, A., 2019. Understand Regression Performance Metrics [WWW Document]. Becoming Human: Artificial Intelligence Magazine. URL https://becominghuman.ai/understand-regression-performance-metrics-bdb0e7fcc1b3 (accessed 7.4.19).

H3 [WWW Document], n.d. URL https://uber.github.io/H3/#/ (accessed 6.16.19).

H3 [WWW Document], n.d. URL https://uber.github.io/H3/#/documentation/core-library/overview (accessed 6.22.19).

Hastie et al., 2009. The elements of statistical learning : data mining, inference, and prediction / Trevor Hastie, Robert Tibshirani, Jerome Friedman. Second., New York: Springer Verlag.

Hexagonal hierarchical geospatial indexing system. Contribute to uber/H3 development by creating an account on GitHub, 2019. . Uber Open Source.

Ho, T.K., 1995. Random Decision Forests, in: Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1, ICDAR '95. IEEE Computer Society, Washington, DC, USA, pp. 278–.

Huang, K., Li, G., Wang, J., 2018. Rapid retrieval strategy for massive remote sensing metadata based on GeoHash coding. Remote Sensing Letters 9, 1070–1078. https://doi.org/10.1080/2150704X.2018.1508907

Hühn, A.E., Khan, V.-J., Ketelaar, P., van 't Riet, J., Konig, R., Rozendaal, E., Batalas, N., Markopoulos, P., 2017. Does location congruence matter? A field study on the effects of location-based advertising on perceived ad intrusiveness, relevance & value. Computers in Human Behavior 73, 659–668. https://doi.org/10.1016/j.chb.2017.03.003

Identifying outliers with the 1.5xIQR rule [WWW Document], n.d. . Khan Academy. URL https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/identifying-outliers-iqr-rule (accessed 7.4.19).

Imgur, 2016. How Pokemon GO understands the Faces of the Earth [WWW Document]. Imgur. URL https://imgur.com/gallery/pI5vv (accessed 7.6.19).

Importance of good place (POI) data | Mobile Marketer [WWW Document], n.d. URL https://www.mobilemarketer.com/ex/mobilemarketer/cms/opinion/columns/21636.html (accessed 6.30.19).

Information & Examples – STCL, n.d. URL https://discreteglobalgrids.org/information/ (accessed 6.16.19).

Introduction to Axial Coordinates for Hexagonal Tile-Based Games [WWW Document], n.d. . Game Development Envato Tuts+. URL https://gamedevelopment.tutsplus.com/tutorials/introduction-to-axial-coordinates-for-hexagonal-tile-based-games--cms-28820 (accessed 7.6.19).

Isard, Walter. Methods of Regional Analysis : An Introduction to Regional Science /by Walter Isard in Association with David F. Bramhall [et Al.]. Cambridge, Mass.: Massachusetts Institute of Technology, 1960. Print. Regional Science Studies 4.

Journal of Mobile Marketing, Vol. 3 No. 2, pp. 68-75.

Leszczynski, A., Crampton, J., 2016. Introduction: Spatial Big Data and everyday life. Big Data & Society 3. https://doi.org/10.1177/2053951716661366

Li, A.H., Martin, A.P., 2017. Forest-type Regression with General Losses and Robust Forest, in: ICML.

Li, X., 2013. Storage and addressing scheme for practical hexagonal image processing. JEI 22, 010502. https://doi.org/10.1117/1.JEI.22.1.010502

Lloyd, C.D., 2014. Exploring spatial scale in geography. John Wiley & Sons, Chichester, West Sussex, ; Hoboken, NJ.

Longley, P., 2005. Geographical information systems: principles, techniques, management, and applications / edited by Paul A. Longley ... [et al.]., 2nd ed., abridged. ed. John Wiley & Sons, Hoboken, N.J.

Longley, P.A., Tobón, C., 2004. Spatial Dependence and Heterogeneity in Patterns of Hardship: An Intra-Urban Analysis. Annals of the Association of American Geographers 94, 503–519. https://doi.org/10.1111/j.1467-8306.2004.00411.x

BIA/Kelsey, 2017. Location-Targeted Mobile Ad Spend to Reach over $32 Billion in 2021, BIA Advisory Services. URL http://www.biakelsey.com/location-targeted-mobile-ad-spend-reach-32-billion-2021/ (accessed 8.3.19).

Luo, S., Ng, Y., Lim, T.Z.W., Tan, C.C.H., He, N., Manai, G., Li, Y., 2018. Improved Localisation Using Spatio-Temporal Data from Cellular Network, in: 2018 19th IEEE International Conference on Mobile Data Management (MDM). Presented at the 2018 19th IEEE International Conference on Mobile Data Management (MDM), pp. 56–65. https://doi.org/10.1109/MDM.2018.00022

Luo, X., Andrews, M., Fang, Z., Phang, C.W., 2014. Mobile Targeting. Management Science 60, 1738–1756. https://doi.org/10.1287/mnsc.2013.1836

Mahdavi-Amiri, A., Alderson, T., Samavati, F., 2016. Data Management Possibilities for Aperture 3 Hexagonal Discrete Global Grid Systems. http://dx.doi.org/10.11575/PRISM/30988

Maniaka, M., 2019. Interviewed by Eleftherios Sergios for the dissertation of MSc in the Centre for Advanced Spatial Analysis, Bartlett Faculty of the Build Environment, UCL., 25 May.

Marc Langheinrich author, 2019. Privacy in mobile and pervasive computing / Marc Langheinrich, Florian Schaub., Synthesis digital library of engineering and computer science. Morgan & Claypool, San Rafael, California.

Maurya, P., 2018. Lesser known things about Google's S2. Pramod Maurya. URL https://medium.com/@self.maurya/lesser-known-things-about-googles-s2-fea42f852f67 (accessed 6.20.19).

MondayMap: Yes. A Magnetic, Foldable Dymaxion Map | theDiagonal, 2016 URL http://thediagonal.com/2016/10/31/mondaymap-yes-a-magnetic-foldable-dymaxion-map/ (accessed 7.6.19).

Ng, Y., Pei, Y., Jin, Y., 2017. Footfall Count Estimation Techniques Using Mobile Data, in: 2017 18th IEEE International Conference on Mobile Data Management (MDM). Presented at the 2017 18th IEEE International Conference on Mobile Data Management (MDM), IEEE, Daejeon, South Korea, pp. 307–314. https://doi.org/10.1109/MDM.2017.49

NYC Taxi Data Prediction [WWW Document], n.d. URL https://sdaulton.github.io/TaxiPrediction/ (accessed 7.1.19).

OpenStreetMap [WWW Document], n.d. . OpenStreetMap. URL https://www.openstreetmap.org/ (accessed 6.30.19).

Overview [WWW Document], n.d. . S2Geometry. URL http://s2geometry.io/about/overview.html (accessed 6.20.19).

Paul E. Ceruzzi author, 2018. GPS / Paul E. Ceruzzi., The MIT Press Essential Knowledge series. The MIT Press, Cambridge, Massachusetts.

Peterson, P.R., 2017. Discrete Global Grid Systems, in: Richardson, D., Castree, N., Goodchild, M.F., Kobayashi, A., Liu, W., Marston, R.A. (Eds.), International Encyclopedia of Geography: People, the Earth, Environment and Technology. John Wiley & Sons, Ltd, Oxford, UK, pp. 1–10. https://doi.org/10.1002/9781118786352.wbieg1050

POI [WWW Document], n.d. URL https://www.ordnancesurvey.co.uk/business-and-government/help-and-support/products/points-of-interest.html (accessed 6.30.19).

Points of Interest [CSV geospatial data], Scale 1:1250, Items: 708329, Updated: 5 March 2019, Ordnance Survey (GB), Using: EDINA Digimap Ordnance Survey Service, https://digimap.edina.ac.uk, Downloaded: 2019-05-12 09:25:26.149

Prelec, D., Loewenstein, G., 1991. Decision Making over Time and under Uncertainty: A Common Approach. Management Science 37, 770–786.

PubNum. (2014). What is Geohashing?. [Online Video]. 9 May 2014. Available from: https://www.youtube.com/watch?v=T5q_zLk_4s8. [Accessed: 17 June 2019].

Purss, M.B.J., Gibb, R., Samavati, F., Peterson, P., Ben, J., 2016. The OGC® Discrete Global Grid System core standard: A framework for rapid geospatial

integration, in: 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). Presented at the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 3610–3613. https://doi.org/10.1109/IGARSS.2016.7729935

Ramasubramanian, L., Albrecht, J., 2018. Placemaking: Why Everything Is Local, in: Urban Book Series. pp. 87–110. https://doi.org/10.1007/978-3-319-68041-5_5

Raposo, P., Robinson, A., Brown, R., 2019. A Virtual Globe Using a Discrete Global Grid System to Illustrate the Modifiable Areal Unit Problem. Cartographica 54, 51–62. https://doi.org/10.3138/cart.54.1.2018-0015

Red Blob Games: Hexagonal Grids [WWW Document], 2019. URL https://www.redblobgames.com/grids/hexagons/ (accessed 6.22.19).

Red Blob Games: Hexagonal Grids [WWW Document], n.d. URL https://www.redblobgames.com/grids/hexagons/ (accessed 7.6.19).

Region Coverer [WWW Document], n.d. URL https://s2.sidewalklabs.com/regioncoverer/ (accessed 7.6.19).

Rogerson, P.A., 2006. Statistical Methods for Geography: A Student's Guide, Second edition. ed. SAGE Publications Ltd, Thousand Oaks, CA.

Rosenkrans, G., Myers, K., 2018. Optimizing Location-Based Mobile Advertising Using Predictive Analytics. Journal of Interactive Advertising 18, 43–54. https://doi.org/10.1080/15252019.2018.1441080

Roy, A., Pebesma, E., 2017. A Machine Learning Approach to Demographic Prediction using Geohashes, in: Proceedings of the 2nd International Workshop on Social Sensing - SocialSens'17. Presented at the the 2nd International Workshop, ACM Press, Pittsburgh, PA, USA, pp. 15–20. https://doi.org/10.1145/3055601.3055603

S2 Cells [WWW Document], n.d. . S2Geometry. URL http://s2geometry.io/devguide/s2cell_hierarchy.html (accessed 6.20.19).

Sahr, K., 2019. Central Place Indexing: Hierarchical Linear Indexing Systems for Mixed-Aperture Hexagonal Discrete Global Grid Systems. Cartographica: The International Journal for Geographic Information and Geovisualization 54, 16–29. https://doi.org/10.3138/cart.54.1.2018-0022

Sahr, K., White, D., Kimerling, A.J., 2003. Geodesic Discrete Global Grid Systems. Cartography and Geographic Information Science 30, 121–134. https://doi.org/10.1559/152304003100011090

scikit-learn: machine learning in Python — scikit-learn 0.21.2 documentation [WWW Document], n.d. URL https://scikit-learn.org/stable/ (accessed 7.3.19).

scikit-learn: machine learning in Python — scikit-learn 0.21.3 documentation [WWW Document], n.d. URL https://scikit-learn.org/stable/ (accessed 8.4.19).

Security, R., 2019. Origin of Global Positioning System (GPS) I Rewire Security. Rewire Security. URL https://www.rewiresecurity.co.uk/blog/gps-global-positioning-system-satellites (accessed 6.29.19).

Spatial Modelling Tidbits: Honeycomb or Fishnets? – Towards Data Science [WWW Document], n.d. URL https://towardsdatascience.com/spatial-modelling-tidbits-honeycomb-or-fishnets-7f0b19273aab (accessed 7.6.19).

Sundararaj, R., 2017 Mobility Data in the Retail Industry [WWW Document]. URL https://blog.datastreamx.com/mobility-data-in-the-retail-industry (accessed 6.27.19).

Taxi Demand Prediction System [WWW Document], n.d. URL http://athena.ecs.csus.edu/~shahhj/csc219/docs/Progress %20Report.pdf(accessed 7.1.19).

The Modifiable Areal Unit Problem and GIS, 2018. . GIS Lounge. URL https://www.gislounge.com/modifiable-areal-unit-problem-gis/ (accessed 6.22.19).

Timašjov, D., 2016. Human Mobility Mining Using Spatio-Temporal Data 62. Master’s Thesis. University of Tartu

Uber Engineering, 2018a. H3: Tiling the Earth with Hexagons.

Uber Engineering, 2018b. [Uber Open Summit 2018] Spatial Intelligence Using Hex.

Uber Engineering, 2018c [Uber Open Summit 2018] Hexagon Convolution for Data Smoothing & Forecasting.

Uber’s surge pricing secrets revealed as customers brace for price hikes over Christmas | Daily Mail Online [WWW Document], 2015. URL https://www.dailymail.co.uk/news/article-3352167/The-secrets-Uber-s-surge-pricing-revealed-customers-told-brace-extra-price-hikes-Christmas-season.html (accessed 7.6.19).

Uher, V., Gajdoš, P., Snášel, V., Lai, Y.-C., Radecký, M., 2019. Hierarchical Hexagonal Clustering and Indexing. Symmetry 11, 731. https://doi.org/10.3390/sym11060731

Unix time, 2019. . Wikipedia.

Upton, G., Cook, I., 1996. Understanding Statistics. OUP Oxford.

Van Meter, E.M., Lawson, A.B., Colabianchi, N., Nichols, M., Hibbert, J., Porter, D.E., Liese, A.D., 2010. An evaluation of edge effects in nutritional accessibility

and availability measures: a simulation study. International Journal of Health Geographics 9, 40. https://doi.org/10.1186/1476-072X-9-40

Van ’t Riet, J., Hühn, A., Ketelaar, P., Khan, V.-J., Konig, R., Rozendaal, E., Markopoulos, P., 2016. Investigating the Effects of Location-Based Advertising in the Supermarket: Does Goal Congruence Trump Location Congruence? Journal of Interactive Advertising 16, 31–43. https://doi.org/10.1080/15252019.2015.1135089

Viegas, J.M., Martinez, L.M., Silva, E.A., 2009. Effects of the Modifiable Areal Unit Problem on the Delineation of Traffic Analysis Zones. Environ Plann B Plann Des 36, 625–643. https://doi.org/10.1068/b34033

Wang, N., Vlachokostas, A., Borkum, M., Bergmann, H., Zaleski, S., 2019. Unique Building Identifier: A natural key for building data matching and its energy applications. Energy and Buildings 184, 230–241. https://doi.org/10.1016/j.enbuild.2018.11.052

Werner, M., 2015. BACR: set similarities with lower bounds and application to spatial trajectories, in: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS ’15. Presented at the the 23rd SIGSPATIAL International Conference, ACM Press, Bellevue, Washington, pp. 1–10. https://doi.org/10.1145/2820783.2820802

Zhao, L. et al., 2016. Geographical information system parallelization for spatial big data processing: a review. Cluster Computing, 19(1), pp.139–152.

Zheng, Y., 2015. Trajectory Data Mining: An Overview. ACM Trans. Intell. Syst. Technol. 6, 1–41. https://doi.org/10.1145/2743025

Zubcsek, P.P., Katona, Z. & Sarvary, M., 2017. Predicting Mobile Advertising Response Using Consumer Colocation Networks. Journal of Marketing, 81(4), pp.109–126.