TECHNISCHE UNIVERSITÄT WIEN

DIPLOMARBEIT

# Analysis of Movement Data Using ArcGIS in the Cloud

zur Erlangung des akademischen Grades

**Diplom-Ingenieur/in**

im Rahmen des Studiums

**Geodäsie und Geoinformation**

eingereicht von

**Julian Smidek**
Matrikelnummer 01126986

ausgeführt am Department für Geodäsie und Geoinformation
der Fakultät für Mathematik und Geoinformation der Technischen Universität Wien

Betreuung
Betreuer/in: Privatdoz. Dipl. Ing. Dr.techn. Gerhard Navratil

Wien, 09.05.2018

(Unterschrift Verfasser/in) (Unterschrift Betreuer/in)

# Acknowledgements

First, I would like to thank my supervisor Gerhard Navratil for the delightful support he gave when I first conceived this topic and all the feedback on the thesis draft.

I would also like to thank Gernot Tutsch, Philip Glasner, Marko Einspieler and Christina Treschnitzer of *SynerGIS* for the cooperation and support throughout my thesis and Thomas Burgess and Matej Krejci of *indoo.rs* for providing assistance and the possibility to work with their data.

Special thanks go to my parents who gave me the opportunity to study and assisted and inspired me in so many ways. Finally, I want to thank my brother and all my close friends who remained at my side, supported and encouraged me in recent years.

# Abstract

Spatial data mining is a highly emerging field as a consequence of tremendous growth in spatial data collection. Such growth has been made possible through various applications, such as: remote sensing, GIS, environmental assessment, planning, web-based spatial data sharing, and location-based services. Through advanced spatial data mining methods and analysis, valuable knowledge can be extracted. The gained knowledge is used to support decision making based on spatial data. As data based decision making is becoming more and more important and a large proportion of data includes significant spatial components the use of spatial algorithms is becoming an important part of modern data mining.

For this thesis the used dataset is based on user data of a smartphone application for indoor navigation. This smartphone application was developed and designed for a fashion trade show in Copenhagen. This thesis evaluates, if it is possible to analyse this movement data to gain beneficial knowledge with the provided toolset of commercial GIS software. The functions that were provided by this software were embedded and adjusted in several scripts to automatically process datasets in post-processing. By testing the feasibility of these methods in post-processing the possibility of future real-time analysis can be evaluated as well. Furthermore, a comparison shall be made how processing large amounts of data differ from smaller datasets and if the use of cloud computing can improve possible issues.

In conclusion the study found that it is possible to extract valuable knowledge from the provided movement data despite certain limitations. However, such limitations are primarily related to the aspects of data acquisition rather than the data analysis methods. Firstly, in order to analyse some phenomena, for example detecting movement patterns, large amounts of data are necessary in a dense temporal structure. The weight of this limitation is even more severe for real-time applications. Secondly, a relatively high spatial accuracy is necessary in order to yield high quality results. Lastly, some issues related to pre-processing tasks could be observed, especially concerning coordinate transformations.

# Kurzfassung

Räumliches Data Mining ist ein schnell wachsendes Anwendungsgebiet durch das immense Wachstum in der Verfügbarkeit von räumlichen Daten. Der Grund für dieses Wachstum ist gegeben durch diverse Anwendungen wie: Fernerkundung, GIS, städtische Planungsvorhaben, Social Media und Location Based Services. Durch Anwendung von Methoden des räumlichen Data Minings und Datenanalyse kann wertvolles Wissen extrahiert werden. Dieses Wissen kann dazu verwendet werden um Entscheidungen auf Basis der räumlichen Datengrundlage zu treffen. Diese Datenbasierten Entscheidungen werden immer wichtiger. Da ein Großteil der verfügbaren Daten einen signifikanten räumlichen Anteil hat, ist auch die Anwendung räumlicher Algorithmen ein essenzieller Bestandteil von modernen Data Mining.

Der in dieser Diplomarbeit verwendete Datensatz basiert auf Benutzerdaten einer Smartphone Applikation für Indoor Navigation. Diese Applikation wurde für eine Messe in Kopenhagen entwickelt. In dieser Studie soll evaluiert werden, ob es möglich ist diese Bewegungsdaten mit den Werkzeugen einer kommerziellen GIS Software zu analysieren um wertvolles Wissen zu erlangen. Die von der Software bereitgestellten Funktionen wurden in eigens erstellte Python Skripte implementiert um als Post Processing Datensätze automatisch analysieren zu können. Indem die Möglichkeiten dieser Methoden im Post Processing getestet wurden ist es möglich eine Vorhersage über eine potentielle Anwendung für Echtzeitanwendungen zu treffen. Es wurde außerdem evaluiert welchen Einfluss die Prozessierung der Daten in der Cloud hat.

Die Ergebnisse dieser Studie sind, dass es durchaus möglich ist aus dem vorhandenen Datensatz, wertvolles Wissen zu extrahieren, obwohl gewisse Limitierungen beobachtet werden konnten. Diese Limitierungen, beziehen sich vorwiegend auf Aspekte der Datenakquisition. Erstens, um gewisse Phänomene wie Bewegungsmuster detektieren zu können, sind große Datenmengen in einem dichten Zeitrahmen notwendig. Diese Limitierung trifft umso mehr zu falls Echtzeitanwendungen angedacht werden. Es ist außerdem eine relativ hohe Positionierungsgenauigkeit erforderlich um qualitativ hochwertige Ergebnisse liefern zu können. Es konnten außerdem einige Erschwernisse im Pre-Processing beobachtet werden, welche vor allem auf Schwierigkeiten der Koordinatenübertragung von dem gegebenen lokalen Koordinatensystem auf ein globales Koordinatensystem zurückzuführen sind.

# Table of Content

# List of Figures

# List of Tables

# List of Equations

# 1 Introduction

Data Analysis is "*the process of examining information, especially using a computer, in order to find something out, or to help with making decisions*".[1] In this thesis it shall be discussed, which questions can be answered by using methods of geospatial data analysis on movement data and to what extent. In particular, the analysis is performed in a commercial GIS (Geographic Information System) software environment. This thesis will outline the requirements on necessary data quality and what knowledge can be extracted from a particular dataset. Differences will be compared, that can be found when the analysis is performed conventionally or in the cloud in a big data context. As a result, this thesis will discuss, what is the actual value hidden in a movement data and by what means it can be unveiled.

## 1.1 Purpose and Motivation

In recent years the availability of geospatial data has increased tremendously due to manifold reasons. With the recent growth of Volunteered Geographic Information (*VGI*), Open Government Data (*OGD*) and the everyday use of way finding with GPS, enormous amounts of geospatial data is created every day. Nowadays there is also enough capacity available to store this data, use comprehensive analysis methods and visualize results in great detail on the fly. This can lead to the building of a foundation of knowledge base which can be a great benefit when it comes to decision making. Recently in this context the term Business Intelligence (BI) is becoming more and more popular, which means *'methods and technologies that gather, store, report, and analyse business data to help people make business decisions'*.[2] Business, government and science organizations are increasingly moving toward decision-making processes that are based on information. In parallel, the amount of data representing the activities of organizations that is stored in databases is also growing. Therefore, the pressure to extract as much useful information as possible from this data is very strong.

This recent development is also impacting the field of GIS. The use of geospatial data mining techniques introduces new ways of extracting knowledge. Much data include significant spatial components (estimates range between 50% and 85%) which make it even more important to handle methods that are specifically focused on processing spatial information. Analysing movement data is an interesting challenge in the context of geospatial data mining. Valuable information about the behaviour of an individual can be provided by analysing the features of his trajectory and the places this person visited.

[1] https://dictionary.cambridge.org/dictionary/english/data-analysis last access 2018 - 03
[2] http://www.dictionary.com/browse/business-intelligence last access 2018 - 03

The analysis of movement data consists of various steps, including a data assessment, pre-processing, actual data analysis tasks and an evaluation of results. This data assessment is an essential first step in order to understand the characteristics of an available dataset and determine possible ways to mine data. Pre-processing is usually a necessary step before data analysis tasks to enrich a dataset with additional fields, change values of fields for easier processing and coordinate transformations, specifically for working with geospatial data.

Nowadays, an automation of such data analysis tools is becoming more and more popular. The aim is to process similar datasets by scripted tools which decrease the necessary processing times by depleting manual processing steps. Cloud computing is introducing many benefits for storing, handling and processing data, especially big datasets. This thesis shall also evaluate how this computing technique can improve data analysis tasks, especially concerning processing times.

The tools used in the scope of this thesis are mostly commercial software developed by *ESRI*. A wide range of tools for storing, handling and analysing geospatial data is provided by the software *ArcGIS Pro*. An aspect of the usage of these tools is that they can be modified and used for individual scripts with the programming language *Python*. This ability introduces ways to automatically process datasets in many desired ways. *ESRI* is also providing a new software product for cloud computing called *ArcGIS Enterprise*. The tools provided by this cloud computing solution introduce new ways for a flexible data management and analysis, and also a flexible way to distribute data and maps through the internet.

The hypothesis that shall be answered in this thesis is if knowledge can extracted from movement data by using the commercial GIS software *ArcGIS* and more specifically, if the provided tools allow to detect movement patterns. Proving this hypothesis can be considered as a first step to realize future real-time detection of movement patterns by means that are applied within this thesis.

## 1.2 Structure of Thesis

The thesis will be finished with the following structure.

**Chapter 1 - Introduction:** This chapter gives a brief introduction of this project, the motivation and the research objectives.

**Chapter 2 - Theoretical Foundation:** In this chapter, the research background, the previous work achieved and related scientific papers are introduced. It shall give an overview of Geographic Information, Geographic Information Systems, methods of Spatial Data Mining, Big Data and Indoor Navigation techniques.

**Chapter 3 - Case Evaluation:** This chapter introduces the dataset that was used for this data analysis. It will explain how and where data acquisition was performed and what knowledge could be extracted.

**Chapter 4 - Methodology:** This chapter starts by giving detailed information about the dataset and leads then to the different steps of the proposed workflow. The workflow consists of pre-processing, data analysis and finally a comprehensive comparison of performing data analysis on conventional systems and cloud computing.

**Chapter 5 - Conclusion:** In the end, the findings are summarized and possible future work and limitations of the thesis are stated.

# 2 Theoretical Foundation

## 2.1 Geographic Information

There is a wide range of ambiguous and conflicting definitions for describing the term Information. At first, the relationship of Information to the terms Data, Knowledge and Wisdom needs to be clarified. This relation can be explained with the DIKW hierarchy, also known as the knowledge hierarchy, the information hierarchy and the knowledge pyramid (Fig. 1). Rowley (2007) gathered some established definitions of these terms to give an overview of how they are connected to each other (Table 1). (Baskarada, 2013)


**Fig. 1: The DIKW hierarchy - developed from (Rowley, 2007, p. 163)**

Geographic Information can be defined as information that link some properties, characteristics or phenomena to a location on or near the Earth's surface. This information usually consists of three components – space, time and attribute. The human perception of space is generally three dimensional. Every object can be described by length, width and height, and is located at some distance and direction from the others. When we are dealing with the scientific disciplines of GIScience and Cartography we mostly get in touch with large geographic regions, where the interest can be limited to a thin shell of the earth's surface. For much mapping and GIS this space is mainly two dimensional.

Time often plays a silent role in maps by implicit or explicit temporal reference. A map commonly works like a snapshot – valid for a specific moment in time. The interplay of time and space can be described by time geographers by a diagram (Hägerstrand 1970). In GIS applications the time component will usually have more significant role than in mapping. When visualizing geographic information with GIS applications rather than with maps, changing geometries or attributes of an object can be updated. Such application can also be realised as a real-time implementation.

The third component of geographic information, the attribute, can range from observable physical quantities to any arbitrary information that is linked to a location. Information extracted from the time and space component, such as velocity, can also be treated as an attribute.

| | |
|---|---|
| **Wisdom** | Wisdom is accumulated knowledge, which allows you to understand how to apply concepts from one domain to new situations or problems (Jessup and Valacich, 2003).<br>Wisdom is the highest level of abstraction, with vision foresight and the ability to see beyond the horizon (Awad and Ghaziri, 2004, p. 40).<br>Wisdom is the ability to act critically or practically in any given situation. It is based on ethical judgement related to an individual's belief system (Jashapara, 2005, pp. 17-18). |
| **Knowledge** | Knowledge is the combination of data and information, to which is added expert opinion, skills, and experience, to result in a valuable asset which can be used to aid decision making (Chaffey and Wood, 2005, p. 223).<br>Knowledge is data and/or information that have been organised and processed to convey understanding, experience, accumulated learning, and expertise as they apply to a current problem or activity (Turban et al., 2005, p. 38).<br>Knowledge builds on information that is extracted from data ... While data is a property of things, knowledge is a property of people that predisposes them to act in a particular way (Boddy et al., 2005, p. 9). |
| **Information** | Information is data which adds value to the understanding of a subject (Chaffey and Wood, 2005, p. 233).<br>Information is data that have been shaped into a form that is meaningful and useful to human beings (Laudon and Laudon, 2006, p. 13).<br>Information is an aggregation of data that makes decision making easier (Awad and Ghaziri, 2004, p. 36). |
| **Data** | Data has no meaning or value because it is without context and interpretation (Jessup and Valacich, 2003, Bocij et al., 2003, Groff and Jones, 2003).<br>Data are discrete, objective facts or observations, which are unorganised and unprocessed, and do not convey any specific meaning (Awad and Ghaziri, 2004, Chaffey and Wood, 2005, Pearlson and Saunders, 2004, Bocij et al., 2003).<br>Data items are an elementary and recorded description of things, events, activities and transactions (Laudon and Laudon, 2006, Turban et al., 2005, Boddy et al., 2005). |

**Table 1: Defining Data, Information, Knowledge, Wisdom: ambiguous and/or conflicting definitions developed from (Rowley, 2007, pp. 170-174)**

Each of the three components of geographic information is measured with respect to a particular reference system. This reference system provides rules to interpret and generalize individual observations and to be able to repeat and compare results. The clearest reference systems are those established by explicit standards.

The spatial reference system of large geographic regions will most likely be a geodetic framework, while some local mapping project might still be performed using an isolated, planar reference system. As a result of the complexity of the actual shape of the earth, the geoid, a model is used instead. This model is usually realized by an ellipsoid. There are dozens of ellipsoids in use, each chosen to provide an optimal fit to the shape of the earth in the particular region. Furthermore, a geodetic datum must be established by specifying points through astronomical surveying. Some frequently used geodetic reference systems

are for instance WGS 84, especially because of its use as reference system used by GPS, International Terrestrial Reference Frame (ITRF88, 89, 90, 91, 92, 93, 94, 96, 97, 2000, 2005, 2008, 2014 – realizations of the International Terrestrial Reference System) and PZ-90, the current geodetic reference used by GLONASS.

While the use of a geodetic reference system is most appropriate for geographic information, most maps adopt a simpler geometric model. Through a process called projection, the geodetic coordinates can be transformed into Cartesian coordinates. There are hundreds of possible projections, though in practice only a small number will be used. Projections can be classified by the properties they preserve as well as by their geometry.

- Conformal or orthomorphic projections: A projection that preserves the correct shapes of small areas and maintains all angles at each point. Important conformal projections are Mercator, Transverse Mercator and Lambert conformal conic
- Equal-area projection: These projections preserve area measure but leads to distorting shapes in order to do that. Examples are Lambert cylindrical equal-area and Mollweide.
- Equidistant projection: These projections preserve distance from some standard point or line, for example an azimuthal equidistant projection centered on Chicago shows the correct distance between Chicago and any other point on the projection, but not between any other two points.
- Gnomonic projection: This projection uses the earth's center as its perspective point. All great circles are straight lines, regardless of the aspect. This is a useful projection for navigation purposes because great circles determine routes with the shortest distance. [3]

| Level of Measurement | Information Required |
|---|---|
| Nominal | Definitions of categories |
| Graded membership | Definition of categories plus degree of membership or distance from prototype |
| Ordinal | Definitions of categories plus ordering |
| Interval | Unit of measure plus zero point |
| Extensive ratio | Unit of measure (additive rule applies) |
| Cyclic ratio | Unit of measure plus length of cycle |
| Derived ratio | Unit of measure (ratio of units; weighting rule) |
| Counts | Definition of objects counted |
| Absolute | Type (probability, proportion, etc.) |

**Table 2: List of Attribute Reference Systems (Chrisman, 2001, p.33)**

[3] https://support.esri.com/en/other-resources/gis-dictionary last access 04 - 2018

Since the technology of temporal reference systems is quite ancient, the most common used systems, calendar and clocks, are much more established compared to spatial reference systems.

In contrast to the spatial and temporal reference systems, each particular attribute require its own reference system. In cartography some general rules are commonly used for attribute measurements and attribute scales. A list of attribute reference systems is illustrated in Table 2.

### 2.1.1 Geographic Information Systems

A Geographic Information System (*GIS*) can be described as a computer application capable of performing operations on geographic data. These operations include tasks such as analyzing, querying, visualizing, sharing and archiving. Modern GIS application can be used in any area of science dealing with phenomena distributed over Earth. Research on global climate change, patterns of disease and crimes or the distribution of plants and animals show the wide range of possibilities. (Goodchild, 1998)

Besides the use of GIS in research fields it is also being increasingly used in business as a powerful tool to analyse data that describes location (e.g. addresses, zip codes, geographic or cartesian coordinates, etc.). As most of the available data includes a geographical component the user is provided with the opportunity to acquire greater benefits from the data. Hereby it is possible to unleash the wealth of information that is locked up in the data. (Mennecke, 1996)

Nowadays GIS applications are not only limited to research and professional use. With the recent growth of Volunteered Geographic Information (*VGI*), Open Government Data (*OGD*) and the everyday use of way finding with GPS it is evident that modern GIS are also aiming for the general public as users.

### 2.1.2 GIS Software

To represent a geographic object in GIS a data representation needs to be established first. There are usually two different approaches to represent a geographical phenomenon: raster representation and vector representation. In a raster representation a regular grid of cells is used, where every cell represents the value of the attribute that represents the phenomenon – like Red/Green/Blue values in a digital image. Raster representation is usually used to represent attributes that are continuous over space, such as land cover or elevation. A vector model is commonly used to store spatially discrete objects. Every object is represented by a vector geometry – generally point, line or polygon and value fields that describe the non-spatial object properties, the so-called ‘attributes’, in a table.

GIS Software encompasses a wide range of applications with different functionalities. Fig. 2 summarizes commonly used GIS software categories. Desktop GIS usually serves all GIS tasks

and is sometimes classified into three functional categories: GIS Viewer, GIS Editor, and GIS Analyst. The main purpose of Spatial Database Management Systems (DBMS) is to store the data but often also provides data manipulation functionality. WebMap Servers are used to distribute maps over the internet, similar to WebGIS Clients that are used to display, access and query data. Libraries and Extensions are providing additional functionality to analyse data that are not part of the basic GIS software. For instance functions for network and terrain analysis, or functions to read specific data formats might be added. Finally, Mobile GIS are often used for data acquisition in the field.

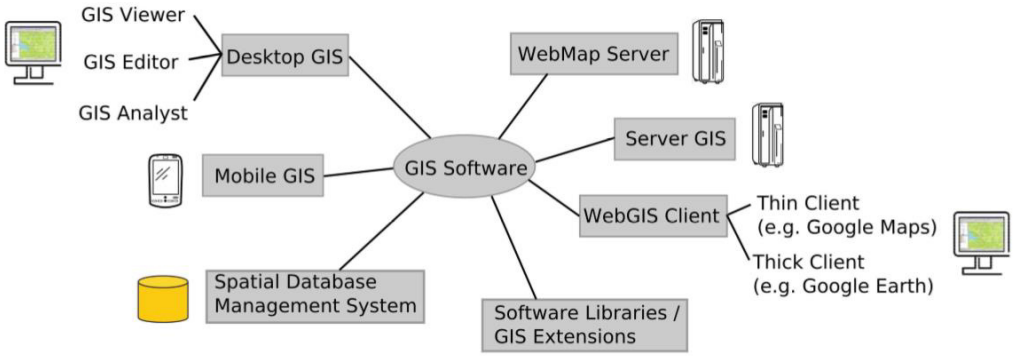



**Fig. 2: Categories of GIS Software (Steiniger & Weibel, 2009)**

GIS software is not only provided by companies but increasingly also by free and open source software projects. While commercial vendors usually offer products for all of software categories, open software projects often concentrate on a single category, e.g. desktop GIS or WebMap server. The key players in the GIS software market today are Autodesk, Bentley, ESRI, GE (Smallworld), Pitney Bowes (MapInfo), and Intergraph. GIS software companies tend to target specific application domains. For instance, ESRI's ArcGIS product tends to be mainly used for business analysis, planning, and environmental applications, while Autodesk, GE and Bentley products are rather used in utility and facility management. Competitive GIS software that is developed by free software projects exists as well - especially with respect to server applications (MapServer, GeoServer) and spatial DBMS (PostGIS). Free desktop GIS projects, such as Quantum GIS and gvSIG, currently experience growing user communities. Such free GIS software rather complements the set of proprietary software instead of competing with it. (Steiniger & Weibel, 2009)

It has been shown that for analysing data with a desktop GIS customization of functions is becoming more popular. A common open source programming language that is frequently used in this context is *Python*. *Python* was created by Guido van Rossum with the first release in 1991.[4] It features a dynamic type system and supports multiple programming paradigms, including object-oriented, imperative, functional and procedural. Some GIS software products also provide *Python* packages to be able to perform functions, such as

[4] http://python-history.blogspot.co.at/2009/01/brief-timeline-of-python.html last access 2018 - 03

geographic data analysis, data conversion, data management, and map automation with *Python*. A widely used package is a*rcpy,* distributed by ESRI, which will be used for analysis purposes in this thesis. This enables other GIS specialists to review, modify or develop scripts for their own purposes.

Another programming language that is very commonly used in GIS is the *Structured Query Language* – SQL. This language was designed to query and manipulate data in a relational database management system (RDBMS). The SQL standard is defined by *The American National Standards Institute* (ANSI). In the context of the geodatabase, SQL can be used to access, create, and update simple data; in other words, data that does not participate in any geodatabase functionality such as networks, topology, terrains, parcel fabrics, schematics, relationship classes, geodatabase domains, or geodatabase replication.[5]

### 2.1.3 GIS for Decision Making

For several years GIS have been used in the natural resources, forestry and environmental industries. Only recently they have begun to be used for a broader array of business and management functions such as logistics, site and facilities management, marketing, decision making, and planning. The fact that businesses have begun to use GIS is not surprising, when considering the fact that much of the data that organizations are typically using includes significant spatial components (estimates range between 50% and 85%). Because of these reasons, an increasing number of businesses have begun to make substantial use of GIS for a variety of routine decision support and analysis applications such as market and demographic analyses (Mennecke, 2000).

Mennecke (2000) proposed the model shown in Fig. 3 to describe the unique features which are facilitated by GIS. In this model, the various characteristics of GIS, in comparison to a non-spatial Decision Support System (DSS), are highlighted by specifically noting the spatial data, spatial data models, and spatial query and reporting features that are part of GIS. This proposed model shows that while non-spatial Decision Support Systems and Geographic Information Systems possess many analogies, there are also important distinctions that must be made between these two types of systems.

A conceptual model of GIS (Fig. 4) portrays four GIS functions and related applications. These four functions are related to four unique activities of GIS in order to address the needs of business. These applicable GIS functions are spatial visualization, imaging, database management, decision modelling, as well as design and planning. Spatial imaging refers to the fundamental GIS means of representing data and information within a spatially defined coordinate system. The database management function represents the capability of GIS to store, manipulate, query and provide access to data. The decision modelling function

[5] http://desktop.arcgis.com/en/arcmap/10.3/manage-data/using-sql-with-gdbs/what-is-sql.htm last access 2018 - 03

represents the capability to provide support for analysis and decision-making. Lastly, the design and planning function represents the capability of GIS to be used to create, design, and plan. In addition to these unique functions, the model also describes several specific GIS applications towards these functions can be applied: spatial data collection and automated mapping, facility management, market analysis, transportation, logistics, strategic planning, decision-making, design and engineering. (Mennecke, 2000)

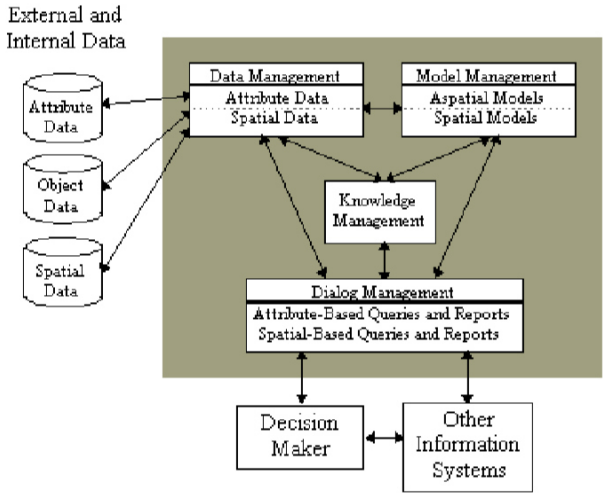



**Fig. 3: Model of a Geographic Information System Used For Decision Support (Mennecke, 2000)**

The ability to capture spatial data in order to generate maps automatically was one of the first applications of GIS. Computer systems which are designed to create digital maps represent powerful tools for business applications because of the capability to generate spatial data in-house. In addition, remote sensing and global navigation satellite systems (GNSS), such as GPS or GALILEO, allow more accurate map production. (Goodchild, 1992)

For a long time, GIS have been used extensively for facilities management (FM) in the public and private sector, for example, utility firms which represents one of the largest private-sector groups for GIS end-users. The key functions of GIS used in FM are visualizing spatial data and database management functions. They also rely heavily on the imaging capabilities of GIS to represent the spatial arrangement of data elements. The digital mapping functions of GIS are often combined with other FM functions to provide organizations with a system for generating, handling and utilizing maps and other spatial data that can be used to manage an organization's facility.

GIS is a powerful tool for market analysis applications because it provides a platform for representing the spatial relationship between the components of the market: that is, the customers, suppliers, and competitors. Strategies such as target marketing, micro marketing, and relationship marketing all require that firms capture and maintain detailed information about their customers. The ultimate goal of all of these efforts is usually to bring a product or service to someone, somewhere. This is why an understanding of the geo-demographic

characteristics of a companies customers is critical to a successful marketing strategy. In most cases, market analysis applications use historical or real-time data in combination with decision modeling and support tools to analyze marketing environment of an organization. Moreover, GIS is a powerful tool in market analyses because of the ability to provide a way to combine data from multiple sources and link them based on spatial attributes.

GIS and other spatial technologies are also becoming more and more essential tools for addressing logistics and transportation problems. For such applications these technologies are used as a platform for supporting decision modeling activities as well as a tool for displaying the results of these performed analyses. A variety of specific tools fit into this category of GIS. These tools include vehicle routing and navigation systems, intelligent vehicle highway systems, dispatch systems, production control systems, and inventory systems. Each of these functions represent useful applications that can be used by manager to develop tactics to reduce waste, lower personnel and fuel costs, and provide better customer service. (Azaz, 2011)

Computer aided design (CAD) and other design systems are widely used by engineering companies to create and archive architectural drawings. These technologies have been widely used for many years.

Similar to CAD systems, GIS technology can be used to design plans, layouts, and maps eventhough GIS is showing disparities in comparison to CAD systems. CAD systems have rudimentary links to databases, they deal with relatively small quantities of data, they do not usually allow users to assign symbology automatically based on user defined criteria, and they have limited analytical capabilities. (Maguire & Goodchild, 1991, p. 13)

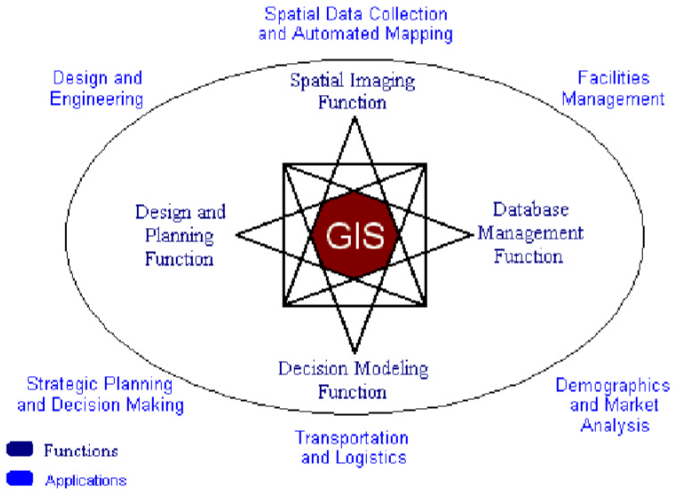



**Fig. 4: Functions and Applications (Mennecke, 2000)**

## 2.2 Time Geography

According to Kwan (2004), the origin of Time Geography leads back to a group of Swedish geographers – including, Torsten Hägerstrand, Tommy Carlstein, Bo Lenntorp and Don Parkes. Hägerstrand (1970) proposed a framework where he emphasized the importance of a time-variable in a geographical context. In this framework the relationships between various constraints and human activities on a spatiotemporal scale were examined. By establishing an integrated space-time system, time geography introduces a space-time path to describe a trajectory of an individual's movement in physical space over time. Furthermore, the concept of a space-time prism was introduced to depict the space-time extent that an individual can access (Fig. 5). Human activities and interactions are generally performed in physical space. While the representation of an individual's trajectory by a space-time path seems to be sufficient in order to describe activities and interactions of an individual, the recent growth of the Internet and smartphone technologies have changed the way of human communication. The information and communication technologies make it necessary to enable another space – the virtual space or cyberspace – where people can transmit information or connect and communicate electronically. As Geographic Information Systems are specifically designed to manage spatial data, they are considered a powerful tool to study human activities. Attempts have been made to investigate travel and activity data with GIS (e.g. Shaw and Wang, 2000) with focusing on human activities in physical space only and considering activities in both physical and virtual space (Kwan, 2000).

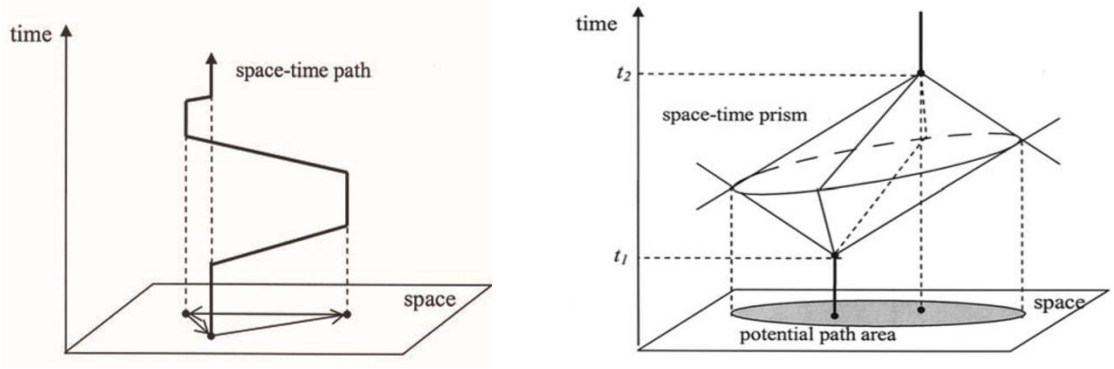



**Fig. 5: Illustration of space-time path and space-time prism (Yu, 2006)**

With the recent increase of location aware devices, the availability of trajectory information has grown as well. Various types of moving objects can be described, such as vehicles, animals and bank notes, as well as living things like athletes, visitors at mass events or tourists. This advanced ability of tracking information has led researchers from various fields the spatio-temporal data sets for the purpose of knowledge discovery.

## 2.3 Spatial Data Mining

Mining knowledge from large amounts of spatial data – spatial data mining – is a highly emerging field, as a tremendous amount of spatial data has been collected by various applications, such as remote sensing, GIS, environmental assessment, planning, web-based spatial data sharing as well as location-based services. The collected data is not only provided by various applications but also by a wide range of sources. These sources include public institutions, private companies, universities or general public. As a result of the huge amount of collected data, human's ability to analyse this data is by far exceeded. By applying spatial data mining methods on large spatial databases, interesting knowledge can be extracted. In particular, they can be used for understanding spatial data, discovering relationships between spatial and non-spatial data, construction of spatial knowledge-bases, query optimization, data reorganization in spatial databases and capturing general characteristics of datasets. This has wide applications in GIS, remote sensing, image database exploration and other areas where spatial data is used. Knowledge discovered from spatial data can be of various forms, like characteristic and discriminant rules, extraction and description of structures or clusters, spatial associations and others. With the recent growth of data mining, researchers proposed various methods for discovering knowledge from large databases. These methods rely on the use and the combination of already well developed areas like machine learning, databases and statistics. (Koperski et. al, 1997)

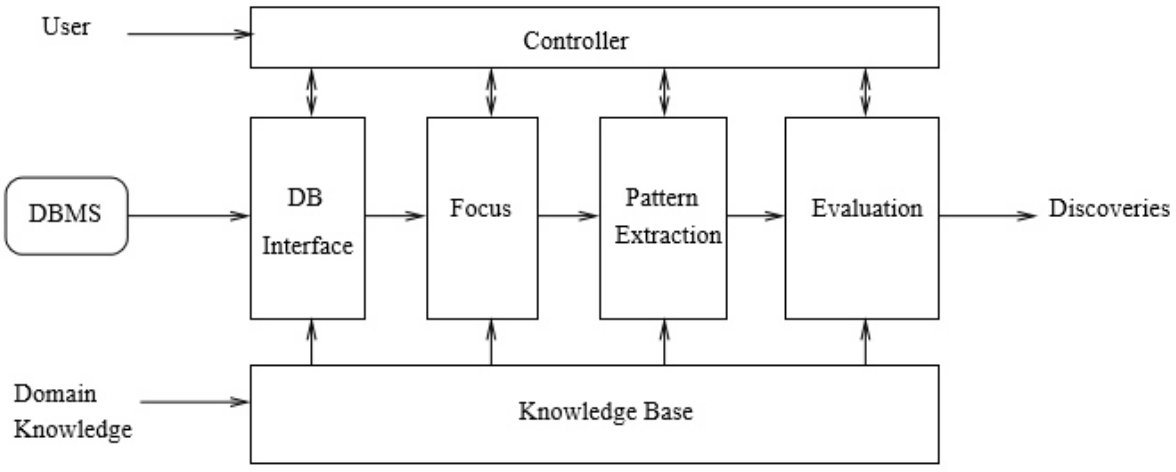



**Fig. 6: Spatial Data Mining Architecture (Koperski et. al. 1997)**

Data mining for the purpose of knowledge discovery is an iterative process that consists of multiple steps:

- Data selection, cleaning, pre-processing and transformation
- Incorporate prior knowledge

- Analysis with computational algorithms and visual assessment
- Evaluation and interpretation of results
- Formulation or modification of hypotheses and theories
- Adjustment to data and analysis methods
- Re-evaluation of results

(Fayyad et. al, 1996)

Various tasks can be tackled by spatial data mining approaches, and for each task there are a number of potential methods. Typical tasks include supervised classification, unsupervised classification, association rule mining and geovisualization.

### 2.3.1 Spatial classification and prediction

Spatial classification means grouping data items into categories depending on particular attribute values. Supervised classification needs a training dataset to configure the classification model, a validation dataset to validate the configuration, and a test dataset to evaluate the performance of the trained model. Classification methods include, for example, decision trees, artificial neural networks (ANN), maximum likelihood estimation (MLE), linear discriminant function (LDF), support vector machines (SVM), nearest neighbour methods and case-based reasoning (CBR). Spatial classification methods extend the general purpose of classification methods to consider not only attributes of the object to be classified but also the attributes of neighbouring objects and their spatial relations.

### 2.3.2 Spatial association rule mining

The original intention of association rule mining was to investigate correlations between items in large databases. Similar to mining of association rules in transactional or relational databases, spatial association rules can be mined in spatial databases by considering spatial properties and predicates. One example for a spatial association rule might be *"is_a(X,university)⇒inside(X,city)"*. Many different spatial predicates can be used in spatial association rules, e.g. intersect, close to, within, far away.

### 2.3.3 Spatial Clustering

The principle of cluster analysis is widely used in the field of data analysis. A set of items is organized into groups so that items in the same group are similar to each other and different to items in other groups. Many different clustering methods have been discovered in different research fields such as, statistics, pattern recognition, data mining, machine learning and spatial analysis. Two groups of clustering methods can be defined: partitioning clustering and hierarchical clustering. Partitioning clustering methods, such as k-means divides a set of data items into non-overlapping clusters. Depending on proximity or dissimilarity measures the data item is assigned to the closest cluster. Hierarchical clustering

organizes data in a set of nested partitions and groupings. Popular hierarchical clustering methods are Ward's method (Ward, 1963), single-linkage clustering, average-linkage clustering, and complete-linkage clustering (Gordon, 1996; Jain & Dubes, 1988).

### 2.3.4 Geovisualization

Geovisualization is dealing with theories and methods to create knowledge through visual exploration and geospatial analysis. The main difference between the scientific disciplines of traditional cartography and geovisualization is that cartography focuses on the design and use of maps as a device for information communication and geovisualization emphasizes the development of highly interactive maps and associated tools for data exploration, hypothesis generation and knowledge construction. (Guo et. al. 2009)

## 2.4 Big Data

The term "Big Data" appeared for the first time in 1998 in a Silicon Graphics (SGI) slide deck by John Mashey with the title "Big Data and the Next Wave of InfraStress" (Diebold 2012). The origin of this term is due to the fact that we are creating a huge amount of data every day. The amount of data produced nowadays is in the order of zettabytes and is growing around 40% per year. There is no need to distinguish the terms of "Big Data Analytics" and "Data Analytics" as the amount of data that is being collected and processed will keep growing.

In the recent past we have witnessed a massive increase of capabilities to create, process and store large amounts of data. As an example for this growth in the quantity of data we can consider the internet data. While in 1998 the web pages indexed by Google were around one million, it quickly reached one billion in 2000 and surpassed one trillion in 2008. Another big impact that we can consider is the dramatic increase in the acceptance of social network applications, such as Facebook, Twitter, Instagram, Weibo, etc. These applications allow the user to create content freely and might also be used by social network providers to create data based on the user actions when using them. With mobile phones becoming the sensory gateway to get real-time data on people from different aspects the rapid expansion in the growth of data is accelerated even further. It can be foreseen that Internet of things (IoT) applications will raise the scale of data to an unprecedented level. People and devices (from home coffee machines to cars, buses, railway stations, airports, etc.) are all loosely connected. These connected components will create an unforeseeable data ocean. When processing this data, the goal must be to discover valuable information that can be beneficial to the majority of users.

To deal with this data we need new algorithms and tools. The "five V's" of big data management can give as a summary of the most important aspects:

- Volume: There is more data than ever before and its size is still increasing.
- Variety: A huge diversity of data types can be observed, such as text, audio, video and all kinds of sensor data (Location, Time, Temperature, etc.).
- Velocity: Data is arriving as a continuous stream and should be processed in a proper timely manner – for some applications also in real time.
- Variability: The structure of data is changing repeatedly over time with the introduction of new technologies and services. The other aspect of variability is how users want to interpret data.
- Value: The value of processing big data for a user or an organization by receiving the advantage of making decisions based on answering the questions that were previously considered beyond reach.

Gartner[6] summarizes this in his definition of Big Data as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. As Big Data is a new and emerging hot topic, much controversy has been generated and many issues that come along with this new technology needs to be discussed. In future these issues might intensify what makes it even more important to raise this discussion.

Concerning the aspect of velocity of Big Data Management, the currency is more important than the size of data. Especially in real time analytics, where data is changing frequently, this is an important issue.

When looking on a large amount and a wide variability of data some misleading correlations might appear. For example, Leinweber (Leinweber, 2007.) showed that the S&P 500 stock index was correlated with butter production in Bangladesh, and other strange correlations.

It is important to keep in mind how the data was collected and how representative it can be considered. For example, sometimes Twitter users are assumed to be representative of the global population, when this is not always the case.

There are ethical concerns about the issue that people that are using smartphone applications, social networks or some internet platforms are being analysed without knowing it. These privacy issues need to be discussed and possibly legal actions need to be enforced to equalize the disparity between users and application providers.

If there is a limited access to Big Data technologies a new digital divide will be created. There might be a digital divide between persons and organizations that are able to use Big Data technologies and those who are not. This also applies to organizations with access to the

[6] http://www.gartner.com/it-glossary/bigdata last access 2018-03

opportunity to analyse Big Data and are able to extract knowledge that other without access have not. This might create a division between Big Data rich and poor organizations. (Fan et. al 2013)

## 2.5 Cloud Computing

The rapid development of processing, storing and Internet technologies has paved the way to a new computing model called cloud computing. In a cloud computing model, resources such as CPU and storage, are provided as general utilities that can be leased and released by users through the Internet in an on-demand fashion. In a cloud computing environment, the role of the service provider is divided into two: the infrastructure providers who manage cloud platforms and lease resources according to a usage-based pricing model, and service providers, who rent resources from infrastructure providers to serve the end users. A typical cloud computing architecture is illustrated in Fig. 7.

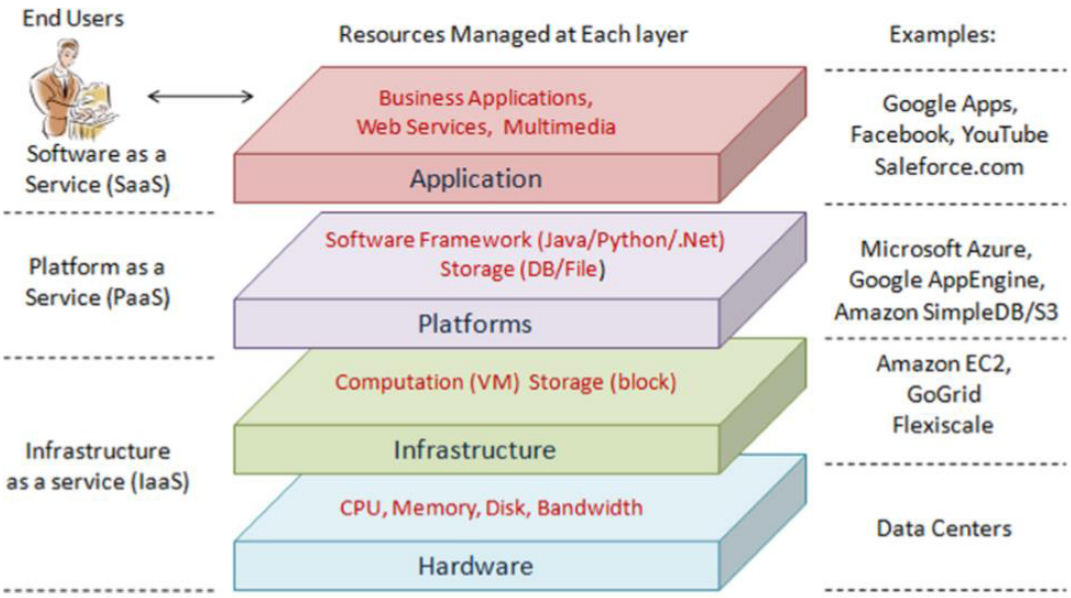



**Fig. 7: Cloud Computing Architecture (Zhang et. al 2010)**

The emergence of cloud computing technologies has made a big impact on the IT industry as the implementation provides attractive benefits to business aspects:

- No up-front costs: A service provider does not need to invest in the infrastructure, as cloud computing works on a pay-as-you-go basis. This means that it rents all necessary resources from the cloud according to the current needs and pay for the usage.

- Lowering operating cost: The needed resources can be allocated and re-allocated in a quick manner when working in a cloud environment. Hence, the service provider does not need to cover capacities according to the peak load. This allows huge savings on operating costs because the resources can be restrained when service demand is low.
- High scalability: Infrastructure providers are holding a large pool of resources that can be allocated if needed. Thereby service providers can easily expand their services to a large scale in order to meet the current demand on service tasks.
- Accessibility: Services hosted in the cloud are generally web-based. They are easily accessible by various devices that are able to connect to the Internet. These devices not only include desktop and laptop computers, but also cell phones and other mobile devices.
- Reducing business risk and maintenance expenses: Outsourcing the infrastructure into the cloud leads to a reduced business risk, as all maintenance activities (e.g. hardware failure) are no longer necessary.

Apart from the impressive advantages that a cloud computing architecture brings into the IT industry, there are still a number of open research questions concerning the disadvantages of this technology. There are many concerns about data security and how to detect a possible attack. In particular, the service provider must establish confidentiality, for secure data access and transfer, as well as auditability, to verify whether security requirements have been tampered or not. (Zhang et. al 2010)

# 2.6 Indoor Navigation

## 2.6.1 Introduction to Indoor Navigation

Navigation Systems can be classified by their usage environment into two categories, indoor and outdoor. Outdoor navigation is usually based on satellite techniques such as GPS, GLONASS and Galileo to locate an object in any outdoor area. Satellite based navigation systems are showing good performance in open spaces with a clear line of sight to the satellites. In contrast these techniques do not perform well in an indoor environment, as the signals get scattered and attenuated by physical objects. Even if the GPS signals are not blocked or obscured, the reception of GPS signals inside most buildings is not reliable. The wide availability of smartphones with multiple communication protocols such as Wi-Fi, Bluetooth, NFC, etc. has given way to new types of indoor navigation techniques. These Radio Frequency based localization techniques are widely used for indoor navigation. However, such a system needs comprehensive hardware infrastructure support.

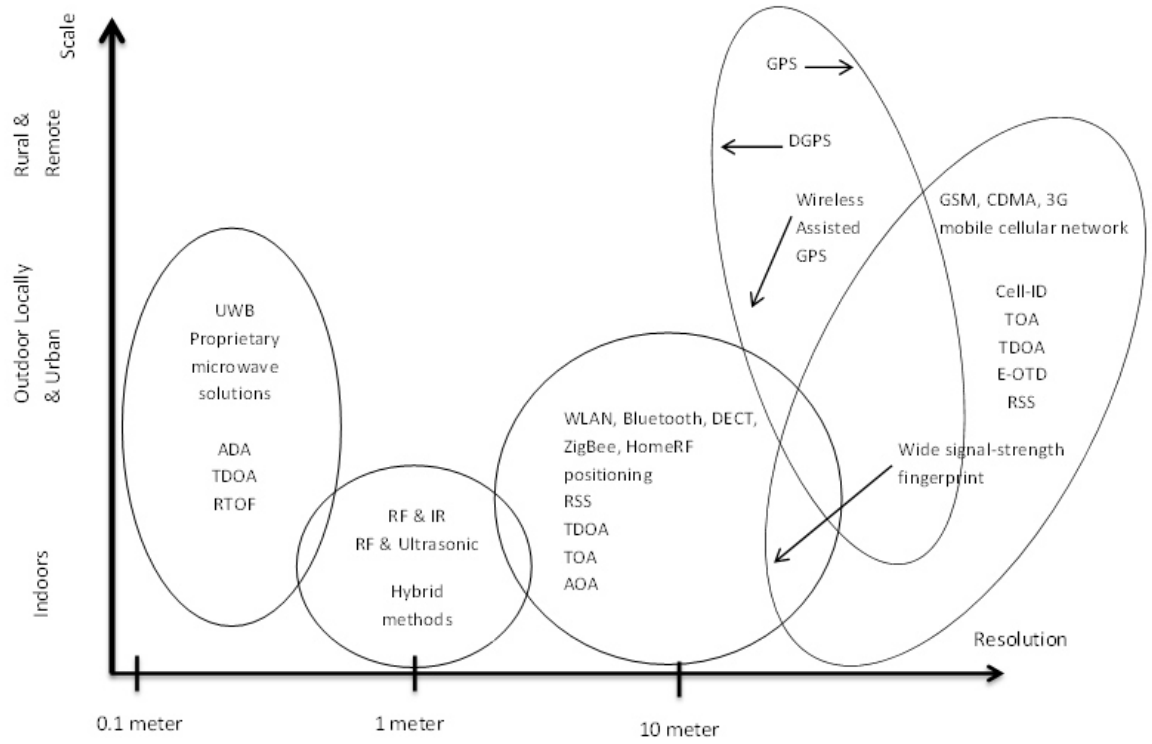



**Fig. 8: Resolution of wireless-based positioning systems (Brown et. al. 2011)**

Indoor Positioning is defined by Retscher as *"any system which attempts to provide an accurate positioning inside of a covered structure using radio waves, acoustic signals, or other sensory information collected by mobile devices. It is primarily used for real-time location of people or objects in large buildings and in closed areas/spaces. Several types of location-sensing systems exist in which each have its own strengths and limitations."* (Retscher 2016)

Following Li and Rizos (2014), indoor positioning technologies can be classified into three categories: technologies based on signal transmission infrastructure that is set up for the purpose of positioning, technologies based on "signal of opportunity" and not signal-based technologies. The first category includes systems using infrared or ultrasonic signals, magnetic fields, ultra-wideband (UWB), or other Radio Frequency-based systems (RF-based systems). Technologies based on "signal of opportunity" include all systems where the observed RF-signals are not intended for positioning, for instance wireless fidelity (Wi-Fi), digital television, mobile telephony, FM radio, and others. Inertial sensors, such as accelerometer and gyroscope, and vision camera systems fall into the third category. Each positioning technology has its own weaknesses, strengths and accuracy range. The characteristics and application area of the most important technologies are shown in Table 3 and Fig. 8.

| | Technology | Approx. accuracy | Coverage | Cost | | Strengths | Weaknesses |
|---|---|---|---|---|---|---|---|
| | | | | IC | UC | | |
| Technologies with signal encoding | Infrared | 57 cm - 2.3 m | Room | H | L | Cheap for user | Sunlight interference |
| | VLC | 10 cm | Building (ML) | H | L | Cheap for user, unintrusive | Expensive infrastructure |
| | Ultrasonic | 1 cm - 2 m | Room | H | H | Good precision | Cost, interference |
| | Audible Sound | Meters | Room | L | L | Low cost | Low precision |
| | Wi-Fi | 1.5 m | Building | L | L | Low cost, good precision | access point changes |
| | Bluetooth | 30 cm - meters | Building | L | L | Low cost, good precision | Intrusive, needs signal mapping |
| | ZigBee | 25 cm | Building | L | H | Could reuse infrastructure | Low precision, user needs special equip. |
| | RFID | 1 - 5 m | Room | H | L | Very low cost | Very low precision |
| | UWB | 15 cm | Building | H | H | High precision | High cost |
| Passive technologies without signal encoding | Geomagnetic | 2 m | - | L | L | No need for infrastructure, good precision | Requires Mapping |
| | Inertial | 2 m | - | L | L | Low cost, private | Accumulates error |
| | Ambient sound | Meters | - | L | L | Cheap, not intrusive | Not accurate, sensitive to changes |
| | Ambient light | 10 cm - meters | - | L | L | Cheap | Sensitive to sunlight and changes such as a bulb and a window |
| | Computer vision | 1 cm - 1 m | - | L | L | Low cost, privacy if cellphone camera is used | Sensitive to light conditions |

**Table 3: Comparison of Indoor Positioning Systems (Brena et. al. 2017)**

In the following chapter indoor positioning with Bluetooth is explained in further detail. This is because of the fact that the dataset which is processed in this thesis was obtained by an indoor navigation system based on Bluetooth technology.

## 2.6.2 Indoor Positioning with Bluetooth

Bluetooth is a wireless technology standard for exchanging data over short distances. It is a widely used technology among mobile devices, especially to establish a wireless connection between smartphones and other devices, and is also used to establish wide area personal networks (WPAN). Bluetooth as an indoor positioning technology has become increasingly popular as it is a cost-effective and easy-to-deploy solution. Furthermore, the system is energy efficient compared to other techniques such as Wi-Fi, GSM or GPS so that the beacons can run on a small battery for several months. A recently proposed technology is

Apple's *iBeacon*, which uses Bluetooth Low Energy (BLE) as signal transmission technique. This *iBeacon* acts as an emitter continuously broadcasting Bluetooth signals. Each signal contains a Universally Unique Identifier (UUID) and a Received Signal Strength Indicator (RSSI). The central peripheral relationship is one of the primary reasons that BLE consumes low energy. The beacons are acting as a peripheral device, so that they simply broadcast their information, while the central device (usually a smartphone) collect and process that data. (Lin et. al. 2015)

UUID is a unique identity for each beacon. Each location has the unique RSSI values and transmitting power values in order to evaluate the distance of a particular position. A typical positioning algorithm based on mapping RSSI values and transmitting power values that evaluates the distance in meter between BLE beacons and a Smartphone is described by Gast (2014):

1. Collect the RSSI and Transmitting power value for each location.

2. If RSSI=0, then it cannot determine accuracy and return to -1.0.

3. Calculate distance, D= rssi*1.0/Tx power.

4. Calculate best fit curve to measured data point.

The accuracy depends on nature of signal, signal attenuation and noise, which increases with signal strength. With Bluetooth technologies for indoor positioning a spatial accuracy of around 30 centimetres to several meters can be achieved. For most common indoor navigation tasks an accuracy of a few meters is most likely to be sufficient. If data of these navigation trajectories is recorded with the aim to analyse it a higher accuracy is favourable. Furthermore, in this case other data related issues may need to be addressed, for example recording rate, error estimation or outlier detection.

# 3 Case Evaluation

In the scope of this thesis spatial data mining and data analysis techniques are applied on a certain dataset for the purpose of knowledge extraction. In particular, they can be used for understanding spatial data, discovering relationships between spatial and non-spatial data, construction of spatial knowledge-bases and capturing general characteristics of datasets. The acquisition of data that is used for this analysis is described in the following chapter. In chapter 4.1 the provided dataset is specified and interpreted comprehensively. In this case evaluation data that was recorded with a smartphone application for the purpose of indoor navigation is analysed. The application was designed for a trade fair based on fashion with a focus on contemporary Scandinavian designers. This fashion trade show is called *Revolver Copenhagen Int.* which was inaugurated in February 2015.

## 3.1 Data Acquisition

The data was obtained by a smartphone application provided by the Austrian company *indoo.rs*. This application was developed to provide a free real-time indoor navigation service exclusively for the fashion trade show and is therefore named *Revolver*. As the system is based on an active indoor navigation basis certain steps needs to be followed in order to be able to use this service. This includes things such as setting up the hardware infrastructure and fingerprinting (see 2.5 Indoor Navigation). The navigation solution of *indoo.rs* provides the following characteristics:

- Primarily used for real-time indoor navigation
- High processing quality and processing speed
- Low battery consumption as positioning is only active when needed
- Due to the fact that positioning is only active when needed, there is only a small timespan when data is collected
- Positioning provides 2D - Coordinates with additional floor information
- Accuracy ranges from 3 - 5 m
- Recording frequency: 1 s
- *iPhone* and *Android* platforms are supported

The data was recorded in the years 2016 and 2017 on specific days when the exhibition was taking place. In total four exhibitions were considered for the data analysis. The origin of the dataset is located in an exhibition hall in Copenhagen, Denmark. Although all exhibitions took place in the same exhibition hall, it should be mentioned that differences between these events are to be expected. These differences might include issues such as changing locations of exhibitors and general adaptions of the exhibition hall. The navigation system is based on a number of *iBeacons* assembled in the venue. In this case around 300 Bluetooth

beacons were mounted across the exhibition hall and most of them were placed in Section 1 and Section 2.

The illustrations in Fig. 9 and Fig. 10 are based on the last event, which took place between 09-08-2017 and 11-08-2017. The exhibition site can be divided into five buildings that are connected to each other.

- Section 1: This part of the building is located in the southern part of the site. It is equipped with the main entrance, the main helpdesk, a wardrobe, one coffee shop and 71 booths for exhibitors. A main aisle connects this building to section 2.
- Section 2: This part of the exhibition site consists of 172 booths of exhibitors, two coffee shops and two toilets. It is connected to Section 1 by the main aisle and to Section 3 and Section 4 by two side aisles.
- Section 3: A restaurant and smoking lounge is located here. Side aisles are leading to Section 2, Section 4 and Section 5. In this section the positioning system works only to a little degree. Therefore, a small number of records are provided here.
- Section 4: This is the smallest building and features a showroom. Side aisles are leading to Section 2, Section 3 and Section 5. No beacons were mounted in this area so that no records can be found here.
- Section 5: The western part of the exhibition site consists of two floors and is harbouring 9 booths of exhibitor's. The other floors can be reached by lift or a staircase.

The main entrance to the site is located in the southern part together with a helpdesk and a wardrobe. The entrance area is rather narrow and can potentially lead to the emergence of a crowd in combination with the close-by facilities. On both sides of the main aisle the exhibitor's booths are allocated. The booths are segregated into groups of between three and thirteen and are accessible through side aisles. Beverages are served in four places – three coffee shops, distributed in Section 1 and 2 and the restaurant in Section 3.

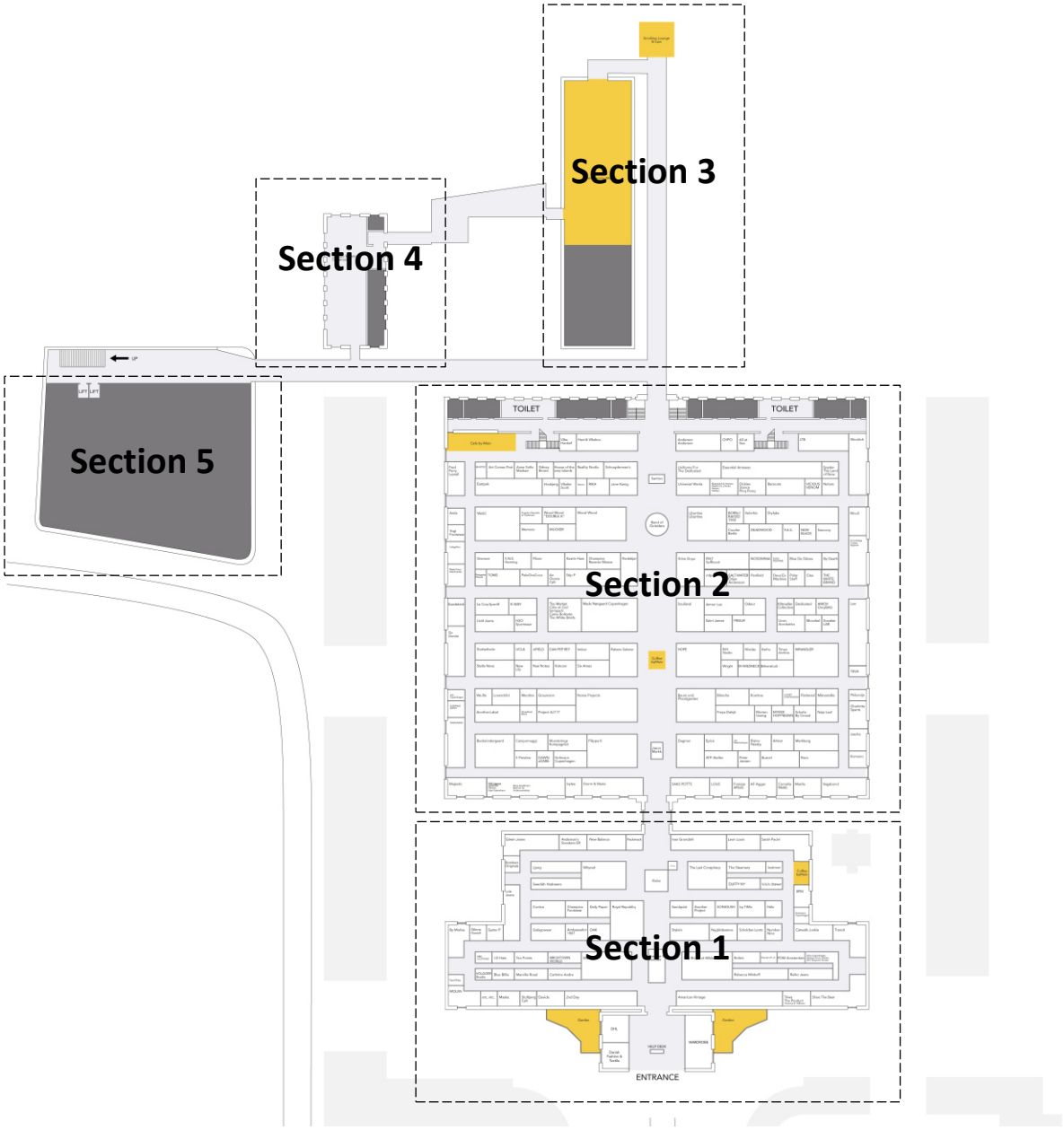



**Fig. 9: Map of exhibition site – main floor**

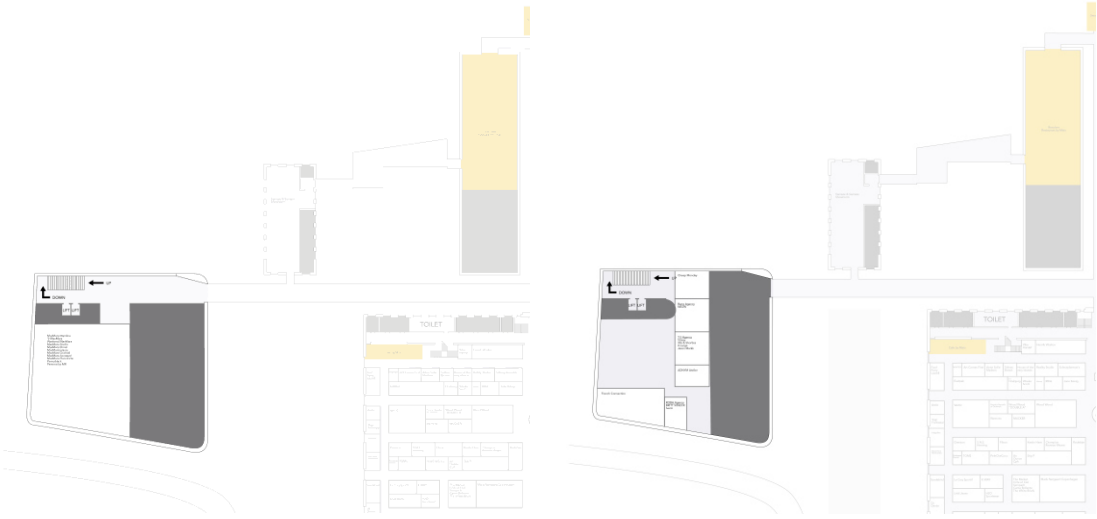


**Fig. 10: Map of exhibition site – first floor & second floor**

## 3.2 Research Questions

The main hypothesis that shall be answered in the scope of this thesis is if it is possible to extract valuable knowledge in an ArcGIS environment when a particular dataset is provided. A brief description of the used dataset was provided in Chapter 3.1 and is characterized comprehensively in Chapter 4.1. The proposed hypothesis opens up a broad range of possible ways for verification although a relevant geospatial context of research questions shall always be the main focus. In the following, several approaches are described how the hypothesis could be verified. If these approaches are also feasible with the provided dataset will be assessed in Chapter 4.

- Obstacle Detection

Detecting obstacles in an environment like the available exhibition site can provide valuable information. Detectable obstacles might be non-fixed objects (e.g. barriers in case of repair work) or fixed objects (e.g. walls). While the detection of non-fixed objects prove to be more valuable in real-time applications, detection of fixed objects can also be worthwhile in case of a not well-known environment. Regardless of the kind of application, respectively the actual purpose of obstacle detection, it is important that if an obstacle was detected, the type of obstacle should also be identified afterwards.

It shall also be considered that various obstacles will impact the spatial accuracy of positioning when using Bluetooth technology. These objects can attenuate the signals which are emitted by Bluetooth beacons and therefore will distort the calculated position. This leads to an decreased spatial accuracy.

- Crowd Detection

Detecting a crowd is a very similar problem definition as detecting an obstacle. In fact, a crowd can also be seen as an obstacle and therefore this task can be indicated as a subtask of obstacle detection, so that in a first step an obstacle is discovered and in a second step a classification of the type of obstacle will be made.

- User Classification

Valuable information of the behaviour of a user can be provided by analysing the features of their trajectory, e.g. dwell-time, pauses, velocity or direction of movement. Users can be classified with this information, for example into visitors or personnel of the exhibition. Another example for a classification might be to distinguish visitors who systematically visit the booths they are interested in from the start and visitors who prefer to “wander around”. Furthermore, a user’s trajectory could also be divided into parts of systematically visiting booths and randomly visiting booths.

Another way to classify users is to examine the booths they have visited. This information is especially beneficial because it provides an opportunity to describe the actual interest of a visitor. When a database that provides information about the exhibitors is linked to the visited booths, additional prospects are possible. Combining this method with the trajectory analysis might allow a relatively accurate user description.

- Geovisualization

Geovisualization communicates geospatial information in ways that, when combined with human comprehension, perception of complex information is eased. This allows easier ways to explore data for decision making processes. Geovisualization techniques could be applied in a number of different ways on this occasion, for example calculation of point density or interpolation of velocity fields with subsequent visualization. Geovisualization methods might be especially interesting when the temporal dimension is also considered for the processing. Investigating how point densities and velocity fields change over time can show valuable insights in the characteristics of the exhibition site. With the information of how point densities are changing over time it is also possible to determine certain trends.

- Temporal Distribution

Analysing the temporal distribution of records might bring valuable insight and could be used in many ways. This temporal distribution can either refer to a specific booth of an exhibitor and also to the whole exhibition site. Investigating the change of visitors over time might help to detect some incidents or let us evaluate the usefulness of marketing activities like live-shows.

# 4 Methodology

The main idea is to create a workflow to automatically process similar datasets. The proposed workflow is illustrated in Fig. 11 and consists of an implementation phase and an application phase. In the implementation phase all necessary tools for pre-processing and data analysis are developed and tested. The developed tools are scripted with *Python* and are mainly using functions of the *arcpy* package provided by *ESRI ArcGIS*. Additionally a series of tests will be executed in order to compare differences when processing data on a desktop basis or in the cloud and also when processing Datasets with a small or medium size and big datasets. The results of data analysis and comparison of processing concepts will eventually be evaluated. In the application phase the evaluated tools can be used automatically or semi-automatically if the datasets show similar characteristics as the ones used in the implementation phase. With pre-processing tasks datasets can be aligned to some degree to fit to the desired shape for further analysis.

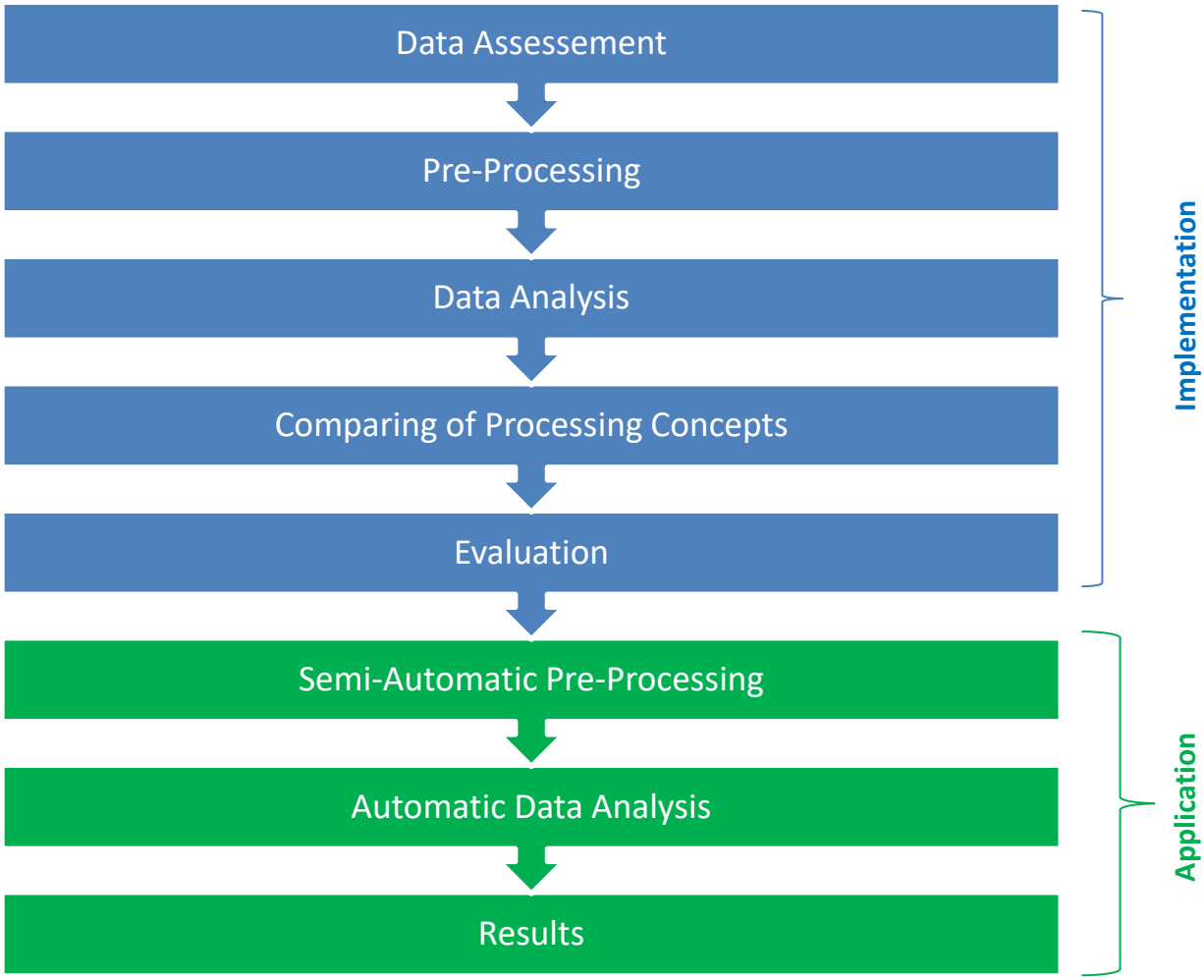



**Fig. 11: Workflow**

The first step in this workflow is data assessment. It is necessary to get a clear image on how your data looks like in the first place, in order to know what exact tasks needs to be done in all subsequent steps.

Data pre-processing is an important step in every data mining task. The dataset that is used for this analysis is provided by a single company and generally shows a homogenous character, which reduces the necessity and the effort of pre-processing operations. The most important pre-processing step in this case is a coordinate transformation. If data is provided in different file formats homogenization is a necessity for further processing.

The main processing task lies within the data analysis step. Spatial data mining approaches are used to extract knowledge from the data. A big challenge is posed by assessing which analysis tasks are in the first place possible and in the second place reasonable to yield useful results. This assessment can be done by taking a closer look into the data, even though there cannot be an absolute certainty that the results are meaningful, which makes it necessary to evaluate the outcome. All tasks that are included in this step should not be considered as a final and immutable process but rather as an initial version that can be extended and improved in future.

After developing tools for analysing data, a comparison of processing concepts shall be conducted. There is a general understanding that for Big Datasets cloud computing is an appropriate solution, but there is no precise definition when a dataset actually becomes *big*. By investigating the influence of the size of a dataset to processing runtimes in a desktop and a cloud environment, the issue on when to use which processing environment might be clarified.

Evaluating the results of all previous steps will show if it is necessary to make any improvements. This could include changes in the pre-processing, by transform coordinates into another coordinate system or present values in another shape. For data analysis and comparison of processing concepts the main point is to verify if the results are reasonable and significant.

## 4.1 Data Assessment

The dataset is provided as a CSV-file (comma separated value). Each line of the file is a data record which consists of one or more fields, separated by commas. Although the CSV file format is not standardized it is widely used by consumer, business and scientific applications as a data exchange format. CSV formats are best used to represent sets or sequences of records in which each record has an identical list of fields. This corresponds to a single relation in a relational database, or to data in a typical spreadsheet.

In the case of this dataset one record consists of the following fields:

- *timestamp* in format [YYYY-MM-DD hh:mm:ss.ms]. This value gives temporal information about when the line was recorded.
- *accuracy* [mm]. An estimation of the expected spatial accuracy of the record. This value is either 3000 or 5000 which corresponds to a spatial accuracy of 3 m or 5 m.
- *ipv4* (Internet Protocol version 4) addresses are expressed as a integer number.
- *x* and y coordinates are provided in an arbitrary coordinate system with unknown scale. For further analysis a transformation to a well-known coordinate system shall be performed.
- The *mobile_id* is used to identify each user of the smartphone application by an integer with the length of 9 or 10 characters. The identification uses and arbitrary number system to anonymise users for privacy reasons.
- *application_id* represents the version of the smartphone application by an integer with the length of 9 characters.
- b*uilding_id* identifies the used base map. As the map is changing with each event another *building_id* is assigned as well. The values are represented by an integer with the length of 9 or 10 characters.
- The field *floor_id* acts similar to the field *building_id* with the difference that each floor of a building is represented by an integer. The values are represented by an integer with the length of 9 or 10 characters even though some records are missing values.
- *level* is used to provide information about the level on which the record was recorded on. The values are either -1, 0 or 1.

In the following table (Table 4) an example of records is shown:

| timestamp [YYYY-MM-DD hh:mm:ss.ms] | accuracy [mm] | ipv4 | x | y |
|---|---|---|---|---|
| **2016-08-10 08:11:13.8108** | 5000 | 169977955 | 47423 | 12373 |
| **2016-08-10 08:11:15.394** | 5000 | 169977955 | 47650 | 12373 |
| **2016-08-10 08:11:19.459** | 5000 | 169977955 | 47657 | 12373 |
| **2016-08-10 08:11:24.478** | 5000 | 169977955 | 47681 | 12373 |
| **2017-08-10 08:29:53.875** | 5000 | 1408106486 | 125834 | 103596 |
| **2017-08-10 08:01:02.278** | 3000 | 1408106486 | 124452 | 145140 |
| **2016-02-01 12:17:43.116** | 3000 | 170735495 | 197114 | 101890 |

| mobile_id | application_id | building_id | floor_id | level |
|---|---|---|---|---|
| **595344168** | 605003540 | 824226007 | 824226876 | 0 |
| **595344168** | 605003540 | 824226007 | 824226876 | 0 |
| **595344168** | 605003540 | 824226007 | 824226876 | 0 |
| **595344168** | 605003540 | 824226007 | 824226876 | 0 |
| **1021446377** | 605003540 | 1021791096 | | 0 |
| **1021965522** | 605003540 | 1021791096 | | 0 |
| **596151942** | 605003540 | 618362149 | 618362628 | 0 |

**Table 4: Example of records**

In order to get a better overview of the available data the most important fields for further analysis were investigated. First of all, an assessment of the field *timestamp* was made. More precisely, more information on which dates data is available and therefore on which dates each event had happened was conducted (Table 5). This also includes how many records are available for each day and for each event. Furthermore, hourly timelines of the records of each event were created (Fig. 12). With these two assessments a determination of the temporal composition was concluded.

| Name | Dates | Number of Records (Day) | Number of Records (Event) | Number of Records (Total) |
|---|---|---|---|---|
| **Event 0** | 01-02-2016 | 14460 | 135251 | 688778 |
| | 02-02-2016 | 13194 | | |
| | 03-02-2016 | 58357 | | |
| | 04-02-2016 | 37628 | | |
| | 05-02-2016 | 11612 | | |
| **Event 1** | 10-08-2016 | 98897 | 232464 | |
| | 11-08-2016 | 94901 | | |
| | 12-08-2016 | 39476 | | |
| **Event 2** | 31-01-2017 | 54 | 179735 | |
| | 01-02-2017 | 51103 | | |
| | 02-02-2017 | 79952 | | |
| | 03-02-2017 | 48626 | | |
| **Event 3** | 09-08-2017 | 51118 | 137328 | |
| | 10-08-2017 | 53160 | | |
| | 11-08-2017 | 33050 | | |

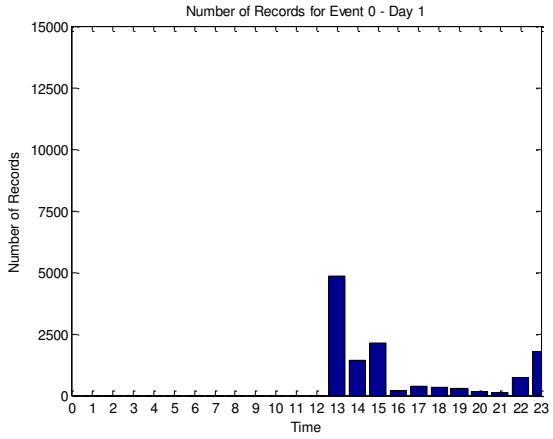
**Table 5: Number of Records on daily basis and per event**

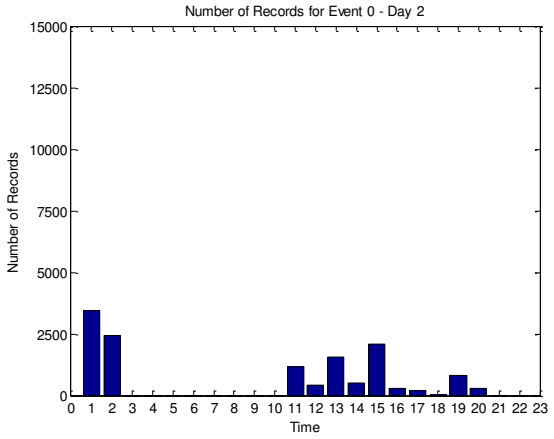
As shown in Table 5, the number of total records is close to 700000, which means there are close to 700000 points with coordinates, which can be represented as a point cloud. It is assumed that Event 0 lasted for five days (01-02-2016 till 05-02-2016) while Event 1 (10-08-2016 till 12-08-2016), Event 2 (01-02-2017 till 03-02-2017) and Event 3 (09-08-2017 till 11-

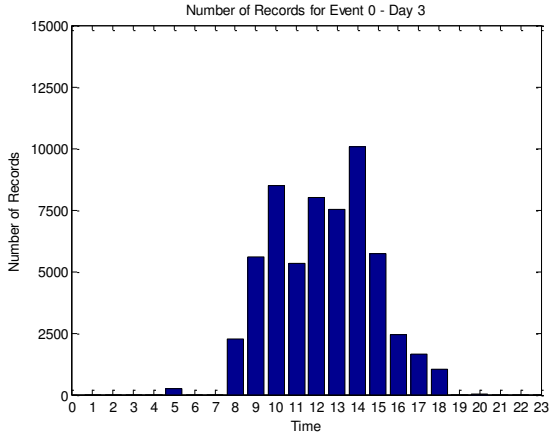
08-2017) lasted for three days. Even though Event 2 shows data for four days, the first day of the event is ignored due to the low number of records. The data recorded on this day (31-01-2017) was presumably collected for test purposes.

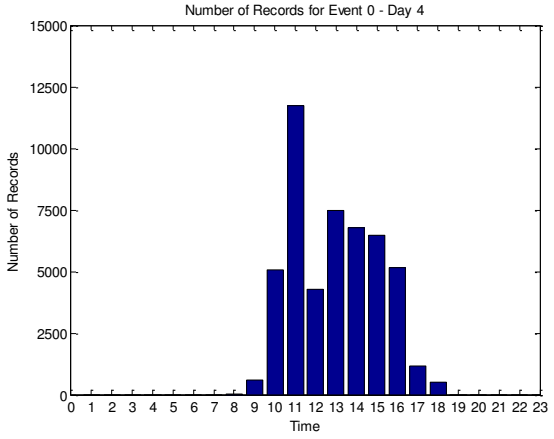
When investigating the number of records per day a relatively wide range of numbers can be observed. For example, at Event 1 the number of records on the third day (03-02-2016) is three times higher than on the first, second and fourth day (01-02-2016, 02-02-2016, 04-02-2016). A second example is that when daily number of records are compared to other days of one particular event, the last day always shows the smallest number of records. These two abnormalities might lead to some wrong assumptions. We have to keep in mind that the number of records is not necessarily proportional to the number of visitors of the event, as the data is recorded by voluntarily use of the indoor navigation smartphone application. When considering this fact we need to be extremely cautious so we are not following a wrong reasoning. The second example showed us that on the last day of each event the lowest number of records is available. This could lead us to the conclusion that on the last day there are fewer visitors than on the previous days. Another explanation for this fact might be that those visitors already have a better image of the exhibition hall and therefore require the use of indoor navigation in a lesser amount. Obviously we cannot explain such issues with full certainty, as there is a lack of additional information like number of visitors per day or user/visitor - ratio. The number of records per event also shows variations, with a minimum at Event 0 (135251) and a maximum at Event 1 (232464). As already mentioned, we cannot be certain that this numbers also express a higher number of visitors.

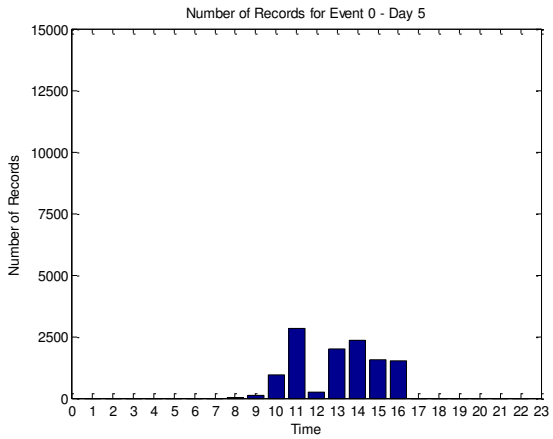
Fig. 12 shows the aggregated number of records within one hour for each event. This illustration can give an idea of when the exhibition hall is opening and closing on a particular day. We have to consider that not necessarily every user of the smartphone application is a visitor of the exhibition but might as well be personnel of the exhibiting companies or employees of the trade show. These figures generally show the expected characteristics. A maximum can be observed in the morning, between 9 a.m. and 12 a.m., then leads to a decrease around lunchtime, between 12 a.m. and 2 p.m., and another maximum in the afternoon, between 2 p.m. and 4 p.m. After the second maximum the number of records is decreasing until around 6 p.m. where the exhibition hall is presumptively closing. Some of the records are scattered before 8 a.m. and after 6 p.m. These records are not assumed to be conducted by visitors of the exhibition. Event 0 – Day 1 and Event 0 - Day 2 are showing a different behaviour as other days. On those two days the event seem to start later (Event 0 – Day 1 at 1 p.m.; Event 0 – Day 2 at 11 a.m.) and also include less records than most other days. Another abnormality can be observed between Event 0 - Day 1 - 10 p.m. and Event 0 - Day 2 – 2 a.m. There is a relatively high amount of records in this period which could be caused by a late night introductory meeting.

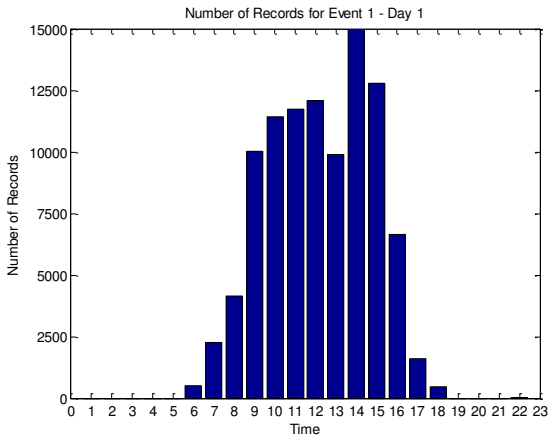

Number of Records for Event 0 - Day 1
Number of Records
Time

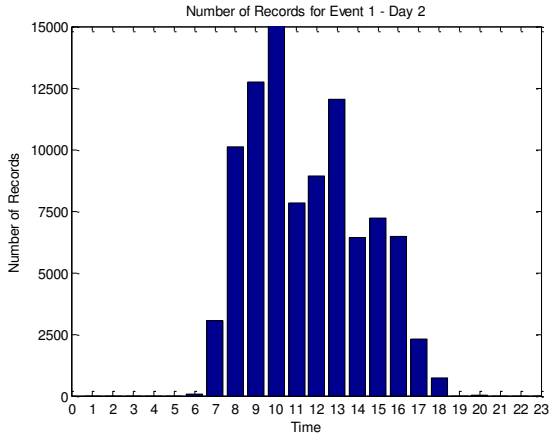

Number of Records for Event 0 - Day 2
Number of Records
Time

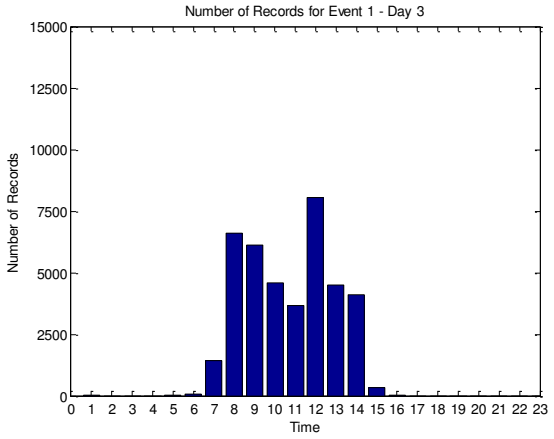

Number of Records for Event 0 - Day 3
Number of Records
Time


Number of Records for Event 0 - Day 4
Number of Records
Time


Number of Records for Event 0 - Day 5
Number of Records
Time


Number of Records for Event 1 - Day 1
Number of Records
Time


Number of Records for Event 1 - Day 2
Number of Records
Time


Number of Records for Event 1 - Day 3
Number of Records
Time

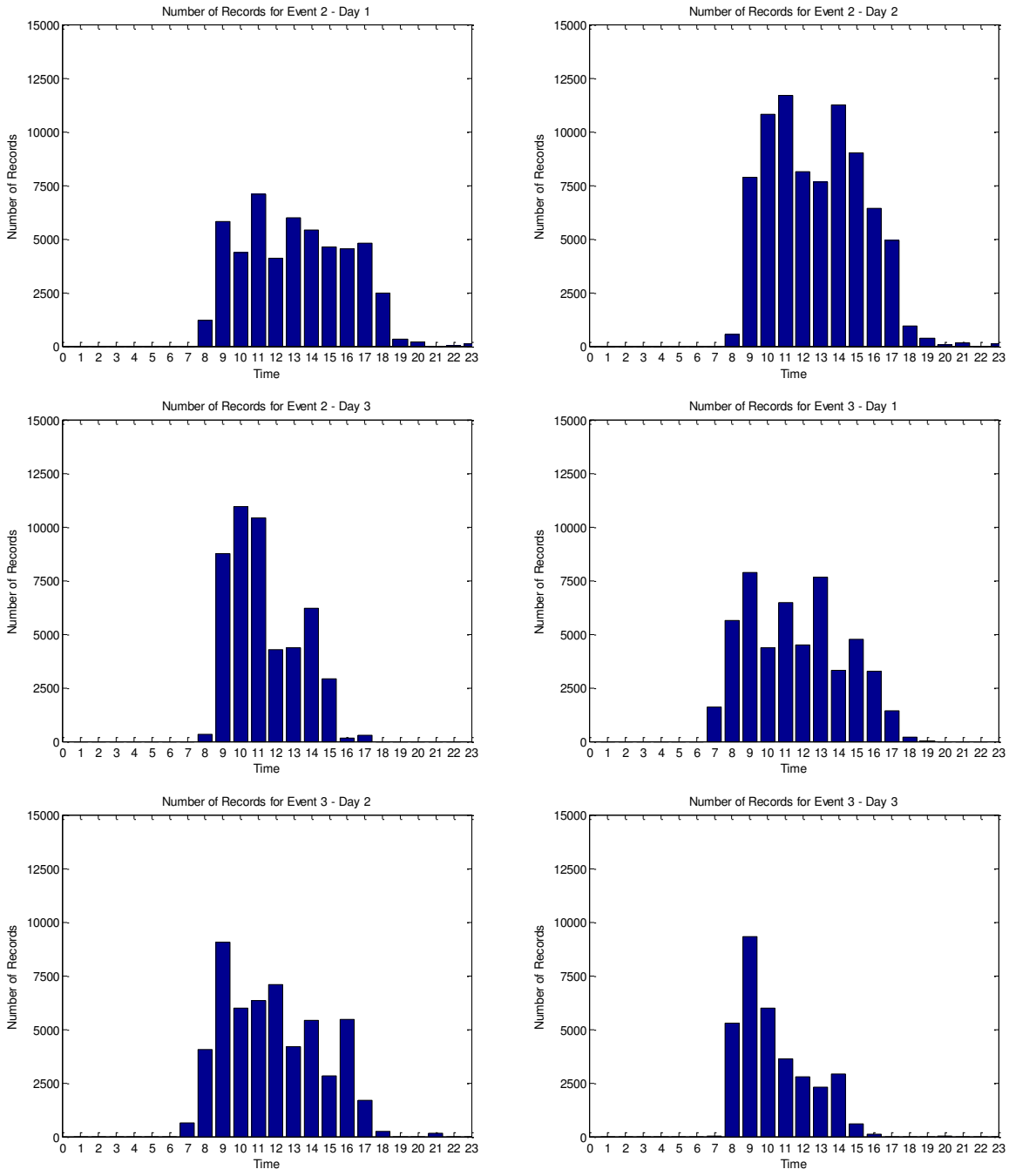



**Fig. 12: Illustrates the aggregated number of records within one hour for each event**

The temporal accuracy of smartphones is expected to be more than sufficient, as it is estimated to be clearly below one second, which is the recording frequency. Contrary to positioning with GNSS technologies, indoor navigation with Bluetooth technologies does not require such precise time measurements, as it is based on signal fingerprinting.

The accuracy field of the CSV-file presents an estimation of the expected accuracy of each record. This field contains either the value 5000, which is equal to an accuracy estimation of 5 meters, or 3000, which equals to 3 meters. Table 6 illustrates the percentage of records of accuracy values for each event. The ratio ranges between 3:1 and 2:3.

| Name | Number of Points | Accuracy 5000 [%] | Accuracy 3000 [%] |
|---|---|---|---|
| **Event 0** | 135251 | 65,6 | 34,4 |
| **Event 1** | 232464 | 45,7 | 54,3 |
| **Event 2** | 179735 | 39,8 | 60,2 |
| **Event 3** | 137328 | 44,5 | 54,5 |

**Table 6: Composition of Estimated Accuracy of Records**

As the x and y coordinates are in an unknown coordinate system a coordinate transformation is necessary. Fig. 13 illustrates the point cloud of the raw data. The point cloud partly shows an expected, distinctive shape similar to the boundaries of the exhibition hall. When considering the whole shape of the point cloud it does not fit well into the boundaries of the exhibition hall. It seems that the coordinate system of each event differs and that therefore a different coordinate transformation of the records for each event is required.

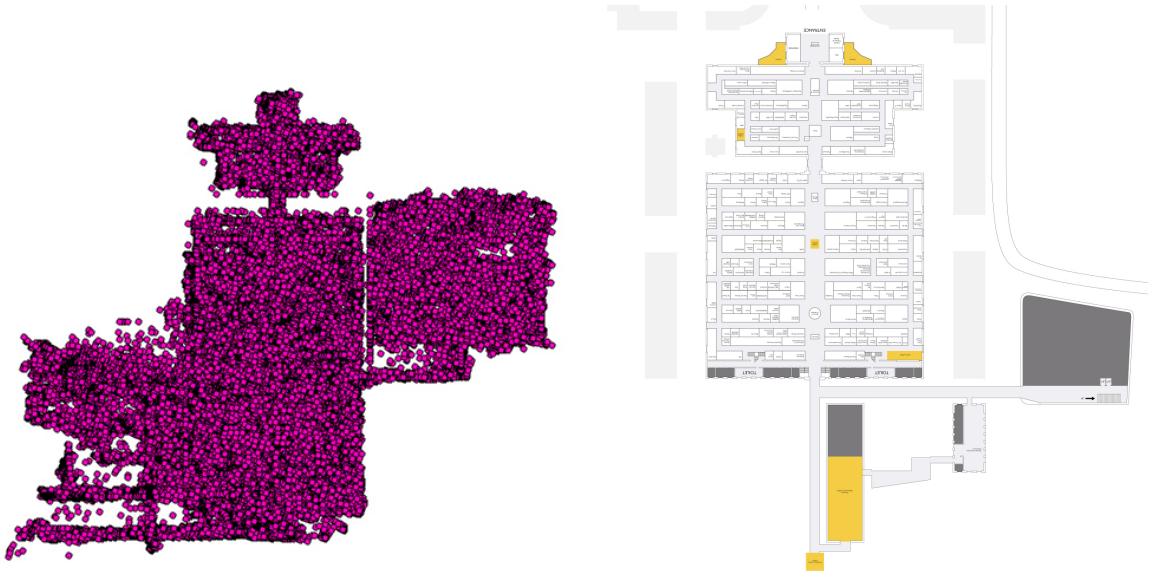


**Fig. 13: Illustration of Point Cloud and exhibition hall**

The whole exhibition hall consists of three different levels. The percentage of records of total number of points on every level is shown in Table 7. There are only a very small number of records on Level 1 and Level 2 which consists of less than one percent of the dataset, as almost the entire data has been recorded on Level 0. Due to the lack of data on Level 1 and Level 2, these records cannot be used for a meaningful analysis and therefore will be ignored.

| Name | Number of Points | Level 0 [%] | Level 1 [%] | Level 2 [%] |
|---|---|---|---|---|
| **Event 0** | 135251 | 100 | 0,00 | 0,00 |
| **Event 1** | 232464 | 99,91 | 0,02 | 0,07 |
| **Event 2** | 179735 | 99,26 | 0,24 | 0,50 |
| **Event 3** | 137328 | 98,81 | 0,22 | 0,98 |

**Table 7: Percentage of Records on each Level**

The number of users and the average number of records per user for each event is shown in Table 8. The number of users is acquired by counting the different values of the field *mobile_id* for every event. The number of users of the smartphone application ranges from around 400 to around 1800 and the average number of records per user from around 100 to 300.

| Name | Number of Points | Number of Users | Average Number of Records per User |
|---|---|---|---|
| **Event 0** | 135251 | 442 | 306 |
| **Event 1** | 232464 | 1817 | 128 |
| **Event 2** | 179735 | 1751 | 102 |
| **Event 3** | 137328 | 1203 | 114 |

**Table 8: Number of Users & Average Number of Records per User**

## 4.2 Pre-processing

To facilitate analysis of movement data, we perform initial pre-processing in the database, which enriches the data with additional fields, change values of fields for easier processing and transform coordinates. In an initial step the data is filtered by their date to aggregate the records to the corresponding event. There are two reasons why this step is necessary. First, a joint data analysis of different events will appear to be not meaningful in most cases. Generally an exhibition hall will be subject to several changes. The exhibitors and their locations might change, and also the whole exhibition hall might change as well. Processing data which was obtained under different, or in worst case unknown, circumstances together always poses a difficult challenge. The second reason is that as mentioned in the previous chapter, the coordinate system of the point clouds differs from each event to each other. These different coordinate systems make it necessary that the records are filtered to their corresponding event and a separate coordinate transformation is performed to match all records to a well-known mutual coordinate system. In case of the provided dataset, filtering the records by their event can be easily done by a simple script and therefore is an obvious first step. Filtering the data, which formally belongs to pre-processing, was already done before the data assessment in chapter 4.1 due to the fact that it provides a better understanding of the data, if the data is filtered for their belonging to each event.

A coordinate transformation was performed in two steps. First, a 4-Parameter-Helmert transformation from an arbitrary coordinate system into a local, Cartesian coordinate system was performed and second, a 7-Parameter-Helmert transformation into global geographic coordinate system – WGS84.

$$\begin{bmatrix} X \\ Y \end{bmatrix}^B = \begin{bmatrix} c_x \\ c_y \end{bmatrix} + m * \begin{bmatrix} 1 & -r_z \\ r_z & 1 \end{bmatrix} * \begin{bmatrix} X \\ Y \end{bmatrix}^A$$

**Equation 1: 4-Parameter-Helmert transformation**

$$\begin{bmatrix} X \\ Y \end{bmatrix}^B = \begin{bmatrix} X \\ -Y \end{bmatrix}^B$$

**Equation 2: Horizontal tilt**

The first transformation was performed with Equation 1 and Equation 2 and transformation parameter from Table 9. The necessary transformation parameters were acquired in an iterative process by adjusting the extent of the point cloud to the extent of the exhibition hall, as the corner points of the exhibition hall were known in a local, Cartesian coordinate system. Slightly different parameters had to be used for the transformation of point clouds for each event in order to align them into the same coordinate system. The illustration in Fig. 14 shows that the different point clouds harmonize well after the performed transformation. Within this local, Cartesian coordinate system most data analysis approaches would be possible. Nevertheless another transformation was performed with the aim to present the dataset in a global coordinate system. With data in this global system it is easily possible to visualize it in combination with other available data, for example satellite images or global maps. This provides better ways to represent results and gives more possibilities for additional use of results and also provides opportunities for using the dataset for other purposes.

| Name | $c_x$ | $c_y$ | m | $r_z$ [rad] |
|---|---|---|---|---|
| **Event 0** | -89.5 | 223 | 0.00105 | 0 |
| **Event 1** | -4 | 151 | 0.00105 | 0 |
| **Event 2** | -15.5 | 174 | 0.0012 | 0 |
| **Event 3** | -5 | 217 | 0.0015 | 0 |

**Table 9: Parameters for 4-Parameter-Helmert transformation**

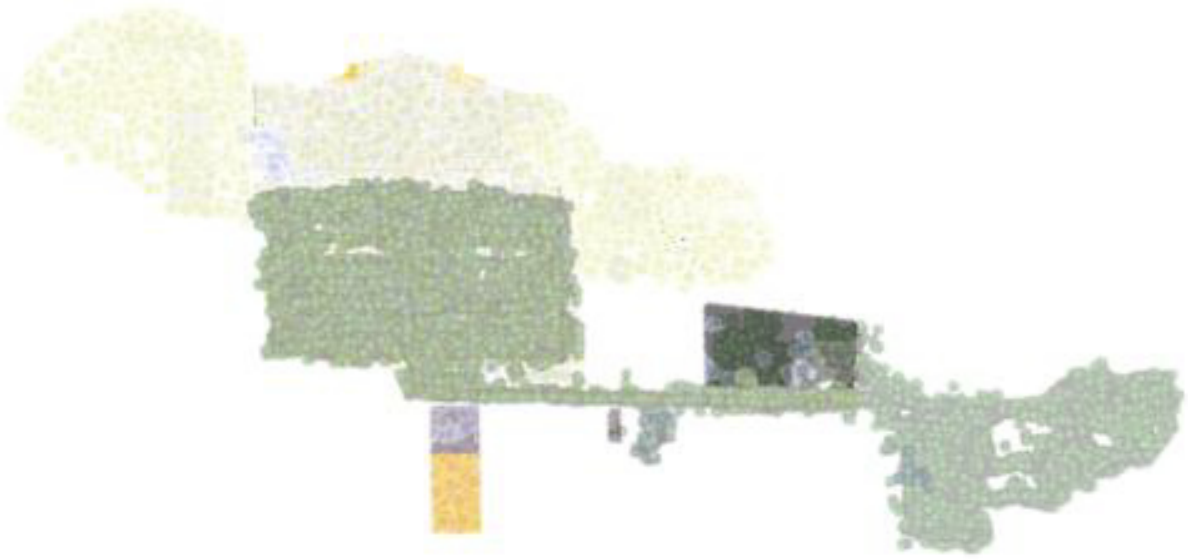


**Fig. 14: Point Cloud after first transformation**

The second transformation, a 7-Parameter-Helmert transformation, was performed with equation 3 and equation 4 and the parameters from Table 10. The transformation is called 7-Parameter-Helmert transformation, because 7 parameters are used: 3 parameters for rotation, 3 parameters for translations and 1 parameter to determine the scale. When the rotations are performed it is highly important to consider the correct sequence of rotations, as it is an intrinsic rotation, which means that rotations about the axes of the rotating coordinate system changes its orientation after each elemental rotation. If a different sequence of rotations is used other rotation parameters need to be used in order to get the same result. The results of the second transformation are coordinates in a geographic, geodetic coordinate system – WGS84.

Strictly speaking, such a coordinate transformation is not correct, as it is only valid to transform coordinates from a projected coordinate system into another projected coordinate system. The divergence of coordinates due to earth's curvature is not significant for this dataset because of the rather small extent of the point cloud and the given spatial accuracy of three to five meters. This is demonstrated by equation 5 and therefore earth's curvature will not be considered any further in this transformation.

$$R_x(rot_x) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(rot_x) & -\sin(rot_x) \\ 0 & \sin(rot_{x)} & \cos(rot_x) \end{bmatrix}$$

$$R_y(rot_y) = \begin{bmatrix} \cos(rot_y) & 0 & \sin(rot_y) \\ 0 & 1 & 0 \\ -\sin(rot_y) & 0 & \cos(rot_y) \end{bmatrix}$$

$$R_z(rot_z) = \begin{bmatrix} \cos(rot_z) & -\sin(rot_z) & 0 \\ \sin(rot_z) & \cos(rot_z) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$R = R_z R_y R_x$$

**Equation 3: Basic Rotation Matrix about x-, y-, and z-axis**

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}^C = \begin{bmatrix} c_x \\ c_y \\ c_z \end{bmatrix} + m * R * \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}^B$$

**Equation 4: 7-Parameter-Helmert transformation**

$$\Delta = \frac{L^2}{2R} = \frac{(247 + 77)^2}{2 * 6371000} = 0{,}008\, m$$

**Equation 5: Divergence due to Earth's curvature**

With:

Δ … Divergence due to Earth's curvature

L … Length of the point cloud

R … Earth's radius

| $c_x$ | $c_y$ | $c_z$ | $rot_x$ [°] | $rot_y$ [°] | $rot_z$ [°] | m |
|---|---|---|---|---|---|---|
| 12.56338 | 55.67032 | 0 | 157 | -63 | 19 | $14.4 * 10^{-6}$ |

**Table 10: Parameters for 7-Parameter-Helmert transformation**

| Extent | x [m] | y [m] |
|---|---|---|
| **Top** | 0 | 203 |
| **Bottom** | 0 | -15 |
| **Left** | -77 | 0 |
| **Right** | 247 | 0 |

**Table 11: Extent of Point Clouds after first transformation**

After performing the second transformation the coordinates are expressed in a geographical coordinate system by latitude and longitude and can therefore be combined with commonly available basemaps, e.g. *OpenStreetMap*, as illustrated in Fig. 15.

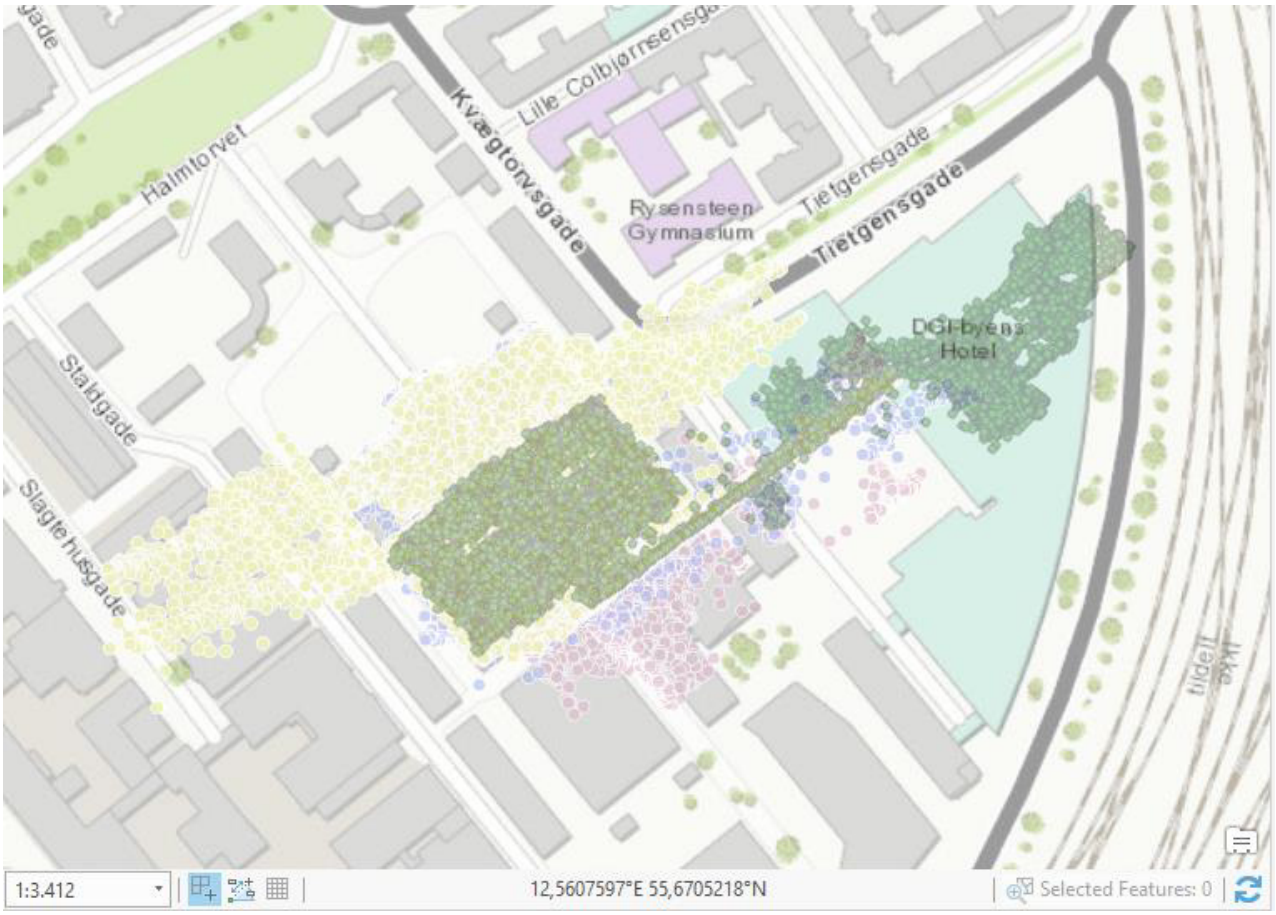



**Fig. 15: Point Clouds after 7-Parameter-Helmert transformation**

The performed coordinate transformations are a rather inconvenient and time intensive process. For future applications it is advisable to include pass points in order to simplify this process. Furthermore the inclusion of well-known pass points can improve spatial accuracy as there is likely to be a significant systematic error present if rough coordinate transformations are performed.

## 4.3 Analysis 1: Basic Data Insight

The first tool included in the data analysis section is called *Basic Data Insight.* The main idea is to provide basic information about the dataset in order to get a better overview. Necessary input parameters for this tool are the CSV-file that contains the records and a shape-file that consists of polygon features, which describes the locations of present exhibitors. It needs to be considered that the polygon features are based on the basemap of the corresponding event. Output parameters for this tool are:

- Filename
- Data type of input file
- Number of Records per Event

- Number of Users per Event
- Number of Records per Day
- Number of Users per Day
- Average Numbers of Records per Track
- Occurring Dates
- Extent of Point Cloud
- Spatial Reference
- Timetable of hourly Records per Day
- Timetable of hourly Users per Day

```
#Define workspace and filename
arcpy.env.workspace = "C:/Users/Julian/PycharmProjects/Testdaten/Kopenhagen.gdb"
filename = 'event' + str(number) + '_trans'

# Part 1: Information about event-file
cursor = arcpy.SearchCursor(filename)
number_of_points = int(arcpy.GetCount_management(filename).getOutput(0))  # Count number of points in event-file

for user in cursor:
    ID = user.getValue('mobile_id')
    date = str(user.getValue('date'))[0:10]
    time = user.getValue('time_time')
```

**Fig. 16: Implementation of *ArcPy* Package**

The included *arcpy* package functions *SearchCursor*, *UpdateCursor* and *InsertCursor* are mainly used to accomplish the required tasks implemented as a *Python* code. A cursor is a data access object that can be used either to iterate through a set of rows in order to search or update rows in a table or to insert new rows into a table. A code snippet of the *arcpy* package and *SearchCursor* is illustrated in Fig. 16.

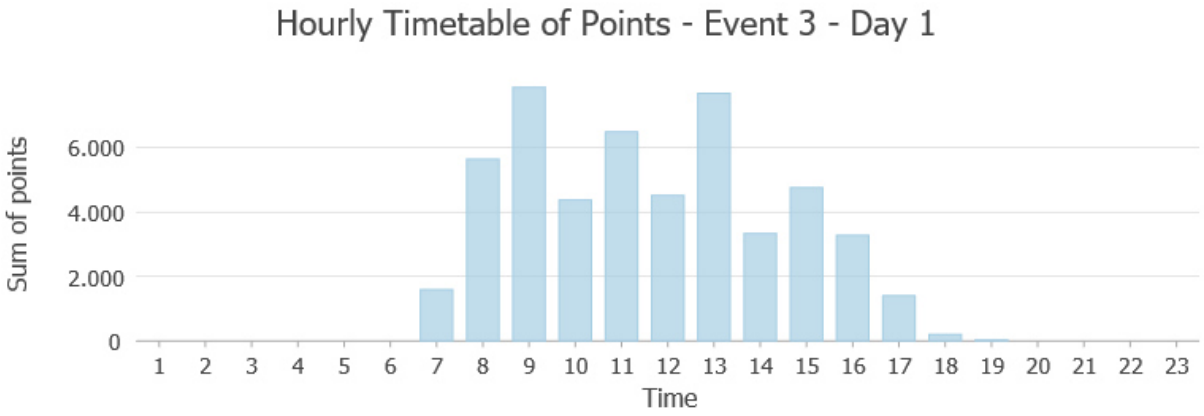
- The *SearchCursor* function provides a read-only cursor on a feature class or table. *SearchCursor* can be used to iterate through Row objects and extract field values. The search can optionally be limited by a where clause or by field and optionally sorted.
- The *UpdateCursor* function creates a cursor that allows you to update or delete rows on the specified feature class, shapefile or table. The cursor places a lock on the data that will remain until either the script completes or the update cursor object is deleted.
- The *InsertCursor* inserts rows into a feature class, shapefile or table.

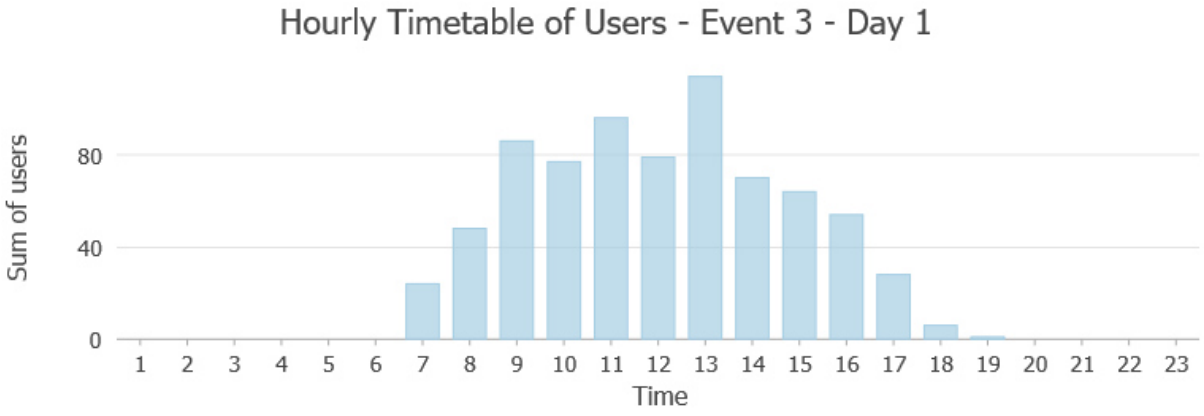
Results of this tool on basis of event 3 are shown in Fig. 17 and Fig. 18.

```
The name of the file is event3_trans
The number of points is: 137328
The number of users is: 1203
The average number of points per track is: 114.15461346633417
The dates of the events are: 2017-08-09 2017-08-10 2017-08-11
The number of points on date 2017-08-09 is 51118
The number of users on date 2017-08-09 is 747
The number of points on date 2017-08-10 is 53160
The number of users on date 2017-08-10 is 842
The number of points on date 2017-08-11 is 33050
The number of users on date 2017-08-11 is 373
The number of exhibitors is: 248
Extent:
  XMin: 12.561567242234503, XMax: 12.564339922249644, YMin: 55.66874694851293, YMax: 55.66980071742921
Spatial reference name: GCS_WGS_1984:
Datatype: FeatureLayer
On average one user visits 6.461928934010152 exhibitors
```

**Fig. 17: Text Output of Analysis 1 on Basis of Event 3**

Fig. 18 shows hourly timetables of points and of users. It would seem logical that with these two pieces of information the average duration of stay of visitors could be derived. In fact, this is not possible because of the used Bluetooth technology. Points are only recorded if the navigation system is actively used. That means the information that can be derived is the average duration of use of the navigation system of users, which is around 114 seconds in this case.



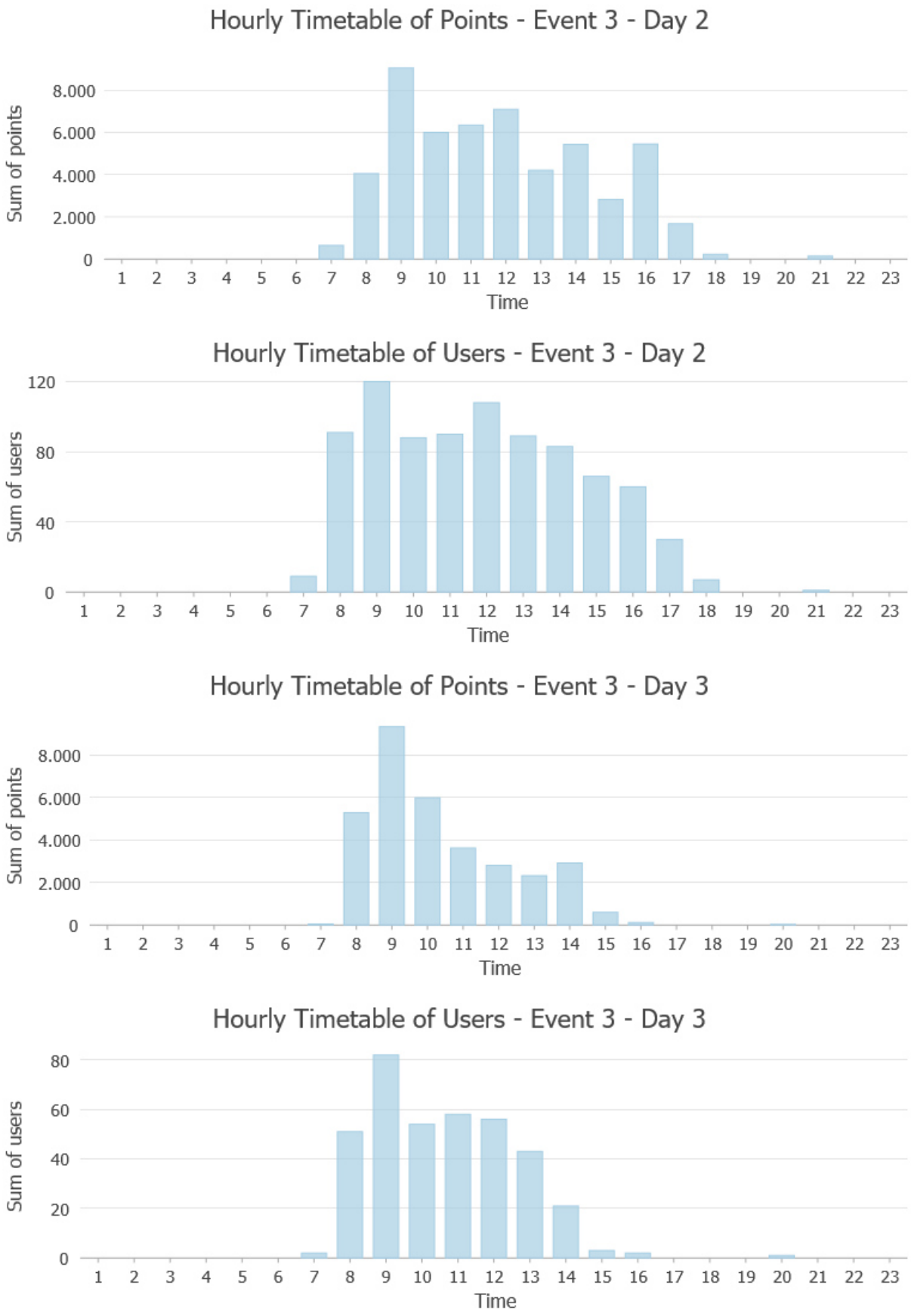



**Fig. 18: Graphical Output of Analysis 1 – Hourly Timetable of Points & Users**

## 4.4 Analysis 2: Estimation of User Interests and User Classification

The second analysis tool focuses on describing the interests of users. By assessing the exhibitors one particular user visited, an estimation of their individual interests can be made. A very simplistic approach is used which can be developed further in future. Furthermore, by applying subsequent association rule mining algorithms the results could be improved even more. Necessary input for this analysis is the record file, a shapefile which determines the location of each exhibitor by a polygon and a table which describes exhibitors by tags (Fig. 19). The used tags mainly focus on presented products of exhibitors which can be described more easily than other attributes, such as style or appearance. Characterizing exhibitors by those attributes might be more beneficial in order to classify users more accurate but is a rather difficult challenge. Defining style or appearance in an objective way will most likely prove to be an impossible task but might not be necessary anyway as the characterization of exhibitors could be performed by expected expert opinions. The table also contains fields to describe the origin of the exhibitor by country and estimated number of results of a search query on *Google.* Both parameters were not included in any analysis as the number of query results turned out to be unsuitable to describe the popularity of a brand appropriately and no meaningful analysis for origin of exhibitors could be found.

| OBJECTID | NAME | TAG | GOOGLE | COUNTRY |
|---|---|---|---|---|
| 1 | Transit | Avantgarde, Leather, Men, Women | 9390000 | Italy |
| 2 | Catwalk Junkie | Women | 1390000 | Netherlands |
| 3 | Karmamia Copenhagen | Women, Middle East | 58700 | Denmark |
| 4 | 8PM | Women | 30200000 | Italy |
| 5 | Coffee by Maio 1 | Coffee | | |
| 7 | Sarah Pacini | Women | 284000 | Belgium |
| 8 | Leon Louis | Casual, Men, Women | 2650000 | Denmark |
| 9 | Ivan Grundahl | Avantgarde, Women | 352000 | Denmark |
| 10 | Shoe The Bear | Leather, Shoes, Men, Women | 4110000 | Denmark |
| 11 | Triwa The Product Veronica B. Vallenes | Watches, Accessories, Sunglasses | 3220000 | Sweden |
| 12 | American Vintage | Men, Women, Basic, T-Shirt | 855000000 | France |
| 13 | Wardrobe | | | |
| 14 | Danish Fashion & Textile | | | |
| 15 | 2nd Day | Women, Leather, Denim | 13800000 | Denmark |
| 16 | Davida | Women, T-Shirt | | US |
| 17 | Stolbjerg Cph | Women, Accessories, Bag | 17300 | Denmark |
| 18 | Maska | Women | 7280000 | |
| 19 | etc. etc. | Women, Shirt, Dress, Skirt, Trouser | 438000000 | Denmark |
| 20 | MOLIIN | Women, Casual, Blouse | 98200 | Denmark |
| 21 | Fjord Ruby | Garments, Jewellry | 537000 | |
| 22 | VOLGGER Studio | Men, Women, Leather | 68000 | Austria |
| 23 | Blue Billie | Jewellry | 160000 | Sweden |
| 24 | Marville Road | Women, Dress, Skirt, Trouser, Coat | 198000 | UK |
| 25 | Catherine Andre | Women, Coat, Dress, Skirt, T-Shirt, Accessories | 30600000 | France |

**Fig. 19: Table of tags to characterize exhibitors**

The algorithm has the following structure:

1. Attach every record with an additional field that indicates if the coordinates of the record lies within the polygon of an exhibitor and if so, the name of the exhibitor. Also the number of records where this condition is true is indicated.
2. Create a new table where every line represents one user-ID and all exhibitors the user visited.

3. Connect the table that was created in the previous step with the table that indicates the exhibitor's tags so that one line represents all exhibitors a particular user visited and all tags that are used to describe those exhibitors. A few example results are shown in Table 12.

```
arcpy.SelectLayerByAttribute_management(Polygon_File, "NEW_SELECTION", '"OBJECTID" = %s' % exhibitor_id)
arcpy.SelectLayerByLocation_management(event_File, "Intersect", Polygon_File)
```

**Fig. 20: Code snippet of Analysis 2**

Important tools for realizing this analysis are the functions included in the data management toolbox *SelectLayerByAttribute_mangement* and *SelectLayerByLocation_management.* These functions make it possible to select features in a layer based on an attribute query or spatial relationship to features in another layer.

- The function *SelectLayerByAttribute_mangement* applies an attribute query on a feature layer or table. The query is a SQL expression used to select a subset of records and follows the common SQL syntax.
- The function *SelectLayerByLocation_management* evaluates each feature in the input feature layer against the features in the selecting features layer or feature class; if the specified relationship is met, the input feature is selected. A number of different relationship options are available to evaluate their relationship, for example: *intersect, within a distance, contains, within, completely within, are identical to*. The definition of relationships might differ depending on the type of geometry. [7]

The output of this tool should be rather seen as an intermediate step for further applications than as a final outcome. As the data acquisition is based on a smartphone application the estimation of one's interests can turn out to be of great value. The information that is extracted can be used in several ways, for example targeted advertisement via smartphone, prediction of single user actions, prediction of visits of exhibitors – just to name a few.

As already mentioned before, the used algorithm is rather simple. By implementing adaptions to the algorithm a performance improve might be achieved. The current algorithm only considers coordinates that are within one polygon, respectively one exhibitor. It is possible that customers will examine an exhibitor without being located inside this polygon. Furthermore, there are issues with uncertainty in the data. These issues include the spatial accuracy – with a spatial accuracy between three meters and five meters the certainty if coordinates are actually within a polygon can be doubted – but also uncertainty about the coordinates of the polygon itself. The introduction of a buffer zone around an exhibitor might be considered to overcome this issue. Another concern might be that at the current version, the condition that a particular user is assigned to have visited an exhibitor, is that

[7] http://pro.arcgis.com/en/pro-app/tool-reference/data-management/select-by-location-graphical-examples.htm last access 2018-03

one record lies within this polygon. It is arguable if this condition is sufficient to ensure that a user has visited, or has interest in the exhibiting company. This condition could be adapted to require a higher number of records inside a polygon or the implementation of a time requirement. These are only two possible adjustments of many to improve the algorithm. If these or other adaptions are implemented a comparison of results is necessary in order to ensure their usefulness.

| User ID | Visited exhibitors | Interests |
|---|---|---|
| 1021274176 | RIKA - Band of Outsiders - Jane Konig - Schnaydermans | Men, Trouser, Jacket, Shirt, T-Shirt, Jewellry, Women, Hat |
| 1021274189 | None | None |
| 1021274241 | Packmack - Storm & Marie - Royal Republiq - Rains - Samsoe & Samsoe - Styleln - Dagmar - Andersen Andersen - Stine Goya - Soulland - Libertine Libertine - CHPO - Kjaergaard & Stampe Atelier De LAmee Hemen Atalaye - All at Sea - LTB | Men, Women, Shoes, Bag, Accessories, Jacket, Belt, Dress, Skirt, Coat, Hoodies, Hat, Men, Kids, Jeans, Swearshirt, Trouser, Shirt, Blouse, Knitwear, Swimwear, Sweaters, T-Shirt, Top, Minimalistic, Watch, Sunglasses, Silk, Luxury |
| 1021446431 | Rodebjer - Henrik Vibskov - Coffee by Maio 2 | Coffee, Women, Jacket, Blazer, Knitwear, Dress, Skirt, Top, Blouse, Trouser, Denim, Shoes, Men, Accessories, Hat, Keychain, Scarf, Sunglass, Wallet, Bag, Backpack |
| 1021446893 | Silverblack Blank | Jewellry |

**Table 12: Example results of estimation of user's interests**

Based on the results of this analysis statistics can be created to gain a deeper understanding of the behaviour of users. Table 13 shows some examples of numbers and ratios of users and exhibitors that were assigned with a particular tag. That means by describing the example on the tag *Dress* that 41.3 % of the users were assigned to have interest and 16 % of exhibitors were offering dresses and for the tag *Sports,* 2.4 % of the users were assigned to have interest and 2.0 % of exhibitors were offering sportswear.

| Tag | User Tags Total | User Tags Ratio [%] | Exhibitor Tags Total | Exhibitor Tags Ratio [%] |
|---|---|---|---|---|
| **Coffee** | 70 | 5.9 | 2 | 1.6 |
| **Dress** | 497 | 41.3 | 41 | 16.0 |
| **Jewellery** | 84 | 7.0 | 8 | 3.1 |
| **Shoes** | 381 | 31.7 | 40 | 15.6 |
| **Sports** | 29 | 2.4 | 5 | 2.0 |
| **Socks** | 37 | 3.1 | 4 | 1.6 |
| **Luxury** | 46 | 3.8 | 5 | 2.0 |
| **Women** | 697 | 58.0 | 141 | 55.1 |

**Table 13: Total number and ratio of User Tags and Exhibitor Tags**

The research question that was posed in chapter 3.2 about user classification by characterising their trajectories was rejected. Due to the relatively low spatial accuracy and the fact that trajectories were only recorded for a short amount of time, such analysis was not able to be performed.

## 4.5 Analysis 3: Point Density Maps & Trend Detection

The third analysis is designed to describe the distribution of users by point density maps and detect positive and negative trends in the distribution. Density mapping as a GIS tool is a relatively simple way to show locations where points or lines may be concentrated in a given area. Often, such maps utilize interpolation methods to estimate, across a given surface, where concentration of a given feature might be. Kernel density measures are sometimes used to smooth point estimates to create a surface of density estimates in a given area. Trend analysis can help to identify trend when comparing different datasets or subsets of one dataset.

The most important tool that was used in this analysis is the *arcpy* function *CalculateDensity*. A density map based on point or line features is created by spreading known quantities of some phenomenon (represented as attributes of the points or lines) across the map. The result is a layer of areas classified from least dense to most dense (Fig. 21). A higher density value in a new location means that there are more points near that location.

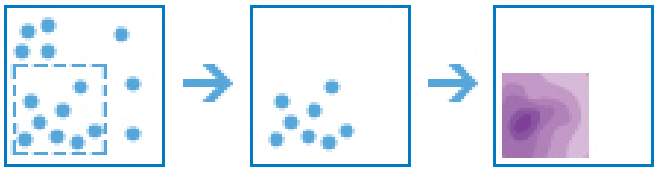


**Fig. 21: Principle of *CalculateDensity* function**[8]

Similar to *CalculateDensity* is the function *PointDensity*. The Point Density tool calculates the density of point features around each output raster cell. Conceptually, a neighbourhood is defined around each raster cell center, and the number of points that fall within the neighbourhood is totalled and divided by the area of the neighbourhood (Fig. 22). (Silverman 1986)

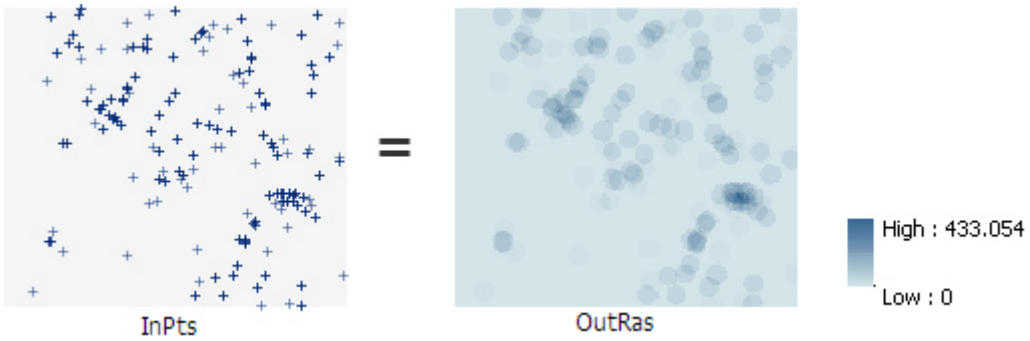



**Fig. 22. Principle of *PointDensity* function**[9]

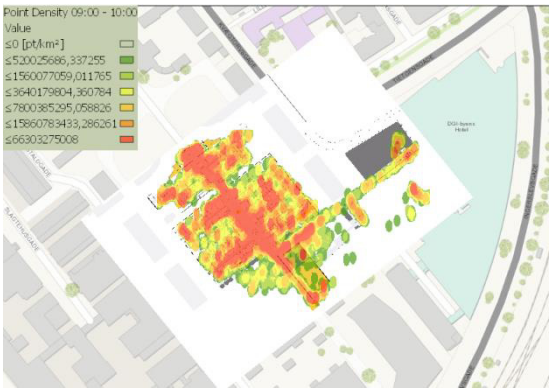
This tool was realized based on the functions *SelectLayerByAttribute_management* and *CalculateDensity*. The idea is to filter the records by their time and create density maps of the filtered records. The created output files can then be visualized as a sequence of maps or shown as an animation featured by *ArcGIS* or second party animation software. The created density maps also serve for the trend detection analysis. By subtracting chronological adjacent density maps a detection of areas which show a significant increase or decrease in density can be performed. If the difference in density is above a certain threshold this area is showing a positive or negative trend.

Fig. 23 illustrates the point density as a sequence of maps. In order to compare the classification of density values the same class breaks needs to be chosen for each map. A high density can be identified by red colour and a low density by green colour. Some areas could not be assigned with a point density value as the number of records was too low to calculate density. In the first and last map of this sequence low densities can be observed because the number of records at this time is relatively low because at this time the trade fair is about to start (first map) and close (last map). In the middle of the exhibition hall the

[8] https://pro.arcgis.com/en/pro-app/tool-reference/feature-analysis/calculate-density.htm last access 2018-03
[9] http://pro.arcgis.com/en/pro-app/tool-reference/spatial-analyst/point-density.htm last access 2018-03

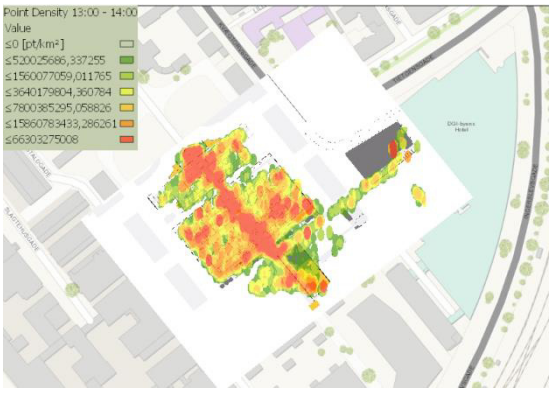
main aisle is located. Naturally along the main aisle a high density can be observed as most visitors will move along this way. The highest density is located at the main entrance in the northern part of the exhibition hall.











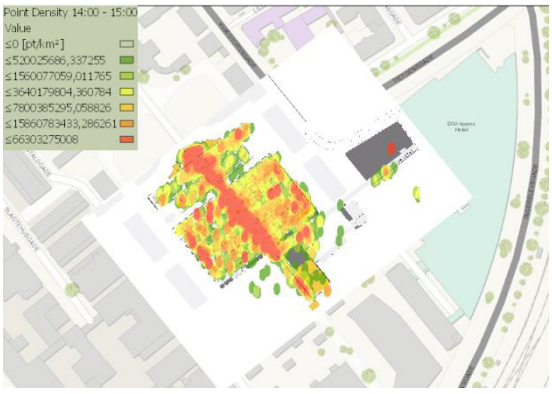




**Fig. 23: Hourly Density Maps of Event 3**

A comparison of different density maps can bring more information of differences on the distribution of points. The differences are extracted by subtraction of density distributions. Such comparisons can be made for example of hourly density maps, daily density maps or density maps of a whole event. In the following example a comparison of density distributions of whole events was performed, namely Event 2 and Event 3. First, density maps of both events needs to be created in order to extract information of the differences between them (Fig. 24). At first glance, both density distributions seem to be quite similar. The highest densities can be observed at the entrance and the main aisle and low to mediocre densities across the rest of the exhibition hall. There are also some low densities along the path located in the south-east and in the near surroundings outside of the building. Second, both density distributions are subtracted. The result of the subtraction is illustrated in Fig. 25. The result shows predominantly negative values. This can be explained by the fact that Event 3 consists of less records than Event 2 (Event 2: 179735; Event 3 137328). In the most southern part of the main aisle that leads to the restaurant area only positive values are present. These positive values can be explained due to the lack of records in this area in Event 2. Besides these facts, areas with positive and negative trends can be observed.


**Fig. 24: Point Density Maps of Event 2 and Event 3**


**Fig. 25: Subtraction of Densities of Event 3 and Event 2**

In order to determine areas that show a significant trend a certain threshold needs to be defined. All areas that show a value above the defined upper threshold respectively below the lower threshold are considered to show a significant trend. The thresholds are defined by a fraction of the maximum and minimum density values. Fig. 26 shows where a significant trend was detected. Red colour represents a negative trend while green colour represents a positive trend. The first illustration shows the real trend as it was obtained by the subtraction of densities and comparison with the defined thresholds. The second illustration relates the trend to the location of exhibitors in order to determine exhibitors that show a positive or negative trend. The result can be illustrated as a map and also as a table.


**Fig. 26: Exhibitors showing significant trend**

At the current stage the algorithm differentiates only between positive, negative and no significant trend due to the relatively small number of recorded points. This could be extended in future versions to detect different levels of trends.

## 4.6 Analysis 4: Analysis of Timetables

With this tool tables are created that contain information on how many users have visited each exhibitor at every hour. With this information predictions on the number of visitors and their temporal distribution specified for a particular exhibitor can be made. Furthermore, in the case of known activities, such as marketing events, statistical tests can be conducted to verify if a particular activity has led to an increase in the number of visitors. Dixon's Q-Test can be performed to detect outliers in order to check a significant increase of visitors. Fig. 27 illustrates the timetables of four exhibitors as an example to show the results. Due to the fact that only a small fraction of visitors of the exhibition were using the smartphone app, the number of visitors in the graph seems to be rather small. We have to consider that this number does not represent the total number of visitors that were visiting the fashion fair. Nevertheless, we expect that the data still represents the behaviour of all visitors.

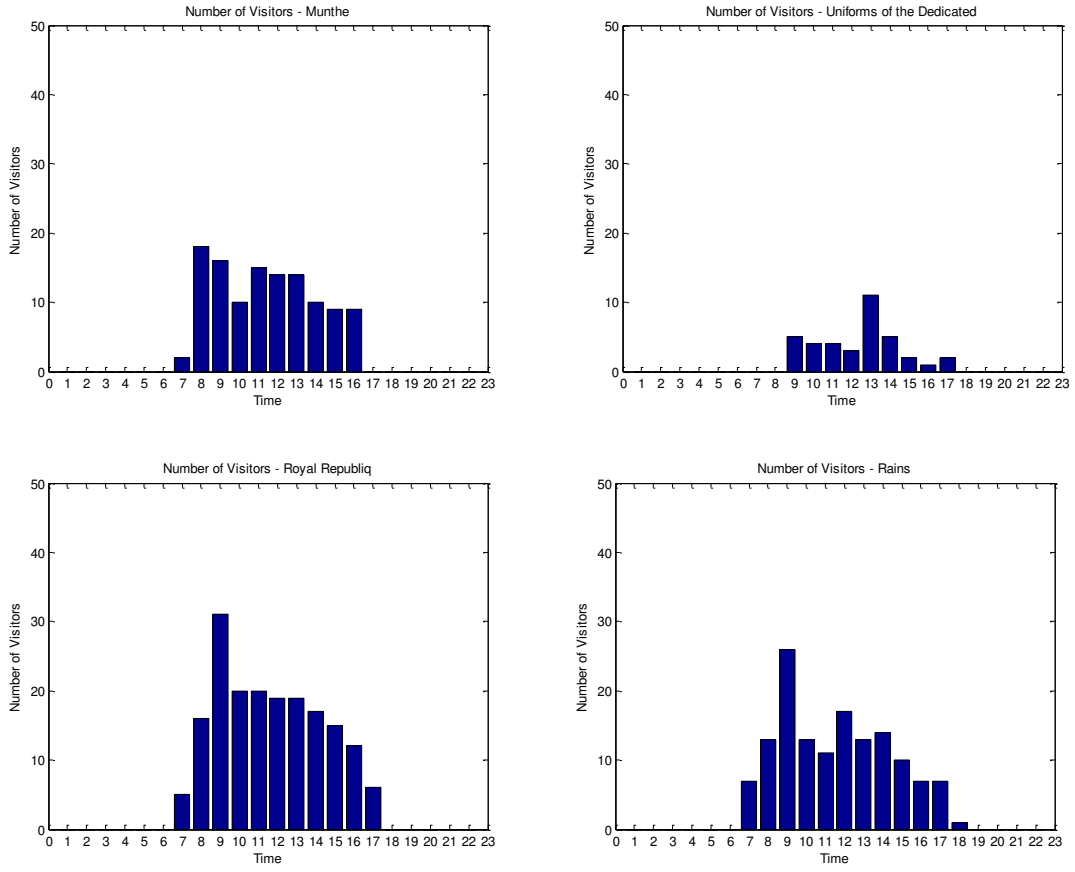



**Fig. 27: Number of Visitors of four example exhibitors**

This test is carried out by calculating the experimental value of the corresponding statistic parameter followed by its comparison with a critical value located at the appropriate position of the relevant statistical table (see Appendix, A1). Although the use of Q-test is increasingly discouraged in favour of other robust methods, it remains a fact that Q-test is the simplest test for the objective rejection or acceptance of a grossly deviant value within a small set of observations. (Efstathiou, 2006)

$$Q_1 = \frac{x_2 - x_1}{x_N - x_1}$$

$$Q_N = \frac{x_N - x_{N-1}}{x_N - x_1}$$

**Equation 6: Dixon's Q-Test for testing the smallest value ($x_1$) and the largest value ($x_N$) (Efstathiou, 2006)**

In order to prove the practical use of this method Dixon's Q-Test is applied on the timetable of one exhibitor assuming that a marketing activity was performed at 9 a.m. The exhibitor *Royal Republiq* was used because there is a significant increase in frequency which is reason to suspect some incident has happened. Dixon's Q-Test requires rearranging the data in an increasing order. Further on, all zeros are eliminated at the time the trade fair is closed because no data can be expected. The result of the Q-value is $Q_N$ = 0.423. With 10 observations and at 90% confidence, $Q_N$ = 0.423 > 0.412 = $Q_{Table}$, so we conclude that this value is an outlier. Following this conclusion, the exhibitor *Royal Republiq* is showing a significant increase of frequency at 9 a.m.

It might be problematic to use this method to automatically check all exhibitors. A significant increase or decrease in frequency does not certainly mean that some particular incident or activity happened at this moment. A high increase in frequency could be caused by a considerable number of reasons, for example a vast increase of visitors at a neighbouring booth could lead to a waiting crowd that also influences nearby exhibitors.

## 4.7 Comparison of Processing Methods

The storing, processing and visualization of big datasets is becoming more and more essential these days. In this chapter an experiment is made to investigate the possibilities and downsides of cloud computing techniques compared to conventional processing techniques. In particular, functions included in the *Python* package *arcpy* are used to represent conventional processing and the *Python* package *geoanalytics* is used to represent cloud processing. Both packages were developed by ESRI and provide a very similar toolset for analysing geospatial data. The *geoanalytics* tools are based on the capabilities of the *ArcGIS GeoAnalytics Server*, which makes it possible to analyse big data or accelerate

traditional ArcGIS Desktop analysis workflows through ArcGIS Pro and Portal for ArcGIS. The following functions are available in both packages and therefore suitable for a comparison[10]:

- **Aggregate Points**

The Aggregate Points tool uses area features to summarize a set of point features. The boundaries from the area feature are used to collect the points within each area and use them to calculate statistics. The resulting layer displays the count of points within each area using graduated symbols.

**Fig. 28: Workflow – Aggregate Points**

- **Calculate Density**

This function works as described in Chapter 4.5

- **Calculate Field**

The Calculate Field tool calculates field values on a new or existing field.

- **Other functions:**

Copy To Data Store, Create Buffers, Detect Incidents, Find Hot Spots, Find Similar Locations, Geocode Locations from Table, Join Features, Reconstruct Tracks, Summarize Attributes, Summarize Within

In order to use to your *GeoAnalytics* server a connection needs to be established first. This can be conducted following the code snippet in Fig. 29. An established connection enables the possibilities to manage, analyse and visualize data in the cloud.

```
from arcgis.gis import GIS

gis = GIS("http://machinename.domain.com/webadapter", "username", "password")
```

**Fig. 29: Establishing connection to your *Geoanalytics* server with *Python***

[10] http://enterprise.arcgis.com/en/server/latest/get-started/windows/perform-big-data-analysis.htm last access 04 - 2018

### 4.7.1 Outline

In order to investigate the characteristics of the mentioned processing methods a number of alterable variables are defined to examine changes in processing time. Five different files were used. These files are based on the original file of event 3 and were enlarged by copying all records by a particular number of times and modifying the coordinates by applying a uniformly distributed error of ± 5 meters in accordance with the spatial accuracy. For a comparison of processing times the function *PointDensity* of the *arcpy* package and the function *calculate_density* of the *arcgis.geoanalytics* package were used. It was also tested if the applied bin size of these functions is influencing the processing time. Four different bin sizes were used, 0.01 meter, 0.1 meter, 1 meter and 10 meters. The neighbourhood radius for this operation is the bin size multiplied by two. Only points that fall within this specified neighbourhood of a bin are considered when calculating the density.

The files that were used are stated in Table 14.

| Filename | event3 | event3_x5 | event3_ x10 | event3 _x20 | event3_x50 |
|---|---|---|---|---|---|
| Number of Records | 137,328 | 686,640 | 1,373,280 | 2,746,560 | 6,866,400 |
| Size[11] | 3,472 kB | 20,855 kB | 42,529 kB | 86,003 kB | 175,326 kB |

**Table 14: Files used for the experiment**

### 4.7.2 Computing Specifications

The CPU, central processing unit, is an electronic circuit which carries out the instructions of a computer program by performing the basic arithmetic, logical, control and input/output (I/O) operations specified by the instructions. CPUs are usually microprocessors which means that they are contained on a single integrated circuit (IC) chip. Most modern computers employ a multi-core processor, which is a single chip containing two or more CPUs called "cores". The clock speed of a CPU typically refers to the frequency at which a chip is running and is used as an indicator of the processor's speed. The frequency of the clock pulses determines the rate at which a CPU executes instructions and, consequently, the faster the clock, the more instructions the CPU will execute each second. In computing benchmarks are commonly used to assess performance characteristics of computer hardware or software. A CPU benchmark is a series of tests designed to measure the performance of a COU by comparing the performance of different systems, using the same methods and circumstances. RAM, Random-access memory, is a form of computer data storage that stores data and machine code currently being used.

[11] Size of file when exported as shapefile and extracted to zip

| | Desktop Computing | Cloud Computing |
|---|---|---|
| **Processor** | Intel(R) Core(TM) i7-2670QM CPU @ 2.20 GHz | Intel(R) Xeon (R) CPU E5-2686 v4 @ 2.30 GHz |
| **Average CPU Benchmark**[12] | 5908 | 19255 |
| **Number of Cores** | 4 | 18 |
| **Installed Memory (RAM)** | 8 GB | 32 GB |
| **System Type** | 64-Bit Operating System, x64-based processor | 64-Bit Operating System, x64-based processor |

**Table 15: Specifications for Desktop Computing and Cloud Computing**

### 4.7.3 Results

First processing times are compared with respect to their file size and the bin size of *PointDensity* function. This test was only performed with desktop processing. The results are showing an interesting behaviour of processing times, which is shown in Table 16. While for the original file the average processing times are showing a consistent behaviour, the other four files are showing a peak in processing time for 0.01 m bin size but consistent numbers for 0.1 m, 1 m and 10 m bin size. The cause for this rather unexpected behaviour could not be evaluated. The results of point density maps with different bin sizes are shown in Fig. 30. The choice of an appropriate bin size is an important decision to visualize data in a desired way. While for the smallest and the largest bin size not a great deal of knowledge can be extracted, both bin sizes 0.1 m and 1 m shows a pleasant result where areas of interest can be detected at first glance.

| Average processing time [s] | 0.01 m | 0.1 m | 1 m | 10 m |
|---|---|---|---|---|
| **event3_x1** | 8.6 | 8.9 | 8.3 | 8.6 |
| **event3_x5** | 82.5 | 22.2 | 22.3 | 23.2 |
| **event3_x10** | 98.6 | 38.3 | 40.2 | 41.8 |
| **event3_x20** | 137.7 | 73.1 | 76.5 | 80.0 |
| **event3_x50** | 146.1 | 101.8 | 106.6 | 107.0 |

**Table 16: Average processing times for different bin sizes and number of records**

The expected behaviour that processing times are increased with increasing records is confirmed. The assumption that processing times are linearly increasing with a smaller bin size could not be confirmed, although a correlation can be observed.

[12] According to https://www.cpubenchmark.net/ last access 04-2018

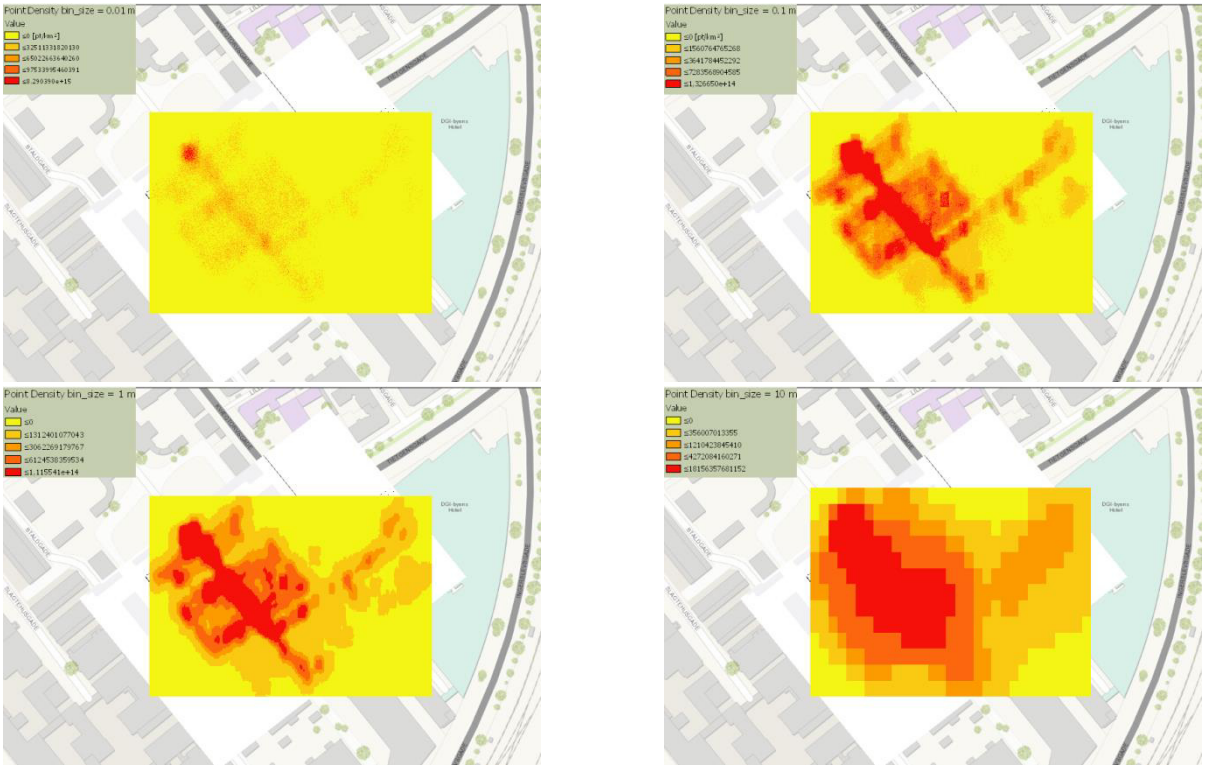
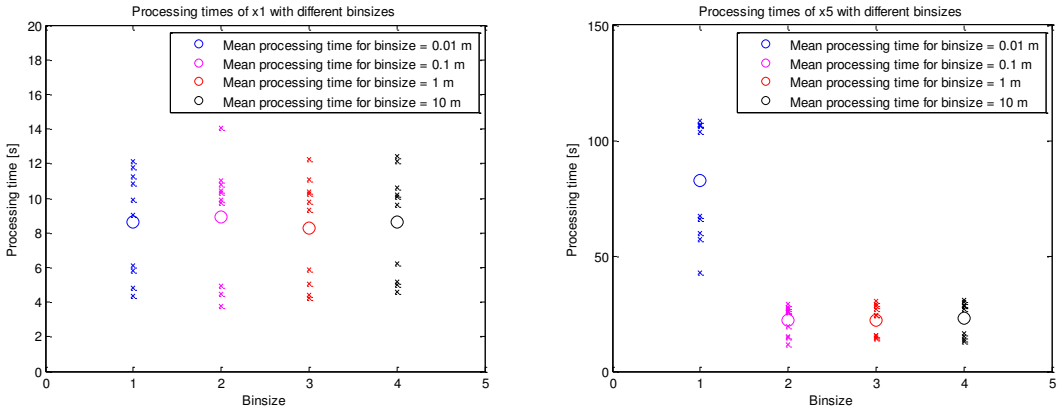



**Fig. 30: Point Density calculation with different bin sizes [0.01 m, 0.1 m, 1 m and 10 m]**

Fig. 31 shows a graphical illustration of the results. X-markers show the result of a single process, while o-marker show the mean of single processes of one file size and one bin size. Processing times are showing a relatively consistent behaviour. The variation when processing the smallest file is rather big in relative numbers (the longest process takes three times longer than the shortest) but might not be significant, as it is only a difference of around 8 seconds and the calculation of point density is usually not a task that is performed repeatedly in high numbers.



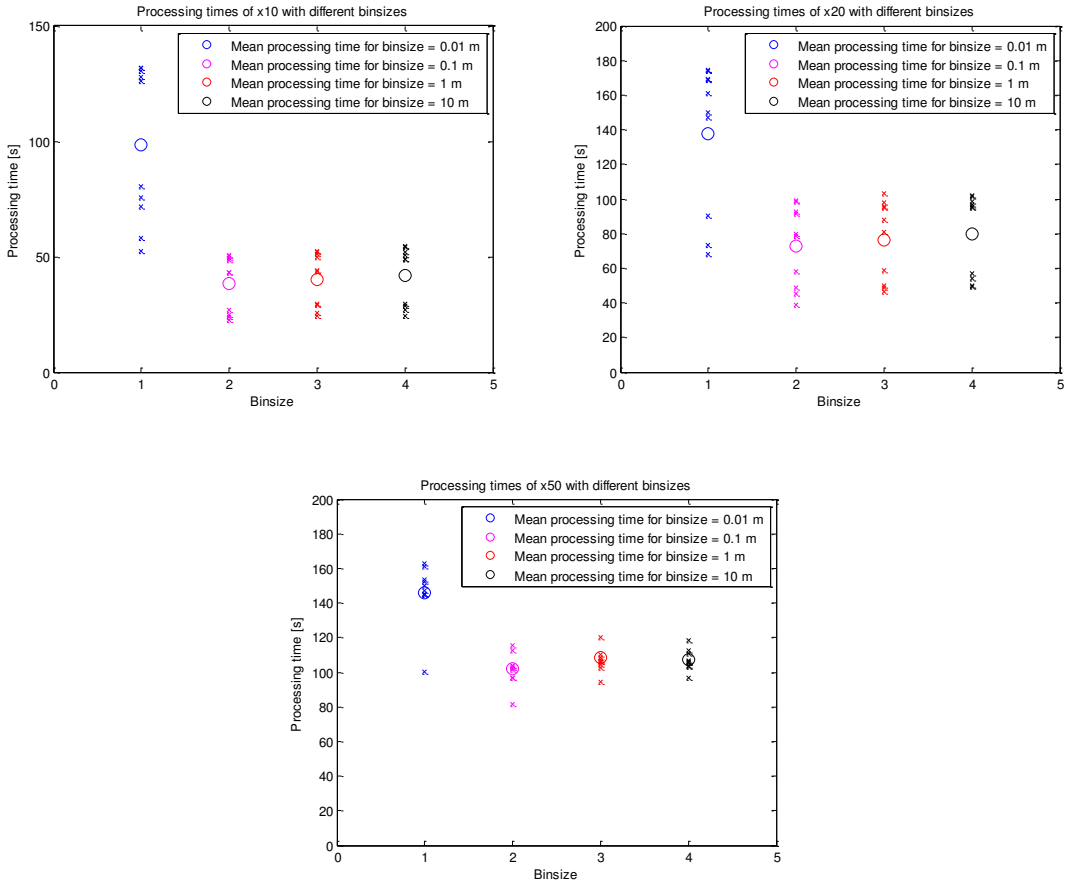



**Fig. 31: Processing times of Point Density calculations with different bin sizes**

Lastly, processing times are compared for desktop processing and cloud processing. The outcome is a rather surprising result. For all different file sizes the desktop processing was multiple times faster, most of the times by the factor of 10. Under these circumstances, data analysis in the cloud cannot be considered as a valuable alternative to desktop processing. Besides the observed high processing times, the process of uploading data into the cloud was also taking a relevant amount of time. Despite these problems the cloud solution also provides various advantages for mutual management and processing of data and distributing and publishing maps.

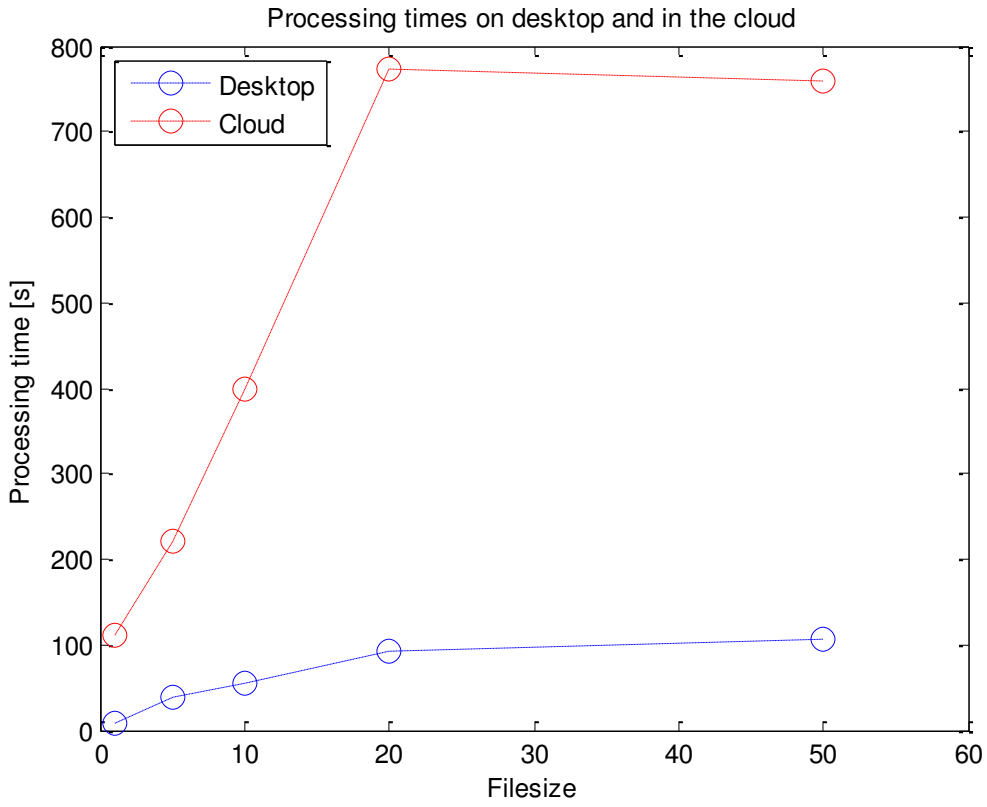



**Fig. 32: Comparison of Desktop and Cloud Processing**

These low processing times for cloud processing in contrast to desktop processing are shown in Fig. 32. The cause for this observed plunge of processing times is probably because of network connectivity issues which led to this distortion. For future applications these issues might be solved which potentially leads to an improved usage of the proposed cloud solution.

# 5. Conclusion

## 5.1 Summary

As suspected, it is possible to extract valuable knowledge from movement data, although several restrictions come with it. Those restrictions are mainly related to the dataset and not to the possible analysis methods or computational capacities. Several algorithms can be applied that are either already included *ESRI's ArcGIS* software or use parts of their provided tools. The possibility to modify available scripts or create own scripts and combine them with the included tools is a powerful way for data analysis and is applicable on a wide number of assignments.

Processing this dataset has shown where issues might arise when movement data is being analysed. First, it is important to get a clear image of the scope and characteristics of a dataset. This includes aspects of data quality, in particular spatial accuracy as well as a proper estimation of spatial accuracy, but also the temporal component (temporal accuracy, recording frequency) and other aspects, such as completeness and consistency. In this case a lack of completeness and consistency could lead to wrong conclusions. For example, we observed that there were no records available in some areas of the exhibition site, for instance the restaurant area. We need to keep in mind that this is because of a lack of completeness of records, due to unavailable beacons in this area. Otherwise it could lead us to the conclusion that no visitors have accessed the restaurant or this area was closed during that time. Another aspect arises from the fact that not the entirety of visitors is recorded but only a number of visitors that used the smartphone application for indoor navigation. Furthermore, the trajectories of the users of this application were not recorded continually, but only when they required a navigation or positioning task. These issues lead to one of the most challenging problems for further analysis, which is a low number of records, respectively a small dataset. For such analysis tasks, more records, and especially more complete and consistent records might improve the results tremendously.

The developed tools for data analysis provide some interesting results. It was proven that it is possible to estimate which exhibitors an individual user has visited and on that basis in what topics this individual might be interested in. This acquired information might prove to be of great benefit for various uses, for example targeted advertisement via smartphone, prediction of single user actions, prediction of visits of exhibitors – just to name a few. The creation of point density maps can introduce an image of general visitor behaviour. It can show which areas are very attractive to visitors or areas that are subject to a possible emergence of crowds. Point density maps can also be split in temporal segments in order to show how density distribution is changing over time. Furthermore, these density distributions can be used to detect significant positive and negative trends when comparing different events, different days of one event or different time segments of one day. Lastly,

timetables of exhibiting companies were created. These timetables indicate how many users have visited each booth. This gives not only valuable information about the temporal distribution of user's visits but also enables to detect significant increases in the frequency of user's visits.

Subsequent to the analysis tasks a comparison of processing methods was conducted. The processing times of desktop processing with the *Python* package *arcpy* and cloud processing with the *GeoAnalytics* package - both provided by *ESRI* - were compared. The results of this comparison are rather surprising. It has been shown that although the processing capability of the cloud computing solution is multiple times higher, its processing times are higher by at least the factor of 10 for different file sizes, for the creation of point density maps. Also a number of problems occurred for uploading datasets to *GeoAnalytics* server. It turned out to be a very time consuming task and many times it was not possible to upload files. Despite these problems the cloud solution also provides various advantages for mutual management and processing of data and distributing and publishing maps.

## 5.1 Outlook

It was proven that with the available tools it is possible to gain benefit and improve ways for decision making, through analysing movement data. There are still many ways to improve the results. These improvements are either data-sided aspects or software-sided. It is expected that with an increase in the amount of data and improved data quality not only the existing methods can yield better results but also many new methods can be developed to gain beneficial knowledge. An easy way to generate more data is to adapt the policy that data is only recorded when the smartphone application is actively used. This would result in a higher battery usage of the smartphone but will increase the amount of generated data by many times. Another way is to increase the number of users that are using the smartphone application, for instance by advertising the app on - or offline. Another limiting factor is the relatively low spatial accuracy of between three to five meters. If the spatial accuracy is increased it will not only increase certainty of the results but also other methods of data analysis are possible, for instance analysing trajectories. At this stage the spatial accuracy is not sufficient for this task to yield proper results.

Another way of improving the results is a refinement of the algorithms that are used for the analysis. For most tools a very simple algorithm was used to proof if the tools are applicable and if it is possible to get meaningful results. Those algorithms can be improved in many ways as a modification of already developed scripts is an easily performable task.

By developing tools that specifically aim to analyse movement data provided by this smartphone application for indoor navigation, there are many ways these tools can be used for future applications. In particular, the dataset that was used in this thesis was conducted at a bi-yearly trade fair. That means the already developed pre-processing steps and analysis

tools can be easily applied to future datasets and therefore make an analysis easily performable. The company *indoo.rs* which developed the smartphone application, the indoor navigation systems and provided the dataset also provide indoor navigation systems at different venues, so that with some adaption the developed tools are also applicable there.

# References


**Azaz, L., 2011.** The use of Geographic Information Systems (GIS) in Business. *International Conference on Humanities, Geography and Economics (ICHGE'2011) Pattaya Dec. 2011*

**Baskarada, S., Koronios, A., 2013.** Data, Information, Knowledge, Wisdom (DIKW): A Semiotic Theoretical and Empirical Exploration of the Hierarchy and its Quality Dimension. *Australasian Journal of Information Systems, Vol. 18, No. 1, 2013.*

**Brena, R. F., et. al., 2017.** Evolution of Indoor Positioning Technologies: A Survey. *Journal of Sensors, vol. 2017, Article ID 2630413, 21 pages, 2017. doi:10.1155/2017/2630413*

**Brown, DW. R., Dunn, D. B., 2011.** Classification Schemes of Positioning Technologies for Indoor Navigation. *2011 Proceedings of IEEE, 978-1-61284-738-2/11*

**Chrisman, N., 2001.** Exploring Geographic Information Systems, *2nd Edition. p. 12-31, Chapter 1: Reference Systems for Measurements, John Wiley & Sons Inc, ISBN: 978-0-471-31425-7*

**Efstathiou, C. E., 2006.** Estimation of type I error probability from experimental Dixon's "Q" parameter on testing for outliers within small size data sets. *Laboratory of Analytical Chemistry, Department of Chemistry, University of Athens, University Campus, Athens 15771, Greece, https://doi.org/10.1016/j.talanta.2005.12.031*

**Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. 1996.** From data mining to knowledge discovery - a review. *Advances in knowledge discovery (pp. 1–33). Cambridge, MA: AAAI Press/The MIT Press*

**Gartner.** *http://www.gartner.com/it-glossary/bigdata.*

**Gast, M. S., 2014.** Building Application with IBeacon: Proximity and Location Services with Bluetooth Low Energy. *Pages-80, Publisher- O'Reilly Media, Inc. ©2014 O'Reilly*

**Goodchild, M.F., 1992.** Geographical Information Science. *International Journal of Geographical Information Systems, 6:1, 31-45.*

**Goodchild, M.F., 1998.** 081: Geographic Information Systems. *Center for Spatial Studies and Department of Geography, University of California, Santa Barbara*

**Gordon, A. D., 1996.** Hierarchical classification. *In P. Arabie, L. J. Hubert, & G. D. Soete (Eds.), Clustering and classification (pp. 65–122). River Edge, NJ, USA: World Scientific Publisher.*

**Guo, D. S., Mennis, J., 2009.** Spatial data mining and geographic knowledge discovery—An introduction. *Computers, Environment and Urban Systems 33 (2009) 403–408*

**Diebold, F., 2012.** On the Origin(s) and Development of the Term ”Big Data”. *Pier working paper archive, Penn Institute for Economic Research, Department of Economics, University of Pennsylvania, 2012.*

**Fan, D., Bifet, A., 2013.** Mining Big Data: Current Status, and Forecast to the Future. *ACM SIGKDD Explor Newslett. 2013;14(2):1–5.*

**Hägerstrand, T., 1970.** What about people in regional science? *Papers of the Regional Science Association 24 (1), 6-21*

**Jain, A. K., Dubes, R. C., 1988.** Algorithms for clustering data. *Englewood Cliffs, NJ: Prentice Hall.*

**Kwan, M. P. 2000.** Human extensibility and individual hybrid-accessibility in space–time: A multi-scale representation using GIS. *Information, Place, and Cyberspace: Issues in Accessibility. Berlin, Germany: Springer-Verlag. pp. 241-56*

**Kwan, M. P., 2004.** GIS Methods in Time-Geographic Research: Geocomputation and Geovisualization of Human Activity Patterns. *Geogr. Ann., 86 B (4): 267–280.*

**Koperski, K., Adhikary, J., Han., J. W., 1997.** Spatial Data Mining: Progress and Challenges Survey paper. *School of Computing Science, Simon Fraser University*

**Maguire, D.J., Goodchild, M.F., 1991.** Geographical information systems.

**Mennecke, B. E., 1996.** Geographic Information Systems: Applications and Research Opportunities for Information Systems Researchers. *Proceedings of the 29th Annual Hawaii Intemational Conference on System Sciences*

**Mennecke, B. E., 2000.** Understanding the Role of Geographic Information Technologies in Business: Applications and Research Directions. *Journal of Geographic Information and Decision Analysis, vol.1, no.1, pp. 44-6*

**Leinweber, D. J. 2007.** Stupid Data Miner Tricks: Overfitting the S&P 500. *The Journal of Investing, 16:15–22.*

**Li, B., Rizos, C., 2014.** *Editorial: Special Issue International Conference on Indoor Positioning and Navigation 2012, Part 2. Journal of Location Based Services, 8(1), 1–2.*

**Lin, X. Y., Ho, T. W., Fang, C. C., Yen, Z. S., Yang, B. J., Lai, F., 2015.** A Mobile Indoor Positioning System Based on iBeacon Technology. *Conf Proc IEEE Eng Med Biol Soc. 2015;2015:4970-3. doi: 10.1109/EMBC.2015.7319507.*

**Retscher, G., 2016.** Indoor Navigation. *Springer International Publishing Switzerland 2016 E.W. Grafarend (ed.), Encyclopedia of Geodesy, DOI 10.1007/978-3-319-02370-0_9-1*

**Rowley, J., 2007.** The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science, 33, 163–180*

**Shaw, S. L., Wang, D., 2000.** Handling disaggregate spatiotemporal travel data in GIS. *GeoInformatica 4(2): 161-78*

**Silverman, B. W., 1986.** Density Estimation for Statistics and Data Analysis. *New York: Chapman and Hall.*

**Steiniger, S. & Weibel, R., 2010.** GIS Software - A description in 1000 words. *10.5167/uzh-41354.*

**Ward, J. H., 1963.** Hierarchical grouping to optimise an objective function. *Journal of the American Statistic Association, 58, 236–244.*

**Yu, H.B., 2006.** Spatio-temporal GIS Design for Exploring Interactions of Human Activities. *Cartography and Geographic Information Science, 33:1, 3-19, DOI: 10.1559/152304006777323136*

**Zhang, Q., Cheng, L. & Boutaba, R., 2010.** Cloud computing: state-of-the-art and research challenges. *J Internet Serv Appl (2010) 1: 7. https://doi.org/10.1007/s13174-010-0007-6*

1. https://dictionary.cambridge.org/dictionary/english/data-analysis *last access 2018 - 03*

2. http://www.dictionary.com/browse/business-intelligence *last access 2018 - 03*

3. https://support.esri.com/en/other-resources/gis-dictionary *last access 04 - 2018*

4. http://python-history.blogspot.co.at/2009/01/brief-timeline-of-python.html *last access 2018 - 03*

5. http://desktop.arcgis.com/en/arcmap/10.3/manage-data/using-sql-with-gdbs/what-is-sql.htm last access 2018 - 03

6. http://www.gartner.com/it-glossary/bigdata *last access 2018-03*

7. http://pro.arcgis.com/en/pro-app/tool-reference/data-management/select-by-location-graphical-examples.htm *last access 2018-03*

8. https://pro.arcgis.com/en/pro-app/tool-reference/feature-analysis/calculate-density.htm *last access 2018-03*

9. http://pro.arcgis.com/en/pro-app/tool-reference/spatial-analyst/point-density.htm *last access 2018-03*

10. http://enterprise.arcgis.com/en/server/latest/get-started/windows/perform-big-data-analysis.htm *last access 04 - 2018*

12. https://www.cpubenchmark.net/ *last access 04-2018*

# Appendix

## A1: Limit Values of Dixon's Q-Test

| Number of Values | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| $Q_{90\%}$ | 0.941 | 0.765 | 0.642 | 0.560 | 0.507 | 0.468 | 0.437 | 0.412 |
| $Q_{95\%}$ | 0.970 | 0.829 | 0.710 | 0.625 | 0.568 | 0.526 | 0.493 | 0.466 |
| $Q_{99\%}$ | 0.994 | 0.926 | 0.821 | 0.740 | 0.680 | 0.634 | 0.598 | 0.568 |

# A2: Python Scripts

## A2.1: Analysis 1

```
#Analysis 1: This Tool prints out general information about the file

import arcpy
from datetime import datetime
import timeit
import sys

start = timeit.default_timer()
#start timer

number = 3
#indicates which event file is processed

#Define workspace and filename
arcpy.env.workspace =
"C:/Users/Julian/PycharmProjects/Testdaten/Kopenhagen.gdb"
filename = 'event' + str(number) + '_trans'

# Part 1: Information about event-file
cursor = arcpy.SearchCursor(filename)
#Assigns cursor to the file
number_of_points = int(arcpy.GetCount_management(filename).getOutput(0))
# Count number of points in event-file

#initialize variables
all_ids =[]
#list with all ids total
all_ids_hourly = []
#list with all ids each hour
for i in range(0, 23):
#for loop to initialize array
    all_ids_hourly.append([])
number_of_visitors = 0
all_dates=[]
count=[0] * 24
count_users_hourly = [0] * 24
#ids and points for whole event
for user in cursor:
    ID = user.getValue('mobile_id')
#read the file and get value for id, date and time
    date = str(user.getValue('date'))[0:10]
    time = user.getValue('time_time')
    if ID not in all_ids:
#if id has not been processed before it is appended to all_ids array
        all_ids.append(ID)
        number_of_visitors += 1
# Count number of visitors in event-file
    if date not in all_dates:
        all_dates.append(date)
#change to string?
    for j in range(0, 23):
        if time[0:2] == str(j) or time[0:2] == '0' + str(j):
            count[j] = int(count[j]) + 1
        if ID not in all_ids_hourly[j] and (time[0:2] == str(j) or
time[0:2] == '0' + str(j)):
```

```
            count_users_hourly[j] += 1
            all_ids_hourly[j].append(ID)

#ids and points each day
count_daily=[0] * 24
#initialize variables as in previous step
count_users_hourly_daily= [0] *24
count_daily_final=[]
count_users_hourly_daily_final =[]
for d in range(0, len(all_dates)):
    all_ids_hourly_dayly = []
    for i in range(0, 23):
        all_ids_hourly_dayly.append([])
    cursor = arcpy.SearchCursor(filename)
    for user in cursor:
        ID = user.getValue('mobile_id')
        date = str(user.getValue('date'))[0:10]
        time = user.getValue('time_time')
        if date == all_dates[d]:
            for j in range(0,23):
                if time[0:2] == str(j) or time[0:2] == '0' + str(j):
                    count_daily[j] = int(count_daily[j]) + 1
                if ID not in all_ids_hourly_dayly[j] and (time[0:2] ==
str(j) or time[0:2] == '0' + str(j)):
                    count_users_hourly_daily[j] += 1
                    all_ids_hourly_dayly[j].append(ID)
    del cursor

arcpy.CreateTable_management("C:/Users/Julian/PycharmProjects/Testdaten/Kop
enhagen.gdb", "timetable_event" + str(number) + '_' + str(d))
    arcpy.AddField_management("timetable_event" + str(number) + '_' +
str(d), "points", "LONG")
    arcpy.AddField_management("timetable_event" + str(number) + '_' +
str(d), "users", "LONG")
    k=0
    with arcpy.da.InsertCursor("timetable_event" + str(number) + '_' +
str(d), ['points', 'users']) as inCursor:
        for x in range(0, 23):
            inCursor.insertRow((count_daily[x],
count_users_hourly_daily[x]))
    del inCursor
    count_daily_final.append(count_daily)
    count_users_hourly_daily_final.append(count_users_hourly_daily)
    count_daily = [0] * 24
    count_users_hourly_daily = [0] * 24

timetable = ''.join(str(count))
#prepare variables for print
timetable_users = ''.join(str(count_users_hourly))
all_dates_join = ' '.join(all_dates)
points_per_visitor = number_of_points / number_of_visitors

#Part 2: Information about Polygon-file
number_of_Polygons = int(arcpy.GetCount_management('Polygon').getOutput(0))
# Count number of points in event-file

print('The name of the file is ' + filename)
print('The number of points is: ' + str(number_of_points))
print('The number of users is: ' + str(number_of_visitors))
print('The average number of points per track is: ' +
str(points_per_visitor))
```

```
print('The dates of the events are: ' + all_dates_join)
for x in range (0, len(all_dates)):
    print('The number of points on date ' + all_dates[x] + ' is ' +
str(sum(count daily final[x])))
    print('The number of users on date ' + all_dates[x] + ' is ' +
str(sum(count_users_hourly_daily_final[x])))

print('The number of exhibitors is: ' + str(number_of_Polygons))

desc = arcpy.Describe(filename)

# Print dataset properties
print(("Extent:\n  XMin: {0}, XMax: {1}, YMin: {2}, YMax: {3}".format(
    desc.extent.XMin, desc.extent.XMax, desc.extent.YMin,
desc.extent.YMax)))
print(("Spatial reference name: {0}:".format(desc.spatialReference.name)))
print(('Datatype: {0}'.format(desc.dataType)))

# Calculate how many exhibitors one user is visiting on average
cursor  = arcpy.SearchCursor('users_interests')
total_exhibitors_visited = 0
total_users = 0
for row in cursor:
    exhibitors = row.getValue('exhibitor')
    exhibitors_visited = exhibitors.count('-') - 1
    if exhibitors_visited > 1:
        total_exhibitors_visited += exhibitors_visited
        total_users += 1
average_exhibitors_visited = total_exhibitors_visited/total_users
print('On average one user visits ' + str(average_exhibitors_visited) + '
exhibitors')

#Timetable for exhibitors
all_ids_hourly = []
for i in range(0, 23):
    all_ids_hourly.append([])
fieldname = 'Time_'
arcpy.AddField_management("exhibitor_timetable", fieldname, "LONG")
#add field to describe timetable for exhibitors
count = [0] * 24
with arcpy.da.UpdateCursor('exhibitor_timetable', [fieldname, 'Name',
'OBJECTID']) as upCursor:
    for exhibitor in upCursor:
#Update field "Class" to match exhibitor name
        count = [0] * 24
        all_ids = []
        all_ids_hourly = []
        for i in range(0, 24):
            all_ids_hourly.append([])
        exhibitor_name = exhibitor[1]
        exhibitor_id = exhibitor[2]
        arcpy.SelectLayerByAttribute_management("Polygon", "NEW_SELECTION",
'"OBJECTID" = %s' % exhibitor_id)       #Select one Polygon (exhibitor) in
every instance
        arcpy.SelectLayerByLocation_management(filename, "Intersect",
'Polygon')                                  #Select points within Polygon
        cursor2 = arcpy.SearchCursor(filename)
        for row2 in cursor2:
            time_time = str(row2.getValue('time_time'))
            id = row2.getValue('mobile_id')
            for t in range(0, 24):
```

```
                if (time_time[0:2] == str(t) or time_time[0:2] == '0' +
str(t)) and id not in all_ids_hourly[t]:
                    count[t] += 1
                    all_ids_hourly[t].append(id)
        #print('There are ' + str(count) + ' Points at Time ' + str(t) +'
in ' + exhibitor_name)
        #exhibitor[0] = count[0]
        upCursor.updateRow(exhibitor)
        arcpy.SelectLayerByAttribute_management('Polygon',
"CLEAR_SELECTION")                                    #Clear selections
        arcpy.SelectLayerByAttribute_management(filename,
"CLEAR_SELECTION")
        print(exhibitor_name + ': ' + str(count))

stop = timeit.default_timer()
#stop timer
print(stop - start)
#print timer
```

## A2.2: Analysis 2

```
#This Tool creates a file that describes the interests of every user

import arcpy
import timeit

start = timeit.default_timer()                          #start timer

arcpy.env.workspace =
"C:/Users/Julian/PycharmProjects/Testdaten/Kopenhagen.gdb"
Polygon_File = 'Polygon'                                #assign Polygon and
event file
event_File = 'event3_trans'

cursor = arcpy.SearchCursor(Polygon_File)           #Polygon_File is a
shapefile that describes the location of exhibitors as a polygon feature
i=0
for exhibitor in cursor:
    print(exhibitor.getValue('Name'))
    exhibitor_name = exhibitor.getValue('Name')     #get values for name
and id
    exhibitor_id = exhibitor.getValue('OBJECTID')
    arcpy.SelectLayerByAttribute_management(Polygon_File, "NEW_SELECTION",
'"OBJECTID" = %s' % exhibitor_id)       #Select one Polygon (exhibitor) in
every instance
    arcpy.SelectLayerByLocation_management(event_File, "Intersect",
Polygon_File)                                   #Select points within
Polygon
    result = arcpy.GetCount_management(event_File)
#Count number of points within Polygon
    print("There are {0} Points within the Polygon".format(result))
    with arcpy.da.UpdateCursor(event_File, 'Class') as upCursor:
        for row in upCursor:
#Update field "Class" to match exhibitor name
            row[0] = exhibitor_name
            upCursor.updateRow(row)
            i += 1
        print(exhibitor.getValue('Name'))

    arcpy.SelectLayerByAttribute_management(Polygon_File,
"CLEAR_SELECTION")                                          #clear selection
    arcpy.SelectLayerByAttribute_management(event_File, "CLEAR_SELECTION")
print('Process finished. ' + str(i) + ' Points found.')

#Create New Table. This Table stores the visited exhibitors for every user
arcpy.CreateTable_management("C:/Users/Julian/PycharmProjects/Testdaten/Kop
enhagen.gdb", "users_interests")
arcpy.AddField_management("users_interests", "mobile_id", "LONG")
arcpy.AddField_management("users_interests", "exhibitor", "TEXT")

cursor = arcpy.SearchCursor(event_File)
x = 0
all_ids =[]
#list with all ids
m = 0
user_visit = []                                                         #
list with all exhibitors one id visited
cursor_for_insert =
arcpy.da.InsertCursor("C:/Users/Julian/PycharmProjects/Testdaten/Kopenhagen
.gdb/users_interests",('OBJECTID', 'mobile_id', 'exhibitor'))
```

```
for user in cursor:
    #cursor_users_interests = arcpy.SearchCursor('users_interests')
    ID = user.getValue('mobile_id')
    exhibitor = user.getValue('Class')
    if ID not in all_ids:
        cursor_for_insert.insertRow((m, ID, ''))
        all_ids.append(ID)
        m += 1
        user_visit = []
del cursor_for_insert
print('IDS created')
cnt=0
for i in all_ids:
    finalString=""
    whereClause=str(i)+'= mobile_id'
    cursor = arcpy.da.SearchCursor(event_File, 'Class',whereClause)
    for j in cursor:
        string=str(j)
        string=string[2:-3]
        if string not in finalString and len(finalString) + len(string) <
252:
            finalString+=' - '+string

updateCursor=arcpy.da.UpdateCursor("C:/Users/Julian/PycharmProjects/Testdat
en/Kopenhagen.gdb/users_interests", 'exhibitor',whereClause)
    for j in updateCursor:
        j[0]=finalString
        updateCursor.updateRow(j)
del updateCursor

arcpy.AddField_management("users_interests", "Tag", "TEXT")
with arcpy.da.UpdateCursor('users_interests', ['exhibitor', 'Tag']) as
upCursor:
    for row in upCursor:  # Update field "Class" to match exhibitor name
        tags = []
        tags_string = ''
        cursor = arcpy.da.SearchCursor('Polygon_File', ['NAME', 'TAG'])
        for row2 in cursor:
            if str(row2[0]) in row[0]:
                tags.append(row2[1])
        if len(tags) != 0:
            for x in tags:
                x = str(x)
                x = x.split(',')
                for y in x:
                    if str(y) not in tags_string and y != 'None' and
len(tags_string) == 0:
                        tags_string += str(y)
                    elif str(y) not in str(tags_string) and y != 'None' and
len(tags_string) + len(str(y)) < 254:
                        tags_string += ', ' + str(y)

        print(tags_string)
        row[1] = tags_string
        upCursor.updateRow(row)

stop = timeit.default_timer()                #stop timer
print(stop - start)                          #print timer
```

## A.2.3: Analysis 3

```
#Analysis3: TREND DETECTION

env.workspace = "C:/Users/Julian/PycharmProjects/Testdaten/Kopenhagen.gdb"
arcpy.AddField_management("Polygon", "Trend", "Double")
#Create new field in table


pdensOut = PointDensity("event3_trans.shp", "NONE")
#create point density for one event
pdensOut.save("event3_trans_pointdensity")
pdensOut = PointDensity("event2_trans.shp", "NONE")
#create point density for second event
pdensOut.save("event2_trans_pointdensity")
arcpy.Minus_3d("C:/Users/Julian/PycharmProjects/Testdaten/Kopenhagen.gdb/ev
ent3_trans_pointdensity",
"C:/Users/Julian/PycharmProjects/Testdaten/Kopenhagen.gdb/event2_trans_poin
tdensity",
"C:/Users/Julian/PycharmProjects/Testdaten/Kopenhagen.gdb/event3_minus_even
t2")

max_result =
arcpy.GetRasterProperties_management("C:/Users/Julian/PycharmProjects/Testd
aten/Kopenhagen.gdb/event3_minus_event2", "MAXIMUM")
#calculate high values for positive trend
max = max_result.getOutput(0)
max = float(max) *0.1
outConstRaster = CreateConstantRaster(max, "FLOAT", 2, Extent(0, 0, 90,
90))
outGTE =
GreaterThan("C:/Users/Julian/PycharmProjects/Testdaten/Kopenhagen.gdb/event
3_minus_event2", outConstRaster)
arcpy.RasterToPolygon_conversion(outGTE,
"C:/Users/Julian/PycharmProjects/Testdaten/Kopenhagen.gdb/event3_minus_even
t2_OutPolygon", "NO_SIMPLIFY")
arcpy.SelectLayerByAttribute_management("event3_minus_event2_OutPolygon",
"NEW_SELECTION", '"gridcode" = %s' % 1)
arcpy.SelectLayerByLocation_management('Polygon', "Intersect",
'event3_minus_event2_OutPolygon')                                    #Select
points within Polygon

updateCursor = arcpy.da.UpdateCursor('Polygon', 'Trend')
#update field
for j in updateCursor:
    j[0] = 1
    updateCursor.updateRow(j)

del updateCursor

min_result =
arcpy.GetRasterProperties_management("C:/Users/Julian/PycharmProjects/Testd
aten/Kopenhagen.gdb/event3_minus_event2", "MINIMUM")             #calculate
low  values for negative trend
min = min_result.getOutput(0)
min = float(min) *0.05
print(min)
outConstRaster = CreateConstantRaster(min, "FLOAT", 2, Extent(0, 0, 90,
90))
outLTE =
LessThan("C:/Users/Julian/PycharmProjects/Testdaten/Kopenhagen.gdb/event3_m
```

```
inus_event2", outConstRaster)
arcpy.RasterToPolygon_conversion(outLTE,
"C:/Users/Julian/PycharmProjects/Testdaten/Kopenhagen.gdb/event3_minus_even
t2_OutPolygon", "NO_SIMPLIFY")
arcpy.SelectLayerByAttribute_management("event3_minus_event2_OutPolygon",
"NEW_SELECTION", '"gridcode" = %s' % 1)
arcpy.SelectLayerByLocation_management('Polygon', "Intersect",
'event3_minus_event2_OutPolygon')                                    #Select
points within Polygon

arcpy.AddField_management("Polygon", "Trend", "Double")


updateCursor = arcpy.da.UpdateCursor('Polygon', 'Trend')
#update field
for j in updateCursor:
    j[0] = -1
    updateCursor.updateRow(j)
```

## A.2.4 Comparison of Processing Methods

```
import arcpy
from arcpy import env
from arcpy.sa import *
import timeit
import sys
import arcpy
from arcgis.gis import GIS
from arcgis.geoanalytics.analyze_patterns import calculate_density
from IPython.display import display
import random
import arcgis

ids =
['0847418430b4483db8a682bc595bf0d5','8c93ee08e2e343a78996d4eaac0ea601',
'86ef139c52fd41789e100f55886d000e','88ac576f0010441ea47a7b5c87fdd661','9c9b
f5a23c534d6a9f089a113e6509a7','f1dacbac75ab497fb1481edf3ad813a1']
#original, x5, x10, x20, x50, x100

start = timeit.default_timer()

arcpy.env.workspace =
"C:/Users/Julian/PycharmProjects/Testdaten/Kopenhagen.gdb"
size=[1,5,10,20,50,100]   #original file multiplied size by [size]
gis = GIS("portal", "user", "password")
binsize = [0.01, 0.1, 1, 10]
loops = 10

for j in range(0, loops):
    for multiplier in binsize:
        for i in range(0,len(size)):
            # file stored local - processing local
            if i == 1:
                start = timeit.default_timer()
                PointDensity(

in_point_features="C:/Users/Julian/PycharmProjects/Testdaten/Kopenhagen.gdb
/event3_trans",
                    population_field=None, cell_size=0.00000920636,
neighborhood=NbrCircle(2, "CELL"),
                    area_unit_scale_factor='SQUARE_METERS')
                stop = timeit.default_timer()
                print('Processing time for I - size ' + str(size[i]) + '
binsize ' + str(multiplier) + ' ' + str(stop - start))
            else:
                start = timeit.default_timer()
                cellsize = 0.00000920636
                PointDensity (in_point_features =
"C:/Users/Julian/PycharmProjects/Testdaten/Kopenhagen.gdb/event3_trans_x" +
str(size[i]),
                              population_field = None, cell_size = cellsize
* multiplier, neighborhood = NbrCircle(2, "CELL"),
                              area_unit_scale_factor = 'SQUARE_METERS')
                stop = timeit.default_timer()
                print('Processing time for I - size ' + str(size[i]) + '
binsize ' + str(multiplier) + ' ' + str(stop - start))


            #file stored in the cloud (rds) - process in the cloud
            analysis_item = gis.content.get(ids[i]) #event3_rds
```

```
            start = timeit.default_timer()
            myradius = 2 * multiplier

arcgis.geoanalytics.analyze patterns.calculate density(analysis item.layers
[0], fields=None, weight='Kernel',

bin_type='SQUARE', bin_size=multiplier, bin_size_unit='Meters',

time_step_interval=None, time_step_interval_unit=None,

time_step_repeat_interval=None,

time_step_repeat_interval_unit=None,

time_step_reference=None, radius=multiplier, radius_unit='Meters',

area_units='SquareKilometers', output_name='event3_trans_pd' +
str(size[i]))
            stop = timeit.default_timer()
            search = gis.content.search('title:event3_trans_pd' +
str(size[i]))
            search[0].delete()
            print('Processing time for II - size ' + str(size[i]) + ' ' +
str(stop - start))
```