Track & Know

Big Data for Mobility Tracking Knowledge Extraction in Urban Areas

D1.1 Track & Know Observatory

Document Summary Information

Grant Agreement No	780754	**Acronym**	TRACK & KNOW
Full Title	Big Data for Mobility Tracking Knowledge Extraction in Urban Areas		
Start Date	01/01/2018	**Duration**	36 months
Project URL	https://trackandknow.eu		
Deliverable	D1.1 Track & Know Observatory		
Work Package	WP1		
Contractual due date	30.6.2018	**Actual submission date**	30.6.2018
Nature	Report	**Dissemination Level**	Public
Lead Beneficiary	UPRC		
Responsible Author	Yannis Theodoridis (UPRC)		
Contributions from	Ibad Kureshi (Inlecom), Angelos Liapis (KT), Mirco Nanni (CNR), Nikos Katzouris (NCSRD), Luk Knapen (UHASSELT), Gennady Andrienko (FRHF), Rob Weibel (UZH), Nicolas Baskiotis (UPMC), Marios Logothetis, Ioannis Daskalopoulos (Intrasoft), Eva Chondrodima, Christos Doulkeridis, Harris Georgiou, Nikos Pelekis, Akrivi Vlachou (UPRC), Toni Staikova (CEL), Ian Smith (PAP), Leonardo Longhi (SIS), Panagiotis Livanos (ZEL), Dimitrios Flokos (ZEL), Anagnostis Delkos (ZEL), Athanasios Koumparos (ZEL), Konstantionos Koufogiannis (ZEL)		

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No 780754.

Revision history (including peer reviewing & quality control)

Version	Issue Date	% Complete[1]	Changes	Contributor(s)
0.1	23.03.2018	5%	TOC	Yannis Theodoridis (UPRC)
0.2	04.04.2018	5%	TOC revised	Yannis Theodoridis (UPRC)
0.3	15.05.2018	70%	1st draft (integrated, revised contributions per partner), sent to internal reviewers	Ibad Kureshi (Inlecom), Angelos Liapis (KT), Mirco Nanni (CNR), Nikos Katzouris (NCSRD), Luk Knapen (UHASSELT), Gennady Andrienko (FRHF), Rob Weibel (UZH), Nicolas Baskiotis (UPMC), Marios Logothetis, Ioannis Daskalopoulos (Intrasoft), Eva Chondrodima, Christos Doulkeridis, Harris Georgiou, Nikos Pelekis, Akrivi Vlachou (UPRC), Toni Staikova (CEL), Ian Smith (PAP), Leonardo Longhi (SIS), George Kaptsis (ZEL)
0.5	26.06.2018	100%	2nd draft (revised, according to reviewers' comments), sent to PM and QM for final check	(same as above)
1.0	29.06.2018	100%	Final (revised, according to PM's and QM's comments), sent to PM for submission to the EC	Marios Logothetis (Intrasoft), Ibad Kureshi (Inlecom)

[1] According to TRACK&KNOW's Quality Assurance Process:
Due Date – 6 Weeks: Peer Review (Reviewers)
Due Date – 2 Weeks: Quality Manager Review (INTRA)
Due Date – 2 Days: Sent to Project Coordinator for Submission to the EC after addressing all comments by Quality Manager and Peer Reviewers

Disclaimer

Copyright message

Table of Contents

List of Figures

List of Tables

Glossary of terms and abbreviations used

Abbreviation / Term	Description
BDVA	Big Data Value Association
CER	Complex Event Recognition
FLP	Future Location Prediction
GDPR	General Data Protection Regulation
HDFS	Hadoop Distributed File System
HLAS	Health, Liveability, Adaptability and Sustainability
IMN	Individual Mobility Network
IoT	Internet of Things
TP	Trajectory Prediction
VA	Visual Analytics

1 Deliverable Purpose and Structure

The purpose of deliverable **1.1 – Track & Know Observatory** is to provide a single point of reference on the literature review and market analysis for competing and complementing tools and technologies. This allows for the clear identification of the R&D work that needs to be conducted to meet the desired KPIs during the project's lifetime. Accompanying to this report, an online observatory, implemented as an integral part of the project website, will identify and monitor the current state-of-the-art throughout the project.

The structure of this deliverable is as follows: Section 2 presents the overall Track & Know concept, according to the BDVA reference model. Sections 3, 4, 5, and 6 provide an in-depth literature review on state-of-the-art big data platforms and architectures, big data management and processing techniques, big data analytics methods, and big data visualization and visual analytics methods and tools, respectively, which are highly relevant to the project. Sections 7, 8, and 9 conduct a market analysis on the three industries that will be addressed in Trach & Know project: car insurance, healthcare services, and fleet management, respectively. Section 10 presents the structure and functionality of the online observatory, implemented as an integral part of the project website, as already mentioned. Section 11 concludes the report.

2 The Track & Know concept

As the world's population living in metropolitan areas increases, so increases the need for effective and sustainable interventions and services to inject mobility intelligence and improve the quality of life in large urban environments. Technological developments, in particular the extended and expanding use of ICT, have resulted the collection of unprecedented volumes of data across systems operating in the transport, mobility and the urban applications domains. Moreover, the influence of digital evolution is changing the experience of consumers of services in these domains and is driving the expectations that will shape the demand in the coming years. Although certain markets have been radically changed by the influence of technology, the transport, mobility and urban services sectors are changing at a slower pace and both commercial providers and public operators are very slowly adapting to the current technology offerings.

The existing accumulated large volumes of data, known also as "big data", are generating a strong interest in the research communities, the relevant industries and among policy makers. The adoption of digital services is expected to enable service providers to deliver secure and efficient services across intelligent infrastructures with higher automation capacity. As a result, the demand for efficient and scalable smart services facilitating personalized, adaptable, environmental and sustainable capabilities impose new requirements for the improved exploitation of the immense and continuously rising amounts of data generated by industrial operations, sensors and devices (Internet of Things - IoT), social media and often aggregated Open Data sources.

Increasingly, in smart cities cloud-based infrastructures combined with behavioural, institutional and policy-related data sources create a critical pathway in achieving the four-partite goals of Health, Liveability, Adaptability and Sustainability (HLAS). Elaborating on the cross-domain Big Data Value generation, so far, the most common approaches e.g. for smart mobility focus on tracking vehicles' spatial and temporal information and have not been related with health, adaptable insurance services and life quality indices (i.e. driving and cognitive capabilities, driving skills deterioration, prevention, symptoms early-detection, prediction and control, etc.).

Unfortunately, in today’s information society, the ability of retrieving knowledge from mobility and contextual data is becoming more and more critical for the competitiveness of all the economic, political and cultural entities. Companies produce within their individual activities and along with synergies (i.e. traffic monitoring, fleet and healthcare management, emergency response and/or adaptable insurance services, etc.) a huge flow of data coming from diverse domains, different devices, processes, markets, user generated content/feedback or external sources. The rise of ubiquitous connectivity combined with IoT (Internet of Things) is the key enabler for the rise of “Industry 4.0” and already allows acquiring data from an evolving mobile and stationary ecosystem. The more data are available, the more the tools to manage them have advanced. Databases, software, computing power and in general technical infrastructure have grown and improved in the past years, thanks to investments and research. However, there is a gap between the availability of data and the potential information or knowledge that can be extracted from them: not every enterprise has now the means to analyse such an amount of data. In addition to this, Data Science imposes the tight collaboration among different stakeholders coming from diverse domains (i.e. ICT, Transport, Insurance, Health, etc.) to add value to this kind of data in an intelligent, efficient and scalable manner. But still, this amount of data most of the time remains decoupled and isolated within private infrastructures. Many companies can benefit from data management and analysis in order to infer knowledge by effectively blending data. Knowledge from Big Data is the key for success.

In this direction, Track&Know, by injecting computational thinking capabilities in the context of smart mobility services, aims to address the challenges of the emerging Big Data Value ecosystem, including the autonomous, connected and shared vehicles technologies, the self-enablement configuration

features (such as e.g. in Intelligent Transportation Systems - ITS, self-parking, speed and/or lane control), addressing the questions of what type of information is needed, and with whom, when and how it is used. More specifically, Track&Know integrates multidisciplinary research teams from Mobility Data management, Complex Event Recognition, Geospatial Modelling, Complex Network Analysis, Transportation Engineering and Visual Analytics to develop new models and applications. An insight of Track&Know concept is represented in Figure 1 (source: www.bdva.eu), providing an overview on how the aforementioned challenges are going to be addressed conjunctively and towards providing a coherent Big Data Framework and related solutions.

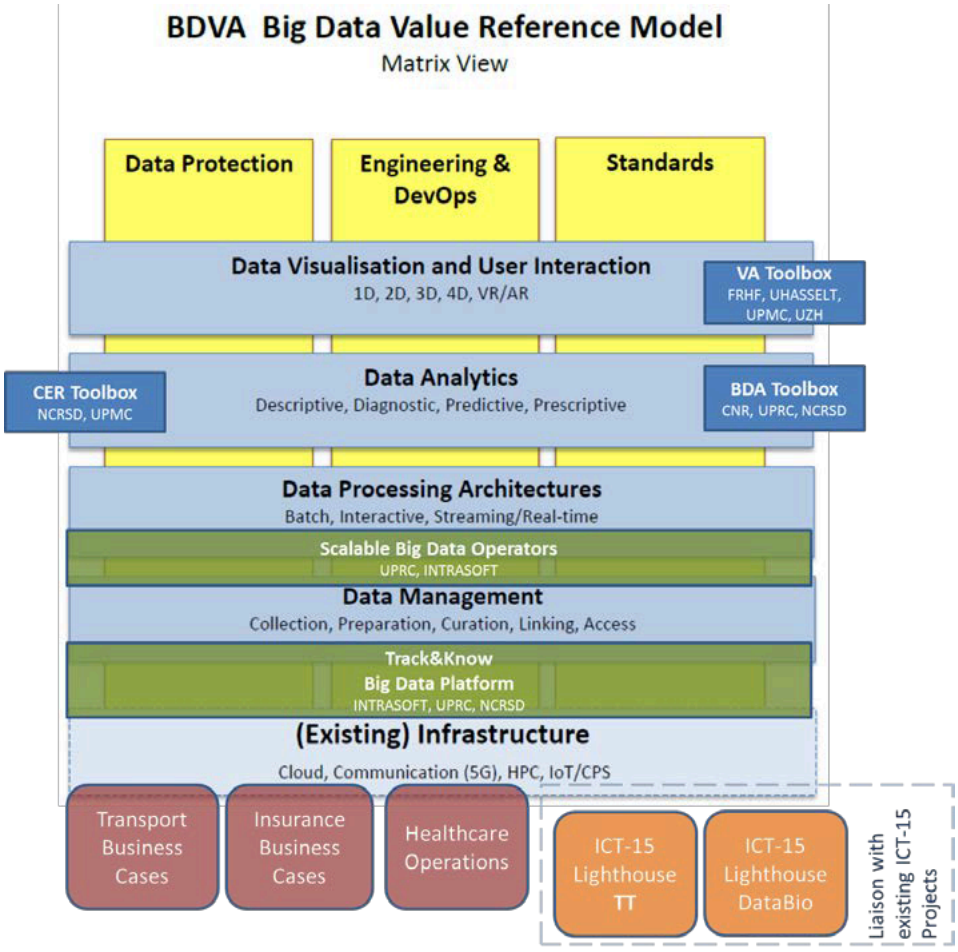


Figure 1: The Track & Know components mapped to the BDVA reference model.

2.1 Scope and Approach

Track&Know project brings together interdisciplinary partners from the transport, insurance, emergency healthcare industries, academia and research along with users and data-provision partners focusing on real-life and user-defined challenges to address the open issues arising from the automotive transportation in modern metropolitan areas and increase the contextual awareness in urban mobility by delivering intelligent information and predictive analytics to user-interest groups, stakeholders and city managers.

Track&Know is a research and innovative collaborative project targeting at developing and exploiting novel technologies and methods in order to increase the efficiency of Big Data domains such as

Transport, Insurance and Healthcare, aiming in the same time at the applicability in other European industries. The overall Big Data architecture is shown below (Figure 2).

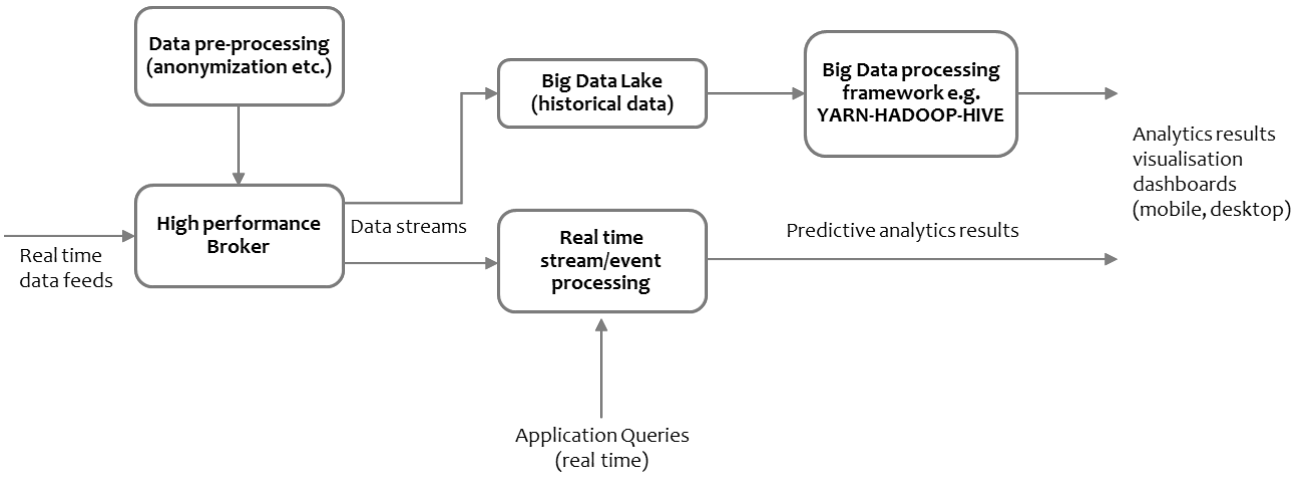


Figure 2: The Track&Know Real time Processing Flow

The Data Sources are multiple streaming and heterogenous data sources, as well as archival and contextual data of huge volumes.

Big Data Toolboxes for intelligent and integrated services with predictive safety capabilities (i.e. for collision avoidance, optimized emergency response and/or accident management, driving skills deterioration, adaptable insurance services, etc.) are considered critical by people and researchers within the spectrum to reshape the way that visualization techniques make data accessible in ways humans understand. Advanced data analytics support more efficient decision-making and scalable, iterative processes generate trusted insights in the automotive transportation landscape.

Track&Know incorporates novel methodologies for real-time detection and prediction of individuals and mobility patterns, enabling for risk assessment and crisis management, inference of useful knowledge and complex events related to car drivers, people's transportation activities, together with advanced visual analytics methods, over multiple heterogeneous, voluminous, unlabelled, fluctuating, and noisy data streams, correlating them with archival data exhibiting behavioural and social characteristics, demographics, health and life quality indices, geographical information, mobility analytics and intentional data (e.g. commuting, daily/weekly regularities) etc., in a timely manner. Track&Know applies Big Data driven innovations for Operations and Tasks Planning improvements as well as the configuration of new disruptive business models in the domains of Mobility, Insurance and Health.

More specifically, Track&Know will provide a streamlined, Big Data Platform that will be endowed with the Big Data frameworks, software stacks and Toolboxes that the project will research, innovate and demonstrate. The Track&Know Big Data Platform will be the hub that will be integrated with existing industrial systems that are made available to the project (from Consortium partners, namely INTRASOFT, ZEL, SIS, PAP) and will be used within the pilot applications and will be validated during the demonstration cases. Furthermore, the Track&Know Big Data Platform will aim at securing the sustainability of the produced results, engaging the vast spectrum of research and software development communities that participate in the project, and also at engaging global communities of data providers and end users to promote pilot activities not only in the Transport domain, but also to other domains that could benefit by the produced tools. The Big Data Platform will be developed by INTRASOFT, a core member within BDVA and is thus aligned with the BDVA reference model (BDV SRIA, 2016), as shown at a high level earlier in Figure 1.

In particular, the BDVA Reference Model covers the most important Big Data technical areas (shown horizontally). The BDVA Reference Model also covers key cross-cutting concerns, such as data protection, cyber security, development and operations, and standards (shown vertically). Track&Know architecture will capitalize on the existing infrastructure that is provided by participating industrial partners in the corresponding demonstration Pilots.

The Track&Know Big Data Platform will integrate the components and Toolboxes developed during the project. The design of the proposed developments is based on the BDVA reference architecture and the respective guidelines aim at supporting Big Data application developers at improving their applications performance and efficiency.

2.2 References

BDV SRIA (2016) Big Data Value Strategic Research and Innovation Agenda. Available online at: http://www.bdva.eu/sites/default/files/EuropeanBigDataValuePartnership_SRIA__v2.pdf

3 Big Data Platforms and Infrastructure

The scope of this section is to describe the available Big Data Platforms, the Big Data File / Storage Systems, as well as the existing Big Data Batch / Stream Processing approaches and the connectors in order to present a system in its entirety which we can store, send and analyze data.

3.1 Big Data Platforms and Infrastructure

Cloudera Enterprise - Cloudera Enterprise (Cloudera) includes CDH, an open source Hadoop-based platform. CDH is Cloudera's 100% open source platform distribution, including Apache Hadoop and built to meet enterprise demands. By integrating Hadoop with more than a dozen other critical open source projects, Cloudera has created a functionally advanced system that helps in performing end to end Big Data workflows. Cloudera Enterprise is available in three editions, each offering varying levels of services management capabilities. The basic edition provides management capabilities to support cluster running core CDH services that include HDFS, Hive, MapReduce, Oozie, YARN and ZooKeeper.

Cloudera's Benefits

- One integrated system, bringing diverse users and application workloads to one pool of data on common infrastructure; no data movement required
- Perimeter security, authentication, granular authorization, and data protection
- Enterprise-grade data auditing(control), data lineage, and data discovery
- Native high-availability, fault-tolerance and self-healing storage, automated backup and disaster recovery, and advanced system and data management -
- Apache-licensed open source to ensure your data and applications remain yours, and an open platform to connect with all of your existing investments in technology and skills
- One massively scalable platform to store any amount or type of data, in its original form, for as long as desired or required
- Integrated with your existing infrastructure and tools
- Flexible to run a variety of enterprise workloads - including batch processing, interactive SQL, enterprise search and advanced analytics

Cloudera's Disadvantages

- Not fully Open Source, couple of components of the distributions are privately owned, meaning with public contributions are not welcome
- More Up to date technologies
- Improvements to Cluster Management tool is required, which are already available to its competitors

Hortonworks - Architected, developed, and built completely in the open, Hortonworks Data Platform (Hortonworks) provides Hadoop designed to meet the needs of enterprise data processing. HDP is a platform for multi-workload data processing across an array of processing methods - from batch through interactive to real-time - all supported with solutions for governance, integration, security and operations. Furthermore, Hortonworks is a massive contributor to the open-source Hadoop community focused on evolving it into a broadly capable data-management platform. Everything in the Hortonworks Data Platform (HDP) is freely available as open-source software. Hortonworks distribution has master-slave architecture and it can support among others MapReduce, YARN, Spark, Kafka, Flink and HBase. Furthermore, Hortonworks has no proprietary software, uses Ambari for management and Stinger for handling queries, and Apache Solr for searches of data.

Hortonworks's Benefits

- Completely Open: HDP is the only completely open Hadoop data platform available. All solutions in HDP are developed as projects through the Apache Software Foundation (ASF). There are no proprietary extensions or add-ons required.
- Fundamentally Versatile: At its heart HDP offers linear scale storage and compute across a wide range of access methods from batch to interactive, to real time, search and streaming. It includes a comprehensive set of capabilities across governance, integration, security and operations.
- Wholly Integrated: HDP integrates with and augments your existing applications and systems so that you can take advantage of Hadoop with only minimal change to existing data architectures and skillsets. Deploy HDP in-cloud, on-premise or from an appliance across both Linux and Windows.
- Hortonworks provides you with the flexibility to run the same industry-leading, open source platform to gain data insights in the data center as well as on the public cloud of choice.

Hortonworks's Disadvantages

- Installation can be complex and needs to be streamlined
- There exist minor stability issues with the platform. As a remedy for that someone may choose not to follow the latest releases to make sure more stable versions are used.
- Upgrading from lower versions is at the moment feasible but demands effort.
- There is room for improvement in monitoring. The Ambari Management interface on HDP is just a basic one and does not have many rich features.

MapR - The MapR Distribution (MapR) including Apache Hadoop provides an enterprise-grade distributed data platform that can reliably store and process big and fast data. MapR Distribution gives a good foundation for running batch, interactive, and real-time applications. With an open choice approach to open source, MapR gives you a broad range of technologies (multiple projects for SQL-on-Hadoop, NoSQL databases, execution engines such as Spark, etc) to choose from, so that the right tool is employed for a specific need. Furthermore, MapR M7 Hadoop distribution addresses weakness in HBase by doing away with region servers, table splits and merges, and data-compaction steps. MapR has also implemented its own architecture for snapshotting, high availability, and system recovery. With M7, MapR also introduced optional LucidWorks Search software on top of Hadoop for building out recommendation engines, fraud-detection, and predictive applications.

The three platform services offered (MapR-FS, MapR-DB, and MapR Streams), are unified by common core capabilities built into the underlying platform such as high availability, real-time access, unified security, multi-tenancy, disaster recovery, a global namespace, self-healing, and management and monitoring. The MapR Converged Data Platform allows you to quickly and easily build breakthrough, reliable, real-time applications by providing:

- Single cluster for streams, file storage, database, and analytics.
- Persistence of streaming data, providing direct data access to batch and interactive frameworks, eliminating data movement.
- Unified security framework for data-in-motion and data-at-rest, with authentication, authorization, and encryption.
- Utility-grade reliability with self-healing and no single point-of-failure architecture.

MapR's Benefits

- Unified Big Data Platform: Capable of creating a complete picture of all data, including high-velocity, real-time data, to find previously unidentifiable insights. Process more data types with the schema-less flexibility and the high-velocity read/write capabilities of the integrated in-Hadoop online database platform.

- Proven Production Readiness: Ability to get continuous value from your data with the technology proven in production to meet strict service level agreements. Deploy 24x7 online applications with enterprise-grade capabilities to achieve zero downtime. Run HBase-compatible applications with zero database administration.
- Consistent High Performance at Any Scale: Ability to get faster results on larger data sets to respond more quickly to more complete data. Achieve quicker application responsiveness for an enhanced user experience. Easily load and process high volumes and high velocities of incoming data.

Apache Ambari - Ambari (Apache Ambari) is a completely open source management platform for provisioning, managing, monitoring and securing Apache Hadoop clusters. Apache Ambari can help in taking the guesswork out of operating Hadoop. Apache Ambari, as part of the Hortonworks Data Platform, allows enterprises to plan, install and securely configure HDP making it easier to provide ongoing cluster maintenance and management, irrespective to the size of the cluster. Ambari makes Hadoop management simpler by providing a consistent, secure platform for operational control with an intuitive Web UI as well as a robust REST API, which is particularly useful for automating cluster operations. The tools allow Hadoop operators get the following core benefits:

- Simplified Installation, Configuration and Management
- Centralized Security Setup.
- Full Visibility into Cluster Health
- Highly Extensible and Customizable

3.2 Big Data File / Storage Systems

In general, Data stores are grouped according to their data model, i.e. SQL vs. NoSQL (Cattell et al. 2011):

- Key-value Stores: These systems store values and an index to find them, based on a programmer defined key.
- Document Stores: These systems store documents, as just defined. The documents are indexed, and a simple query mechanism is provided.
- Extensible Record Stores: These systems store extensible records that can be partitioned vertically and horizontally across nodes. Some papers call these “wide column stores”.
- Relational Databases: These systems store (and index and query) tuples.

Key Value Stores: The simplest data stores use a data model similar to the popular memcached distributed in-memory cache, with a single key-value index for all the data. These systems are called key-value stores. Unlike memcached, these systems generally provide a persistence mechanism and additional functionality as well: replication, versioning, locking, transactions, sorting, and/or other features. The client interface provides inserts, deletes, and index lookups. Like memcached, none of these systems offer secondary indices or keys. Some noticeable Key Value Stores include:

- Project Voldermort
- Riak
- Redis
- Tokyo Cabinet
- Document Stores

Document stores support more complex data than the key-value stores. Although termed “document store” and these systems could store “documents” in the traditional sense (articles, Microsoft Word files, etc.), a document in these systems can be any kind of “pointerless object”. Unlike the key-value stores, these systems generally support secondary indexes and multiple types of documents (objects)

per database, and nested documents or lists. Like other NoSQL systems, the document stores do not provide ACID transactional properties. Here are some of the Document Stores:

- SimpleDB
- CouchDB
- MongoDB
- Terrastore

Extensible Record Stores: The extensible record stores seem to have been motivated by Google's success with BigTable. Their basic data model is rows and columns, and their basic scalability model is splitting both rows and columns over multiple nodes:

- Rows are split across nodes through sharding on the primary key. They typically split by range rather than a hash function. This means that queries on ranges of values do not have to go to every node.
- Columns of a table are distributed over multiple nodes by using "column groups". These may seem like a new complexity, but column groups are simply a way for the customer to indicate which columns are best stored together.

Extensible Record Stores include:

- Hadoop Distributed File System(HDFS)
- HBase
- Cassandra

Hadoop Distributed File System - The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets.

Hardware failure is the norm rather than the exception. An HDFS instance may consist of hundreds or thousands of server machines, each storing part of the file system's data. The fact that there are a huge number of components and that each component has a non-trivial probability of failure means that some component of HDFS is always non-functional. Therefore, detection of faults and quick, automatic recovery from them is a core architectural goal of HDFS.

HDFS has been designed to be easily portable from one platform to another. This facilitates widespread adoption of HDFS as a platform of choice for a large set of applications. An HDFS has a master/slave architecture and an HDFS cluster consists of a single NameNode, a master server that manages the file system namespace and regulates access to files by clients. In addition, there are several DataNodes, usually one per node in the cluster, which manage storage attached to the nodes that they run on. HDFS exposes a file system namespace and allows user data to be stored in files. Internally, a file is split into one or more blocks and these blocks are stored in a set of DataNodes. The NameNode executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes. The DataNodes are responsible for serving read and write requests from the file system's clients. The DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode.

HDFS is part of the Apache Hadoop Core project. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

HBase - Apache HBase (Jiang 2012) is the Hadoop database, is a type of NoSQL database, a distributed, scalable, big data store. When there is a need for random, real-time read/write access to your Big Data, Apache HBase is befitting. This project's goal is the hosting of very large tables (billions of rows X millions of columns) using clusters of commodity hardware. Apache HBase is an open-source, distributed, versioned, non-relational database modeled. Some key features include the following:

- Linear and modular scalability.
- Strictly consistent reads and writes.
- Automatic and configurable sharding of tables
- Automatic failover support between RegionServers.
- Convenient base classes for backing Hadoop MapReduce jobs with Apache HBase tables.
- Easy to use Java API for client access.
- Block cache and Bloom Filters for real-time queries.
- Query predicate push down via server side Filters
- Thrift gateway and a REST-ful Web service that supports XML, Protobuf, and binary data encoding options
- Extensible jruby-based (JIRB) shell
- Support for exporting metrics via the Hadoop metrics subsystem to files or Ganglia; or via JMX

Cassandra - Apache Cassandra (Rabl et al. 2012) is a free and open-source distributed NoSQL database management system designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure. Cassandra offers robust support for clusters spanning multiple datacenters, with asynchronous masterless replication allowing low latency operations for all clients.

Data is automatically replicated to multiple nodes for fault-tolerance. Replication across multiple data centers is supported. Failed nodes can be replaced with no downtime. There are no network bottlenecks. Every node in the cluster is identical. Every node in the cluster has the same role. There is no single point of failure. Data is distributed across the cluster (so each node contains different data), but there is no master as every node can service any request. Read and write throughput both increase linearly as new machines are added, with no downtime or interruption to applications.

Cassandra is not row level consistent, meaning that inserts and updates into the table that affect the same row that are processed at approximately the same time may affect the non-key columns in inconsistent ways. One update may affect one column while another affects the other, resulting in sets of values within the row that were never specified or intended.

Relational Databases

Unlike the other data stores, relational DBMSs have a complete pre-defined schema, a SQL interface, and ACID transactions. Traditionally, RDBMSs have not achieved the scalability of the some of the previously described data stores. It appears likely that some relational DBMSs will provide scalability comparable with NoSQL data stores, with two provisions:

- Use small-scope operations: As we've noted, operations that span many nodes, e.g. joins over many tables, will not scale well with sharding.
- Use small-scope transactions: Likewise, transactions that span many nodes are going to be very inefficient, with the communication and two-phase commit overhead.

It should be noted that NoSQL systems avoid these two problems by making it difficult or impossible to perform larger scope operations and transactions.

Relational Databases are:

- MySQL Cluster
- VoltDB
- Clustric

- ScaleDB
- ScaleBase
- NimbusDB

SQL vs NoSQL: Key Differences (SQL vs. NoSQL)

- One of the key differentiator is that NoSQL supported by column-oriented databases where RDBMS is row oriented database.
- NoSQL seems to work better on both unstructured and unrelated data. The better solutions are the crossover databases that have elements of both NoSQL and SQL.
- RDBMSs that use SQL are schema–oriented which means the structure of the data should be known in advance to ensure that the data adheres to the schema. For example, predefined schema-based applications that use SQL include Payroll Management System, Order Processing and Flight Reservations.
- SQL Databases are vertically scalable – this means that they can only be scaled by enhancing the horse power of the implementation hardware, thereby making it a costly deal for processing large batches of data.
- NoSQL databases give up some features of the traditional databases for speed and horizontal scalability. NoSQL databases on the other hand are perceived to be cheaper, faster and safer to extend a preexisting program to do a new job than to implement something from scratch.
- More importantly, data Integrity is a key feature of SQL based databases. This means, ensuring the data is validated across all the tables and there's no duplicate, unrelated or unauthorized data inserted in the system.
- Advantages of SQL databases are that they are typically more performant when dealing with more complex queries. Users cite the relational nature of SQL DBs encourages a well-structured database.

3.3 Big data Batch processing

Apache Flink - Apache Flink (Apache Flink) is an open source stream processing framework developed by the Apache Software Foundation. The core of Apache Flink is a distributed streaming dataflow engine written in Java and Scala. Flink executes arbitrary dataflow programs in a data-parallel and pipelined manner. With its pipelined runtime system, it enables the execution of bulk/batch and stream processing programs. Furthermore, Flink's runtime supports the execution of iterative algorithms natively.

Flink provides a high-throughput, low-latency streaming engine as well as support for event-time processing and state management. Flink applications are fault-tolerant in the event of machine failure and support exactly-once semantics. Programs can be written in Java, Scala, Python, and SQL and are automatically compiled and optimized into dataflow programs that are executed in a cluster or cloud environment. It is worth mentioning that Flink does not provide its own data storage system and provides data source and sink connectors to systems such as Amazon Kinesis, Apache Kafka, HDFS, Apache Cassandra, and ElasticSearch.

In general, Flink:

- provides results that are accurate, even in the case of out-of-order or late-arriving data;
- is stateful and fault-tolerant and can seamlessly recover from failures while maintaining exactly-once application state;
- performs at large scale, running on thousands of nodes with very good throughput and latency characteristics.

Apache Hadoop - Apache Hadoop (Hadoop) is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data

and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model. The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS) and a processing part which is a MapReduce programming model. Hadoop splits files into large blocks and distributes them across nodes in a cluster. It then transfers packaged code into nodes to process the data in parallel. This approach takes advantage of data locality, where nodes manipulate the data they have access to. This allows the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking.

Apache Spark - Apache Spark (Spark) has as its architectural foundation the resilient distributed dataset (RDD), a read-only multiset of data items distributed over a cluster of machines, that is maintained in a fault-tolerant way. RDD is a fundamental data structure of Spark. This processing tool is a programming abstraction that represents an immutable collection of objects that can be split across a computing cluster. Operations on the RDDs can also be split across the cluster and executed in a parallel batch process, leading to fast and scalable parallel processing. RDDs can be created from simple text files, SQL databases, NoSQL stores (such as Cassandra and MongoDB) among others. Much of the Spark Core API is built on this RDD concept, enabling traditional map and reduce functionality, but also providing built-in support for joining data sets, filtering, sampling, and aggregation.

Hadoop MapReduce - Hadoop MapReduce (Dean and Ghemawat, 2004) is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. A MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically, both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks.

Typically, the compute nodes and the storage nodes are the same, that is, the MapReduce framework and the Hadoop Distributed File System are running on the same set of nodes. This configuration allows the framework to effectively schedule tasks on the nodes where data is already present, resulting in very high aggregate bandwidth across the cluster. The MapReduce framework consists of a single master JobTracker and one slave TaskTracker per cluster-node. The master is responsible for scheduling the jobs' component tasks on the slaves, monitoring them and re-executing the failed tasks. The slaves execute the tasks as directed by the master.

At their minimal, applications specify the input / output locations and supply map and reduce functions via implementations of appropriate interfaces and/or abstract-classes. These, and other job parameters, comprise the job configuration. The Hadoop job client then submits the job (jar/executable etc.) and configuration to the JobTracker which then assumes the responsibility of distributing the software/configuration to the slaves, scheduling tasks and monitoring them, providing status and diagnostic information to the job-client. Although the Hadoop framework is implemented in Java, MapReduce applications need not be written in Java.

3.4 Big data Stream processing

Spark Streaming - Spark Streaming (Spark Streaming) makes it easy to build scalable fault-tolerant streaming applications. Spark Streaming brings Apache Spark's language-integrated API to stream processing, letting you write streaming jobs the same way you write batch jobs. It supports Java, Scala and Python. Stateful exactly-once semantics out of the box. Combine streaming with batch and interactive queries.

Spark Streaming extended the Apache Spark concept of batch processing into streaming by breaking the stream down into a continuous series of microbatches, which could then be manipulated using the Apache Spark API. In this way, code in batch and streaming operations can share (mostly) the same code, running on the same framework, thus reducing both developer and operator overhead. A criticism of the Spark Streaming approach is that microbatching, in scenarios where a low-latency response to incoming data is required, may not be able to match the performance of other streaming-capable frameworks like Apache Storm, Apache Flink, and Apache Apex, all of which use a pure streaming method rather than microbatches.

Instead of processing the streaming data one record at a time, Spark Streaming discretizes the streaming data into tiny, sub-second micro-batches. In other words, Spark Streaming's Receivers accept data in parallel and buffer it in the memory of Spark's workers nodes. Then the latency-optimized Spark engine runs short tasks (tens of milliseconds) to process the batches and output the results to other systems. Note that unlike the traditional continuous operator model, where the computation is statically allocated to a node, Spark tasks are assigned dynamically to the workers based on the locality of the data and available resources. This enables both better load balancing and faster fault recovery.

Kafka Streaming - Kafka Streams (Kafka Stream) is a client library for building applications and microservices, where the input and output data are stored in a Kafka cluster. It combines the simplicity of writing and deploying standard Java and Scala applications on the client side with the benefits of Kafka's server-side cluster technology.

Some benefits of Kafka include:

- Elastic, highly scalable, fault-tolerant
- Deploy to containers, VMs, bare metal, cloud
- Equally viable for small, medium, & large use cases
- Fully integrated with Kafka security
- Write standard Java applications
- Exactly-once processing semantics
- No separate processing cluster required

Apache Storm - Apache Storm (Apache Storm, 2016) is a free and open source distributed real-time computation system. Storm makes it easy to reliably process unbounded streams of data, doing for real-time processing what Hadoop did for batch processing. Storm is simple, can be used with any programming language. Storm has many use cases: real-time analytics, online machine learning, continuous computation, distributed RPC, ETL, and more. Storm is fast: a benchmark clocked it at over a million tuples processed per second per node. It is scalable, fault-tolerant, guarantees your data will be processed, and is easy to set up and operate.

Storm integrates with the queueing and database technologies you already use. A Storm topology consumes streams of data and processes those streams in arbitrarily complex ways, repartitioning the streams between each stage of the computation however needed.

3.5 Connectors

Kafka Connect - Kafka Connect (Kafka Connect) is a framework for scalably and reliably streaming data between Apache Kafka and other data systems. Connect makes it simple to use existing connector implementations for common data sources and sinks to move data into and out of Kafka. Kafka Connect's applications are wide ranging. A source connector can ingest entire databases and stream table updates to Kafka topics or even collect metrics from all of your application servers into Kafka topics, making the data available for stream processing with low latency. A sink connector can deliver data from Kafka topics into secondary indexes like Elasticsearch or into batch systems such as Hadoop for offline analysis.

Kafka Connect is focused on streaming data to and from Kafka. This focus makes it much simpler for developers to write high quality, reliable, and high-performance connector plugins and makes it possible for the framework to make guarantees that are difficult to achieve in other frameworks.

The main benefits of using Kafka Connect are:

- Data Centric Pipeline – use meaningful data abstractions to pull or push data to Kafka.
- Flexibility and Scalability – run with streaming and batch-oriented systems on a single node or scaled to an organization-wide service.
- Reusability and Extensibility – leverage existing connectors or extend them to tailor to your needs and lower time to production.

Spark Connectors - Both Spark and HBase are widely used, but how to use them together with high performance and simplicity is a very challenging topic. Spark HBase Connector (SHC) provides feature rich and efficient access to HBase through Spark SQL. It bridges the gap between the simple HBase key value store and complex relational SQL queries and enables users to perform complex data analytics on top of HBase using Spark. SHC implements the standard Spark data source APIs, and leverages the Spark catalyst engine for query optimization. To achieve high performance, SHC constructs the RDD from scratch instead of using the standard HadoopRDD. With the customized RDD, all critical techniques can be applied and fully implemented, such as partition pruning, column pruning, predicate pushdown and data locality. The design makes the maintenance easy, while achieving a good tradeoff between performance and simplicity.

Furthermore, Spark is collaborating with MongoDB, Cassandra, Sorl, Elasticseaerch etc. in creating connectors between them (Spark packages). There is also the ability to define custom and purpose-built connectors depending on the needs (Connector Devel.). The community offers a big selection of implementations that work with specific technologies and tools. There also exists a Connect Developer Guide so as to create connectors if necessary although in general it is preferable to use already existing connectors and it is a solution when something specific is required.

3.6 Industry-related Benchmarks

In this section, some applications of big data technologies and solutions are presented by looking into specific industries which have adopted some big data approaches in order to achieve their goals. In the following paragraphs the Nissan motor company, Octo and Novartis are highlighted:

Nissan Motor Company Ltd (Nissan): Connected cars that leverage driving data are a vision that automobile manufacturers are aggressively pursuing. Nissan Motor Company Ltd (Nissan) is on this journey also. The company was experiencing a variety of business challenges, namely the need for infrastructure capable of storing huge volumes of vehicle driving data and product quality data on a long-term basis. Also, it had a need for a Hadoop platform capable of deploying a variety of data cross-functionally. Nissan turned to Hadoop as the solution for its Big Data problem, relying on Hortonworks Data Platform (HDP). The open source model appealed to the company due to the large numbers of engineering talent in the market and the flexibility to pivot if circumstances changed down the road.

OCTO (Octo): Octo Telematics is provider of telematics and data analytics solutions for the auto insurance industry. By collecting and analyzing data from connected cars, Octo Telematics gives insurers insights to more effectively assess driver risk, deliver crash and claim services, and manage customer relationships. “We utilize every type of data—contextual data, driving data, behavioral data, and crash data—to forecast driving habits, improve crash notifications and response, evaluate crash dynamics, and detect fraud,” said Gianfranco Giannella, COO, Octo Telematics. As Octo Telematics grew, executives sought to replace a custom-made data platform with a more scalable, next generation data management platform. “We wanted to rapidly expand the footprint of our services,” said Giannella. “We needed a platform that would support a growing volume of telematics and IoT data and enable us to prototype services and products much faster.” Octo Telematics today powers its telematics Internet of Things (IoT) solution with Cloudera Enterprise. The platform stores,

processes, and analyzes data on more than 170 billion miles of driving and approximately 400,000 severe crashes from five million connected cars. In all, Octo adds over 11 billion new data points from connected cars daily to the platform. Internal and external data sources, such as traffic and weather data, are also incorporated to provide additional context. Using machine learning, the company can make more accurate predictions and risk models.

NOVARTIS (Novartis): The MapR-based flexible workflow tool is now being used for a variety of different projects across Novartis, including video analysis, proteomics, and meta-genomics. The combined Spark and MapR-based workflow and integration layers allow the company's life science researchers to meaningfully take advantage of the tens of thousands of experiments that public organizations have conducted, which gives them a significant competitive advantage**.** One of their areas of drug research, Next Generation Sequencing (NGS) data, requires heavy interaction with diverse data from external organizations such as 1000 Genomes, NIH's GTEx (Genotype-Tissue Expression) and The Cancer Genome Atlas—paying particular attention to clinical, phenotypical, experimental and other associated data. Integrating these heterogeneous datasets is labor intensive, so they only want to do it once. To solve the first part of this NGS big data problem, the Novartis team built a workflow system that allows them to process NGS data while being responsive to advances in the scientific literature. Although NGS data requires high data volumes that are ideal for Hadoop, a common problem is that researchers rely on many tools that simply don't work on native HDFS. Since these researchers previously couldn't use systems like Hadoop, they have had to maintain complicated 'bookkeeping' logic to parallelize for optimum efficiency on traditional High Performance Computing (HPC). This workflow system uses the MapR Distribution for Hadoop for its performance and robustness and to provide the POSIX file access that lets bioinformaticians use their familiar tools. Additionally, it uses the researchers' own metadata to allow them to write complex workflows that blend the best aspects of Hadoop and traditional HPC.

3.7 References

Apache Ambari. URL: https://hortonworks.com/apache/ambari/

Apache Cassandra. URL: http://cassandra.apache.org/

Apache Flink. URL: https://flink.apache.org/

Apache Flink features. URL: http://flink.apache.org/features.html

Apache Hadoop. URL: http://hadoop.apache.org

Apache Hadoop HDFS. URL: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

Apache Hadoop MapReduce. URL: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

Apache HBase. URL: https://hbase.apache.org/

Apache Spark. URL: https://spark.apache.org/releases/spark-release-2-0-0.html

Apache Spark packages. URL: https://spark-packages.org/?q=tags%3A%22Data%20Sources%22

Apache Spark Streaming. URL: https://spark.apache.org/streaming/

Apache Storm. URL: http://storm.apache.org/

Cattell, R. (2011) Scalable SQL and NoSQL data stores. SIGMOD Rec. 39, 4, 12-27.

Chintaballi, S., Dagit, D., Evans, B., et al. (2016) Benchmarking Streaming Computation Engines: Storm, Flink and Spark Streaming. IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), Chicago, IL, pp. 1789-1792. IEEE. DOI: 10.1109/IPDPSW.2016.138.

Cloudera CDH. URL: https://www.cloudera.com/products/open-source/apache-hadoop/key-cdh-components.html

Connector Devel. URL: https://docs.confluent.io/current/connect/devguide.html

Das, T., M. Zaharia, and P. Wendell (2015) Diving into Apache Spark Streaming's Execution Model. Databricks Engineering Blog. URL: https://databricks.com/blog/2015/07/30/diving-into-apache-spark-streamings-execution-model.html

Dean, J., and Ghemawat, S. (2008) MapReduce: Simplified Data Processing on Large Clusters. Commun. ACM 51, 1, 107-113. DOI: https://doi.org/10.1145/1327452.1327492

Hortonwork HDP. URL: https://hortonworks.com/products/data-platforms/hdp/

Jiang, Y. (2012). HBase Administration Cookbook. ISBN: 9781849517140. Packt Publishing Ltd.

Kafka Connect. URL: https://docs.confluent.io/current/connect/intro.html

Kafka Stream. URL: https://docs.confluent.io/current/streams/index.html

MapR Distr. URL: https://mapr.com/products/mapr-distribution-including-apache-hadoop/

Nissan. URL: https://hortonworks.com/customers/nissan/

Novartis. URL: https://mapr.com/resources/novartis-relies-mapr-flexible-big-data-solutions-drug-discovery/

Octo. URL: https://www.cloudera.com/more/customers/octo-telematics.html

Pointer, I. (2017) What is Apache Spark? The big data analytics platform explained. InfoWorld Analytics blog. URL : https://www.infoworld.com/article/3236869/analytics/what-is-apache-spark-the-big-data-analytics-platform-explained.html

Rabl, T., Sadoghi, M., Jacobsen, H.-A., Gomez-Villamor, S., Muntes-Mulero, V., Mankovskii, S. (2012) In Proceedings of VLDB, Istanbul, Turkey.

Wodehouse, C. (2016) SQL vs. NoSQL Databases: What's the Difference? Upwork Hiring Headquarters blog. URL: https://www.upwork.com/hiring/data/sql-vs-nosql-databases-whats-the-difference/

Yang, W. and M. Tang (2017) Apache Spark—Apache HBase Connector: Feature Rich and Efficient Access to HBase through Spark SQL. URL: https://databricks.com/session/apache-spark-apache-hbase-connector-feature-rich-and-efficient-access-to-hbase-through-spark-sql

4 Big Data Management and Processing

Big data management raises numerous research challenges (Jagadish et al. 2014) in different phases of the big data processing and analysis pipeline, including: (a) data acquisition, (b) information extraction and cleaning, (c) data integration, aggregation, and representation, (d) modeling and analysis, and (e) interpretation. The modern trend for scalable storage of massive datasets is by means of a NoSQL store (Catell, 2010) (Davoudian et al., 2018). The exact choice depends on numerous parameters, including the type of data, the data access patterns, the purpose of data processing (read/write, read-only, etc.), as well as any special requirements with respect to the consistency, availability, and partition-tolerance (also known as CAP). Also, the current landscape of big data management comprises multiple frameworks targeting different aspects of big data. One major separating line is drawn between frameworks for batch and real-time processing, although lately some systems have been designed to tackle both cases. In the batch processing domain, Spark (Zaharia et al. 2016) is one of the most popular solutions nowadays with a large and growing user-base, however other solutions, such as Flink (Carbone et al. 2015), are also applicable with success. In particular, Spark has successfully addressed many of the limitations (Doulkeridis & Nørvåg, 2014) of Hadoop, and operates in main-memory by its core abstraction: RDDs (Resilient Distributed Datasets) (Zaharia et al. 2012). In the real-time processing domain, the most notable systems in use today are Storm (Toshniwal et al. 2014) and Flink (Carbone et al. 2015).

This section provides an *overview of the state-of-the-art in big data storage and processing*, focusing primarily on **scalable solutions for mobility data**, i.e., spatial but most importantly spatio-temporal data, which is the core topic of Track&Know. Despite the rich literature on management of spatio-temporal and mobility data, only a limited number of research prototypes attempt to address this problem in the context of Big Data, and most of them focus solely on big spatial data (Eldawy & Mokbel, 2016) (Hagedorn et al. 2017), rather than spatio-temporal data. In fact, the majority of developed prototypes extend Hadoop or Spark in order to be applicable for spatial data.

4.1 Big Data Storage and Indexing

Systems that extend scalable storage solutions for multidimensional data have been proposed, most notably MD-HBase (Nishimura et al. 2011), but also solutions tailored specifically for spatio-temporal data, such as Pyro (Li et al. 2015), as well as spatio-textual data, such as ST-HBase (Ma et al. 2013). In all these storage systems the main underlying challenge is to *map* spatial or spatio-temporal data (2D or 3D) to 1-dimensional values, which are used as keys for storage in key-value based NoSQL storage systems. The mapping is typically achieved using variants of space-filling curves, such as Z-order, Hilbert, or Moore encoding.

Essentially, this mapping is necessary in order to bridge the gap between mobility data and (1-dimensional) key-based NoSQL stores. Based on this *mapping* to keys, data is distributed, replicated and stored based on partitioning techniques that operate at the level of 1-dimensional key. The challenge is then to translate spatial and spatio-temporal queries to multiple 1-dimensional range scans and discover efficient and scalable processing algorithms.

As several big data storage systems also include a processing engine, they are reviewed in the following subsection on Big Data Processing. Instead, in this subsection, we mainly focus on design choices for the storage layer.

MD-HBase: MD-HBase (Nishimura et al. 2011) encodes multidimensional data in 1-dimensional values using Z-order encoding. This 1-dimensional representation is then used by an *index layer* as a key for storing data in HBase (the *storage layer*). In this way, standard multidimensional index structures, such as k-d trees and Quad trees, can be implemented on top of a distributed key-value store. By using the properties of a technique called longest common prefix naming scheme, this mapping of

multidimensional indexes to 1-dimensional ranges is achieved, offering, in turn, the fundamental mechanism for answering point, range, and nearest-neighbor queries.

Pyro: Pyro (Li et al. 2015) employs the Moore encoding algorithm, inspired from the Moore space-filling curve, in order to transform (map) spatio-temporal data to 1-dimensional values. Then, range queries are translated to multiple 1-dimensional range scans, which are processed efficiently by means of different optimizations introduced at the storage layer of HDFS, resulting in PyroDFS, and at an extension of HBase, named PyroDB (Figure 3). In addition, a multi-scan optimizer is used to find the best reading strategy from HBase while considering multiple range scans. Also, a new block grouping algorithm is introduced at the level of the Distributed File System, which preserves data locality and improves the efficiency of dynamic load rebalancing. Pyro is shown to outperform MD-HBase by one order of magnitude for rectangular range queries.

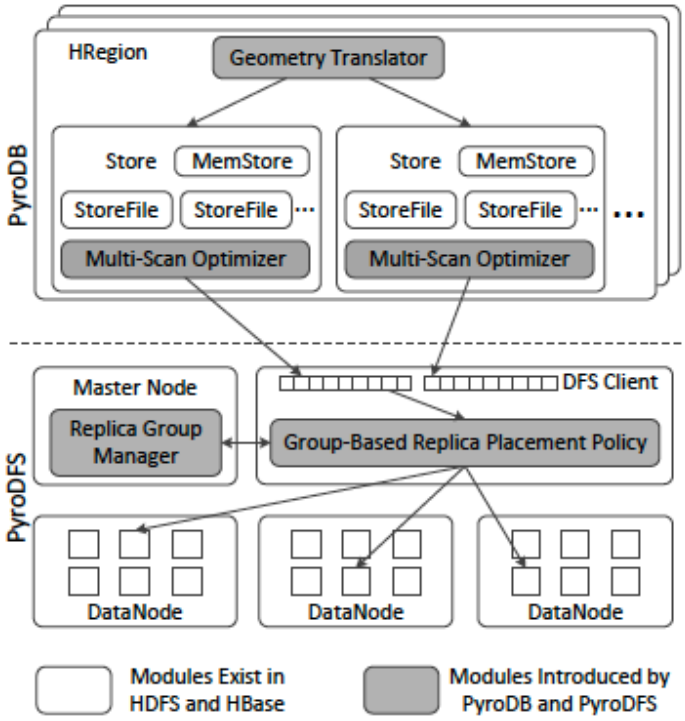


Figure 3: The architecture of Pyro (Li et al. 2015)

ST-HBase: ST-HBase (Ma et al. 2013) focuses on spatio-textual data, namely data that combines spatial location with textual description. Typical examples of spatio-textual data include geo-tagged objects, for instance tweets, images, etc. ST-HBase resembles the approach followed by MD-HBase, since it also exploits Z-order to transform spatial data to 1-dimensional values. However, it goes one step further to support combined spatial and textual retrieval, by introducing the functionality of an inverted index and representing keywords along with 1-dimensional values as key in HBase. In this way, textual filtering is supported together with spatial filtering.

4.2 Big Data Processing

Lately, several research projects have extended popular parallel data processing platforms. such as Hadoop or Spark, in order to provide customized solutions for big spatial or spatio-temporal data. The most prominent prototypes and systems in this field include Hadoop GIS (Aji et al. 2013), Parallel SECONDO (Lu & Güting, 2013), SpatialHadoop (Eldawy & Mokbel, 2016), AQWA (Aly et al. 2015), ST-Hadoop (Alarabi et al. 2017) (Alarabi & Mokbel, 2017), SpatialSpark (You et al. 2015), GeoSpark (Yu et al. 2016), LocationSpark (Tang et al. 2016), Simba (Xie et al. 2016a) (Xie et al. 2016b), STARK (Hagedorn et al. 2017a), which are reviewed in the following subsections. We also refer to (Hagedorn et al. 2017b)

for a comparative evaluation of big spatial data processing systems. In addition, a brief overview of recent systems built for parallel processing of big trajectory data is presented as a separate subsection.

Table 1: Overview of spatial and spatio-temporal parallel processing frameworks.

	Framework	Partitioning	Indexing	Queries
Spatial	**Hadoop GIS**	N/A	Global/local indexing (global region indexes, on demand local indexing)	Range queries (box), spatial joins
	Parallel Secondo	N/A	Local indexing using full-featured Secondo DBMS	All those offered by Secondo
	SpatialHadoop	Space partitioning (grid, quad tree), data partitioning (STR, STR+, k-d tree), space-filling curves (Z-order, Hilbert)	Global/local indexing (R-trees, grid files)	Range queries (box), kNN queries, spatial join
	AQWA	Adaptive (based on k-d tree)	N/A	Range queries, KNN queries
	SpatialSpark	Fixed grid partitioning, binary space partitioning, tile partitioning	Pre-built local indexes on HDFS	Range queries, spatial join
	GeoSpark	Grid-based partitioning	Local indexes (R-tree and quad tree)	Range queries, KNN query, spatial join,
	LocationSpark	Data partitioning e.g. using quad-tree (based on sampling)	Global/local indexing (global: grid and region quad-tree, local: grid, R-tree, quad-tree, IR-tree)	Range queries, KNN query, spatial join, KNN join, spatio-textual queries
	Simba	STRPartitioner (sampling and STR)	IndexRDD	Range queries, KNN query, distance join, kNN join
Spatio-temporal	**ST-Hadoop**	Multi-level temporal partitioning	Temporal hierarchy of spatial indexes at multiple levels of temporal resolution	Spatio-temporal range queries and joins
	STARK	Spatial-only	R-trees	Spatio-temporal range queries and joins

	STJoins@ESRI	Data (re-)partitioning based on quad-tree decomposition	Equi-sized splitting of complete data set and local quad trees	Spatio-temporal join
Trajectory	**UlTraMan**	Supports a repartition operator to support different partitioning strategies (including STR)	In-memory: random access RDD using on-heap arrays or using ChronicleMap, an embedded, key-value store	Range query, KNN, aggregation, co-movement pattern queries
	DITA	Grouping of trajectories based on first and last point, and use of STR for partitioning	Global/local indexing: (global: two R-trees built on MBR of first and last points respectively, local: trie-like index on selected points)	Similarity search, similarity join

4.2.1 Spatial and Spatio-temporal Frameworks

Hadoop GIS (Aji et al. 2013) is a large-scale spatial data warehousing system for executing spatial queries in parallel. It is available both as a library and as an integrated package in Hive, thus facilitating ease of use. To support indexing, global indexes are built and replicated on all nodes using Hadoop's Distributed Cache. Thus, each node can efficiently determine the regions of the space that contain relevant results for the spatial query at hand. Local indexes are dynamically constructed on demand, using main memory. Regarding query types, Hadoop GIS supports range queries and spatial joins.

Parallel Secondo (Lu & Güting, 2013) is a hybrid system that is built using Hadoop in order to efficiently process mobility data. It combines Hadoop with a set of single node instances of Secondo database, which has been built for mobility data management and processing. This hybrid coupling is inspired by an earlier attempt, namely HadoopDB, to couple Hadoop with relational DBMSs. Parallel Secondo offers the data types and execution language as a front-end, thus enabling users to express their parallel queries in the same way as sequential queries, while using the features of the execution language.

SpatialHadoop (Eldawy & Mokbel, 2015) is an extension of the basic Hadoop implementation developed by the University of Minnesota. It is designed for efficient processing of spatial data, and achieves this by supporting spatial indexing, a feature missing from basic Hadoop. SpatiaHadoop utilizes a two-layered spatial index which enables selective access to data by spatial operations. Implemented indexes include R-trees and Grid files. In more detail, SpatialHadoop uses a single *global index* and several *local indexes*. The global index maintains information about the data partitions across cluster nodes. The local indexes organize data stored on single nodes. Different partitioning techniques have been studied and evaluated (Eldawy et al. 2015) in the context of SpatialHadoop, including grid and quad tree as space partitioning techniques, STR, STR+ and k-d trees as data partitioning techniques, and Z-order and Hilbert curve as partitioning based on space-filling curves. Also, a spatial MapReduce language called Pigeon (Eldawy & Mokbel, 2014) is also provided as part of SpatialHadoop, thus easing the development of scalable applications that process vast-sized spatial data.

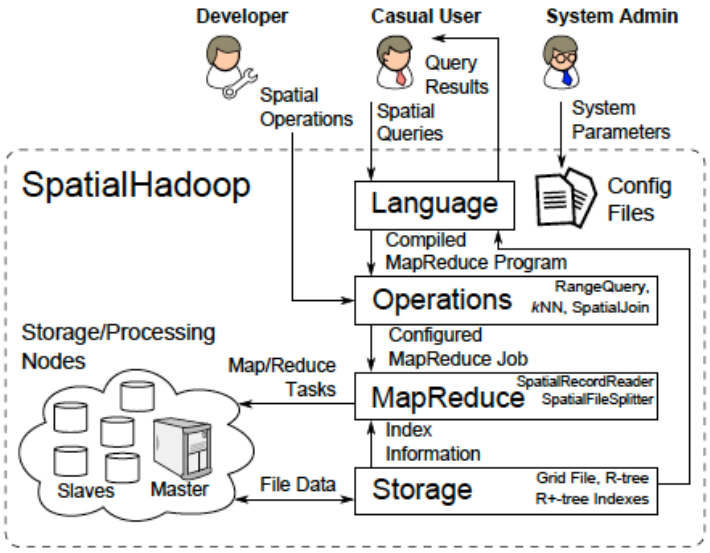


Figure 4: SpatialHadoop system architecture (Eldawy & Mokbel, 2015)

AQWA (Aly et al. 2015) is a research prototype system that focuses on adaptive partitioning for big spatial data, with a strong emphasis on query-workload-aware partitioning. In contrast to SpatialHadoop that uses static partitioning, AQWA incrementally updates the partitions based on data changes and the distribution of queries. Experiments demonstrate up to one order of magnitude performance gain compared to SpatialHadoop.

ST-Hadoop (Alarabi et al. 2017) (Alarabi & Mokbel, 2017) is an open-source MapReduce extension of Hadoop tailored for spatio-temporal data processing, also developed by the University of Minnesota. Support for spatio-temporal indexing is a core feature of ST-Hadoop. It is achieved by means of a multi-level temporal hierarchy of spatial indexes. Each level corresponds to a specific time resolution (e.g., day, month, etc.). Also, in each level the entire data set is replicated and spatio-temporally partitioned based on the temporal resolution of that particular level. ST-Hadoop supports spatio-temporal range queries, aggregations and spatio-temporal joins.

SpatialSpark (You et al. 2015) is a prototype implementation that focuses mainly of efficient processing of spatial join in parallel, although range queries are also supported. For partitioning data to machines, data partition strategies such as fixed grid or kd-tree are employed. SpatialSpark has implemented several spatial indexing and spatial filtering techniques, and it reuses (at the local level) the popular JTS (https://sourceforge.net/projects/jts-topo-suite/) for spatial refinement, i.e., testing whether two geometric objects satisfy a certain spatial relationship (e.g., point-in-polygon) or calculating a certain metric between two geometric objects (e.g., Eucledian distance).

GeoSpark (Yu et al. 2016) is a framework for processing large spatial data. Essentially, it offers a spatial layer built on top of Apache Spark, aiming at providing efficient support for spatial data processing. GeoSpark uses JTS (https://sourceforge.net/projects/jts-topo-suite/) to create and process geometries in order to support different query types: Range Queries, k-nearest neighbor (kNN), and Spatial Join. It provides a new abstraction named Spatial Resilient Distributed Datasets (SRDDs). Spatial RDDs, such as PointRDD and RectangleRDD, are used in order to effectively partition spatial data to different machines. Partitioning is achieved using a standard, uniform grid partitioning mechanism, and spatial objects that intersect more than one grid cells are duplicated to all cells. Each RDD partition can be indexed local using QuadTree and R-tree indexes. However, global indexing is not supported.

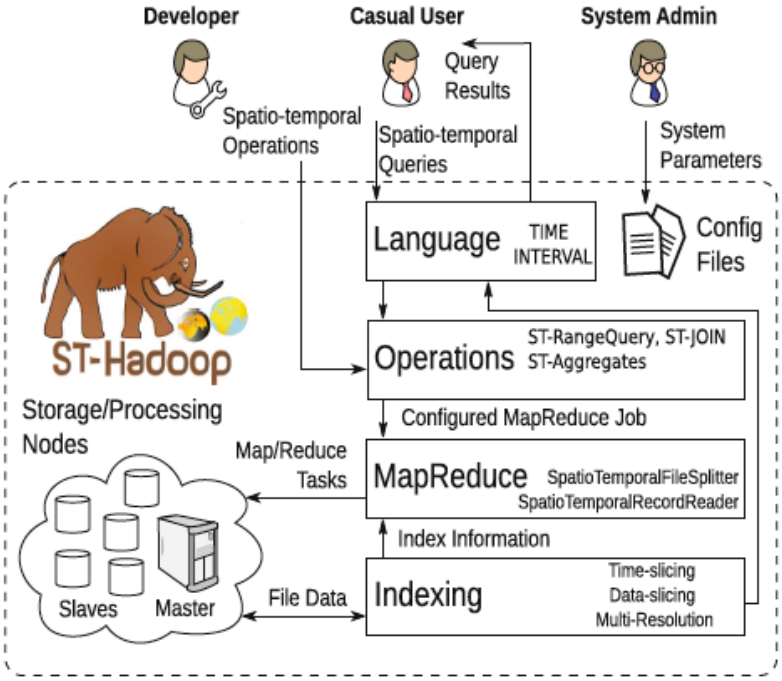


Figure 5: ST-Hadoop system architecture (Alarabi et al. 2017)

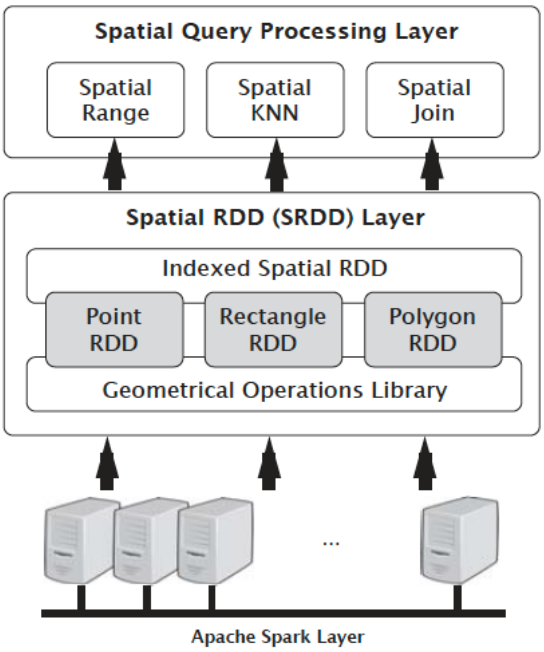


Figure 6: The architecture of GeoSpark (Yu et al. 2015)

LocationSpark (Tang et al. 2016) is a spatial data processing system developed on top of Spark that supports different spatial operators (e.g., Range, KNN, Spatial Join, KNN Join). It follows the global-local indexing approach, where a global index is used (based on sampling) to partition data to cluster nodes, while local indexes are built for each partition. Different options are implemented in terms of global and local indexes. Global indexing of data partitions is achieved by sampling the data and creating equi-sized partitions. Each partition is locally indexed on each machine using a local index of choice, including grid index, R-tree, quad-tree, or an IR-tree. In this way, data skew can be effectively

addressed. Also, the authors address query skew, by means of a query scheduler that identifies data partitions that are queried by many queries and chooses to reallocate partitions when this cost is affordable. Interestingly, processing of range queries is performed by exploiting a Spatial Bloom Filter that efficiently determines whether a point is contained in a spatial range, thus avoiding the overheads of typical cases for parallel range query processing: (a) replicating points to neighboring partitions, or (b) directing a range query to all overlapping partitions. Experiments report one order of magnitude improvement in performance compared to GeoSpark.

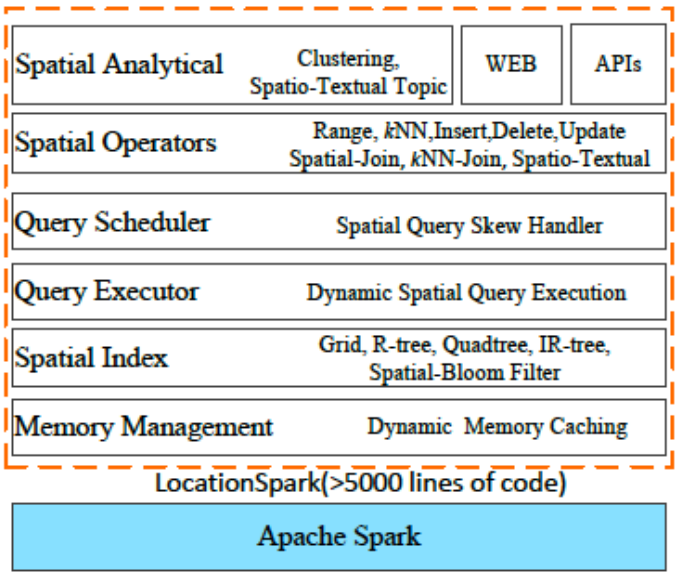


Figure 7: The architecture of LocationSpark (Tang et al. 2016)

Simba (Xie et al. 2016a; Xie et al. 2016b) is a system for in-memory spatial analytics implemented in Spark. It extends the Spark SQL engine to support spatial query processing and develops an optimizer that can exploit indexes in order to improve the performance of query processing. At a technical level, Simba introduces the concept of *IndexRDDs*, thus allowing efficient random access in large datasets in memory, thereby avoiding the limitation of linear (in-memory) scan of Spark when accessing RDDs. Simba supports a new partitioning type, named STRPartitioner, which performs random sampling on the input and then runs one iteration of the Sort-Tile-Recursive (STR) algorithm (Leutenegger et al. 1997) in order to determine the partition boundaries. The computed partition boundaries need to be extended in order to cover the space of the complete data set.

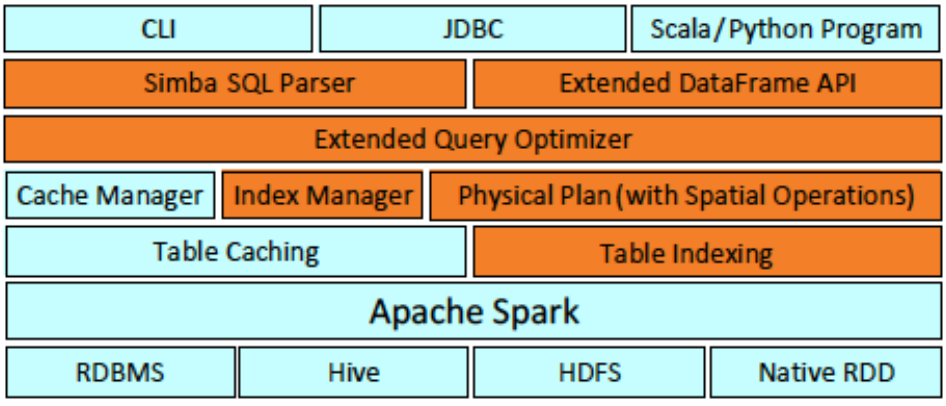


Figure 8: The architecture of Simba (Xie et al. 2016b)

In terms of query operators, Simba supports range queries, kNN, distance join, and kNN joins, and introduces new physical execution plans to Spark SQL, in order to efficiently process such spatial queries. This is a notable difference to other systems, such as GeoSpark and SpatialSpark, which are libraries implemented on top of Spark, whereas Simba introduces changes to the kernel of Spark SQL. In this way, cost-based optimization of spatial queries is also provided in Simba. Also, Simba supports multiple dimensions, in contrast to most other systems that are constrained to 2 dimensions. Simba is evaluated against SpatialHadoop and Hadoop GIS and is considerably faster, due to the in-memory processing. Also, Simba is shown to be more efficient than in-memory parallel processing systems, such as GeoSpark and SpatialSpark, because of its indexing and query optimizer which are built inside the query engine of Spark.

STARK (Hagedorn et al. 2017a) is of the few existing solutions targeting big spatio-temporal data. STARK addresses query processing of spatio-temporal data in Spark, whereas other approaches only consider the spatial dimensions. STARK supports spatio-temporal partitioning and indexing using R-trees. Thus, it supports spatio-temporal filtering and join operations. However, the temporal dimension is not treated equally to the spatial dimensions. For example, partitioning in (Hagedorn et al. 2017a) is performed solely based on spatial criteria, and the temporal part of a query is used to filter out triples that do not satisfy the temporal constraint. In essence, the temporal dimension is treated as yet another dimension that can be queried, and it cannot be used for eager pruning of data in the case of a very selective temporal constraint. In summary, STARK handles separately the temporal from the spatial dimension, thus not fully exploiting spatio-temporal correlations present at the data and performs data filtering separately for time and space.

STJoins@ESRI (Whitman et al. 2017) presents an algorithm for spatio-temporal join over large spatio-temporal data sets. It is not a complete system of a framework that supports different functionalities, rather the focus is on a specific operation. In the case that one of the inputs is relatively small and fits in memory of cluster nodes, broadcast join is employed, where the small data set is sent to all nodes, whereas the other one is partitioned to the nodes. In the more generic case where both inputs are large, a repartition join algorithm is employed, which is called bin join.

4.2.2 Trajectory Management Frameworks

Although several systems and prototypes for the management of trajectory data exist, there is limited work on the management and analysis of big trajectory data.

Only recently, **UlTraMan** (Ding et al. 2018) proposes a unified platform for the complete management cycle of big trajectory data. It provides both storage and processing layer for trajectory data. In the storage layer, ChronicleMap is used, an embedded key-value store, which is integrated in the block manager of Apache Spark. In the processing layer, UlTraMan employs an enhanced MapReduce paradigm that provides flexible APIs to applications. Interestingly, this is one of the few approaches that target the entire lifecycle of big trajectory data, from data loading and indexing, to processing and analytics. Supported query operators include range queries, KNN queries, aggregation queries. In addition, co-movement pattern mining on trajectory data is also supported, demonstrating the trajectory analytics capabilities of UlTraMan.

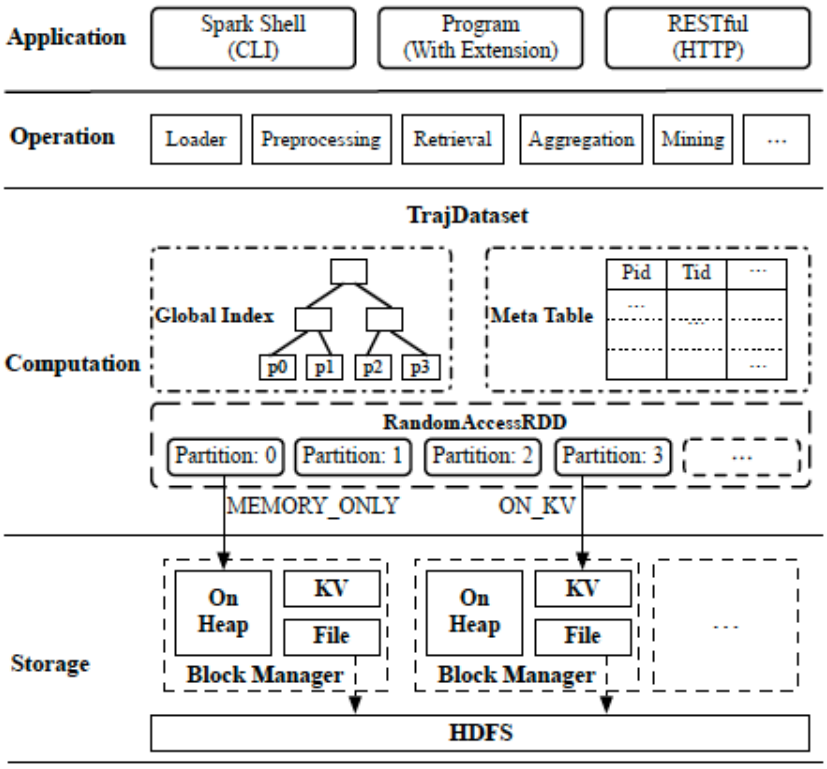


Figure 9: The architecture of UlTraMan (Ding et al. 2018)

DITA (Shang et al. 2018) is another recent research prototype that targets in-memory trajectory analytics, also extending Apache Spark. It offers an extended Spark SQL language that facilitate the declarative specification of queries, but also index construction. Furthermore, DITA extends the Catalyst optimizer of Spark SQL in order to optimize trajectory similarity queries, using cost-based optimization. At the indexing level, DITA uses local/global indexes and proposes an approximate representation technique for trajectories based on pivot points. For data partitioning the STR algorithm is used, operating on selected points of trajectories, namely the first and last points of each trajectory. The trajectories are grouped based on their first points, and then each subgroups are created by grouping based on the last points. Then, the global indexing mechanism consists of two R-trees, one constructed on the MBRs of first points and another one constructed on the MBRs of last points. The local indexing is a variant of trie-based indexing which is built on top of the pivot points of trajectories. At the algorithmic/processing level, DITA adopts the filter-and-verification paradigm, in order to efficiently process similarity search and similarity joins. It is shown that pruning can be achieved by specific conditions that can be checked on the global and local indexes.

Distributed trajectory similarity join is also investigated in (Shang et al. 2017), where a two-phase algorithm is proposed that is parallelized and computes for each trajectory other similar trajectories in its first phase. Then, during the second phase, it performs result merging in order to deliver the final result.

4.3 References

Aji, A., Wang, F., Vo, H., Lee, R., Liu, Q., Zhang, X., Saltz, J.H. (2013) Hadoop-GIS: A High Performance Spatial Data Warehousing System over MapReduce. PVLDB 6(11): 1009-1020.

Alarabi, L., Mokbel, M.F. (2017) A Demonstration of ST-Hadoop: A MapReduce Framework for Big Spatio-temporal Data. PVLDB 10(12): 1961-1964.

Alarabi, L., Mokbel, M.F., Musleh, M. (2017) ST-Hadoop: A MapReduce Framework for Spatio-Temporal Data. Proceedings of SSTD.

Aly, A.M., Mahmood, A.R., Hassan, M.S., Aref, W.G., Ouzzani, M., Elmeleegy, H., Qadah, T. (2015) AQWA: Adaptive Query-Workload-Aware Partitioning of Big Spatial Data. PVLDB 8(13): 2062-2073.

Carbone, P., Katsifodimos, A., Ewen, S., Markl, V., Haridi, S., Tzoumas, K. (2015) Apache Flink™: Stream and Batch Processing in a Single Engine. IEEE Data Eng. Bull. 38(4): 28-38.

Cattell, R. (2010) Scalable SQL and NoSQL data stores. SIGMOD Record 39(4): 12-27.

Davoudian, A., Chen, L., Liu, M. (2018) A Survey on NoSQL Stores. ACM Computing Surveys, 51(2), June 2018.

Ding, X., Chen, L., Gao, Y., Jensen, C.S., Bao, H. (2018) UlTraMan: A Unified Platform for Big Trajectory Data Management and Analytics. PVLDB (to appear).

Doulkeridis, C., Nørvåg, K. (2014) A survey of large-scale analytical query processing in MapReduce. VLDB J. 23(3): 355-380.

Eldawy A., Mokbel, M.F. (2015) SpatialHadoop: A MapReduce framework for spatial data. Proceedings of ICDE.

Eldawy, A., Alarabi, L., Mokbel, M.F. (2015) Spatial Partitioning Techniques in Spatial Hadoop. PVLDB 8(12): 1602-1605.

Eldawy, A., Mokbel, M.F. (2014) Pigeon: A spatial MapReduce language. Proceedings of ICDE.

Eldawy, A., Mokbel, M.F. (2016) The Era of Big Spatial Data: A Survey. Foundations and Trends in Databases 6(3-4): 163-273.

Hagedorn, S., Götze, P., Sattler, K-U. (2017) Big Spatial Data Processing Frameworks: Feature and Performance Evaluation. Proceedings of EDBT.

Hagedorn, S., Götze, P., Sattler, K-U. (2017) Big Spatial Data Processing Frameworks: Feature and Performance Evaluation. Proceedings of EDBT.

Hagedorn, S., Götze, P., Sattler, K-U. (2017): The STARK Framework for Spatio-Temporal Data Analytics on Spark. Proceedings of BTW.

Jagadish, H.V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R., Shahabi, C. (2014) Big data and its technical challenges. Commun. ACM 57(7), 86-94.

Leutenegger, S.T., Edgington, J.M., López, M.A. (1997) STR: A Simple and Efficient Algorithm for R-Tree Packing. Proceedings of ICDE.

Li, S., Hu, S., Ganti, R.K., Srivatsa, M., Abdelzaher, T.F. (2015) Pyro: A Spatial-Temporal Big-Data Storage System. Proceedings of USENIX Annual Technical Conference.

Lu, J., Güting, R.H. (2013) Parallel SECONDO: Practical and efficient mobility data processing in the cloud. Proceedings of BigData.

Ma, Y., Zhang, Y., Meng, X. (2013) ST-HBase: A Scalable Data Management System for Massive Geo-tagged Objects. Proceedings of WAIM.

Nishimura, S., Das, S., Agrawal, D., Abbadi, A.E. (2011) MD-HBase: A Scalable Multi-dimensional Data Infrastructure for Location Aware Services. Proceedings of MDM.

Shang, S., Chen, L., Wei, Z., Jensen, C.S., Zheng, K., Kalnis, P. (2017) Trajectory Similarity Join in Spatial Networks. PVLDB 10(11): 1178-1189.

Shang, Z., Li, G., Bao, Z. (2018) DITA: Distributed In-Memory Trajectory Analytics. Proceedings of SIGMOD (to appear).

Tang, M., Yu, Y., Malluhi, Q.M., Ouzzani, M., Aref, W.G. (2016) LocationSpark: A Distributed In-Memory Data Management System for Big Spatial Data. PVLDB 9(13): 1565-1568.

Toshniwal, A., Taneja, S., Shukla, A., Ramasamy, K., Patel, J.M., Kulkarni, S., Jackson, J., Gade, K., Fu, M., Donham, J., Bhagat, N., Mittal, S., Ryaboy, D.V. (2014) Storm@twitter. Proceedings of SIGMOD.

Whitman, R.T., Park, M.B., Marsh, B.G., Hoel, E.G. (2017) Spatio-Temporal Join on Apache Spark. Proceedings of SIGSPATIAL.

Xie, D., Li, F., Yao, B., Li, G., Chen, Z., Zhou, L., Guo, M. (2016) Simba: spatial in-memory big data analysis. Proceedings of SIGSPATIAL.

Xie, D., Li, F., Yao, B., Li, G., Zhou, L., Guo, M. (2016) Simba: Efficient In-Memory Spatial Analytics. Proceedings of SIGMOD.

You, S., Zhang, J., Gruenwald, L. (2015) Large-scale spatial join query processing in Cloud. Proceedings of ICDE Workshops.

Yu, J., Wu, J., Sarwat, M. (2016) A demonstration of GeoSpark: A cluster computing framework for processing big spatial data. Proceedings of ICDE.

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauly, M., Franklin, M.J., Shenker, S., Stoica, I. (2012) Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. Proceedings of NSDI.

Zaharia, M., Xin, R.S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M.J., Ghodsi, A., Gonzalez, J., Shenker, S., Stoica, I. (2016) Apache Spark: a unified engine for big data processing. Commun. ACM 59(11): 56-65.

5 Big Data Analytics

This section first explains knowledge discovery in big data. Several topics and techniques are discussed. For each case, research papers covering both moderately sized datasets as well as big data are discussed. First, we discuss trajectory 'clustering' for a given time period and then consider the problem to discover groups of objects moving together. Next, we focus on sequential pattern mining. Pattern growth algorithms are less computationally expensive the Apriori- based algorithms and better suited to be transformed to versions that allow big data being processed in parallel taking advantage of the MapReduce model. Hot spot analysis discretizes space-time and identifies cells for which a particular attribute takes a statistically significant value. In the mobility domain, the number of moving objects in a space-time cell can be counted and analyzed using the classical Getis-Ord statistic. Ongoing and recent research on finding hot spots in big data is discussed. Future location prediction takes a history of movements (visited location sequences) and tries to predict the sequence of visited locations for a given time horizon. 'Pattern-based prediction' is based on the history of sets of moving objects (as opposed to individual histories). We discuss several ways to represent the mined patterns. Predicting object movements on a network from streaming data is of high importance to mobility science (travel guidance). Location prediction is extended to 'Trajectory prediction': most related research applies to aircraft movements and is not network bound. Current challenges include the adaptation of the methods discussed for use on big data. Beyond that, the recent concept of predictive queries applied to network bound movements deserves attention because of its relevance to travel guidance. Finally, data may not be available for specific regions. Hence, 'transfer learning' comes into play when trying to apply models and model parameters sets in regions different from where they were mined. This is a particularly challenging topic for complex phenomena like urban mobility because they depend on spatially dependent habits.

What follows is a review of methods on complex network analytics. The use of complex networks is briefly introduced and examples from community detection and (information) diffusion are presented. Sample real world problems show the need to study network topology dynamics. Mobility data includes several relations which leads graph theoretical and complex network problem formulations. The example of Individual Mobility Networks (IMN) that can be mined from big data is explained in detail.

This chapter concludes with a discussion on Complex Event Recognition (CER) techniques. CER or event pattern matching applies to continuous data flows origination from several sources. Tools and methods to define and represent complex events are discussed along the processes leading to CER. Main research topics are: event pattern specification, uncertainty handling, the challenges posed by CER in big data streams and techniques for both supervised and unsupervised learning of event patterns. Finally, the importance of CER in big data is illustrated for mobility data (both road based and maritime).

5.1 Knowledge Discovery in Big Data

In this section, several issues related to knowledge discovery in Big Data analytics are discussed. Specifically, related works and state-of-the-art are presented for the core tasks of clustering, sequential pattern mining, future location prediction and trajectory prediction. Additionally, a short literature review is presented for other related tasks and methods, including predictive query processing and transfer learning with mobility data analytics.

5.1.1 Clustering

Location aware devices, such as mobile phones, tablets and automobiles carry numerous networked sensors, which create huge amounts of data that represent some kind of mobility. In addition, the massive participation of individuals on location based social networks will continue to fuel exponential

growth in the production of this kind of data. This enormous volume of data has posed new challenges in the world of mobility data management in terms of storing, querying, analyzing and extracting knowledge out of them in an efficient way.

One of these challenges is cluster analysis. The typical approach is to either transform trajectories to vector data, in order for well-known clustering algorithms to be applicable, or to define appropriate trajectory similarity functions, which is the basic building block of every clustering approach. For instance, CenTR-IFCM (Pelekis et al. 2014) builds upon a Fuzzy C-Means variant to perform a kind of time-focused local clustering using a region growing technique under similarity and density constraints. For each time period, the algorithm determines an initial seed region (that corresponds to the sub-trajectory restricted inside the period) and searches for the maximum region that is composed of all sub-trajectories that are similar with respect to a distance threshold *d* and dense with respect to a density threshold ϑ. Subsequently, the growing process begins and the algorithm tries to find the next region to extend among the most similar sub-trajectories. The algorithm continues until no more growing can be applied, appending in each repetition the temporally local centroid. In the same line of research, having defined an effective similarity metric, TOPTICS (Nanni et al. 2006) adapts OPTICS (Ankerst et al. 1999) to enable whole-trajectory clustering (i.e. clustering the entire trajectories), TRACLUS (Lee et al. 2007) exploits on DBSCAN (Ester et al. 1996) to support sub-trajectory clustering, while T-Sampling (Panagiotakis et al. 2012, Pelekis et al. 2010), introduces trajectory segmentation (aiming at temporal-constrained sub-trajectory clustering (Pelekis et al. 2017), by taking into account the neighborhood of a trajectory in the rest of the dataset, yielding a segmentation that is related only on the number of neighboring segments that vote for the line segments of a trajectory as the most representatives. All the above trajectory clustering approaches they are capable of identifying trajectory clusters and their densities but do not tackle the issue of statistical significance in the space-time they take place.

A branch of related works aim to discover several types of collective behavior among moving objects, forming a group of objects that moves together for a certain time period. Among the most related to this work, in (Laube et al. 2005a; Laube et al. 2005b), the authors define various mobility behaviors around the idea of the flock pattern, such as the meeting, convergence and encounter patterns. The discovery of a meeting in a time interval *I* of at least *k* timepoints, consists of at least *m* objects that stay within a stationary disk of radius *r* during *I*. There are two variants of meetings: either the same *m* entities stay together during the entire interval (fixed-meeting), or the entities in the meeting region may change during the interval (varying meeting). On the other hand, the convergence pattern describes trajectories that converge to the same location, coming not necessarily from the same origin. Inspired by this idea, the notion of a moving cluster was introduced in (Kalnis et al. 2005), which is a sequence of clusters $\{c_1, \ldots, c_k\}$, such that for each timestamp *i*, c_i and c_{i+1} share a sufficient number of common objects. There are several related works that emanated from the above ideas, like the approaches of convoys, swarms, platoons, traveling companion, gathering pattern (Zheng et al. 2015). There are several other methods that try to identify frequent (thus, dense) trajectory patterns. In case where moving objects move under the restrictions of a transportation network, (Sacharidis et al. 2008) proposed an online approach to discover and maintain hot motions paths while (Chen et al. 2011) tackled the problem of discovering the most popular route between two locations based on the historical behavior of travelers. In case where objects move without constraints, (Cao et al. 2006) proposed a method to discover collocation patterns.

However, all of the aforementioned approaches are centralized and in order to meet with the challenges posed in the Big Data Era, one should think beyond the centralized paradigm and start examining how solutions to such problems could be implemented in a way that would meet with these challenges. A line of research is to adapt well-known solutions to trajectory datasets. In this context, (Deng et al. 2015) introduces a scalable GPU-based trajectory clustering approach which is based on a scalable density-based clustering approach for point data (POPTICS) (Patwary et al. 2013). As to finding flock patterns in large trajectory databases, (Valladares et al. 2013) presented a GPU-based approach

for finding extremal sets within a family *F* of *k* finite sets, which has no restrictions on the input. (Fort et al. 2014) studied the problem of finding flock patterns in trajectory databases and presented some parallel algorithms based on GPU for reporting all maximal flocks, the largest flock and the longest flock. Moreover, (Jinno et al. 2012) attempts to discover frequent movement patterns from the trajectories of moving objects. More specifically, they propose a MapReduce-based approach to trajectory pattern mining using a hierarchical grid with quadtree search in order to identify complex patterns involving different levels of granularity.

(Moussalli et al. 2015) and (Moussalli et al. 2013) presented FPGA- and GPU-based solutions for parallel matching of variable-enhanced complex patterns by stream-mode (single pass) filtering. Both implementations are able to process the trajectory data in a single pass when handing pattern queries with no more than one variable or no wildcards with two or more variables, but result in false positive matches when two or more variable occur in a pattern query alongside wildcards. The parallel solutions can outperform the current state-of-the-art CPU-based approaches by two or three orders of magnitude at certain circumstances and shows very good scalability with regard to pattern complexity. Similarly, in (Lan et al. 2017) a streaming environment is assumed, however, here, a new concept is proposed, that of evolving group pattern that captures the interesting group patterns over streaming trajectories that cannot be captured by the current group pattern detection techniques.

An approach that defines a new generalized mobility pattern is presented in (Fan et al. 2016). In more detail, the general co-movement pattern (GCMP), is proposed, which models various co-movement patterns in a unified way and can avoid the loose-connection anomaly. Further, the GCMP detector is deployed on a modern MapReduce platform (i.e., Apache Spark) to tackle the scalability issue. On the other hand, in (Ding et al. 2018) an efficient and flexible platform for an open-ended range of trajectory data management and analytics techniques, called UlTraMan, is proposed. Within this system, the GCMP detector is implemented. Moreover, all the necessary preprocessing tasks that are not covered in (Fan et al. 2016) can be supported efficiently in UlTraMan, hence avoiding unnecessary data transfer.

5.1.2 Sequential Pattern Mining

Sequential pattern mining discovers subsequences that appear in a sequence database with frequency no less than a user-specified threshold. A sequence database stores a number of records, where all records are ordered sequences of events, with or without concrete notions of time. Sequential pattern mining is an important data mining problem with broad applications, such as mining customer purchase patterns, identifying outer membrane proteins, automatically detecting erroneous sentences, discovering block correlations in storage systems, identifying copy-paste and related bugs in large-scale software code, API specification mining and API usage mining from open source repositories, and Web log data mining.

This problem was defined as follows: Given a set of sequences, where each sequence consists of a list of elements and each element consists of a set of items, and given a user-specified *min_support* threshold, sequential pattern mining is to find all frequent subsequences, i.e., the subsequences whose occurrence frequency in the set of sequences is no less than *min_support* (Agrawal et al. 2014).

Generally, sequential pattern mining algorithms can be categorized into two major classes: Apriori-based approaches and pattern growth algorithms. The first class of algorithms (i.e., Apriori-based approaches) form the vast majority of algorithms proposed in the literature for sequential pattern mining. They depend mainly on the Apriori property, which states the fact that any super-pattern of an infrequent pattern cannot be frequent, and are based on a candidate generation and- test paradigm proposed in association rule mining (Agrawal et al. 1993). These methods have the disadvantage of repeatedly generating an explosive number of candidate sequences and scanning the database to maintain the support count information for these sequences during each iteration of the algorithm, which makes them computationally expensive.

To alleviate these problems, pattern growth approach for efficient sequential pattern mining adopts a divide-and-conquer, pattern growth paradigm as follows, sequence databases are recursively projected into a set of smaller projected databases based on the current sequential pattern(s), and sequential patterns are grown in each projected database by exploring only locally frequent fragments (Han et al. 2000). The frequent pattern growth paradigm removes the need for the candidate generation and prune steps that occur in the Apriori-based algorithms and repeatedly narrows the search space by dividing a sequence database into a set of smaller projected databases, which are mined separately.

In the era of Big Data, where huge amounts of data are available, algorithm and implementation of sequential pattern mining has to re-designed and re-implemented under a distributed computing framework as traditional approaches are not designed to handle massive amounts of data. In recent years research has been done for finding sequential patterns in parallel and distributed areas like Hadoop, Grid, Cloud, etc.

In Parallel Transaction Decomposed Sequential Pattern Mining (PTDS) (Wang et al. 2010) transactions are decomposed to mine the sequential patterns and pattern growth approach is greatly accelerated to improve the efficiency of large scale data. First, PTDS sorts the sequences and plan the sequences with identical or similar prefix, which is considered as first transaction of each sequence. The input sequence is split in to two parts one is the first transaction and other is the remaining part of transaction in the sequence. PTDS collects sequences with equal prefix, decompose the prefix and applies serial sequential pattern mining method on the set of subsequences; each one contains the remaining transactions of the raw sequence, and finally merges the mining results together. PTDS is implemented using MapReduce framework on Apache Hadoop environment which greatly accelerate pattern growth approach and improves the performance and efficiency of parallel sequential pattern algorithm on large scale data.

Following collaborative pattern mining for distributed information system (CLAP) (Zhu et al. 2011), mining of data is divided into three parts: first, identify locally important patterns on individual database; second, determine major patterns after combining distributed database into single view; third, find patterns which follow special relationship across different data collection. This algorithm makes use of pattern mining for query processing to satisfy user specified query constraints to discover patterns from distributed databases. In existing system pattern pruning is based on single database, so to solve this problem cross-database pruning concept is used for distributed sequential pattern mining. CLAP encourage pattern discovery in distributed approach where each distributed site carries pattern pruning in collaboration with its peers by employing bloom filter-based pattern switching mechanism. A bloom filter is space efficient data structure which contains k hash functions, and binary array of m bits. Patterns like $x_1,x_2,\ldots,x_n$ can be added into the bloom filter to check whether pattern exist in bloom filter or not by using all k hash functions to map x_t to k positions. CLAP system consists of mainly two parts as one construction of FP-tree and bloom filter for each local site and second CLAP cross database pruning and pattern growth. CLAP only focuses on frequent itemset mining.

Recently, many applications are moved to cloud infrastructure. Sequential pattern mining on cloud (SPAMC) (Chen et al. 2013) adapts is developed for mining sequential patterns on MapReduce model on cloud. SPAMC is a cloud-based version of sequential pattern mining algorithm consisting of two phases: scanning phase, and mining phase. In the scanning phase, high performance is achieved by distributing tasks on multiple computers by using MapReduce programming model to proceed in parallel by distributing sub-tasks to independent machines. Each mapper scans and transforms a partitioned database, and reducers are used to count the frequency of each item and eliminate infrequent items. The bitmap information of frequent items will be stored into a distributed hash table (DHT) that can be accessed in the mining phase. After that, in the mining phase, the sequential pattern mining tasks are processed in parallel by distributed machines. Main task of the mining phase is to construct the complete lexical sequence tree, and then all patterns can be derived. Additionally, to achieve better load balancing, depth first search strategy is used to bring out the steps of sequence

and itemset extension with limited sub-tree depth. This strategy effectively improves the situation like mapper may stand and wait for a long time. In such a context, each MapReduce round will complete two levels of lexical sequence sub tree construction. On the other side, reducers efficiently integrate output results from mappers and do the support counting to generate frequent sequential patterns of the current sub-tree.

5.1.3 Hot-spot Analysis

The data wealth, produced by the proliferation of GPS technology, the widespread adoption of smartphones, social networking, as well as the ubiquitous nature of monitoring systems, contributes to the ever-increasing size of what is recently known as Big spatial (or spatio-temporal) data (Eldawy et al. 2016), a specialized category of Big data focusing on mobile objects. Analyzing spatio-temporal data has the potential to discover hidden patterns or result in non-trivial insights, especially when its immense volume is considered. To this end, specialized parallel data processing frameworks (Alarabi et al. 2017a, Alarabi et al. 2017b, Hagedorn et al. 2017, Tang et al. 2016) and algorithms (Doulkeridis et al. 2017, Fang et al. 2016, Whitman et al. 2017, Xian et al. 2016) have been recently developed aiming at spatial and spatio-temporal data management at scale.

In this context, a useful data analysis task is Hot spot analysis, which is the process of identifying statistically significant clusters. However, there is practically no work on hot spot analysis for Big trajectory data. One of the main challenges is focused on discovering hot spots in the maritime domain, as this relates to significant challenging use-case scenarios (Claramunt et al. 2017), such as identifying different types of activities in a region of interest, estimating fishing pressure, environmental fingerprint, etc. Similarly, in the aviation domain the predicted presence of a number of aircrafts above a certain threshold results in regulations in air traffic, while in the urban domain such a presence accompanied with low speed patterns implies traffic congestions. Thus, the effective discovery of such diverse types of hot spots is of critical importance for our ability to comprehend the various domains of mobility.

Hot Spot discovery and analysis is usually based on spatio-temporal partitioning of the 3D data space in cells. The identification of cells that constitute hot spots includes having high concentration of mobile objects and in statistically significant densities. One of these methods is the Getis-Ord statistic (Ord et al. 1995), a popular metric for hot spot analysis, which produces z-scores and p-values. A cell is considered as a hot spot if it is associated with high z-score and low p-value. Unfortunately, the Getis-Ord statistic is typically applicable in the case of 2-D spatial data, and even though it can be extended to the 3-D case, it has been designed for point data.

The problem of Trajectory hot spot analysis can be formulated by taking into account the contribution of a moving object's trajectory to a cell's density, which is proportional to the time spent by the moving object in the cell. To this end, the Getid-Ord statistic can be adapted (Nikitopoulos et al. 2018) to capture this approach for the case of trajectory data and the algorithm can be designed for parallel and scalable processing for computing hot spots in terms of spatio-temporal cells produced by grid-based partitioning of the data space under consideration.

Similar approaches can be adapted and applied in the urban environment, especially designed for Big mobility data. Hot spot analysis will be a very important aspect of detecting points of high density, bottlenecks and points of interest, which can be combined with efficient identification of mobility patterns.

5.1.4 Future Location Prediction

The problem of Future Location Prediction (FLP) can be informally described as follows: Given the recent spatio-temporal history of N previous data points of a moving object, i.e., consisting of its time-stamped locations recorded at N past time instances, and an integer look-ahead value L, predict the anticipated future locations of the object for the next L time instances. The main factors for any FLP

algorithm are size of the history (N), the extent of the prediction window (L) and the way these two are combined together in a predictive model.

The FLP problem finds two broad categories of application scenarios. The first scenario involves cases where the moving entities are traced in real-time to produce analytics and compute short-term predictions, which are time-critical and need immediate response. Short-term FLP can be extremely important in domains where safety, adaptiveness and responsiveness out outmost importance and a decision-making process. The second scenario involves cases where long-term FLP is important to identify cases which exceed regular mobility patterns, detect anomalies, and determine a position or a sequence of positions of special interest at a given time interval in the future. In this case, although response time may not be a critical factor per se, it is still crucial in order to identify correlations between historical mobility patterns and patterns that are expected to appear, e.g. approach to a restricted area.

There are two main directions when dealing with the FLP problem: (a) *vector-based* prediction or the the spatial database management approach and (b) *pattern-based* prediction or the data mining & Machine Learning approach. Each has its own advantages and drawbacks and, most importantly, it is based on different assumptions regarding the data and their organization used as the input.

The vector-based approaches, inspired by the spatial database management domain, aim to model current locations (and perhaps a short history) of objects as *motion functions*, in order to be able to predict future locations by some kind of extrapolation. In practice, they take into consideration space and time and predict future locations of moving objects within a given time interval using a mathematical or probabilistic model, which aims to simulate the anticipated movement. First- or second-degree physics models of movement are commonly used, employing extrapolation with velocity or velocity and acceleration components, respectively, to estimate the evolution of movement, provided that these can be assumed to be constant in a short-term look-ahead time window.

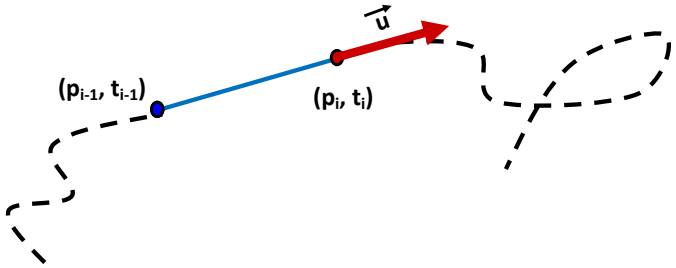


Figure 10: The future position of a moving object as the result of a linear motion function.

The constant-speed assumption is also very useful in the development of proper transformations of the input space that enable time-invariant representations, e.g. via the Hough-X transform (Jagadish et al. 1990). Essentially, the evolving position of a moving object remains a stationary point in dual space as soon as it does not change its velocity vector, thus it can be efficiently indexed in a spatial access method. This is the main concept behind the family of *predictive query processing* techniques for FLP that introduces various state-of-the-art methods including PMR Quadtree (Tayeb et al. 1998; Samet 1990), TPR*-tree (Tao et al. 2003), Bx-tree (Jensen et al. 2004) and STP-tree (Tao et al. 2004).

The pattern-based approaches, inspired by the spatial data mining domain, identify and exploit motion patterns by analyzing historic data of moving objects, i.e., classification models, repetitive patterns, clusters of “similar” movements, etc, based upon a history of movements. An important difference with respect to the vector-based approaches is that in this case the models are built upon the history of movements, not only of the object of interest, but also of the other objects moving in the same

area; therefore, they are able to build better models and use them for addressing the FLP task in a more generic and data-driven way.

Techniques based on Hidden Markov Models (HMM), Neural Networks (NN) and other data-driven approaches have been extensively used to address the FTP problem. (Ishikawa et al. 2004) introduce an algorithm that extracts mobility statistics from indexed spatio-temporal datasets for interactive analysis of huge collections of moving object trajectories. In the maritime domain, (Zorbas et al. 2015) introduce a machine learning model using a NN that exploits geospatial time-series surveillance data generated by sea-vessels, in order to predict future trajectories with real-time constraints with a look-ahead time window of 5 minutes. In a different domain, that of aviation, (Hamed et al. 2013) propose a method for predicting the altitude change of an aircraft within a predefined look-ahead time window of 10 minutes.

There are also pattern-based techniques that are based on association rules or frequent mobility patterns. These include methods like the Mobility Patterns (Yavas et al. 2005), TrajPattern algorithm for pattern groups (Yang et al. 2006), Spatio-Temporal Association Rules (STARs) (Verhein et al. 2006), WhereNext for trajectory patterns (T-patterns)(Monreale et al.2009), as well as state-of-the-art methods in this area like NextLocation (Gomes et al. 2013) and MyWay (Trasarti et al. 2017).

There is also a relatively new category of *semantic-aware* approaches that involves semantics or *enrichments* extracted by the surrounding environment, e.g. stops, hot-spots, etc. Then, patterns are built upon this knowledge of enriched spatio-temporal data and then used for predicting the next location(s). As an example, (Ying et al. 2011) are the first who exploit both geographic and semantic features of trajectories. Their approach is based on a novel cluster-based prediction method, which estimates a mobile user's future location by exploiting frequent patterns in similar users' behavioral activities.

In other works, a set of motion patterns is exploited for optimally designed `codebook' of motion functions that is used to fit the recent history of an object's movement and then extrapolate upon them within a specific look-ahead time window. Such an approach is the LeZi-Update adaptive on-line algorithm (Bhattacharya et al. 1999), incorporating dictionary updates as in the Lempel-Ziv algorithm family (Liv et al. 1978) for lossless compression.

5.1.5 Trajectory Prediction

Typically, the trajectory of a moving object is defined as a sequence of spatio-temporal data points of length N, consisting of its time-stamped locations recorded at N past time instances, chronologically ordered. In principle, the spatial dimension D of the data points is arbitrary, but the most common cases are moving objects on a surface (D=2, e.g. maritime or land) and in a volume (D=3, e.g. aviation). Additionally, in order to simulate continuous movements, we usually make an assumption of interpolation in-between two consecutive data points; the most popular is linear interpolation, although other functions may be used as well (B-splines, etc.).

Similarly to FLP, the Trajectory Prediction (TP) task can be informally described as follows: Given the recent history of S previous trajectories of one or more moving objects, i.e., each consisting of its time-stamped data points recorded in the past, predict the anticipated future trajectory of the same or “similar” objects, based on some common reference initialization (e.g. starting point, time frame, region of interest, etc). The main factors for any TP algorithm are size of the history (N) and how it is exploited by a predictive model.

In principle, the TP problem can be approached as a generalization of the FLP problem (Hamed et al. 2013; Theodoridis et al. 2008; Zheng et al. 2015), which is the task of predicting the next spatio-temporal position(s) of a moving object based on its previous track, most commonly in the short-term context (up to few minutes). On the other hand, the TP problem is to predict the anticipated track of the moving object given a set of constraints and/or historic data. A FLP method could be transformed to address the TP problem, given a specific granularity upon which the same method is applied

iteratively. However, in that case the prediction errors are accumulated with each step (e.g. via multi-step linear regression), thus making the next predicted points increasingly error-prone. In contrast, 'pure' TP methods aim to forecast the trajectory itself from the start, thus making each predicted point equally error-prone.

Recently, there has been plenty of work on location and trajectory prediction in the mobility (Pelekis et al. 2014). The proposed approaches include systems-engineering view (Sip et al. 2003) balancing TP accuracy and processing speed, stochastic approaches other than HMM, splitting the flight phases (Gong et al. 2004), collaborative TP via Conflict Avoidance & Resolution (CA&R) (Chen et al. 2011; Matsuno et al. 2015; Vouros et al. 2018), anomaly detection (Di Ciccio et al. 2016), etc. Not surprisingly, the vast majority of methods are domain-specific (with most of them in the aviation domain) and this is in order to take advantage of the properties of the moving objects under consideration. The issue of exploiting additional data or enrichments in TP have created the notion of semantic-aware TP or Semantic Trajectory Prediction (STP), which enables better estimations for departure and arrival times and, hence, more robust scheduling and logistics, especially in the congestion points.

During the last few years, there is a mainstream trend of using stochastic models for retrieval, with HMM approach being the most popular (Rabiner et al. 1989), as it has proved its efficiency in modeling a wide range of sequences of observations. In general terms, a system is assumed to have the Markovian property if its future situations depend only upon its current state. Exhibiting high accuracy in modeling sequential data, the HMM approach has given rise to a wide range of applications, such as speech recognition, music retrieval, human activity recognition, consumer pattern recognition, etc. Consequently, it is a clear opportunity to apply them in the domain of mobility data analysis. In the context of trajectory prediction, the flight route and all the associated information (weather, semantic data, etc), are usually encoded into discrete values that constitute the HMM states; then, the trajectory itself is treated as an evolution of transitions between these states, using the raw trajectory data of a large set of flights for training, plus spatio-temporal constraints (locality) to reduce the dimensionality of the problem.

(Ayhan et al. 2016) introduce a novel stochastic approach to aircraft trajectory prediction problem, which exploits aircraft trajectories modeled in space and time by using a set of spatio-temporal data cubes. They represent airspace in 4-D joint data cubes consisting of aircraft's motion parameters (i.e., latitude, longitude, altitude, and time) enriched by weather conditions. They use Viterbi algorithm (Viterbi 1967) to compute the most likely sequence of states derived by a HMM, which has been trained over historical surveillance and weather conditions data. The algorithm computes the maximal probability of the optimal state sequence, which is best aligned with the observation sequence of the aircraft trajectory. In their experimental study, they demonstrate that their methodology efficiently predicts aircraft trajectories by comparing the prediction results with the ground truth aligned trajectories, with the error being reasonably low for one-hour flights.

Two of the most widely explored approaches in TP is regression and clustering, separately or in combination, some also exploring the use of weather or other data. These include methods based on Generalized Linear Model (GLM) (de Leege et al. 2013), multi-stage clustering (Yang et al. 2015), typical regression-based short/mid-term TP (Krumm et al. 2003; Tastambekov et al. 2014), combination of clustering and Kalman filters (Song et al. 2012), etc. Neural networks have also been used successfully for the climb/vertical TP (Le Fablec et al. 1999) or in relation to the air traffic flows (Cheng et al. 2003) for Estimated Time of Arrival (ETA).

Regarding en route climb TP, one of the major aspects of ATM decision support tools, (Coppenbarger et al. 1999) discusses the exploitation of real-time aircraft data, such as aircraft state, aircraft performance, pilot intent and atmospheric data for improving ground-based TP. The problem of climb TP is also discussed in (Thipphavong et al. 2013) as it constitutes a very important challenge in ATM. In another work by (Ayhan et al. 2016), the authors investigate the applicability of the HMM for TP on only one phase of a flight, specifically the climb after takeoff. A stochastic approach such as the HMM

can address the TP problem by taking environmental uncertainties into account and training a model using historical trajectory data along with weather observations. There are also numerical approaches to the problem of climb-phase TP, e.g. (Hadjaz et al. 2012).

5.1.6 Other Challenges

As described in the previous sections, both FLP and TP problems have been studied extensively in the last few years. Some of the proposed approaches are compatible with Big data applications and some are not. Mobility data are in the core of various Big data modalities and approaches in addressing analytics and predictive modeling tasks in a wide range of contexts. Thus, it is imperative that such approaches are scalable and parallelizable, in order to handle data of very large volume, velocity, veracity and variety.

A more recent approach for addressing predictive modeling tasks via mobility patterns comes from the area of Predictive Queries (PQ) (Hendawi et al. 2012b, Zhang et al. 2012), which is one of the most exciting research topics in spatio-temporal data management. In many location-based services, including traffic management, ride sharing, targeted advertising, etc., there is a specific need to detect and track mobile entities within specific areas and within specific time frames. In Range Queries (RQ), the task is focused on identifying POIs and mobility patterns related to the current locations of moving objects. Instead, Predictive Range Queries (PRQ) address the same task but for future time frames. This is a typical use case in aviation, when one or more airplanes need to be checked in some spatial context in the future, e.g. for proximity (collision avoidance), scheduling (takeoff/landing), airspace sectorization (avoid overload and/or delays), etc.

In the context of PRQ and most commonly in the RQ task, various approaches can be used for checking arrivals/departures of airplanes to/from specific regions of interest, including optimized k-nearest-neighbour (k-nn) variants that employ spatio-temporal index trees. Similarly, a reverse k-nn query can be used to detect moving objects that are expected to have the query region as their nearest neighbour, e.g. for assigning moving objects to their “nearest” tracking node. Indexing can be implemented by very efficient data management structures, such as R-trees (time-parameterized, a.k.a. TPR/TPR*-trees), variants of B-trees, kd-trees, Quad-trees, etc (Hendawi et al. 2012b, Hendawi et al. 2015b). The predictive model itself can be linear or non-linear and it is most commonly based on historical data in the same spatio-temporal context, in the short- or the long-term w.r.t. time frame. The uncertainty of the prediction is addressed by either model-based approaches, which determine a representative model for the underlying mobility pattern, or pure data-driven approaches, which “learn” and index movements from historic data (Zhang et al. 2009).

Another important aspect especially in FLP is the ability to employ such models in streaming data, i.e., using “live” sources of mobility data as they become available. This task can also be addressed by PRQ approaches, more specifically the continuous PRQ algorithms. The difference between a “snapshot” predictive query and a continuous one is that the second can be continuously re-evaluated with minimal overhead and optimal efficiency. As an example, the Panda system (Hendawi et al. 2012a, Hendawi et al. 2015b), designed to provide efficient support for predictive spatio-temporal queries, offers the necessary infrastructure to support a wide variety of predictive queries that include predictive spatio-temporal range, aggregate (number of objects), and k-nn queries, as well as continuous queries. The main idea of Panda is to monitor those space areas that are highly accessed using predictive queries. For such areas, Panda pre-computes the prediction of objects being in these areas beforehand.

Similar approaches exist in various domains, such as the iRoad (Hendawi et al. 2013), which is employed for tracking vehicles in urban areas. More specifically, the system supports a variety of common PQs including point query, range query, k-nn query, aggregate query, etc. The iRoad is based on a novel tree structure named reachability tree, employed to determine the reachable nodes for a moving object within a specified future time T. By employing spatial-aware pruning techniques, iRoad is able to scale up to handle real road networks with millions of nodes and it can process heavy

workloads on large numbers of moving objects. Since flight routes of civilian and cargo flights are also conditioned by various constraints, e.g. by submitted flight plans (aviation domain) or common ship routes (maritime domain), such road-based approaches can be adapted for a wide variety of problems (Jeung et al. 2010, Hendawi et al. 2015b).

In the context of scalability and the Big data aspect, there are very recent and promising approaches such as the UITraMan (Ding et al. 2018), which addresses the scalability, the efficiency, the persistence and the extensibility of such frameworks. More specifically, it extends Apache Spark w.r.t. data storage and computing by employing a key-value store and enhances the MapReduce paradigm to allow flexible optimizations based on random data access. Another approach for PQs in Big data is presented by Panda* (Hendawi et al. 2017), which is a scalable and generic enhancement of Panda (Hendawi et al. 2012a), applied to traffic management. More specifically, Panda* is a generic framework for supporting spatial PQs over moving objects, introducing prediction function when there is lack of historic data, isolation of the prediction calculation from the query processing and control over the trade-off between low latency responses and use of computational resources. For both UITraMan and Panda*, experimental results on large-scale real and synthetic data sets in other domains, which include comparisons with the state-of-the-art methods in this area, show promising results and hints of successful application to the aviation domain too.

It should be noted that there are also other types of PQs, more advanced than the ones presented above, such as the predictive pattern queries (PPQ), which check conditions muc more complex than simple presence or not of a moving object within a specific spatio-temporal frame. Such advanced PPQs can be considered as a link between data management and data analytics, which can be very valuable in the context of the aviation domain.

5.1.7 Geographical Transfer Learning and Mobility Data

Most machine learning and data mining methods work on the expectation that the context where the models and patterns were extracted is similar (i.e. has the same dependencies between variables) to the one where we want to deploy them. However, in several problems that is not the case, either because the samples in the two contexts are not homogeneous (e.g. the distributions of some variables are different) or because the data available in the second context is poorer. In such cases, transferring the knowledge from one context to the other can be challenging but also extremely useful, since would avoid the set-up of a completely new analysis process, including expensive data collection and labelling. This problem is called transfer learning, or knowledge transfer, and gained a large attention from the research community the latest years.

Transfer learning has been deeply studied in the general context of machine learning (Pan and Yang 2010, Tsung et al. 2017), yet transferring models across different geographical contexts has been only sparsely explored, especially in relation to human mobility.

Some basic, geography-related example of knowledge transfer is given in (Wei, Zhang and Yang 2010), where mobility-based models for estimating air quality are transferred from a city where there exist sufficient multimodal mobility data and labels to cities with insufficient data. Similarly, (Liu at al. 2017) aim to identify the combinations of landscape metrics (inferred from satellite images) that correspond to the presence of urban villages. The technical issue, here, is that the relations between the two phenomena vary in space due to the presence of different geographical factors, and therefore the models must be adapted to the different contexts;

A common problem is the geospatial transfer of models describing physical of social phenomena, such as house prices and seismic movements, across regions having different variable distributions or correlations, as studied in (Bussas et al. 2017). Similarly, the work by (Jun 2010) deals with the problem of classifying spatial data (specifically hyperspectral data) through spatially adaptive model parameters for Gaussian process models, and presents various solutions to infer the parameters locally to each area.

The work in (Wang et al. 2017) considers a slightly different problem: how to transfer models from one set of mobility modes (taxis and buses) to a different one (ridesourcing cars, like Uber and similar services), although in the same geographical area. The main problem, in this case, is to understand how to map (mobility) features across the different modalities.

Finally, various works try to transfer models (i.e. model parameters) for various kinds of recognition tasks from one place to another one that might show slightly different conditions. An example on human activity recognition across different buildings is provided in (Kasteren et al. 2010).

Despite the various examples discussed above, very little has been done so far on the transfer of complex models, such as trajectory patterns, mobility profiles or mobility forecasting models. This is a challenging and very promising direction of research that Track&Know will pursue.

5.2 Complex Network Analysis in Big Data

5.2.1 Complex Networks

Complex Networks (Newman, 2003) are popular mathematical tools commonly used to describe and analyze interaction phenomena that occur in the real world. Social ties formation, economic transactions, face to face communications, the unfolding of human mobility are examples of events usually described by semantic rich Big Data often investigated using instruments borrowed from Graph Theory. Thanks to such heterogeneous analytical context, during the last decades several problems have been modeled and approached leveraging the framework offered by Complex Networks. Among the network related tasks addressed to extract meaningful information from real data, Community Discovery, Link Prediction, Spreading and Epidemic modeling are certainly the most famous ones.

The concept of a “community” in a (web, social, technological, biological or informational) network is intuitively understood as a set of entities that have some latent factors in common with each other, and thus play a specific role in the overall function of the complex system (Fortunato, 2010). Traditional approaches to discover such mesoscale topologies assume that latent factors drive network connectivity; thus, finding sets of nodes with a high edge density among each other and a low edge density with the rest of the network effectively detects the functional modules of the network. Community discovery is then a network variant of data clustering, where proximity is replaced with edge connectivity. Communities are often used as a pre-processing step to enable complex analysis on top of network structures. For instance, they are often used to relate topological structures with external information – as in (Rossetti et al. 2016) where densely connected sets of Skype/Google+/Last.fm users were used to providing a characterization of the overall service usage.

Since network topologies are expected to change as time goes by, forecast the appearance and vanishing of the entities (nodes as well as edges) composing them represents a crucial task to address. In this scenario, Link Prediction (Liben-Nowell et al. 2007, Lu et al. (2011)) focuses on the analysis of network historical data to provide insights on the future evolution of the network topology. Several Link prediction methodologies where proposed with the aim of identifying future friendships in social graphs (Jalili et al. 2017), collaborations in scientific/professional networks, interactions in protein-protein networks as well as future co-locations of individuals (Wang et al. 2011).

Indeed, generally, a dynamic process can describe not only graph topology perturbation but also the diffusion of some kind of content upon such complex structure. Commonly, when we use the word “spreading” we think to contagious diseases caused by biological pathogens, like influenza, measles or sexually transmitted diseases. However, a plethora of phenomena can be linked to the concept of epidemic: the spread of computer viruses (Szor 2004), the spread of mobile phone virus (Wang et al. 2013), the diffusion of knowledge, innovations, products in an online social network, etc. Several network models were designed to approach the complex task of modeling and forecasting diffusive phenomena, often leveraging data-driven analysis of real-world phenomena. As an example, in

(Bakshy et al. 2012) the authors examine the role of information diffusion in the sharing habits of 235 million Facebook users. They study the role of weak and strong ties in information diffusion showing that the propagation of novel information is mostly due to the abundance of weak ties. The authors of (Leskovec et al. 2007) studied a corpora of weblogs (composed by 45,000 blogs and 2.2 million blog-posting) for two months. In their paper, they show that blog posts do not have bursty behavior and that post popularity drops as a function of time. In (Cha et al. 2009) a Flickr dataset of 33 million photos marked as "favorite" from 2.5 million users of the service is analyzed. The authors observed that most of the markings do not spread widely throughout the network: even the more popular photos have limited popularity outside the immediate neighborhood of the original uploader.

Indeed, both network topological dynamics – as the ones studied by Link Prediction approaches – and dynamics that occur on top of network structures are often interdependent. Such dualism has lead in recent years to the rising of the dynamic network analysis field (Holme et al. 2012). In a dynamic scenario, all the network problems defined and studied on top of static data are extended to allow a fine-grained time-aware analysis. Community Discovery, as an example, is revised to tracking network substructures as time goes by (Rossetti et al. 2018): such life-cycle analysis allows not only to profile group of entities involved in a networked structure but also to understand how their profile changes as the phenomenon the network describe evolves.

5.2.2 Mobility Data Analysis with Networks

The massive amount of mobility data available from different sources requires intensive analysis in order to extract useful models and patterns. The challenge is not only the computational aspect, but also the representation of this data in a meaningful and semantically rich way allowing classical and new methodologies algorithms to be applied. In particular the network (or, equivalently, graph) representation of this data gives a flexible way to define relations (edges) among basic concepts (nodes).

In literature we can consider three different approaches considering what the nodes represent.

the first class of works has the **users as nodes** (Hossmann et al. 2011; Wang et al. 2011), in both the cases the edges are weighted links representing the spatio-temporal co-location of them, i.e. the possible contacts, and the authors uses this graph to discover communities of users, connectivity measures and to predict future social ties.

The second approach has **user's locations as nodes** and the edges represent the trips between them (Gonzalez et al. 2002; Rinzivillo et al. 2014) – the link weight being proportional to the frequency of the trip. In this case the main analytical objectives are finding spatio-temporal regularities and patterns in user mobility or classify the purpose of the user's visit.

Finally the third approach, the most used one, is to consider **global locations as nodes**. In this case, current analysis methods in the literature follow various different ways of defining edges between such nodes:

- A link if there is an Infrastructure (e.g. streets, railways, etc.) connecting the locations;
- A link if there is a collective service (i.e. taxi, bus, etc) connecting the locations;
- Weighted links representing the number of users moving between the two locations.

These different ways of building the graph are used for a large variety of analytical objectives, which include: trips simulation (Tian et al. 2002), evaluating the resilience of the road/transport network (Woolley-Meza et al. 2011) and simulating diseases spreading (Brockmann et al. 2009) for the first group; studying network and traffic evolution (Xia et al. 2018) and detecting traffic anomalies (Chawla et al. 2012) for the second group; inferring communities of locations (Brilhante et al. 2012), optimizing traffic (Zhang et al. 2018), comparing the structure of cities (Saberi et al. 2017), inferring new local borders within a country (Thiemann et al. 2010) and nowcasting air quality (Zheng et al. 2013) for the case of weighted links.

A sample analysis framework of particular interest for the study of mobility at the level of single individuals is the work in (Rinzivillo et al. 2014), where the Individual Mobility Networks (IMNs) are defined. IMNs describe the individual mobility of an individual through a graph representation of her locations and movements, grasping the relevant properties of individual mobility and removing unnecessary details. Formally, the Individual Mobility Network of an individual u is a directed graph Gu = (V, E), where V is the set of nodes and E is the set of edges. On nodes and edges the following functions are defined:

- ω : E → N returns the weight of an edge (i.e. the number of travels performed by u on that edge);
- τ : V → N returns the time spent by the individual in a given location;
- pe : E × T → [0, 1] estimates the probability pe (e, t) of observing an individual u moving on edge e at time t;
- pl : V × T → [0, 1] estimates the probability pl(v, t) of observing an individual u at location v at time t.

Nodes represent locations and edges represent movements between locations. We attach to both nodes and edges statistical information by means of structural annotations: edges provide information about the frequency of movements through the ω function; nodes provide an estimation of the time spent in each location through the τ function. To clarify the concept of IMN, let us consider the network in Figure X. It describes the IMN extracted from the mobility of an individual who visited 19 distinct locations. Location "a" has been visited a total of 18 time units (days in the example), i.e. τ(a) = 18. The edge e = (a, b) has weight ω(e) = ω(a, b) = 20, indicating that the individual moved twenty times from location a to location b.

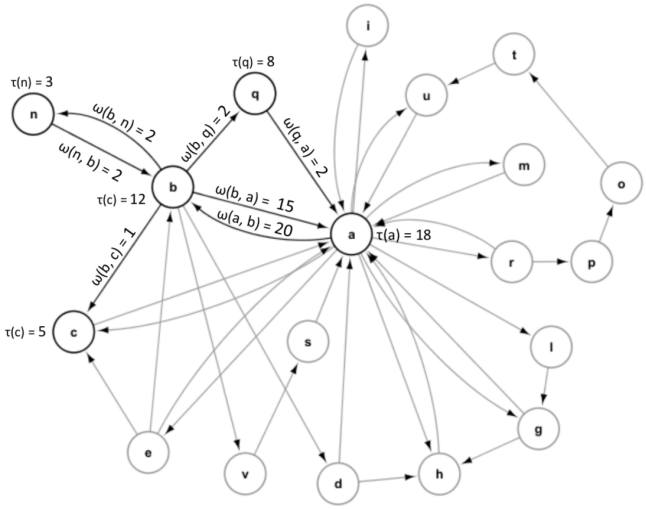


Figure 11: The IMN extracted from the mobility of an individual. Edges represent the existence of a trip between two locations. Function ω(e) is the number of trips performed along edge e, τ(x) the total time spent in location "x".

In (Rinzivillo et al. 2014) the analytical objective is to build a classifier for the purpose of the visits of a user. This work demonstrated that abstracting the mobility data of the user from the geography provided a suitable representation layer for performing a classical data mining task to discover semantically rich models and patterns. Also, the work exploited the explicit relations encoded in each network, which allow, for instance, to propagate information from one node to the others (in the specific application, the activities performed in a location have an impact on the activities performed in adjacent [in terms of network topology] locations).

5.3 Complex Event Recognition in Big Data

Complex Event Recognition (CER) — event pattern matching — applications detect various events of interest in continuous, high-velocity data flows originating from a multitude of distributed sources, by timely providing responses to complex queries. CER plays an important role in Track & Know project, aiming to allow for real-time intelligence in the big data analytics toolbox that will be developed in the project. We review the state-of-the-art in CER with respect to key objectives related to research and development in Track & Know. We begin with an overview of the main CER languages and formalisms, including a brief description of representative systems for each such formalism and their ability to handle the variety of big data. We next discuss uncertainty handling in CER, crucial for addressing the lack of veracity of such streams and continue with important issues related to scaling CER systems to the volume and velocity of big data. We also present some existing techniques for machine learning event patterns from data and conclude with a discussion of CER approaches for mobility applications, which are highly relevant to Track & Know project.

5.3.1 Event Pattern Specification languages

In principle, an event is any time-stamped piece of information. CER systems accept as input simple events, i.e. non-decomposable event occurrences, and they recognize complex events, i.e. event patterns of special significance, which are defined in terms of simple events and potentially other complex events and contextual knowledge. A variety of languages and formal methods for CER have been proposed in the literature - see (Artikis et al. 2012; Cugola and Margara 2012; Artikis et al. 2017) for overviews. Existing approaches have been developed within the database, distributed systems, and artificial intelligence communities. They all have a common goal - express event patterns and match such patterns in the input data - but due to the diversity of their origins, they differ in their architectures, data models, pattern languages, and processing mechanisms.

One family of CER systems relies on automata-based approaches. Event patterns in such systems are compiled into some form of automaton, such as non-deterministic finite automata (NFA) (Mozafari et al. 2013) or finite state machines (FSM) (Schultz-Møller, Migliavacca, and Pietzuch 2009). Such representations are used to provide the semantics of the event pattern language, as well as an execution framework for the event recognition task. Examples of such systems include Cayuga (Brenna et al. 2007), SASE (Mozafari et al. 2013), SASE+ (Zhang, Diao, and Immerman 2014) and TESLA (Cugola and Margara 2010). Some of these approaches use automata both as an event pattern specification formalism and as an execution framework for event recognition (Brenna et al. 2007; Mozafari et al. 2013; Zhang, Diao, and Immerman 2014), while others use an ad-hoc event pattern specification language and then translate such patterns into an automata-based representation, which is eventually used for event recognition (Cugola and Margara 2010). Automata-based methods are well-suited for CER, since they are able to match event sequences in an input string, similarly to strings of characters recognized by regular automata. However, CER automata are more powerful than traditional finite state automata that recognize regular expressions, since they operate on rich event representations consisting of attributes, relations and constraints, they are capable of storing previously observed events in registers, to allow for temporal reasoning between events and they produce output rather than simply deciding whether a string is matched or not.

Another family of CER systems relies on tree-based models. In a tree-based event pattern, leaf nodes in the tree represent event attributes and inner nodes represent event operators, where an operator node is parent to two or more attribute nodes or other operator nodes, thus defining a hierarchy of event operators. Event operators may include e.g. sequencing, negation, conjunction, disjunction, Kleene closure (iteration) etc. Realizations of such models for event pattern specification are ZStream (Mei and Madden 2009) and Esper[2]. The recognition process in tree-based systems is based on

[2] http://www.espertech.com/esper/

assigning buffers to all nodes in the tree. For leaf nodes, these buffers store the input events as they arrive, whereas the buffers of non-leaf nodes store intermediate results that are assembled from sub-tree buffers. To perform even recognition, tree-based CER models start from the leaves of the tree where the input data are loaded, and they traverse the tree in a bottom-up fashion assembling match results based on the semantics of the event operators in the tree.

A third family of CER systems are logic-based. They are characterized by a formal semantics expressed in some form of logic, in contrast to other types of CER systems that often present an informal or procedural semantics (Artikis et al. 2012). In some cases, logic-based CER systems encode rules using logic programming and use inference to detect complex events (Anicic et al. 2011). Prominent logic-based approaches are based on chronicle recognition (Dousson and Maigat 2007) and the event calculus (Artikis, Sergot, and Paliouras 2015). Chronicle recognition relies on temporal logic and encodes event occurrences using logical predicates that define the time of occurrence and the content (attributes) of each event. Complex events are defined starting from primitive ones linked together with contextual and temporal constraints. Event calculus builds on fluents, which are properties that have different values at different points in time. In event calculus-based CER approaches, an event specification consists of rules that define the event occurrences, the effects of events, and the values of fluents.

Logic-based approaches have a number of important advantages as compared to automata-based and tree-based formalisms. In addition to their formal declarative semantics, they also allow to express and reason with complex relations between events and utilize rich domain knowledge in the recognition process. On the other hand, non-logical CER approaches are in general more efficient than logic-based ones. This is not the case with RTEC (Artikis, Sergot, and Paliouras 2015), a recent, event calculus-based CER engine, which relies on reasoning over time intervals, windowing techniques and several other runtime optimizations to scale to massive data volumes and compete in efficiency with non-logical CER approaches.

5.3.2 Uncertainty Handling in Complex Event Recognition

CER systems operate on noisy data streams. In addition to data uncertainty (e.g. missing, or erroneous input), due to the lack of veracity in big data streams, an additional source of uncertainty in CER is pattern uncertainty, i.e. cases where the employed complex event patterns are imprecise or incomplete. The ability to handle erroneous and uncertain input, as well as uncertain event patterns is an important aspect of CER research. A number of CER techniques that can handle uncertainty have been proposed, based mainly on automata, probabilistic graphical models and logic (Alevizos et al. 2017).

Automata-based approaches are usually probabilistic versions of crisp CER systems. For instance, in the probabilistic version of SASE, the goal is to recognize complex events with some probability, via considering alternative “event histories” and calculating a probability for a complex event based on the number of such histories that actually result to the recognition of the complex event and those that do not. Lahar (Ré et al. 2008) is another automata-based approach, in particular, a probabilistic version of the Cayuga CER engine. Lahar handles uncertainty via modelling events by first-order Markov processes, thereby being capable of probabilistic complex event recognition.

Another line of research is based on probabilistic graphical models, with Markov Logic Networks (MLN) being the most prominent example of using such approaches for CER (Alevizos et al. 2017). Complex event patterns in MLN are represented as weighted first-order logic formulae. Patterns with larger weights are “stronger”, while patterns with smaller weights express conditions that are unlikely, but not impossible. Given a set of constants (representing e.g. time-stamps and event attributes) the formulae of an MLN specify a ground Markov network and standard inference methods from the field of probabilistic graphical models may be used to recognize complex events (Tran and Davis 2008; Liu, Deng, and Li 2017; Skarlatidis, Paliouras, et al. 2015). Other probabilistic graphical models-based formalisms have been used in a CER context as well. For instance, in (Cugola et al. 2015), the authors

present an approach where the logical event pattern specification language of the TESLA CER system is embedded into a probabilistic framework based on Bayesian networks. In the field of logic programming, the ProbLog language, a probabilistic version of Prolog has been used as a basis for specifying uncertainty-handling event specifications (Skarlatidis, Artikis, et al. 2015).

5.3.3 Complex Event Recognition in Big Data Streams

Scaling CER systems to massive, high-velocity data streams is an important research topic in the event processing community. A comprehensive survey of related methods and techniques may be found in (Flouris et al. 2017). Such techniques seek to optimize the event recognition task w.r.t. a number of performance metrics, the most important of which are throughput, i.e. the number of events processed by time unit, as well as recognition time. In addition to these metrics, used mainly in cases where the entirety of the data is delivered to a single processing node, approaches based on parallel or distributed CER try to balance the cost of communication between processing nodes and the detection latency, i.e. the time between the occurrence of a complex event and its detection from a central node whose role is to continuously monitor a multitude of geographically distributed streams. Finally, memory management is another important aspect of optimizing CER systems for processing big data streams.

In centralized approaches the goal is to achieve high throughput with low recognition time and a small memory footprint. To this end, a number of techniques are utilized in an attempt to cope with the volume and velocity big data event streams. The most important of these techniques are query rewriting, predicate-related optimizations and memory management. Query rewriting is an optimization technique that allows a suboptimal query expression to be rewritten in a form that is more efficient to execute. The goal is for the rewritten query to produce exactly the same results as the original one, while exhibiting improved performance w.r.t. the optimization objectives. Most approaches to query rewriting use a set of operators that allow to translate an event pattern into a semantically equivalent form, which allows for more efficient execution. Predicate-related optimizations use early event predicate evaluation to optimize the execution of queries for matching event patterns on the input stream. This is achieved by properly partitioning the input stream and filtering the selected events that will actually be part of complex event detection based on the query. Memory management techniques focus on optimizing event buffers by e.g. removing pieces of information stored multiple times across different buffers. We refer to (Flouris et al. 2017) for a detailed presentation of such techniques.

The aforementioned techniques for scaling-up the CER process are generic, i.e. applicable to all CER approaches discussed earlier in this section (i.e. automata-based, tree-based or logic-based). In addition to such generic techniques, different CER approaches use special techniques to further increase their efficiency. For instance, automata-based approaches, which typically use non-deterministic automata, need to store at runtime all possible candidate runs, where each run depends on non-deterministic choices such as ignoring or consuming an event, and different runs result in different outputs. The maintained set of runs rapidly becomes very large (Zhang, Diao, and Immerman 2014), since it grows exponentially with the number of events in the temporal window under consideration. To cope with that, automata-based systems store the set of candidate runs in a compressed form, by e.g. factoring-out commonalities between different runs (Mozafari et al. 2013), or storing only so-called maximal runs from which other runs can be efficiently computed (Zhang, Diao, and Immerman 2014). Logic-based CER systems also resort to specialized techniques to tame the complexity of logical inference mechanisms, by e.g. translating rules into more efficient structures to perform incremental recognition as new events become available. Examples include temporal constraint networks (Dousson and Maigat 2007) and automata (Cugola and Margara 2010). Limiting the scope of inference via windowing techniques is also used in logic-based approaches, such as RTEC (Artikis, Sergot, and Paliouras 2015). Specialized techniques towards enhancing performance are also used to scale-up probabilistic CER systems, where the event recognition task is typically harder than in crisp CER systems. An overview of such techniques may be found in (Flouris et al. 2017).

Distributed CER consists of two main approaches. The first is to centralize the monitoring of the stream and distribute the complex event processing to multiple sites, as proposed in (Schultz-Møller, Migliavacca, and Pietzuch 2009). The second is to distribute the monitoring of the stream to multiple sites (where each cite receives one input stream) and centralize the processing effort, as proposed by (Akdere, Çetintemel, and Tatbul 2008). The first of these approaches seeks to improve throughput, as well as memory management. Optimizing throughput is achieved thanks to the fact that the total number of input events is distributed across multiple nodes, thus overall the system processes more events per time unit. Memory management is also improved in this processing model, since distributed processing allows for dealing with larger time windows. In the second approach to distributed CER (Akdere, Çetintemel, and Tatbul 2008) multiple input event streams are received at multiple sites and a coordinator node communicates with all sites to detect complex events. In this strategy the goal is to optimize the tradeoff between the latency in detecting complex events and the cost of communicating with the coordinator node. An example of such an approach is presented in (Akdere, Çetintemel, and Tatbul 2008), where the authors use pareto-optimality theory to generate monitoring plans for the distributed processing that conform to particular communication cost and latency constraints.

5.3.4 Machine Learning for Complex Event Recognition

Manual authoring of complex event patterns is a difficult task that requires significant effort. Moreover, event patterns need frequent updating to cope with the drifting nature of streaming data. Therefore, machine learning techniques that are able to extract event patterns from data, or revise existing ones as new observations become available are highly desirable. Both supervised and unsupervised techniques have been employed to automatically construct and adapt event definitions. Widely used unsupervised learning techniques include frequency-based analysis of sequences of events (Vautier, Cordier, and Quiniou 2007), or clustering of such sequences (Lee and Jung 2017). Such approaches are promising for discovering unknown events but are limited to propositional learning, therefore they cannot be used to learn complex event patterns expressing relations between events or attributes thereof. Moreover, these techniques are hard to adapt towards learning the structure of complex events that are not frequent in the data – for instance in cases where the goal is to learn event patterns of abnormal behaviour.

A few approaches to supervised learning of complex event patterns have been proposed. In (Margara, Cugola, and Tamburrelli 2014) the authors propose a combination of techniques for learning patterns in the TESLA language (Cugola and Margara 2010), however, their approach is relatively ad-hoc, it is hard to evaluate in more mainstream machine learning settings and has limited support for the incorporation of background knowledge in the learning process. In (Mousheimish, Taher, and Zeitouni 2017) the authors use an existing method for shapelet learning (extracting patterns from time-series data) and propose a technique for temporally combining the extracted shapelets to form event patterns over multiple streams. Patterns learnt with this approach have limited expressive power, while background knowledge is also hard to utilize.

A common feature of all the above-mentioned techniques is that they assume a batch learning setting, where the training data are available before learning begins and the generated models cannot be updated in the face of new data that stream-in. Given the streaming nature of big data flows in CER, machine learning techniques for learning complex event patterns must be capable of learning in an online fashion.

A different line of work towards machine learning of event patterns has been put forth in logic-based CER approaches. Using logical formalisms as a basis for CER allows access to well-established machine learning techniques from the fields of Inductive Logic Programming (ILP) (Raedt 2008) and Statistical Relational Learning (Raedt et al. 2016), which allow to learn patterns expressing arbitrarily complex relations and constraints between events and event attributes, while easily utilizing rich domain knowledge in the process. For instance, in (Carrault et al. 2003) the authors use an off-the-shelf ILP

system to learn complex event patterns in the chronicle formalism (Dousson and Maigat 2007). Moreover, online learning techniques have been proposed in event calculus-based CER approaches (Katzouris, Artikis, and Paliouras 2016; Michelioudakis et al. 2016).

5.3.5 Complex Event Recognition for Mobility Data

CER techniques are becoming increasingly important in a wide range of applications involving mobile objects, where real-time situational awareness is a requirement. Traffic/transport monitoring in intelligent transportation systems (Dasarathy 2011) is a prominent application domain. To give but a few examples, in (Terroso-Saenz et al. 2012) the authors use sensor data from a vehicular network, in addition to environmental and weather data to detect different levels of traffic jams with an event processing methodology, while in (Michelioudakis, Artikis, and Paliouras 2016) data from on-vehicle sensors and sensors mounted on road segments are used to learn complex event patterns for the early detection of traffic jams. In (Terroso-Saenz et al. 2015) a CER-based approach is proposed that allows to detect interesting situations related to the passengers' comfort and security, from data originating from sensors installed in different parts of the vehicle. Related approaches are presented in (Artikis et al. 2013, 2014), where the authors propose CER-based techniques towards the detection of events related to congestion and quality of service in intelligent transport management applications.

Maritime surveillance is another CER application domain related to mobility data. In (Patroumpas et al. 2017) the authors propose a system for online monitoring of maritime activity over streaming positions from numerous vessels sailing at sea. The system employs an online tracking module for detecting important changes in the evolving trajectory of each vessel across time, and thus can incrementally retain concise, yet reliable summaries of its recent movement. In addition, thanks to a CER module, this system is also capable for offering instant notification to marine authorities regarding emergency situations, such as suspicious moves in protected zones, or package picking at open sea. A related approach is put forth in (Boubeta-Puig et al. 2012) where the authors propose a CER-based methodology for detecting vessel communication hijacking or failure, engine malfunction or ship collision. In (Terroso-Saenz, Valdés-Vela, and Skarmeta-Gómez 2016), CER is used to detect illegal and/or dangerous activities in the maritime domain, such as collisions, smuggling or human trafficking.

Distributed processing is of utmost importance in mobility-related applications, such as those addressed in T&K. In such applications, massive data volumes are collected at different sites (e.g. moving vehicles) and much of the processing needs to take place in situ, since moving data around for centralized analysis incurs excessive communication costs. Equally important is the development of machine learning techniques for extracting and updating interesting complex event patterns from data, in order to e.g. discover abnormal mobility patterns, which domain experts have not yet identified. The requirement is for distributed, online machine learning, capable of handling the volume and velocity of data streams in mobility-related applications.

5.4 References

Agrawal, R., Imielinski, T., Swami, A. (1993) Mining association rules between sets of items in large databases. Proceedings of ACM SIGMOD.

Agrawal, R., Srikant, R. (2014) Mining sequential patterns. Proceedings of ICDE.

Akdere, M., U. Çetintemel, and N. Tatbul (2008) Plan-Based Complex Event Detection across Distributed Sources. PVLDB 1 (1): 66–77.

Alarabi, L., Mokbel, M.F. (2017) A Demonstration of ST-Hadoop: A MapReduce Framework for Big Spatio-temporal Data. PVLDB 10(12), 1961–1964.

Alarabi, L., Mokbel, M.F., Musleh, M. (2017) ST-Hadoop: A MapReduce Framework for Spatio-Temporal Data. Proceedings of SSTD.

Alevizos, E., A. Skarlatidis, A. Artikis, and G. Paliouras (2017) Probabilistic Complex Event Recognition: A Survey. ACM Comput. Surv. 50 (5): 71:1–71:31. https://doi.org/10.1145/3117809.

Anicic, D., P. Fodor, S. Rudolph, R. Stühmer, N. Stojanovic, and R. Studer (2011) ETALIS: Rule-Based Reasoning in Event Processing. In Reasoning in Event-Based Distributed Systems, 99–124. Studies in Computational Intelligence. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-19724-6_5.

Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999). OPTICS: ordering points to identify the clustering structure. ACM SIGMOD record, 28(2), 49-60.

Artikis, A., A. Margara, M. Ugarte, S. Vansummeren, and M.s Weidlich (2017) Complex Event Recognition Languages: Tutorial. In Proceedings of the 11th ACM International Conference on Distributed and Event-Based Systems, DEBS 2017, Barcelona, Spain, June 19-23, 2017, 7–10. ACM. https://doi.org/10.1145/3093742.3095106.

Artikis, A., A. Skarlatidis, F. Portet, and G. Paliouras (2012) Logic-Based Event Recognition. Knowledge Eng. Review 27 (4): 469–506. https://doi.org/10.1017/S0269888912000264.

Artikis, A., M. J. Sergot, and G. Paliouras (2015) An Event Calculus for Event Recognition. IEEE Trans. Knowl. Data Eng. 27 (4): 895–908. https://doi.org/10.1109/TKDE.2014.2356476.

Artikis, A., M. Weidlich, A. Gal, V. Kalogeraki, and D. Gunopulos (2013) Self-Adaptive Event Recognition for Intelligent Transport Management. In Proceedings of the 2013 IEEE International Conference on Big Data, 6-9 October 2013, Santa Clara, CA, USA. https://doi.org/10.1109/BigData.2013.6691590.

Artikis, A., M. Weidlich, F. Schnitzler, I. Boutsis, T. Liebig, N. Piatkowski, C. Bockermann, et al. (2014) Heterogeneous Stream Processing and Crowdsourcing for Urban Traffic Management. In Proceedings of the 17th International Conference on Extending Database Technology, EDBT 2014, Athens, Greece, March 24-28, 2014. https://doi.org/10.5441/002/edbt.2014.77.

Ayhan, S., Samet, H. (2016) Aircraft Trajectory Prediction Made Easy with Predictive Analytics. Proceedings of ACM SIGKDD 2016.

Ayhan, S., Samet, H. (2016) Time Series Clustering of Weather Observations in Predicting Climb Phase of Aircraft Trajectories. Proceedings of IWCTS 2016.

Bakshy, E., I. Rosenn, C. Marlow, and L. Adamic (2012) The role of social networks in information diffusion. WWW 2012 - Session: In- formation Diffusion in Social Networks April 16-20, 2012, Lyon, France, pages 519–528.

Bhattacharya, A., & Das, S. (1999) LeZi-update: an information-theoretic approach to track mobile users in PCS networks. Proceedings of ACM MobiCom.

Boubeta-Puig, J., I. Medina-Bulo, G. Ortiz, and G. Fuentes-Landi (2012) Complex Event Processing Applied to Early Maritime Threat Detection. In Proceedings of the 2Nd International Workshop on Adaptive Services for the Future Internet and 6th International Workshop on Web APIs and Service Mashups, 1–4. WAS4FI-Mashups '12. New York, NY, USA. https://doi.org/10.1145/2377836.2377838.

Brenna, L., A. J. Demers, J. Gehrke, M. Hong, J. Ossher, B. Panda, M. Riedewald, M. Thatte, and W. M. White (2007) Cayuga: A High-Performance Event Processing Engine. In Proceedings of the ACM SIGMOD International Conference on Management of Data, Beijing, China. https://doi.org/10.1145/1247480.1247620.

Brilhante, I. R., M. Berlingerio, R. Trasarti, C. Renso, J. A. F. d. Macedo and M. A. Casanova (2012) ComeTogether: Discovering Communities of Places in Mobility Data. Proceedings of IEEE 13th

International Conference on Mobile Data Management, Bengaluru, Karnataka. doi: 10.1109/MDM.2012.17

Brockmann, D. (2009) Human mobility and spatial disease dynamics. Reviews of nonlinear dynamics and complexity, pp. 1–24. Wiley-VCH, Weinheim.

Bussas, M., Sawade, C., Kühn, N. et al. Machine Learning Journal (2017). 106: 1419. https://doi.org/10.1007/s10994-017-5639-3

Cao, H., Mamoulis, N., & Cheung, D. W. (2006). Discovery of collocation episodes in spatiotemporal data. Proceedings of 6th IEEE Conference on Data Mining.

Carrault, G., M.-O. Cordier, R. Quiniou, and F. Wang (2003) Temporal Abstraction and Inductive Logic Programming for Arrhythmia Recognition from Electrocardiograms. Artificial Intelligence in Medicine 28 (3): 231–263. https://doi.org/10.1016/S0933-3657(03)00066-6.

Cha, M., A. Mislove, and K. P. Gummadi (2009) A measurement- driven analysis of information propagation in the flickr social network. In Proceedings of the 18th international conference on World wide web, pages 721–730. ACM.

Chawla, S., Y. Zheng, and J. Hu (2012) Inferring the root cause in road traffic anomalies. In Proceedings of the 12th IEEE International Conference on Data Mining. IEEE, 141-150.

Chen, C.C., Tseng, C.Y., Chen, M.S. (2013) Highly Scalable Sequential Pattern Mining Based on MapReduce Model on the Cloud. Proceedings of IEEE International Congress on Big Data.

Chen, X.W., Landry, S.J., Nof, S.Y. (2011) A framework of enroute air traffic conflict detection and resolution through complex network analysis. Computers in Industry 62, 8 (2011), 787–794.

Chen, Z., Shen, H. T., & Zhou, X. (2011). Discovering popular routes from trajectories. Proceedings of IEEE Conference on Data Engineering.

Cheng, T., Cui, D., Cheng, P. (2003) Data mining for air traffic flow forecasting: a hybrid model of neural network and statistical analysis. Proceedings of ITSC 2003.

Claramunt, C., Ray, C., Camossi, E., Jousselme, A.L., Hadzagic, M., Andrienko, G., Andrienko, N., Theodoridis, Y., Vouros, G., Salmon, L. (2017) Maritime data integration and analysis: recent progress and research challenges. Proceedings of EDBT.

Coppenbarger, R.A. (1999) En Route Climb Trajectory Prediction Enhancement Using Airplane Flight-Planning Information. American Institute of Aeronautics and Astronautics, AIAA-99-4147.

Cugola, G. (2012) Processing Flows of Information: From Data Stream to Complex Event Processing. ACM Comput. Surv. 44 (3): 15:1–15:62. https://doi.org/10.1145/2187671.2187677.

Cugola, G., A. Margara, M. Matteucci, and G. Tamburrelli (2015) Introducing Uncertainty in Complex Event Processing: Model, Implementation, and Validation. Computing 97 (2): 103–44. https://doi.org/10.1007/s00607-014-0404-y.

Cugola, G., and A. Margara (2010) TESLA: A Formally Defined Event Specification Language. In Proceedings of the Fourth ACM International Conference on Distributed Event-Based Systems, DEBS 2010, Cambridge, United Kingdom. https://doi.org/10.1145/1827418.1827427.

Dasarathy, B. V. (2011) A Special Issue on Intelligent Transportation Systems. Information Fusion 12 (1): 1. https://doi.org/10.1016/j.inffus.2010.06.009.

de Leege, A., Van Paassen, M., Mulder, M. (2013) A machine learning approach to trajectory prediction. Proceedings of AIAA GNC 2013.

de Raedt, L. (2008) Logical and Relational Learning. Cognitive Technologies. Springer. https://doi.org/10.1007/978-3-540-68856-3.

de Raedt, L., K. Kersting, S. Natarajan, and D. Poole (2016) Statistical Relational Artificial Intelligence: Logic, Probability, and Computation. Synthesis Lectures on Artificial Intelligence and Machine Learning 10 (2): 1–189. https://doi.org/10.2200/S00692ED1V01Y201601AIM032.

Deng, Z., Hu, Y., Zhu, M., Huang, X., & Du, B. (2015). A scalable and fast OPTICS for clustering trajectory big data. Cluster Computing, 18(2), 549-562.

Di Ciccio, C., var der Aa, H., Cabanillas, C., et al. (2016) Detecting flight trajectory anomalies and predicting diversions in freight transportation. Decision Support Systems 88 (2016), 1–17.

Ding, X., Chenz, L., Gao, Y., Jensenz, C.S., Bao, H. (2018) UlTraMan: A Unified Platform for Big Trajectory Data Management and Analytics. Proceedings of VLDB'18.

Doulkeridis, C., Vlachou, A., Mpestas, D., Mamoulis, N. (2017) Parallel and Distributed Processing of Spatial Preference Queries using Keywords. Proceedings of EDBT.

Dousson, C., and P. Le Maigat (2007) Chronicle Recognition Improvement Using Temporal Focusing and Hierarchization. In IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India. http://ijcai.org/Proceedings/07/Papers/050.pdf.

Eldawy, A., Mokbel, M.F. (2016) The Era of Big Spatial Data: A Survey. Foundations and Trends in Databases 6(3-4), 163–273.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of KDD.

Fan, Q., Zhang, D., Wu, H., & Tan, K. L. (2016). A general and parallel platform for mining co-movement patterns over large-scale trajectories. Proceedings of the VLDB Endowment, 10(4), 313-324.

Fang, Y., Cheng, R., Tang, W., Maniu, S., Yang, X.S. (2016) Scalable Algorithms for Nearest-Neighbor Joins on Big Trajectory Data. IEEE Trans. Knowl. Data Eng. 28(3), 785–800.

Flouris, I., N. Giatrakos, A. Deligiannakis, M. N. Garofalakis, M. Kamp, and M. Mock (2017) Issues in Complex Event Processing: Status and Prospects in the Big Data Era. Journal of Systems and Software 127: 217–236. https://doi.org/10.1016/j.jss.2016.06.011.

Fort, M., Sellarès, J. A., & Valladares, N. (2014). A parallel GPU-based approach for reporting flock patterns. International Journal of Geographical Information Science, 28(9), 1877-1903.

Fortunato S. (2010) Community detection in graphs. Physics Reports. 486, 75–174.

Gomes, J., Phua, C., & Krishnaswamy, S. (2013) Where will you go? Mobile data mining for next place prediction. Proceedings of DaWaK.

Gong, C., McNally, D. (2004) A methodology for automated trajectory prediction analysis. Proceedings of AIAA GNC 2004.

Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L. (2008) Understanding individual human mobility patterns. Nature 453, 779–782.

Goo, J. (2010) Transfer learning for classification of spatially varying data. PhD dissertation at University of Texas at Austin, Department of Electrical and Computer Engineering. http://hdl.handle.net/2152/ETD-UT-2010-08-1962.

Hadjaz, A., Marceau, G., Saveant, P., et al. (2012) Online learning for ground trajectory prediction. CoRR abs/1212.3998.

Hagedorn, S., Räth, T. (2017) Efficient spatio-temporal event processing with STARK. Proceedings of EDBT.

Hamed, M., Gianazza, D., Serrurier, M., & Durand, N. (2013) Statistical prediction of aircraft trajectory: regression methods vs point-mass model. Proceedings of ATM.

Han, J., Pei, J. (2000) Mining frequent patterns by pattern-growth: methodology and implications. ACM SIGKDD Explor. Newsl., 2(2), 14–20.

Hendawi, A.M., Mokbel, M.F. (2012a) Panda: A Predictive Spatio-Temporal Query Processor. Proceedings of ACM SIGSPATIAL GIS'12.

Hendawi, A.M., Mokbel, M.F. (2012b) Predictive Spatio-Temporal Queries: A Comprehensive Survey and Future Directions. Proceedings of ACM SIGSPATIAL MobiGIS'12.

Hendawi, A.M., Bao, J., Mokbel, M.F. (2013) iRoad: A Framework For Scalable Predictive Query Processing on Road Networks. Proceedings of VLDB'13.

Hendawi, A.M., Bao, J., Mokbel, M.F., Ali, M. (2015a) Predictive Tree: An Efficient Index for Predictive Queries on Road Networks. Proceedings of ICDE 2015.

Hendawi, A.M., Ali, M., Mokbel, M.F. (2015b) A Framework for Spatial Predictive Query Processing and Visualization, Proc. of the 16th IEEE International Conference on Mobile Data Management.

Hendawi, A.M., Ali, M., Mokbel, M.F. (2017) Panda ∗: A generic and scalable framework for predictive spatio-temporal queries. GeoInformatica, 21(2), 175-208.

Holme, P., & Saramäki, J. (2012). Temporal networks. Physics reports, 519(3), 97-125.

Hossmann, T., T. Spyropoulos and F. Legendre (2011) A complex network analysis of human mobility. 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Shanghai, pp. 876-881. doi: 10.1109/INFCOMW.2011.5928936

Ishikawa, Y., Tsukamoto, Y., & Kitagawa, H. (2004) Extracting mobility statistics from indexed spatio-temporal datasets. Proceedings of STDBM.

Jagadish, H.V. (1990) On indexing line segments. Proceedings of VLDB.

Jalili, M., Orouskhani, Y., Asgari, M., Alipourfard, N., & Perc, M. (2017). Link prediction in multiplex online social networks. Royal Society open science, 4(2), 160863.

Jensen, C., Lin, D., & Ooi, B. (2004) Query and update efficient B+-tree based indexing of moving objects. Proceedings of VLDB.

Jeung, H., Yiu, M.L., Zhou, X., Jensen, C.S. (2010) Path prediction and predictive range querying in road network databases. VLDB Journal 19 (2010), 585-602.

Jinno, R., Seki, K., & Uehara, K. (2012). Parallel distributed trajectory pattern mining using MapReduce. Proceedings of IEEE Cloud Computing Technology and Science.

Kalnis, P., Mamoulis, N., & Bakiras, S. (2005). On discovering moving clusters in spatio-temporal data. Proceedings of International Symposium on Spatial and Temporal Databases.

Kasteren, T. L. M. van, Englebienne, G., Krose, B. J. A. (2010) Transferring Knowledge of Activity Recognition across Sensor Networks. Pervasive Computing pp 283-300.

Katzouris, N., A. Artikis, and G. Paliouras (2016) Online Learning of Event Definitions. TPLP 16 (5–6): 817–833. https://doi.org/10.1017/S1471068416000260.

Krumm, J., Horvitz, E. (2003) Predestination: inferring destinations from partial trajectories. Proceedings of UbiComp 2003.

Lan, R., Yu, Y., Cao, L., Song, P., & Wang, Y. (2017). Discovering Evolving Moving Object Groups from Massive-Scale Trajectory Streams. Proceedings of IEEE Conference on Mobile Data Management.

Laube, P., Imfeld, S., & Weibel, R. (2005a). Discovering relative motion patterns in groups of moving point objects. International Journal of Geographical Information Science, 19(6), 639-668.

Laube, P., van Kreveld, M., & Imfeld, S. (2005b). Finding REMO—detecting relative motion patterns in geospatial lifelines. Developments in Spatial Data Handling. Springer.

Le Fablec, Y., Alliot, J.M. (1999) Using neural networks to predict aircraft trajectories. Proceedings of ICIS 1999.

Lee, J. G., Han, J., & Whang, K. Y. (2007). Trajectory clustering: a partition-and-group framework. Proceedings of ACM SIGMOD International Conference on Management of Data.

Lee, O.-J., and J. E. Jung (2017) Sequence Clustering-Based Automated Rule Generation for Adaptive Complex Event Processing. Future Generation Comp. Syst. 66: 100–109. https://doi.org/10.1016/j.future.2016.02.011.

Leskovec, J., M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst (2007) Cascading Behavior in Large Blog Graphs. sdm, pages 1–21.

Liben-Nowell D, Kleinberg J (2007) The link prediction problem for social networks. J Am Soc Inform Sci Technol 58(7):1019–1031

Liu, F., D. Deng, and P. Li (2017) Dynamic Context-Aware Event Recognition Based on Markov Logic Networks. Sensors 17 (3): 491. https://doi.org/10.3390/s17030491.

Liu, H., Huang, X., Wen, D., Li, J. (2017) The Use of Landscape Metrics and Transfer Learning to Explore Urban Villages in China. Remote Sens., 9, 365.

Lu L, Zhou T (2011) Link prediction in complex networks: a survey. Phys A Stat Mech Appl 390(6):1150–1170

Margara, A., G. Cugola, and G. Tamburrelli (2014) Learning from the Past: Automated Rule Generation for Complex Event Processing. In The 8th ACM International Conference on Distributed Event-Based Systems, DEBS '14, Mumbai, India. https://doi.org/10.1145/2611286.2611289.

Matsuno, Y., Tachiya, T., Wei, J., et al. (2015) Stochastic optimal control for aircraft conflict resolution under wind uncertainty. Aerospace Science and Technology 43 (2015), 77–88.

Mei, Y., and S. Madden (2009) ZStream: A Cost-Based Query Processor for Adaptively Detecting Composite Events. In Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2009, Providence, Rhode Island, USA. ACM. https://doi.org/10.1145/1559845.1559867.

Michelioudakis, E., A. Artikis, and G. Paliouras (2016) Online Structure Learning for Traffic Management. In Inductive Logic Programming - 26th International Conference, ILP 2016, London, UK. Springer. https://doi.org/10.1007/978-3-319-63342-8_3.

Michelioudakis, E., A. Skarlatidis, G. Paliouras, and A. Artikis (2016) $$\mathtt {OSL}\alpha $$: Online Structure Learning Using Background Knowledge Axiomatization. In Machine Learning and Knowledge Discovery in Databases, 232–47. Lecture Notes in Computer Science. Springer, Cham. https://doi.org/10.1007/978-3-319-46128-1_15.

Monreale, A., Pinelli, F., Trasarti, R., & Giannotti, F. (2009) WhereNext: a location predictor on trajectory pattern mining. Proceedings of ACM SIGKDD.

Mousheimish, R., Y. Taher, and K. Zeitouni (2017) Automatic Learning of Predictive CEP Rules: Bridging the Gap between Data Mining and Complex Event Processing. In Proceedings of the 11th ACM International Conference on Distributed and Event-Based Systems, DEBS 2017, Barcelona, Spain. ACM. https://doi.org/10.1145/3093742.3093917.

Moussalli, R., Absalyamov, I., Vieira, M. R., Najjar, W., & Tsotras, V. J. (2015). High performance FPGA and GPU complex pattern matching over spatio-temporal streams. GeoInformatica, 19(2), 405-434.

Moussalli, R., Vieira, M. R., Najjar, W., & Tsotras, V. J. (2013). Stream-mode fpga acceleration of complex pattern trajectory querying. Proceedings of International Symposium on Spatial and Temporal Databases.

Mozafari, B., K. Zeng, L. D'Antoni, and C. Zaniolo (2013) High-Performance Complex Event Processing over Hierarchical Data. ACM Trans. Database Syst. 38 (4): 21:1–21:39. https://doi.org/10.1145/2536779.

Nanni, M., & Pedreschi, D. (2006). Time-focused clustering of trajectories of moving objects. Journal of Intelligent Information Systems, 27(3), 267-289.

Newman, M. E. J. (2003) The structure and function of complex networks. SIAM Review. 45, 2, 167–256.

Nikitopoulos, P., Paraskevopoulos, A., Doulkeridis, C., Pelekis, N., Theodoridis, Y. (2018) Hot Spot Analysis for Big Trajectory Data. Proceedings of SSDBM'18 (submitted).

Ord, J. K., Getis, A. (1995) Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geographical Analysis 27(4), 286–306.

Pan, S. J. and Yang, Q. (2010) A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, Volume: 22, Issue: 10, pp. 1345-1359.

Panagiotakis, C., Pelekis, N., Kopanakis, I., Ramasso, E., Theodoridis, Y. (2012). Segmentation and sampling of moving object trajectories based on representativeness. IEEE Transactions on Knowledge and Data Engineering, 24(7), 1328-1343.

Patroumpas, K., E. Alevizos, A. Artikis, M. Vodas, N. Pelekis, and Y. Theodoridis (2017) Online Event Recognition from Moving Vessel Trajectories. GeoInformatica 21 (2): 389–427. https://doi.org/10.1007/s10707-016-0266-x.

Patwary, M. M. A., Palsetia, D., Agrawal, A., Liao, W. K., Manne, F., & Choudhary, A. (2013). Scalable parallel OPTICS data clustering using graph algorithmic techniques. Proceedings of IEEE Int. Conf. on High Performance Computing, Networking, Storage and Analysis.

Pelekis, N., Kopanakis, I., Kotsifakos, E. E., Frentzos, E., & Theodoridis, Y. (2011). Clustering uncertain trajectories. Knowledge and Information Systems, 28(1), 117-147.

Pelekis, N., Kopanakis, I., Panagiotakis, C., & Theodoridis, Y. (2010). Unsupervised trajectory sampling. Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases.

Pelekis, N., Tampakis, P., Vodas, M., Doulkeridis, C., & Theodoridis, Y. (2017). On temporal-constrained sub-trajectory cluster analysis. Data Mining and Knowledge Discovery, 31(5), 1294-1330.

Pelekis, N., Theodoridis, Y. (2014) Mobility Data Management and Exploration. Springer.

Rabiner, L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77, 2 (1989), 257–286.

Ré, C., J. Letchner, M. Balazinska, and D. Suciu (2008) Event Queries on Correlated Probabilistic Streams. In Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008, Vancouver, BC, Canada. ACM. https://doi.org/10.1145/1376616.1376688.

Rinzivillo, S., Gabrielli, L., Nanni, M., Pappalardo, L., Pedreschi, D., Giannotti F. (2014) The purpose of motion: Learning activities from individual mobility networks. Proceedings of International Conference on Data Science and Advanced Analytics (DSAA).

Rossetti, G., & Cazabet, R. (2018). Community Discovery in Dynamic Networks: A Survey. ACM Computing Surveys (CSUR), 51(2), 35.

Rossetti, G., Pappalardo, L., Kikas, R., Pedreschi, D., Giannotti, F., Dumas, M. (2016) Homophilic network decomposition: a community-centric analysis of online social services. Social Network Analysis and Mining.

Saberi, M., Mahmassani, H.S., Brockmann, D. et al. A complex network perspective for characterizing urban travel demand patterns: graph theoretical analysis of large-scale origin–destination demand networks. Transportation (2017) 44: 1383. https://doi.org/10.1007/s11116-016-9706-6

Sacharidis, D., Patroumpas, K., Terrovitis, M., Kantere, V., Potamias, M., Mouratidis, K., & Sellis, T. (2008). On-line discovery of hot motion paths. Proceedings of the ACM 11th international conference on Extending database technology: Advances in database technology.

Samet, H. (1990) The Design and Analysis of Spatial Data Structures. Addison-Wesley.

Schultz-Møller, N. P., M. Migliavacca, and P. R. Pietzuch (2009) Distributed Complex Event Processing with Query Rewriting. In Proceedings of the Third ACM International Conference on Distributed Event-Based Systems, DEBS 2009, Nashville, Tennessee, USA. ACM. https://doi.org/10.1145/1619258.1619264.

Sip, S., Green, S.M. (2003) Common Trajectory Prediction Capability for Decision Support Tools. ATM 5th USA/Europa R&D seminar, Budapest.

Skarlatidis, A., A. Artikis, J. Filipou, and G. Paliouras (2015) A Probabilistic Logic Programming Event Calculus. TPLP 15 (2): 213–245. https://doi.org/10.1017/S1471068413000690.

Skarlatidis, A., G. Paliouras, A. Artikis, and G. A. Vouros (2015) Probabilistic Event Calculus for Event Recognition. ACM Trans. Comput. Log. 16 (2): 11:1–11:37. https://doi.org/10.1145/2699916.

Song, Y., Cheng, P., Mu, C. (2012) An improved trajectory prediction algorithm based on trajectory data mining for air traffic management. Proceedings of IEEE ICIA 2012.

Szor P. (2004) Fighting computer virus attacks. USENIX

Tang, M., Yu, Y., Malluhi, Q.M., Ouzzani, M., Aref, W.G. (2016) LocationSpark: A Distributed In-Memory Data Management System for Big Spatial Data. PVLDB 9(13), 1565–1568.

Tao, Y., Faloutsos, C., Papadias, D., & Liu, B. (2004) Prediction and indexing of moving objects with unknown motion patterns. Proceedings of ACM SIGMOD.

Tao, Y., Papadias, D., & Sun, J. (2003) The TPR*-tree: an optimized spatio-temporal access method for predictive queries. Proceedings of VLDB.

Tastambekov, K., Puechmorel, S., Delahaye, D., et al. (2014) Aircraft trajectory forecasting using local functional regression in Sobolev space. Transportation research part C: Emerging technologies 39 (2014), 1–22.

Tayeb, J., Ulusoy, Ö., & Wolfson, O. (1998) A quadtree-based dynamic attribute indexing method. The Computer Journal, 41(3), 185-200.

Terroso-Saenz, F., M. Valdés-Vela, C. Sotomayor Martínez, R. Toledo-Moreo, and A. F. Gómez-Skarmeta (2012) A Cooperative Approach to Traffic Congestion Detection With Complex Event Processing and VANET. IEEE Trans. Intelligent Transportation Systems 13 (2): 914–929. https://doi.org/10.1109/TITS.2012.2186127.

Terroso-Saenz, F., M. Valdés-Vela, F. Campuzano, J. A. Botía, and A. F. Gómez-Skarmeta (2015) A Complex Event Processing Approach to Perceive the Vehicular Context. Information Fusion 21: 187–209. https://doi.org/10.1016/j.inffus.2012.08.008.

Terroso-Saenz, F., M. Valdés-Vela, and A. F. Skarmeta-Gómez (2016) A Complex Event Processing Approach to Detect Abnormal Behaviours in the Marine Environment. Information Systems Frontiers 18 (4): 765–780. https://doi.org/10.1007/s10796-015-9560-7.

Theodoridis, S., Koutroumbas, K. (2008) Pattern Recognition (4/e). Academic Press.

Thiemann, C., F. Theis, D. Grady, R. Brune, D. Brockmann (2010) The Structure of Borders in a Small World. PLoS One. 5(11): e15422. https://doi.org/10.1371/journal.pone.0015422

Thipphavong, D.P., Schultz, C.A., Lee, A.G., et al. (2013) Adaptive Algorithm to Improve Trajectory Prediction Accuracy of Climbing Aircraft. Journal of Guidance, Control and Dynamics (JGCD) 36(1), 15–24.

Tian, J., J. Hahner, C. Becker, I. Stepanov and K. Rothermel (2002) Graph-based mobility model for mobile ad hoc network simulation. Proceedings 35th Annual Simulation Symposium. SS'2002, pp. 337-344.

Tran, S. D., and L. S. Davis (2008) Event Modeling and Recognition Using Markov Logic Networks. In Computer Vision – ECCV 2008, 610–23. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-88688-4_45.

Trasarti, R., Guidotti, R., Monreale, A., & Giannotti, F. (2017) MyWay: Location Prediction via mobility profiling. Information Systems, 64, 350-367.

Tsung, F., Zhang, K., Cheng, L, Song (2017). Statistical transfer learning: A review and some extensions to statistical process control, Quality Engineering, 30:1, 115-128, DOI: 10.1080/08982112.2017.1373810

Valladares Cereceda, I. (2013). GPU parallel algorithms for reporting movement behaviour patterns Proceedings of Spatiotemporal Databases.

Vautier, A., M.-O. Cordier, and R. Quiniou (2007) Towards Data Mining Without Information on Knowledge Structure. In Knowledge Discovery in Databases: PKDD 2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland. Springer. https://doi.org/10.1007/978-3-540-74976-9_29.

Verhein, F., & Chawla, S. (2006) Mining spatio-temporal association rules, sources, sinks, stationary regions and thoroughfares in object mobility databases. Proceedings of DASFAA.

Viterbi, A. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory 13(2), 260–269.

Vouros, G.A., Vlachou, A., Santipantakis, G., et al. (2018) Big data analytics for time critical mobility forecasting: recent progress and research challenges. Proceedings of EDBT 2018.

Wang D., Pedreschi, D., Song, C., Giannotti, F., & Barabasi, A. L. (2011). Human mobility, social ties, and link prediction. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1100-1108). ACM.

Wang, L., Geng, X., Ke, J., Peng, C., Ma, X., Zhang, D., Yang, Q. (2017) Ridesourcing Car Detection by Transfer Learning. Eprint arXiv:1705.08409.

Wang, P., Gonzalez, M.C., Menezes, R., Barabasi, A.L. (2013) Understanding the spread of malicious mobile-phone programs and their damage potential. IJIS

Wang, X., Wang, J., Wang, T., Li, H., Yang, D. (2010) Parallel Sequential Pattern Mining by Transaction Decomposition. Proceedings of 7th Int. Conf. on Fuzzy Systems and Knowledge Discovery.

Wei, Y., Zheng, Y., and Yang, Q. (2016) Transfer Knowledge between Cities. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16).

Whitman, R.T., Park, M.B., Marsh, B.G., Hoel, E.G. (2017) Spatio-Temporal Join on Apache Spark. Proceedings of ACM SIGSPATIAL GIS'17.

Woolley-Meza, O., Thiemann, C., Grady, D., Lee, J., Seebens, H., Blasius, B. and Brockmann, D., (2011), Complexity in human transportation networks: a comparative analysis of worldwide air transportation and global cargo-ship movements, The European Physical Journal B: Condensed Matter and Complex Systems, 84, issue 4, p. 589-600.

Xia, F., J. Wang, X. Kong, Z. Wang, J. Li and C. Liu (2018) Exploring Human Mobility Patterns in Urban Scenarios: A Trajectory Data Perspective. IEEE Communications Magazine, vol. 56, no. 3, pp. 142-149. doi: 10.1109/MCOM.2018.1700242.

Xian, Y., Liu, Y., Xu, C. (2016) Parallel gathering discovery over big trajectory data. Proceedings of IEEE International Conference on Big Data.

Yang, J., & Hu, M. (2006) TrajPattern: Mining sequential patterns from imprecise trajectories of mobile objects. Proceedings of EDBT.

Yang, Y., Zhang, J., Cai, K.Q. (2015) Terminal-area aircraft intent inference approach based on online trajectory clustering. The Scientific World Journal, 671360 (2015).

Yavas, G., Katsaros, D., Ulusoy, Ö., & Manolopoulos, Y. (2005) A data mining approach for location prediction in mobile environments. Data and Knowledge Engineering, 54(2), 121-146.

Ying, J.-C., Lee, W.-C., Weng, T.-C., & Tseng, V. (2011) Semantic trajectory mining for location prediction. Proceedings of ACM SIGSPATIAL.

Zhang, D., He, T., Zhang, F., & Xu, C. (2018). Urban-Scale Human Mobility Modeling With Multi-Source Urban Network Data. IEEE/ACM Transactions on Networking. DOI: 10.1109/TNET.2018.2801598

Zhang, H., Y. Diao, and N. Immerman (2014) On Complexity and Optimization of Expensive Queries in Complex Event Processing. In International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA. ACM. https://doi.org/10.1145/2588555.2593671.

Zhang, M., Chen, S., Jensen, C.S., Ooi, B.C., Zhang, Z. (2009) Effectively Indexing Uncertain Moving Objects for Predictive Queries. Proceedings of VLDB '09.

Zhang, R., Qi, J., Lin, D., Wang, W., Chi-WingWong, R. (2012) A highly optimized algorithm for continuous intersection join queries over moving objects, VLDB Journal 21 (2012), 561-586.

Zheng, Y., F. Liu, and H.-P. Hsieh (2013) U-Air: when urban air quality inference meets big data. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13), New York, NY, USA, 1436-1444. ACM. DOI: https://doi.org/10.1145/2487575.2488188

Zheng, Y. (2015) Trajectory Data Mining: An Overview. Transactions on Intelligent Systems and Technology 6(3), 1–41.

Zhu, X., Li, B., Wu, X., He, D., Zhang, C. (2011) CLAP: Collaborative pattern mining for distributed information systems. Decision Support Systems 52, pp. 40-51.

Ziv, J. & Lempel, A. (1978) Compression of individual sequences via variable-rate coding. IEEE Transactions on Information Theory, 24(5), 530–536.

Zorbas, N., Zissis, D., Tserpes, K., & Anagnostopoulos, D. (2015) Predicting object trajectories from high-speed streaming data. Proceedings of IEEE TrustCom/BigDataSE/ISPA.

6 Big Data Visualization and Visual Analytics

The main idea of Visual Analytics (VA) is to develop knowledge, methods, technologies and practices that exploit and combine the strengths of human and electronic data processing (Keim et al. 2008a). Visualization is the means through which humans and computers cooperate using their distinct capabilities for the most effective results. Visual analytics is "the science of analytical reasoning facilitated by interactive visual interfaces" (Thomas & Cook, 2005a, p. 4), which focuses on developing human-computer methods and procedures for data analysis, knowledge building, and problem solving (Keim et al. 2010a). Visual analytics leverages methods and tools developed in other areas related to data analytics, particularly statistics, machine learning and geographic information science (Andrienko & and Andrienko,2012a). Visual analytics tools address the complex task of presenting multi-dimensional information in several displays of interactive 2D projections. The main components of a VA tool consist of a set of selectors and aggregators in a combination with visual facilities to drill down into the available information and to maintain synchronized displays.

6.1 Visual Analytics for Big Mobility Data

The content of this section is based on (Andrienko & and Andrienko,2012a), Andrienko et al. (2017) (Andrienko et al. 2017a) and (Andrienko et al. 2016c). Most of the text below was taken from these papers and only slightly adapted.

6.1.1 Transportation Data

(Andrienko et al. 2017a) focus on (i) data, (ii) movements and people behavior, and (iii) modeling and planning.

Data and Data transformations

The proposed data typology distinguishes spatial events (bound to a certain location and lasting for a limited time), trajectories (chronologically ordered records describing position of a moving object) spatially referenced time series (chronologically ordered sequences of values of time-variant thematic attributes associated with fixed spatial locations or stationary spatial objects). Trajectories are either quasi-continuous (when it is possible to plausibly estimate intermediate positions) or episodic (in the extreme case only the origin and destination of the trajectory are known). A set of interesting representation methods for the different types can be found in (Andrienko et al. 2017a).

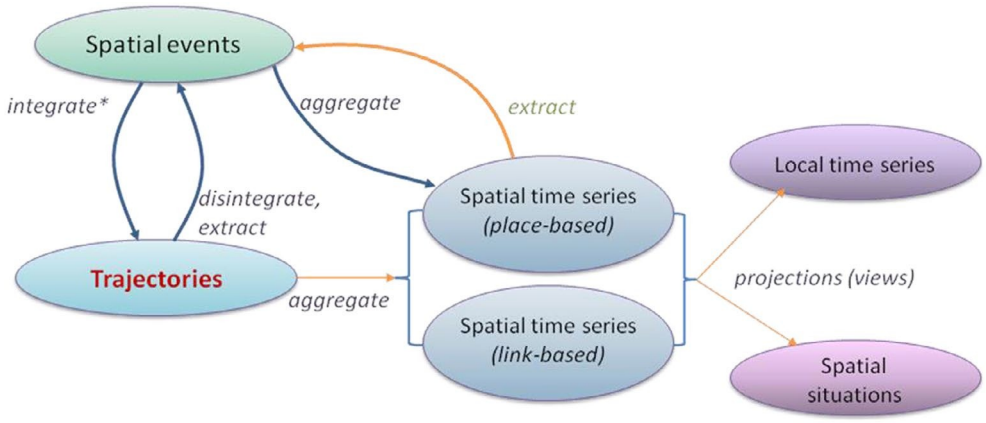


Figure 12: The principal transformations applicable to movement data, depending on the task/analysis goal (taken from (Andrienko et al. 2017a)).

A summary of possible transformations between the spatiotemporal data types is presented in Figure 12. The left part of the diagram shows the tight relationships between spatial events and trajectories.

In fact, trajectories consist of spatial events: each record in a trajectory of an object represents a spatial event of the presence of this object at a specific location at some moment in time.

Other transformations may be beneficial for particular tasks. For example, (Chu et al. 2014a) transform trajectories of taxis into sequences of the names of the traversed streets and apply text mining methods for discovery of "taxi topics", i.e., combinations of streets that have a high probability of co-occurrence in one taxi trip.

Andrienko et al. (2017a; 2017b) present an overview of visual representations and interactive techniques for detailed exploration of following classes of data: (i) individual movements, (ii) sets of taken routes, (iii) movement dynamics along a particular route, (iv) sets of origin-destination pairs, (v) collective movement over a territory, (vi) events (including extraction of events), (vii) contextualized movement (e.g. by integrating with meteorological data), (viii) impacts and risks (e.g. exposure to pollutants which is also a form of contextualizing).

Movement and people behavior

VA provides tools to investigate the use of transportation means by people. The existing techniques analyze the spatial and temporal patterns and trends, reveal behavioral differences between user groups, and relate the use of transport to the spatial and temporal context and people's activities. Human mobility behaviors over public transit systems are commonly explored to identify commute patterns and reveal behavioral differences. For example, (Wood et al. 2011a) and (Beecham & Wood, 2014a) visualize and analyze the dynamic patterns of a bicycle hire scheme in London.

(Laharotte et al. 2015a) used Bluetooth detectors in Brisbane to create B-OD matrices to describe the dynamics of a subpopulation of vehicles to characterize urban networks. van der Hurk et al. (Hurk et al. 2015a) present a methodology for extracting passenger routes based on smart card data from the Netherlands Rail System. (Kieu et al. 2015a) explored the use of smart card data for passenger segmentation.

(Kruger et al. 2013a) develop an interaction technique, TrajectoryLenses. Complex filter expressions are supported by the metaphor of an exploration lens, which can be placed on an interactive map to analyze geospatial regions for the number of trajectories, covered time, or vehicle performance. Another work by Kruger et al. (2015a) enriches the trajectories of the scooter users with semantic information concerning the visited places to infer users' activities and travel purposes. Semantic insights of points of interest are discovered from social media services. The uncertainties in time and space, which result from noisy, imprecise, and missing data, are visually analyzed by the geographic map view and a temporal view of OD patterns.

Mass mobility

The works described in this subsection deal with analyzing people's collective mobility behavior, i.e., mass movements. This includes routine daily and weekly patterns as well as anomalies due to extraordinary events.

Von Landesberger et al. (Landesberger et al. 2016a) present an approach to explore daily and weekly temporal patterns of collective mobility, where the source data are episodic trajectories of people reconstructed from georeferenced tweets or mobile phone use records. The trajectories are aggregated into flows between territory compartments by hourly intervals within the weekly time cycle.

(Beecham et al. 2014b) present a technique for automatically identifying com- muting behavior based on a spatial analysis of cyclists' journeys. They use visual analytics to compare the output of various workplace identification methods to explore data transformations and present insights to analysts in order to develop origin-destination theories of commute patterns.

(Ma et al. 2016a) also develop methods for studying urban flow. This work uses cell phone location records to approximate trajectories across a city, and flow volumes, links, and communities of users are visualized to help analysts identify typical patterns of movement within the city.

Similarly, work by Yang et al. (Yang et al. 2016a) focuses on identifying human mobility hotspots based on mobile phone location data from Shenzhen, China.

Yang et al. applies kernel density estimation and clusters identified hotspots based on the temporal signatures to identify spatial locations with high travel demand.

Work by (Chae et al. 2014a) develops a visual analytics framework for exploring public behavior before, during, and after disaster events. This work utilizes geographically referenced Tweets to create movement trajectories during disasters to identify evacuation flows. Interactions allow users to drill down into the data to also look at the underlying discourse occurring around the movements. Infrastructure data, disaster data (such as hurricane tracks), and Twitter data are all provided as map overlays in order to enable decision support and analysis.

People's activities and interests

In order to understand the current use of transportation systems and plan for expansion and development, it is helpful to understand the reasons why people travel, i.e., the activities and interests related to traveling. Recent work by Andrienko et al. (Andrienko et al. 2016a) presents a procedure for obtaining data similar to personal daily mobility diaries. Such a diary reports what places were visited by a person during a day, at what times, and for what purposes. The presented procedure aims at extracting similar information from long term sequences of spatio-temporal positions of people, which may come from georeferenced tweets or from mobile phone use records. From these sequences, the proposed procedure extracts repeatedly visited personal and public places along with the times these places were visited within the daily and weekly cycles. An interactive interface involving techniques for multi-criteria evaluation and ranking supports assignment of probable meanings ('home', 'work', 'eating', 'shopping', etc.) to subsets of places based on visit times and information about the land use or point of interest categories at these places. The analysis is done in a privacy-respectful manner without accessing individual data.

Modeling and planning

This section reviews research in visual analytics concerned with traffic modeling and transportation planning. This includes the derivation of models from data, applications of traffic forecasting and simulation models, transportation scheduling, and the exploration of decision options.

There is a series of works showing how predictive models of vehicle traffic can be derived from historical data consisting of a large number of vehicle trajectories (Andrienko & Andrienko,2013b), (Andrienko et al. 2015a), (Andrienko et al. 2016b). The approach is based on spatial abstraction and aggregation of the trajectory data into collective movements (flows) of the vehicles between territory compartments. The authors discovered that the dependencies between the traffic intensities and mean velocities in an abstracted transportation network at different levels of abstraction (Figure 13) have the same shapes as in the fundamental diagram of the traffic flow described in traffic theory (Gazis,2002a). While the fundamental diagram refers to links of a physical street network, it turns out that similar relationships also exist in abstracted networks. These dependencies can be represented by formal models (Figure 14).

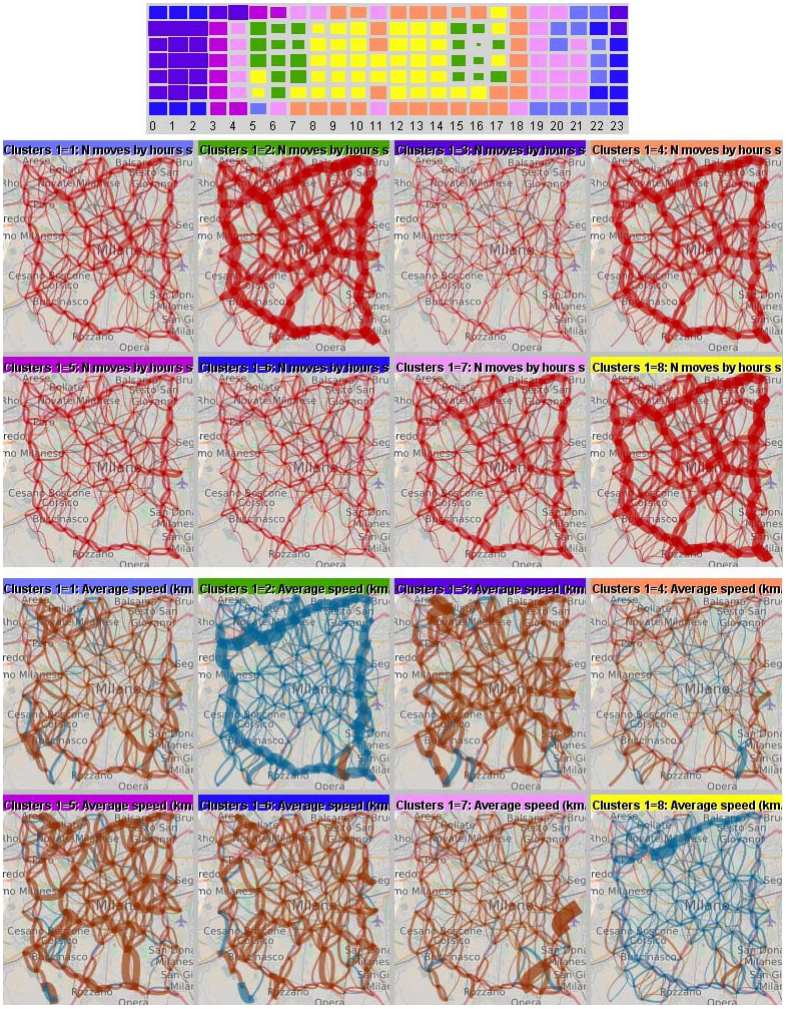


Figure 13: Hourly time intervals over a week have been clustered by the similarity of the spatial situations in terms of the flow magnitudes and average speeds. In a time matrix at the top, the rows correspond to the days from Sunday to Saturday and columns to the day hours. The time intervals are represented by rectangles colored according to the cluster membership; the sizes show the closeness to the cluster centers. Below, representative spatial situations for the clusters are shown by flow maps. In the upper set of 8 maps, the widths of the flow symbols are proportional to the mean flow magnitudes. The lower set of 8 maps represents how the mean speeds in the clusters differ from the median mean speed attained on the links. Positive and negative differences are encoded by proportional widths of flow symbols colored in brown and blue, respectively (Andrienko et al. 2013a).

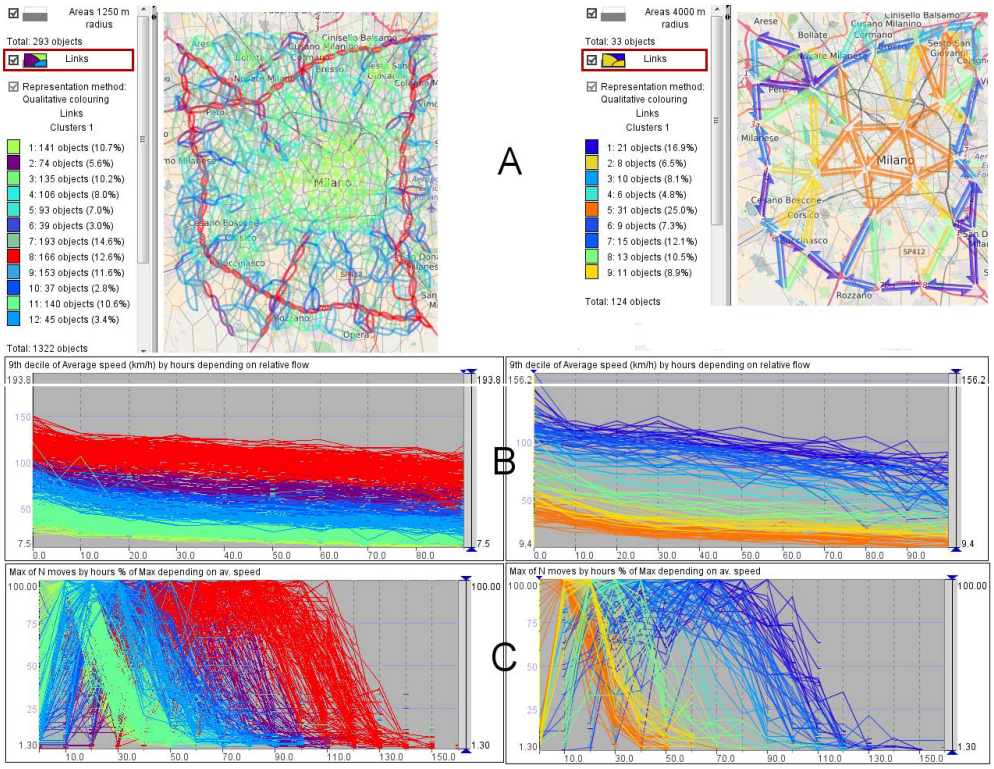


Figure 14: Dependencies between the traffic flow intensities (hourly volumes) and mean velocities on the links of an abstracted transportation network at different levels of abstraction. A: Abstracted networks with the cell radii of about 1250 m (left) and 4000 m (right). The links are clustered and colored according to the similarity of the volume-speed dependencies. B: The dependencies of the mean velocity (vertical dimension) on the traffic flows (horizontal dimension): the velocities decrease as the flows increase. C: The dependencies of the flows (vertical dimensions) on the velocities (horizontal dimension): maximal flows can be achieved for certain velocities and decrease for both lower and higher velocities (Andrienko et al. 2015a).

Historical traffic data can be used not only for predicting future movements under various conditions but also for spatial planning applications. For example, the system SmartAdP (Liu et al. 2017a) uses interactive visual tools to find suitable locations for billboard placement using taxi trajectories.

The task of transportation scheduling is addressed by (Andrienko, 2008a). An example application is planning of evacuation of different groups of people, such as general population, schoolchildren, and hospital patients, from a disaster-affected area. The proposed system consists of a scheduling algorithm and a set of visual displays and interactive tools for exploring scheduling outcomes. The displays al- low the user to detect problems, such as delays, understand their reasons, and find appropriate corrective measures.

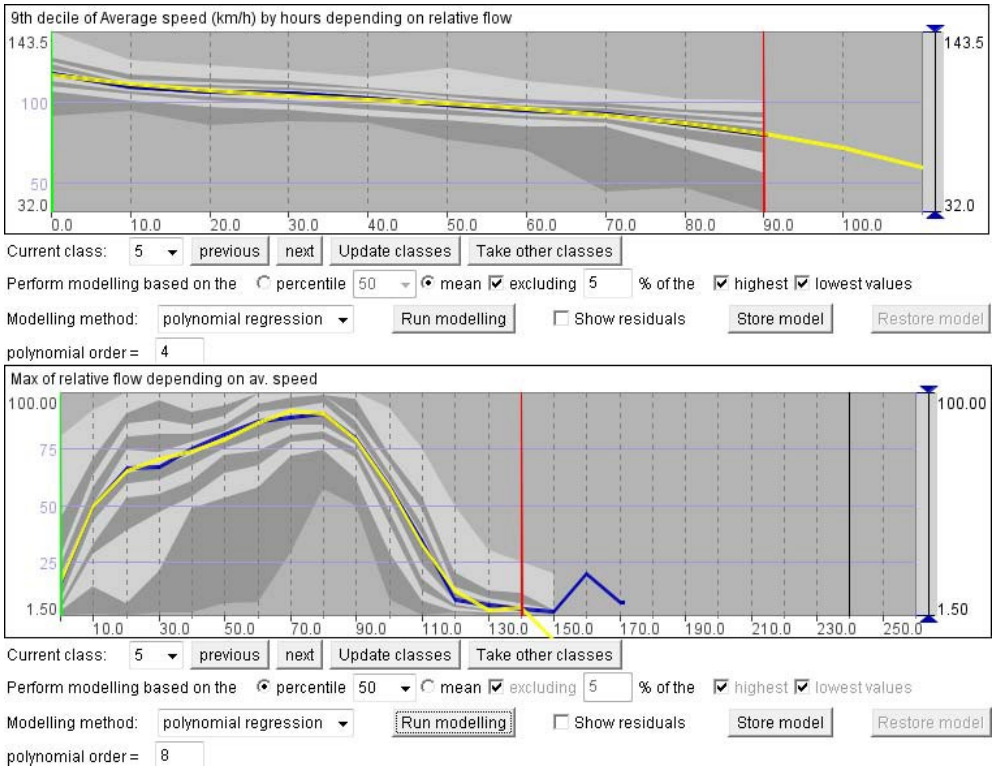


Figure 15: For one of the clusters of links of an abstracted transportation network (see Fig. 2.5), the dependencies flow ⇢ velocity (top) and velocity ⇢ flow (bottom) are being represented by polynomial regression models (Andrienko & Andrienko,2013b).

6.1.2 Assessing data quality

Abstracting from the various specific technologies for collecting movement data, we identify several major methods of position recording (Andrienko,2008c): (i) Location-based, (ii) Time-based, (iii) Change-based, (iv) Event-based and (v) Combinations of these basic approaches. In particular, GPS tracking devices may combine time-based and change-based recording: the positions may be measured at regular time intervals, but recorded only when significant changes of position, speed or direction occur.

As in (Andrienko et al. 2008c) two classes of properties are distinguished:

1. Data structure related properties
   - Mover set properties
   - Spatial properties
   - Temporal properties
2. Data collection procedure related properties
   - position exactness,
   - positioning accuracy,
   - missing positions and
   - meanings of the position absence

Data quality problems are classified using their kind, the data to which the problem applies, the extent of the problem in the dataset and the extent of the problem in the respective value domain. A problem labeling system is introduced in Andrienko et al. (2016) and presented here using a slightly adapted (by adding punctuation) notation.

Problem kind: **M** = missing data, **A** = accuracy problem, P = precision deficiency

Data components: **Mv** = mover identification, **S** = spatial position, **T** = tem- poral reference, **At** = thematic attributes

Extent in the trajectory: **TrE** = elementary (in some elements), **TrI** = inter- mediate (in particular subsequences), **TrO** = overall (in whole trajectory)

Error occurrence can be formulated using formulas similar to M: TrO ∪ MvO ∪ SO ∪ TO which allows for systematic enumeration of error types and the VA tools required to detect and remove problems. Detailed examples are given in Andrienko et al. (2016).

Each kind of problem mentioned in section 3.3 can have several causes and emanations (forms of occurrence) and hence may require multiple types of VA detection tools. Several techniques to validate the data quality w.r.t. several criteria are listed in Andrienko et al. (2016). Such criteria may be problem or dataset specific and hence not expected in advance (e.g. the positional shift of particular traces in space). An inherent property (see main characteristics) of VA is the ad hoc interactive way of the data handling procedures. Therefore, a generic toolbox needs to be created based on the anticipated use cases and prior experience.

The presented examples show cases where the problem was discovered and identified using visual analytics: in big data this will no longer be possible and new methods need to be found based on criteria involving properties of the trajectory, the set of trajectories and the environment. Problem detection by interactive use of VA tools will be much less probable due to the large data size; hence the importance of prior enumeration of potential errors and the availability of a tool to define dataset specific validation criteria.

6.1.3 VA in Transportation Science

In Andrienko et al. (2017), the authors observe that the contribution of transportation scientists in VA projects is rather limited although VA typically requires the help of domain experts. *"Unfortunately, such work has been limited in the transportation domain (Ferreira et al. 2013a), (Fredrikson et al. 1999a), even though visual analytics researchers have intensively worked with transportation relevant data and developed a variety of methods and tools that could be useful for transportation domain researchers and practitioners. . . . The consequences of the insufficient communication are two-fold. On the one hand, visual analytics researchers have only limited understanding of the problems, needs, and constraints of the transportation domain, which may decrease the potential utility and usability of the methods they develop. On the other hand, the transportation community has quite limited awareness of what visual analytics can offer.* In the conclusion the authors observe: *"Both the visual analytics community and transportation community has produced a large body of exploratory research work in analyzing transportation-related data. However, the knowledge acquired and methods developed often lack collaboration between the two communities."*

Clearly the problems solved by the VA community do not sufficiently coincide with the problems encountered by the transportation research community. Potential causes are:

- In (big) data the travel/movement purpose is often missing and needs to be de- rived from the movement and environment properties which requires prediction models because traveling individuals have particular beliefs, attitudes and goals defining their behavior. VA tools can be used to describe the resulting travel behavior but delivering predictive models is difficult.
- Data are extremely expensive (and out of reach of researchers and practitioners in the transportation domain). Pre-processed data can be purchased from the data generators/owners (telecom and ICT companies) in some cases but few information is made available about the pre-processing methods (aggregation, pseudonymisation, etc). Few affordable data are available to researchers in charge to answer specific research

questions related to a given area. On the other hand VA tools are used by industrial companies who have data available (e.g. https://www.be-mobile.com/products/flowcheck/)

This section lists some of the classes of currently open research questions, challenges and opportunities to unite VA with the transportation field needs.

Thereto, we need to keep in mind that travel demand is a derived demand. In most cases, the trip is not a goal by itself but a mean to achieve a goal. In activity-based modeling it is assumed that for a given period of time the goals and intentions of individuals determine their time use and displacements. In this case there is a hidden model (much more complex than a HMM) that controls the observed behavior and that in many cases cannot be derived from mobility big data. Understanding the behavioral model is a prerequisite for the evaluation of intended travel demand measures (TDM). (Andrienko & Andrienko,2017b) explain the integration of state diagrams describing travel behavior in VA.

Research Questions

Problems solved by transportation planners can be classified as follows by the size of the set of travelers involved.

1. problems involving a small set of affected individuals:
    a. Operations research (OR) like solutions to particular problems (short term, static environment): the aim is to provide advice to customers of transportation services. It requires streaming data (parking, expected congestion, incident detection, . . .)
    b. Traffic safety research is interested in hot spots, their physical properties (road infrastructure design) and the usage profile (timing, vehicle properties, movement properties). Incidents are often related to detailed interactions between one actor and the environment (requiring detailed description of local situations) or between multiple actors (requiring detailed description of their behavior in a short period preceding the (near) accident).
2. problems involving a large set of affected people
    a. short term (one week to one year): problems due to planned sudden local changes in a static environment. Traffic guidance measures can be put in place based on flow prediction models. VA can be used to show the recent (1 hour) history and the near future prediction in order to generate advice to the transportation system operation management.
    b. long term: TDM's effect prediction (long term, evolving environment, evolving actors): advice to designers of transportation systems (urban planners, regional development managers, public service, ...). As mentioned above, activity-based modeling is used to this end. Following travel generators can be distinguished:
        i. households: members execute frequently recurring activities (for several purposes) as well as exceptional ones. All activities aim to the achievement of a private goal/purpose (e.g. by maximizing utility de- rived from activity execution).
        ii. business/commercial: trips are executed on behalf of or indirectly triggered by households and aiming to achieve of a goal stated by a third party (example: parcel delivery)

Activity based modelling nearly exclusively focuses on travel generated by households. It is important to estimate the number and properties of business/commercial since modelling travel demand induced by companies turns out to be problematic due to lack of data. Big data and VA may contribute to solve the problem by classifying daily movements of vehicles from revealed trajectories w.r.t the distance driven, the number of trips in a tour (number of visited places),

clustering visited locations etc. Classes may coincide with taxi-like pattern, sales/service-person like pat- tern, packet-delivery like pattern. After finding patterns their share in the flow may be estimated. Particular attention is to be paid to bias in the available datasets.

Data availability challenges

1. In order to solve some of the above questions, *person* traces instead of *car* traces are required. These may be collected by smartphones but induce more privacy problems than car traces and are more difficult to process (because of trip and mode detection).
2. In order to capture behavioral aspects of travel, longitudinal (as opposed to cross-sectional) data and privacy preserving measures are required. These requirements may be mutually incompatible (see the issue of mover identity errors mentioned earlier.
3. Streaming data may originate from parking use or as floating car data; however no such data will be available in the project (none are known to be available for research)

RAM problem

VA tools need to be adapted to the use of large datasets because they typically use RAM based datasets. In (Andrienko & and Andrienko,2012a) the authors propose to use a stepwise approach by which data aggregations are created in advance and loaded into RAM for processing by VA tools.

Dedicated procedures need to be designed for data quality assessment (see also section 3.3 because preprocessing by aggregation-based methods may be insufficient

6.2 Visual Analytics for Complex Event Recognition

Few research works focus specifically on Complex Event Recognition in Visual Analytics: in fact, CER research covers various formal languages and models which are hard to be studied in a same unique framework. However, main ingredients in CER are multivariate time series and event sequences. In these domains, the visualization of data streams have attracted increased interest the last decade. The research work reviewed in the following focus mainly on visualization of multivariate time series and of simple event patterns, without elaborate first-order logic or entanglement of events as CER can deal with. Nevertheless, these techniques are of interest to T&K, providing basic bricks to understand the challenge of CER visualization. The final part of the review is dedicated to dimension reduction techniques allowing to summarize in a few dimensions the original data lying in a high dimensional space. These techniques have been successfully adapted to multivariate time series data and it will be interesting in the T&K project to investigate the hybridation of usual visual analytics techniques and dimension reduction approach, especially in the context of representational learning.

Number of tools have been developed to visualize raw sequences of events. For instance, LifeLines (Plaisant et al. 1996) is dedicated to personal medical histories visualization providing an hierarchical timeline visualization. Events are placed on a time-line accordingly to their occurrences; color encoding and lines with different thickness illustrate the relationship between the events. CloudLines (Krštajic et al. 2011) use a logarithmic time scale to allow the visualization of recent events together with long term patterns at any time scale. ChronoLense (Zhao et al. 2011) emphasizes on the usage of different lenses (possibly with the user interaction) to navigate through the time series. These techniques are well adapted to explore with precision time series and individual traces but do not aggregate or factorize the event streams and thus do not scale with large event collections or many time series.

Other approaches try to summarize multivariate time series in a high dimensionality setting in order to visualize the data. TimeSeer (Dang et al. 2013) proposes the notion of *scagnostics* to capture the characteristics of the data based on a set of measure (density, skewness, …). Scatterplot matrix and charts are used next to interactively explore the data. ThemeRiver (Havre et al. 2000) is another well

know technique to analyze variation and exchange between event streams. RankExplorer (Shi et al. 2012) extends ThemeRiver in a framework for big stream data analysis. In the context of spatio-temporal data analysis, (Tominski et al. 2012) proposes to stack 2D visualization in a 3D representation to analyze trajectories; (Scheepens et al. 2011) uses density fields to capture the multivariate aspect of the time series.

Another way to summarize multivariate time series is to factorize the event streams thanks to a tree structure: LifeFlow (Wongsuphasawat et al. 2011) and EventFlow (Maguire et al. 2013) aggregate time series with respect to common subsequences to provide a tree-like visualization. However, the tree structure grows exponentially with the number of events which hinders the visualization. Graph structures can also be used to analyze time series: Sankey diagram (Riehmann et al. 2005) and derivate techniques (Wongsuphasawat, 2012) are techniques intensively used to analyze the event flow between states of a system. To improve the readability of the Sankey diagram when the graph of states is dense, MatrixFlow (Perer et al. 2012) represents spatio-temporal data with sequences of matrices; MatrixWave (Zhao et al. 2015) improves the use of matrices by apprehending the event flow between the states.

An orthogonal family of approaches is the dimension reduction methods. The objective of this kind of approach is to compress the time series expressed in a high dimensional space into a very few number of dimensions – ideally two – losing the less possible important information. The low dimension projected space can be used to visualize meaningfully the originally high dimensional data. The most used technique, the Principal Component Analysis (PCA) (Candès et al. 2011), which use linear projections to reduce the number of dimensions, has been adapted to temporal data in various works (Yang & Shahabi, 2004) (Liao, 2005). Derived from PCA, Dictionary Learning techniques (Mairal et al. 2009) use combination of weighted atoms to approximate the time series, each atom representing an elementary behavior shared among number of time series. The Non negative Matrix Factorization method (Cichocki et al. 2009) introduces a constraint on the recombination of the atoms, allowing only additive reconstruction. This leads to a better interpretation of the atoms.

Another very active domain in dimension reduction topic is the Representation Learning (Bengio et al. 2013), leaded by the research on deep learning (Lecun et al. 2015). The objective is to learn non linear projection – *embeddings* - of the data in an unsupervised context generally thanks to a deep neural network. A special case is the auto-encoder model, applied successfully to time series (Längkvist et al. 2014). The architecture involves an encoding module and a decoding one. The encoding module projects an input data into a latent space through different neuronal layers with a decreasing number of neurones. The decoder has a mirror architecture, with first layers containing the less neurones and the last layer the same amount as the first layer of the encoding module. The embedding representation in the latent space is learned by a measure of divergence between the input data and the output one: the network learns to reconstruct the original signal with the less distortion possible compressed in a low dimension space. As the layers are connected by non linear functions, the embedding are more expressive than with usual dictionary learning techniques.

Other techniques for dimension reduction uses similarity measures to embed the data in a low dimensional space. A widely used approach is the multidimensional scaling (Cox et al. 2001), organizing the data in a 2D space such as the distances in the projected space are close to the distance in the original space. The T-distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten & Hinton, 2008) is privileged for data in high dimension. The algorithm defines two probabilistic models over the distance between each couple of points, the first one in the original space, the second one in the low dimensional representation space. It learns the representation in order to optimize the minimal divergence between the two probability distributions.

6.3 Cross-scale analysis and dashboards for populations' mobility

6.3.1 Aggregate visualizations and cross-scale analysis

Aggregate visualizations of trajectory data in mobility analytics can be seen to pertain to two main areas: cross-scale analysis, used to adjust the analysis scale to the scale of the mobility pattern(s) studied; and aggregation of mobility-related parameters and trajectory geometries.

Due to the effect of internal and external factors influencing the movement at different spatial and temporal scales, behaviors may manifest different movement patterns at different scales (Nathan et al. 2008); see Figure 16. Therefore, recently, the importance of scale, and thus of cross-scale analysis has been acknowledged in the literature (Keim et al. 2008; Laube & Purves, 2011; Soleymani et al. 2014), and a number of methods and algorithms capable of investigating the relationships between patterns and processes occurring at multiple spatial and/or temporal scales of movement have been developed. Laube & Purves (2011) systematically explored the effects of computing movement parameters across different temporal scales, and thus demonstrated that adjusting the analysis scale is crucial to obtain meaningful results. Recently, more progress has been made, such as developing new multi-scale measures (e.g., multi-scale straightness index (Postlethwaite et al. 2013)), patterns detection using Brownian bridges in low sampling rate movement data (Buchin et al. 2012), measurement of dynamic interactions in movement (Long & Nelson, 2013), or the use of the discrete wavelet transform for movement classification and trajectory segmentation at different spatio-temporal scales (Soleymani at al. 2017). While relevant methods have been developed for cross-scale analysis in the spatial and temporal domain, it should be noted that the issue of scale also applies to semantic aspects of movement, where research is still very limited. To gain a comprehensive understanding of mobility, cross-scale analysis methods should be developed considering spatial, temporal as well as semantic aspects.

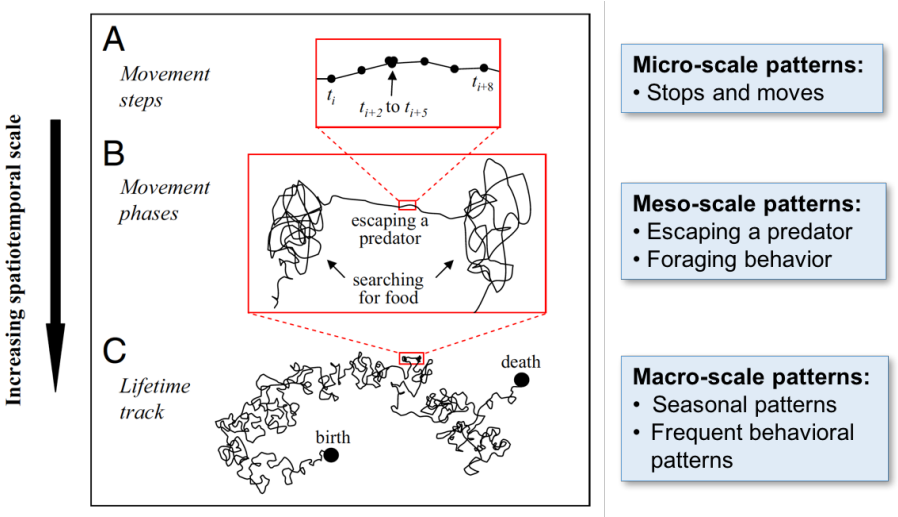


Figure 16: Mobility behaviors may manifest different movement patterns at different scales (i.e., spatial, temporal, or thematic scale). The above shows an example in animal ecology (adapted from Nathan et al. 2008).

Aggregation in support of visualization may affect either the computation of mobility-related parameters or the summarization of the trajectory geometries. Aggregation of mobility-related parameters may take place over different windows of time, different (possibly hierarchical) spatial units, or combination thereof, and may take the form of simple aggregation to advanced data modeling algorithms (Zhang et al. 2012). Andrienko & Andrienko (2008) introduce methods that allow aggregating parameters related to mobility and trajectories in the temporal, spatial and spatio-temporal domains.

Summarizing trajectory geometries typically follows two strands. In the first strand, the geometries (and associated movement parameters) are aggregated to cells of a tessellation. Examples of this approach include Lee et al. (2009), who use a tree data structure for aggregation, as well as Andrienko & Andrienko (2011), where the spatial aggregation units consist of Voronoi cells that capture the essential characteristics of the original trajectories around a subset of 'significant points'. In the second strand of approaches, the geometry of bundles of trajectories is represented by a 'placeholder' that approximates the shape and position of the trajectory bundle. Examples of this group include the TRACLUS clustering algorithm (Lee et al. 2007) and various decedents thereof, algorithms for generating centroid trajectories under positional uncertainty (Pelekis et al. 2011), and algorithms for creating median trajectories (Figure 17; Buchin et al. 2013).

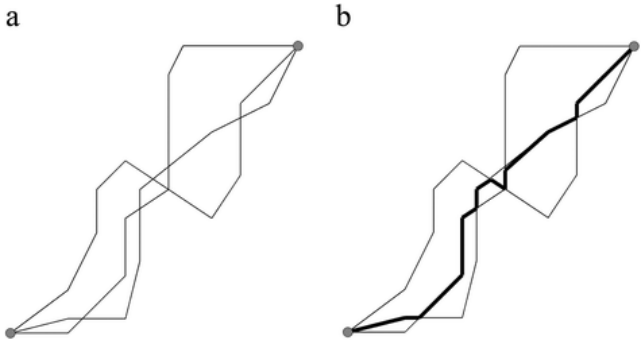


Figure 17: Trajectory aggregation: median trajectory. a) Three trajectories with a common start and end point; b) a median trajectory (bold) representing these three trajectories (Buchin et al. 2013).

6.3.2 Dashboards

Few (2006, p. 26) proposed a commonly accepted definition of a dashboard as "a visual display of the most important information needed to achieve one or more objectives, consolidated and arranged on a single screen so that the information can be monitored at a glance". It often combines text and graphs, with an emphasis on graphics (e.g., maps) to visually present the overview information (Figure 18). Dashboards allow users to explore their data, "not only in terms of spatio-temporal aspects, but also in terms of attribute aspects" (Rahman 2017, p. 1). By nature, dashboards are especially used for providing an overview as they visualize the most essential information at a glance. Very often a dashboard consists of a set of visualizations and controls, allowing interactions such as selection, filtering, and drilling down (Zhang et al., 2012). The use of dashboards has become popular in many fields, such as marketing (Krush et al. 2013), public health (Lechner & Fruhling 2014), construction (Guerriero et al. 2012), urban development (Scipioni et al. 2009), education (Maldonado et al. 2012), and transport monitoring (Rahman 2017).

There are different ways to categorize dashboards. Few (2006) classified them into three groups according to their roles: strategic, analytical and operational. Pappas & Whitman (2011) and Rahman (2017) provided a detailed comparison of these three groups, in terms of supporting scenarios, timeframe, graph presentation, interactivity, and update frequency.

- Strategic dashboards: They are the most popular type. They provide a quick overview of the data with the notion to enable decision makers to monitor the health and opportunities of the business. Very simple graphs that show what is going on without much interactivity work well for this type of dashboards (Rahman 2017). They don't need real-time data, but still need regular update, e.g., daily, weekly or monthly.
- Analytical dashboards: This group of dashboards often offer greater context than just simple overview in strategic dashboards. They show trends or patterns reflected in the data, and enable further explorations, such as drilling down into the underlying details. Like the previous

group, simple graphs work well for analytical dashboards, but with extensive interactivity to allow users to explore the details. They often use historical data.

- Operational dashboards: They represent the most dynamic type compared to the other two groups. They help to monitor situations and act as soon as possible according to particular conditions (Rahman 2017). Highly dynamic graphs such as animated displays work best. They can warn users of outliers or when something goes wrong. Usually, real-time data or near real-time data are required for operational dashboards.

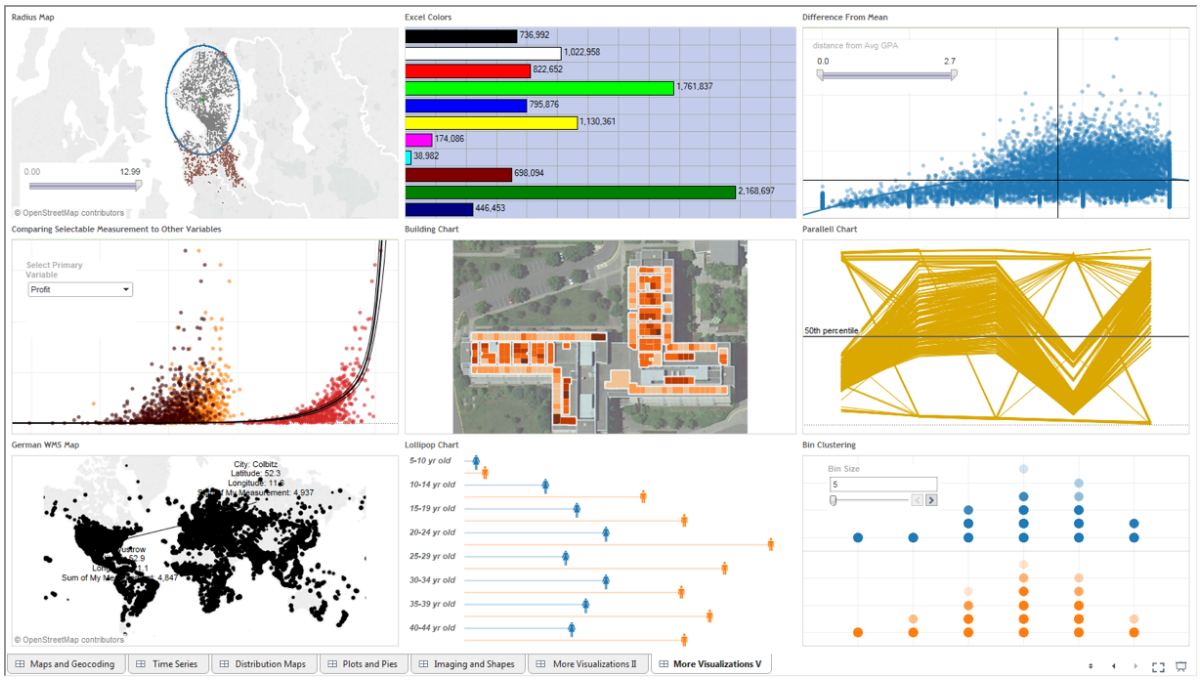


Figure 18: Dashboard examples created in the Tableau software (https://qph.ec.quoracdn.net/main-qimg-000650654f2338d4828bb89ab106ffa8). Different visualization methods can be integrated, such as charts, maps, and even simple texts.

Few (2006) reviewed many existing dashboards, and identified 13 common pitfalls of dashboard design, including: exceeding the boundaries of a single screen, supplying inadequate context for the data, displaying excessive detail, expressing measures indirectly, choosing inappropriate media of display, introducing meaningless variety, using poorly designed display media, encoding quantitative data inaccurately, arranging the data poorly, ineffectively highlighting what's important, cluttering the screen with useless decoration, misusing or overusing color, and designing unattractive visual displays.

To effectively communicate information, a dashboard should be properly designed, particularly using the right visualization (Rahman 2017). Ware (2012) and Few (2006) argued that research on visual perception provides empirical evidences on this aspect, such as those on short-term visual memory, visual encoding of rapid perception, and gestalt theory. Few (2006) further proposed two fundamental principles for appropriate visualizations: 1) it must be the best means that is commonly found, 2) it can be still functional even in a small space. Six types of visualizations were proposed: graphs, images, icons, drawing objects, text, and organizers. Pappas & Whitman (2011) further proposed guidance for choosing the right visualizations for dashboards, such as providing interactivity for strategic and analytic dashboards, and allowing comparisons. Zhang et al. (2012), ActiveWizards (2018) and Machlis (2017) all provide detailed reviews of commercial visualization systems that might be used for the implementation of dashboards.

6.4 Evaluating visual analytics procedures through eye-tracking

Eye tracking is the process of measuring and recording individual's gaze positions and eye movements. This technology is being increasingly used not only in psychology and product design but also in visualization and human-computer interaction sciences for evaluation of visual displays and user interfaces. Researchers employ eye tracking to understand how their designs are actually used and, possibly, even get insights into users' ways of reasoning and problem solving. In evaluating one design, researchers want to check if users' behaviours correspond to the supposed ways of use, see where the users may have difficulties, and understand how the design can be improved. In evaluating two or more alternative designs, researchers want to know not only which design is better in terms of task completion times and error rates but also why it is better: How does the use of this design differ from the use of the others? What is more difficult, confusing, or inconvenient in the other designs?

Eye tracking produces large amounts of data that are quite hard to analyse. The standard tools and methods for analysis of these data have rather limited capabilities. They can show where the users look first and where they look most and can compute basic measures (fixation count, time to the first fixation, statistics of the fixation durations and saccade lengths, etc.). However, these methods are hardly suitable for studying the spatio-temporal structure of eye scan paths, in particular, how the movements change over time while the user carries out a given task. For these purposes, we adopt movement analysis methods (see Andrienko et al. 2012).

6.4.1 Eye tracking vs. geographic movement data

Eye tracking data consist of records about the positions and times of gaze fixations. Each record includes the following components: user identifier, time, position in the display space (x- and y-coordinates), and fixation duration. The records may also include other attributes, e.g., stimulus identifier when different stimuli are used in the data collection. The temporally ordered sequence of records of one user referring to one stimulus is further called eye trajectory or scanpath, as in the literature on eye tracking.

Geographical movement data have the same structure: moving object identifier, time, and position (in geographical space) defined by coordinates; additional attributes may also be present. The structural similarity suggests that both classes of data may be analyzed using the same methods. However, there is a significant difference between eye movements and movements of physical objects governed by inertia: eye movements include instantaneous jumps (saccades) over relatively long distances (Dodge et al. 2009). The intermediate points between the start and end positions of a jump are not meaningful; it cannot be assumed that there exists a straight or curved line between two fixation positions such that the eye focus travels along it attending all intermediate points. This prohibits the use of methods involving interpolation between positions, as in creating movement density surfaces (Willems et al. 2009). Hence, not all movement analysis methods are valid for eye trajectories.

Another concern is whether the tasks for which a method was developed are relevant to eye movement analysis. For example, the methods intended to analyze collective simultaneous movements of multiple objects can hardly be useful in analyzing eye trajectories since simultaneous eye movements of two or more users viewing the same image are usually not tracked. Even if such data were collected, the eye foci of different users are unlikely to interact in the screen space similarly to interactions of material moving objects. Hence, not all movement analysis methods are meaningful for eye trajectories.

6.4.2 Analytical tasks in eye tracking studies

We use the term 'analytical task' to denote possible interests of eye movement analysts, i.e., the questions they may seek to answer. The possible types of tasks have been in part extracted and generalized from the eye tracking-related literature and in part generated during the study, when the

evaluation group posed their questions and the technology group, from their side, applied the methods to the data and looked what could be learned.

The possible tasks can be divided into two major categories: tasks focusing on areas of interest (AOIs) and tasks focusing on movements. The first category deals with the distribution of the user's attention over a display. It can be subdivided into several task types according to the following aspects:

- whether the AOIs are predefined (e.g., certain targets the users are supposed to search for) or need to be extracted from the data (e.g., elements/parts of an image attracting more attention);
- whether the evolution of the attention over time is of interest;
- whether the analyst needs general results for the entire set of users or looks for essential differences between individuals or groups (e.g., experts versus novices);
- whether the study is focused on a single display or compares two or more displays.

Common for these tasks is that only the fixations are analyzed and not the saccades or transitions between the AOIs. For example, Çöltekin et al. (2009) compare two interfaces by analyzing fixation durations and fixation counts for predefined AOIs.

In the second task category, the movements are of primary interest. AOIs are important, but the focus is on transitions between them and their temporal order. Analysts want to discover the users' strategies in visual exploration, search, and performing given tasks. They also want to understand whether and where the users have difficulties. Movement-focused tasks are indispensable in evaluation of information displays and user interfaces. This task category can be subdivided as follows:

- Examine the general characteristics of the movements, e.g., prevalence of long or short movements, presence of sharp turns, path complexity, etc.
- Examine the spatial patterns of the movements, e.g., jumps across large areas or gradual scanning, spatial clustering or dispersion, radial or circular moves, etc.
- Study the relation of the movements to the display content and/or structure, e.g., correspondence to the arrangement of the display elements, movements along available lines or figure boundaries, connections and transitions between the AOIs, etc.
- Understand individual viewing or searching strategies, compare to expected or theoretically optimal strategies.
- Understand general viewing or searching strategies of multiple users, find and interpret different types of activities.
- Find typical paths, e.g., as frequent sequences of attended AOIs.
- Detect and investigate indications of possible users' difficulties: returns to previous points, repeated movements, and cyclic scanning behaviors.

Like the AOI-focused tasks, the movement-focused tasks can be additionally classified according to the following aspects:

- whether the evolution of the eye movements over time is of interest;
- whether different users or groups are compared;
- whether different displays are compared.

Analysis of eye tracking data usually involves many tasks, which may require several analysis methods. The next section briefly reviews the methods that have been previously applied to eye tracking data. It shows that AOI-focused tasks are better supported by the standard methods than movement-focused tasks, which are therefore will be given more attention in our project.

6.4.3 Methods

There are many statistical metrics that can be derived from eye tracking data. Poole and Ball (2006) systemize these metrics and their possible interpretations. For example, high saccade/fixation ratio indicates more processing, large saccade amplitudes indicate more meaningful cues (as attention is

drawn from a distance), etc. However, eye movements cannot be fully understood just from those numbers. Visual analysis is essential for further insight.

The most popular tool to visually analyze eye tracking data is the attention heatmap (Bojko, 2009) showing the distribution of users' attention over the display space. Heatmaps can be easily generated using standard eye tracking software. They can visualize counts of fixations, counts of different users who fixated on different areas, absolute gaze duration, and relative gaze duration (percentage to the total time spent). Attention heatmaps may be useful for AOI-focused tasks. In comparative studies (different time intervals, different users, or different images) several heatmaps are compared. Eye tracking analysts also try to determine users' search strategies by analyzing series of heatmaps generated for consecutive time intervals (Poole & Ball, 2006), which show how the users' attention foci change over time. However, the characteristics of the eye movements, the links between the attention foci, and the paths followed during the search remain unclear.

Another visualization technique provided by standard software is the gaze plot, which represents fixations by circles with sizes proportional to the fixation durations and connects consecutive fixations by lines. Eye movement analysts usually admit that this method is not suitable for large data due to enormous overplotting (Çöltekin et al. 2010).

A common method suitable for movement-focused tasks is scanpath comparison (Duchowski et al. 2010) based on computing the degree of dissimilarity between two scanpaths. The latter are represented as strings where the symbols designate the AOIs and are arranged in the order of attending the AOIs; then a distance function based on string editing is used (Duchowski et al. 2010). The function computes the cost of transforming one string into another by means of deletions, insertions, and substitutions. This can be extended to account for the fixation durations and distances between the AOIs (von der Malsburg & Vasishth, 2011). In analyzing multiple scanpaths, pairwise distances may be averaged (Duchowski et al. 2010) or used to cluster the paths by similarity (Çöltekin et al. 2010). The matrix of pairwise distances can be fed to a projection algorithm, e.g., multidimensional scaling, and the projection can be visualized (Çöltekin et al. 2010) for finding groups of similar scanpaths.

Opach and Nossum (2011) admit that scanpath comparison may be ineffective in case of large variance among eye trajectories. The authors even conclude that the method requires the visual stimuli to be specially designed to minimize the possibilities of different viewing strategies. Thus, this method works well enough in text reading studies (von der Malsburg & Vasishth, 2011) and psychological tests (Duchowski et al. 2010) where the AOIs (words, numbers, letters, etc.) are predefined and supposed to be viewed in a particular order. Çöltekin et al. (2010) represent scanpaths in a generalized way: the possible AOIs are assigned to classes according to their semantics or function; the scanpaths are transformed to sequences of class labels and thereby become more comparable; the analysis is based on these sequences.

The scanpath comparison methodology does not provide a way to see the original scanpaths. The analyst has to deal with the strings, which may be not easy to understand, especially when the symbols represent automatically extracted AOIs and therefore lack semantics. Çöltekin and Kraak (2010) suggest that the space-time cube (STC) (Kraak, 2003) can be used to visualize eye trajectories. It is good for detailed exploration of a single trajectory and even for multiple trajectories when there is not much diversity among them (Çöltekin and Kraak, 2010). Eye trajectories have also been analyzed using the movement summarization method originally developed for geographic data (Fabrikant et al. 2008; Ooms et al. 2012). The successful uses of this method and STC show that geographic movement analysis methods can also be useful in eye movement analysis. Within the project, we are going to perform a systematic investigation of the potential of these and other techniques for eye movement studies.

6.5 References

ActiveWizards, 2018. A Comparative Analysis of Top 6 BI and Data Visualization Tools in 2018. Available at https://activewizards.com/blog/a-comparative-analysis-of-top-6-bi-and-data-visualization-tools-in-2018/

Andrienko, G., Andrienko, N. (2008) Spatio-temporal aggregation for visual analysis of movements. In IEEE Symposium on Visual Analytics Science and Technology, pages 51–58.

Andrienko, N., Andrienko, G. (2011) Spatial generalization and aggregation of massive movement data. IEEE Transactions on Visualization and Computer Graphics, 17(2): 205–219.

Andrienko, N. and Andrienko, G. (2012) Visual analytics of movement: An overview of methods, tools and procedures. Information Visualization, 12(1):3– 24, 2012. doi: 10.1177/1473871612457601.

Andrienko, N. and Andrienko, G. (2013) A visual analytics framework for spatio-temporal analysis and modelling. Data Mining Knowl. Discovery, vol. 27, no. 1, pp. 55–83.

Andrienko, N. and Andrienko, G. (2017) State Transition Graphs for Semantic Analysis of Movement Behaviours. Information Visualization, 17(1):41–65, 2017. doi: 10.1177/1473871617692841

Andrienko, G., Andrienko, N. and Bartling, U. (2008a) Visual analytics approach to user-controlled evacuation scheduling. Inf. Vis., vol. 7, no. 1, pp. 89–103, 2008.

Andrienko G, Andrienko N, Dykes J, et al. (2008b) Geovisualization of dynamics, movement and change: key issues and developing approaches in visualization research. Inf Vis 7(3/4): 173–180.

Andrienko N., G. Andrienko, N. Pelekis, and S. Spaccapietra (2008c). Basic Concepts of Movement Data. In Mobility, Data Mining and Privacy – Geographic Knowledge Discovery, edited by F. Giannotti, D. Pedreschi, 15–38. Heidelberg: Springer Verlag. ch. 1.

Andrienko, G., Andrienko, N. Burch, M., and Weiskopf, D. (2012) Visual Analytics Methodology for Eye Movement Studies. IEEE Transactions on Visualization and Computer Graphics (Proceedings IEEE VAST 2012, best paper award), 18(12): 2889-2898, Dec. 2012.

Andrienko, G., Andrienko, N., Bak, P., Keim, D. and Wrobel, S. (2013) Visual Analytics of Movement. Berlin, Germany: Springer

Andrienko, N., Andrienko, G., and Rinzivillo, S. (2015) Exploiting spatial abstraction in predictive analytics of vehicle traffic. ISPRS Int. J. Geo-Inf., vol. 4, no. 2, pp. 591–606

Andrienko, N., Andrienko, G., Fuchs, G., and Jankowski, P. (2016a) Scalable and privacy-respectful interactive discovery of place semantics from human mobility traces. Inf. Vis., vol. 15, no. 2, pp. 117–153, 2016.

Andrienko, N., Andrienko, G., and Rinzivillo, S. (2016b) Leveraging spatial abstraction in traffic analysis and forecasting with visual analytics, Inf. Syst., vol. 57, no. 1, pp. 172–194, Apr. 2016.

Andrienko, G., Andrienko, N., and Fuchs, G. (2016c) Understanding movement data quality. Journal of Location Based Services, 10(1):31–46, 2016. doi: 10. 1080/17489725.2016.1169322. URL https://doi.org/10.1080/17489725.2016.

Andrienko, G., Andrienko, N., Chen, W., Maciejewski, R., and Zhao, Y. (2017) Visual Analytics of Mobility and Transportation: State of the Art and Further Research Directions. IEEE Transactions on Visualization and Computer Graphics, 24(1):34–44, August 2017. ISSN 1077-2626. doi: 10.1109/TVCG.2017.2744322.

Beecham, B. and Wood, J. (2014) Exploring gendered cycling behaviours within a large-scale behavioural data-set. Transp. Planning Technol., vol. 37, no. 1, pp. 83–97, 2014.

Beecham, R., Wood, J., and Bowerman, A. (2014) Studying commuting behaviours using collaborative visual analytics. Comput., Environ. Urban Syst., vol. 47, pp. 5–15, Sep. 2014.

Bengio, Y., Courville, A., & Vincent, P. (2013) Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8), 1798-1828.

Bojko, A. (2009) Informative or misleading? Heatmaps deconstructed. In J.A. Jacko (Ed.): Human-Computer Interaction, Part I, HCII 2009, LNSC 5610, pp. 30-39.

Buchin, K., Sijben, S., Arseneau, T.J.M., Willems, E.P. (2012) Detecting Movement Patterns Using Brownian Bridges, in: Proceedings of the 20th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '12. ACM, New York, NY, USA, pp. 119–128. https://doi.org/10.1145/2424321.2424338

Buchin, K., Buchin, M., van Kreveld, M., Löffler, M., Silveira, R.I., Wenk, C., Wiratma, L. (2013) Median Trajectories. Algorithmica, 66: 595-614. https://doi.org/10.1007/s00453-012-9654-2

Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011) Robust principal component analysis. Journal of the ACM 58(3), p. 11.

Chae, J., Thom, D., Jang, Y., Kim, S. Y. , Ertl, T. and Ebert, D. S. (2014a) Public behavior response analysis in disaster events utilizing visual analytics of microblog data. Comput. Graph., vol. 38, pp. 51–60, Feb. 2014.

Chu, D. et al. (2014a) Visualizing hidden themes of taxi movement with semantic transformation. in Proc. Pacific Vis. Symp., 2014, pp. 137–144.

Cichocki, A., Zdunek, R., Phan, A. H., & Amari, S. I. (2009) Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation. John Wiley & Sons.

Çöltekin, A., B. Heil, S. Garlandini, S.I. Fabrikant (2009) Evaluating the effectiveness of interactive map interface designs: A case study integrating usability metrics with eye-movement analysis, Cartography and Geographic Information Science, 36(1): 5-17.

Çöltekin, A., S.I. Fabrikant, M. Lacayo (2010) Exploring the efficiency of users' visual analytics strategies based on sequence analysis of eye movement recordings. International Journal on Geographical Information Science, 24(10): 1559-1575.

Cox, T.F. & Cox, M.A.A. (2001). Multidimensional Scaling. Chapman and Hall.

Dang, T N, Anand, A, & Wilkinson, L. (2013) TimeSeer Scagnostics for high-dimensional time series. IEEE Transactions on Visualization and Computer Graphics, 19(3), 470-483.

Dodge, S., R. Weibel, E. Forootan (2009) Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects. Computers, Environment and Urban Systems, 33(6): 419-434.

Duchowski, A.T., J. Driver, W. Tan, A. Robbins, B.N. Ramey, S. Jolaoso (2010) Scanpath comparison revisited. In Proceedings of the Symposium on Eye Tracking Research & Applications (ETRA 2010), pp. 219-226.

Fabrikant, S.I., S. Rebich-Hespanha, N. Andrienko, G. Andrienko, D.R. Montello (2008) Novel method to measure inference affordance in static small-multiple map displays representing dynamic processes. The Cartographic Journal, 45 (3), pp. 201-215.

Ferreira, N., Poco, JVo, ., H. T., Freire, J. and Silva, C. T. (2013a) Visual exploration of big spatio-temporal urban data: A study of New York City taxi trips. IEEE Trans. Vis. Comput. Graphics, vol. 19, no. 12, pp. 2149–2158, Dec. 2013.

Few, S., 2006. Information Dashboard Design. Oreilly & Associates Incorporated.

Fredrikson, A., North, C., Plaisant, C., and Shneiderman, B. (1999a) Temporal, geographical and categorical aggregations viewed through coordinated displays: A case study with highway incident data in Proc. Workshop New Paradigms Inf. Vis. Manipulation (NPIVM), New York, NY, USA, 1999, pp. 26–34.

Gazis, D. C. (2002a) Traffic Theory. Boston, MA, USA: Kluwer.

Greenewald, K., Tsiligkaridis, T., & Hero, A. (2013) Kronecker sum decompositions of space-time data. Proceedings of IEEE CAMSAP.

Guerriero, A., Zignale, D., Halin, G., 2012. A Zoomable Location-Based Dashboard for Construction Management, in: Cooperative Design, Visualization, and Engineering, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 207–210. https://doi.org/10.1007/978-3-642-32609-7_29

Havre, S, Hetzler, B, & Nowell, L. (2000) Themeriver: Visualizing theme changes over time. In Proceedings of the IEEE Symposium on Information Visualization.

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., Melançon, G. (2008) Visual Analytics: Definition, Process, and Challenges, in: Information Visualization, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 154–175. https://doi.org/10.1007/978-3-540-70956-5_7

Keim, D.A., Kohlhammer, J., Ellis, G., and Mansmann, F. (2010) Eds., Mastering the Information Age—Solving Problems With Visual Analytics. Goslar, Germany: Eurographics Association, 2010.

Kieu, L. M., Bhaskar, A., and Chung, E. (2015) Passenger segmentation using smart card data. IEEE Trans. Intell. Transp. Syst., vol. 16, no. 3, pp. 1537–1548, Jun. 2015.

Kraak, M.-J. (2003) The space-time cube revisited from a geovisualization perspective. In Proceedings of the 21st International Cartographic Conference, Durban, South-Africa, pp. 1988-1995.

Krśtajic, M., Bertini, E., & Keim, D. A. (2011) CloudLines Compact display of event episodes in multiple time-series. IEEE Transactions on Visualization and Computer Graphics, 17(12), 2432-2439.

Kruger, R., Thom, D., Worner, M., Bosch, H. and Ertl, T. (2013) TrajectoryLenses—A set-based filtering and exploration technique for long-term trajectory data. Comput. Graph. Forum, vol. 32, pp. 451–460, Jun. 2013.

Kruger, R., Thom, D., and Ertl, T. (2015) Semantic enrichment of movement behavior with foursquare—A visual analytics approach. IEEE Trans. Vis. Comput. Graphics, vol. 21, no. 8, pp. 903–915, Aug. 2015.

Krush, M.T., Agnihotri, R., Trainor, K.J., Nowlin, E.L. (2013) Enhancing organizational sensemaking: An examination of the interactive effects of sales capabilities and marketing dashboards. Industrial Marketing Management, Business Models – Exploring value drivers and the role of marketing 42, 824–835. https://doi.org/10.1016/j.indmarman.2013.02.017

Laharotte, P. A., Billot, R., Come, E., Oukhellou, L., Nantes, A. and El Faouzi, N. E. (2015) Spatiotemporal analysis of Bluetooth data: Application to a large urban network. IEEE Trans. Intell. Transp. Syst., vol. 16, no. 3, pp. 1439–1448, Jun. 2015.

Längkvist, M., Karlsson, L., & Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. Pattern Recognition Letters, 42, 11-24.

Laube, P., Purves, R. (2011) How fast is a cow? Cross-Scale Analysis of Movement Data. Transactions in GIS 15, 401–418. https://doi.org/10.1111/j.1467-9671.2011.01256.x

Lechner, B., Fruhling, A. (2014) Towards Public Health Dashboard Design Guidelines, in: HCI in Business, Lecture Notes in Computer Science. Springer, Cham, pp. 49–59. https://doi.org/10.1007/978-3-319-07293-7_5

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436.

Lee, D.-W., Baek, S.-H., Bae, H.-Y. (2009) aCN-RB-tree: Update Method for Spatio-Temporal Aggregation of Moving Object Trajectory in Ubiquitous Environment. International Conference on Computational Science and Its Applications (ICCSA '09), 10.1109/ICCSA.2009.30.

Lee, J.-G., Han, J., Whang, K.-Y. (2007) Trajectory Clustering: A Partition-and-Group Framework. ACM SIGMOD International Conference on Management of Data, 593-604. doi:10.1145/1247480.1247546

Li, X., A. Çöltekin, M.-J. Kraak (2010) Visual exploration of eye movement data using the space-time cube. In Proceedings of GIScience 2010, pp. 295-309, Springer.

Liao, T. W. (2005). Clustering of time series data—a survey. Pattern recognition, 38(11), 1857-1874.

Liu, D. et al. (2017) SmartAdP: Visual analytics of large-scale taxi trajectories for selecting billboard locations. IEEE Trans. Vis. Comput. Graphics, vol. 23, no. 1, pp. 1–10, Jan. 2017.

Long, J.A., Nelson, T.A. (2013) Measuring Dynamic Interaction in Movement Data. Transactions in GIS 17, 62–77. https://doi.org/10.1111/j.1467-9671.2012.01353.x

Ma, Y., Lin, T., Cao, Z., Li, C., and Chen, W. (2016) Mobility viewer: An Eulerian approach for studying urban crowd flow. IEEE Trans. Intell. Transp. Syst., vol. 17, no. 9, pp. 2627–2636, Sep. 2016.

Machlis, S. (2017). 22 free tools for data visualization and analysis. Computerworld, https://www.computerworld.com/article/2507728/enterprise-applications/enterprise-applications-22-free-tools-for-data-visualization-and-analysis.html

Maguire, E., Rocca-Serra, P., Sansone, S.-A., Davies, J., & Chen, M. (2013) Visual compression of workflow visualizations with automated detection of macro motifs. IEEE Transactions on Visualization and Computer Graphics, 19(12), 2576–2585.

Mairal, J., Bach, F., Ponce, J., & Sapiro, G. (2009). Online dictionary learning for sparse coding. In Proceedings of ICML.

Maldonado, R.M., Kay, J., Yacef, K., Schwendimann, B. (2012) An Interactive Teacher's Dashboard for Monitoring Groups in a Multi-tabletop Learning Environment, in: Intelligent Tutoring Systems, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 482–492. https://doi.org/10.1007/978-3-642-30950-2_62

Nathan, R., Getz, W.M., Revilla, E., Holyoak, M., Kadmon, R., Saltz, D., Smouse, P.E. (2008) A movement ecology paradigm for unifying organismal movement research. PNAS 105, 19052–19059. https://doi.org/10.1073/pnas.0800375105

Ooms, K., G. Andrienko, N. Andrienko, P. De Maeyer, V. Fack (2012) Analysing the spatial dimension of eye movement data using a visual analytic approach. Expert Systems with Applications, v.39 (1): 1324-1332.

Opach, T., A. Nossum. Evaluating the usability of cartographic animations with eye-movement analysis (2011) In Proceedings of the 25th International Cartographic Conference, Paris.

Pappas, L., Whitman, L. (2011) Riding the Technology Wave: Effective Dashboard Data Visualization, in: Human Interface and the Management of Information. Interacting with Information, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 249–258. https://doi.org/10.1007/978-3-642-21793-7_29

Pelekis, N., Kopanakis, I., Kotsifakos, E.E., Frentzos, E., Theodoridis, Y. (2011) Clustering uncertain trajectories. Knowledge and Information Systems, 28: 117-147. https://doi.org/10.1007/s10115-010-0316-x

Perer, A. & Sun, J. (2012) Matrixflow: Temporal network visual analytics to track symptom evolution during disease progression. In Proceedings of AMIA.

Plaisant, C., Milash, B., Rose, A., Widoff, S., & Shneiderman, B. (1996) LifeLines: Visualizing personal histories. Proceedings of CHI.

Poole, A., L.J. Ball (2006) Eye tracking in human-computer interaction and usability research: current status and future prospects. In C. Ghaoui (Ed.), Encyclopedia of Human-Computer Interaction, pp. 211-219, Idea group.

Postlethwaite, C.M., Brown, P., Dennis, T.E. (2013) A new multi-scale measure for analysing animal movement data. Journal of Theoretical Biology 317, 175–185. https://doi.org/10.1016/j.jtbi.2012.10.007

Rahman, A. (2017) Designing a dashboard as geo-visual exploration tool for origin-destination data (Master thesis). ITC.

Riehmann, P., Hanfler, M., & Froehlich, B. (2005) Interactive sankey diagrams. Proceedings of the IEEE Symposium on Information Visualization.

Scheepens, R., Willems N., & van de Wetering, H. (2011) Composite density maps for multivariate trajectories. IEEE Transactions on Visualization and Computer Graphics, 17(12), 2518-2527.

Scipioni, A., Mazzi, A., Mason, M., Manzardo, A., 2009. The Dashboard of Sustainability to measure the local urban sustainable development: The case study of Padua Municipality. Ecological Indicators 9, 364–380. https://doi.org/10.1016/j.ecolind.2008.05.002

Shi, C, Cui, W, Liu, S, Xu, P, Chen, W, & Qu, H. (2012) RankExplorer: Visualization of ranking changes in large time series data. IEEE Transactions on Visualization and Computer Graphics, 18(12), 2669-2678.

Soleymani, A., Pennekamp, F., Petchey, O.L., Weibel, R. (2015) Developing and Integrating Advanced Movement Features Improves Automated Classification of Ciliate Species. PLoS ONE, 10(12): e0145345. DOI: 10.1371/journal.pone.0145345.

Soleymani, A., Pennekamp, F., Dodge, S., Weibel, R. (2017) Characterizing change points and continuous transitions in movement behaviors using wavelet decomposition. Methods in Ecology and Evolution, 8(9): 1113-1123. doi: 10.1111/2041-210X.12755.

Thomas, J. J. and Cook, K. A., Eds. (2005) Illuminating the Path: The Research and Development Agenda for Visual Analytics. Piscataway, NJ, USA: IEEE Press, 2005.

Tominski, C, Schumann, H, Andrienko, G, & Andrienko, N. (2012) Stacking-based visualization of trajectory attribute data. IEEE Transactions on Visualization and Computer Graphics, 18(12), 2565-2574.

van der Hurk, E., Kroon, L., Maro´ti, G., and Vervest, P. (2015) Deduction of passengers' route choices from smart card data. IEEE Trans. Intell. Transp. Syst., vol. 16, no. 1, pp. 430–440, Feb. 2015.

van der Maaten, L.J.P.; Hinton, G.E. (2008) Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research. 9, 2579–2605.

von der Malsburg, T., S. Vasishth (2011) What is the scanpath signature of syntactic reanalysis? Journal of Memory and Language, 65(2): 109-127.

von Landesberger, T., Brodkorb, F., Roskosch, P., Andrienko, N., Andrienko, G. and Kerren, A. (2016) MobilityGraphs: Visual analysis of mass mobility dynamics via Spatio-temporal graphs and clustering. IEEE Trans. Vis. Comput. Graphics, vol. 22, no. 1, pp. 11–20, Jan. 2016.

Ware, C. (2012) Information Visualization, Third Edition: Perception for Design, 3 edition. ed. Morgan Kaufmann, Waltham, MA.

Willems, N., H. van de Wetering, J.J. van Wijk (2009) Visualization of vessel movements. Computer Graphics Forum, 28(3): 959-966.

Wongsuphasawat, K. & Gotz, D. (2012) Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. IEEE Transactions on Visualization and Computer Graphics, 18(12), 2659–2668.

Wongsuphasawat, K., Guerra Gomez, J. A., Plaisant, C., Wang, T. D., Taieb-Maimon, M., & Shneiderman, B. (2011) LifeFlow: Visualizing an overview of event sequences. In Proceedings of SIGCHI.

Wood, Slingsby, J., A., and Dykes, J. (2011) Visualizing the dynamics of Lon- don's bicycle hire scheme. Cartographica, vol. 46, no. 4, pp. 239–251, 2011.

Yang, K. & Shahabi, C. (2004). A PCA-based similarity measure for multivariate time series. In Proceedings of the 2nd ACM international workshop on Multimedia databases.

Yang, X., Zhao, Z., and Lu, S. (2016) Exploring spatial-temporal patterns of urban human mobility hotspots. Sustainability, vol. 8, no. 7, p. 674, 2016.

Zhang, L., Stoffel, A., Behrisch, M., Mittelstadt, S., Schreck, T., Pompl, R., Weber, S., Last, H., Keim, D. (2012) Visual analytics for the big data era — A comparative review of state-of-the-art commercial systems, in: 2012 IEEE Conference on Visual Analytics Science and Technology (VAST). pp. 173–182. https://doi.org/10.1109/VAST.2012.6400554

Zhao, J, Chevalier, F, Pietriga, E, & Balakrishnan, R. (2011) Exploratory analysis of time-series with chronolenses. IEEE Transactions on Visualization and Computer Graphics 17(12), 2422-2431.

Zhao, J., Liu, Z., Dontcheva, M., Hertzmann, A., & Wilson, A. (2015) MatrixWave: Visual Comparison of Event Sequence Data. Proceedings of CHI.

7 Market Analysis – the Insurance Business Case

The aim of this section is to present the insurance business market in terms of utilities given to stakeholders and market leverage growing focal points. We will focus on three geographical macro-areas, two metropolitan cities as London and Rome and one country-urban mixed area such as the region of Tuscany, in Italy. We will investigate two main demanding points: the in-urban transportation from point to point within a metropolitan area and some long-term trips between several urban agglomerations.

In addition to this, we will show also some numbers about electric car mobility business and ride sharing services in the same three geographical areas.

The scope of the whole chapter is providing further information about insurance business, electric car mobility business and car-pooling services, intended to integrate the given data set of the and to be used in junction with the toolbox methods developed in pilot projects, in order to calculate efficient KPIs and achieve three main goals the topics showed above :

1. Develop some telematics tools to insurance stakeholders in order to profile a personal insurance risk giving margins to reduce insurance fees to virtuous drivers without exposing them to increasing insurance risks.
2. Develop some telematics tools and some statistical parameters in order to estimate the convenience in terms of costs and trip timing efficacy of a possible switch to an electric car mobility paradigm in the considered geographical areas.
3. Develop some telematics tools and some statistical parameters in order to estimate the possible impact of a rising car-pooling service in the considered geographical areas.

7.1 About the Car Insurance Industry

Some definitions and state of art

The insurance market has remained unchanged for years, following the line traced by actuarial science since then. However, nowadays insurance companies have to face problems related to the new generation customers, their increasingly poor fidelization trend and their increasingly more tech-innovation demand.

In fact, with the generation Y (anyone's born between 1980 and 1995) (World Insurance Report (2017)) entering in, the insurance market experienced a lowering level of fidelization and an increase in market variability. Trend that will probably be confirmed when the generation Z (anyone's born between 1995 and 2010) (World Insurance Report (2017)) would became the brand-new clients and generation Y the major percentage of the market target.

According to the World Insurance Report 2017 (World Insurance Report (2017)) generation Y and technology-savvy customers demand an increasing level of self-service and personalized offers and an increasing level of technological integration, paying less but not renouncing at excellent personal risk coverage.

The 36% of the generation Y and the 42% of tech-savvy customers say they're likely to buy a new insurance product (compared to 29% of non-GenY and 20% of non tech-savvy customers).

In this panorama we also observe an increasing number of telematic companies offering telematics solutions for new products based on "How much you drive" and "How you drive", that lead to a tailorized risk for each driver.

Telematic companies and insurance companies together have recently created several international Insur-Tech players. The forms of current collaboration between those companies are quite varied: strategic partnerships, acquisitions, venture capital investments, incubators, leading to efficient

sinergies in actuarial science, lowering clients' fees, increasing margins for companies without exposing them to additional risk levels.

For "**Actuarial science**" we use to define the science that nowadays calculated the insurance risk basing on historical data and through information related to the vehicle (Car tipology, Geographical area) and client information (age, gender, previous accident history not related to the pilot but to the vehicle) (Bain et al. 2017).

For "**Pay as much as you drive**" we use to define a model of business also known as UBI (Usage Based Insurance) that spread the risk along the route mileage during the reference time period, leading to variables fees during that period (Bain et al. 2017).

For "**Pay as you drive**" we use to define an evolution of the UBI business model, that takes in count also the customer's driving behaviour scoring calculated through new methods such as advance machine learning and AI-aided computer science. Those methods generally use measured parameters from telematics devices as harsh acceleration/brakes, speed limit exceeding, harsh cornering (Bain et al. 2017).

For this reason, the 79% of insurance companies is investing in blockchain, Robotics and advanced analytics in order to tailor the best client's fees without relevant implications in risk coverage (Bain et al. 2017).

In the following sections, we will introduce some dataset and some aggregate results about the insurance business market in order to integrate the given dataset and to achieve the 1st goal of the pilot project.

7.1.1 Case study: London

United Kingdom, as well known, is one of the most populated European country with a huge circulating car park. To zoom into the London scenario, we can use a free tool www.collisionmap.uk (London Collision Map - 2018), that provides a massive GPS position data set, collecting all the slight, serious and fatal collision occurred in the selected area accompanied by number of vehicles and casualties involved in.

In order to perform our analysis, we collected some statistics about the city, as you can see following.

- **Total amount of collisions in urban areas**: The London collision map provided by UK government (London Collision Map: Fatal and Serious Collision during 2016) (Realated Open Data) clearly shows the situation along years from 2005 to 2016. Taking 2016 as a reference we suppose up to 6000 serious or fatal collision in just one year.
- **Mean insurance annual fees:** According to Association of British Insurers (ABI) reviews the average car insurance premium in the UK for comprehensive cover is spanning between £470 and £490 per year (about 560 Euros). This value definitely increase in London metropolitan area to reach twice or 2.5 times (depending to several factors, i.e.: driver's gender and age, car brand, etc.) the value of the rest of Britain. In particular fees increase of 40% according to driver's age (+40% for drivers younger than 25 years old) without taking in count driver's skills.
- **Size of the circulating park**: According to the BBC analysis (Cars in England's roads), the circulating car park counts about 2.6 millions of vehicles (last update 2015) as we can see in the following figure.

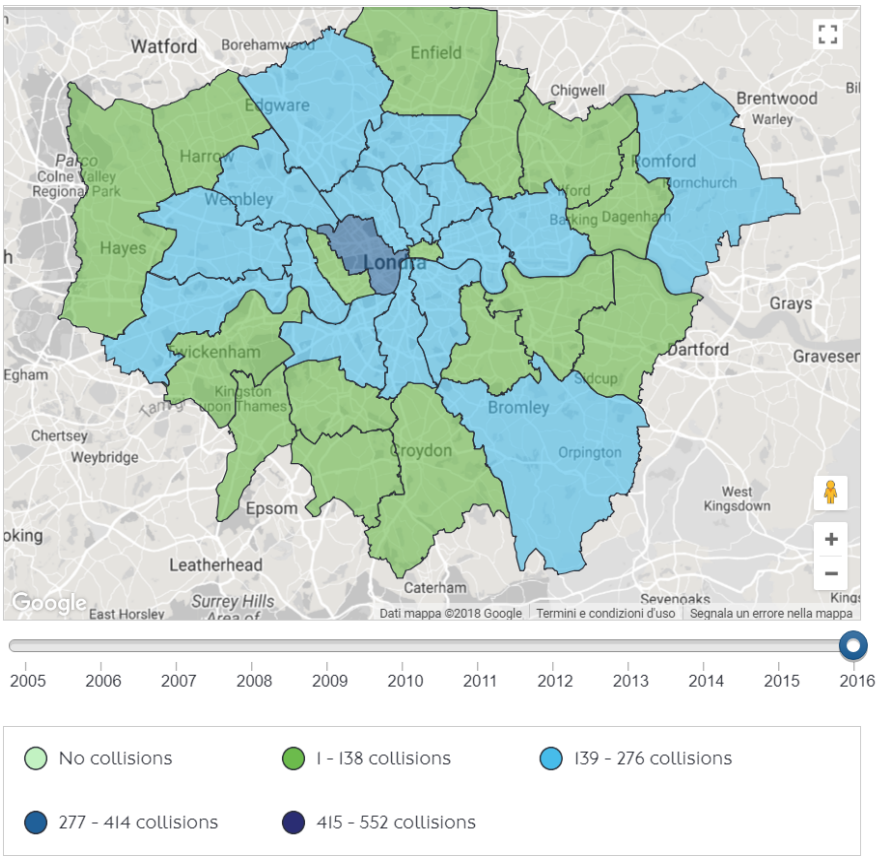


Figure 19: London Collision Map (London Collision Map: Fatal and Serious Collision during 2016).

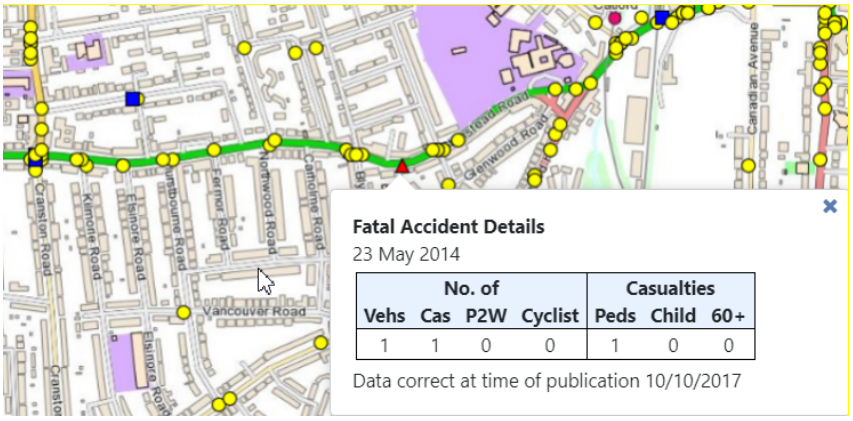


Figure 20: A glimpse of the London Collision Map (London Collision Map).

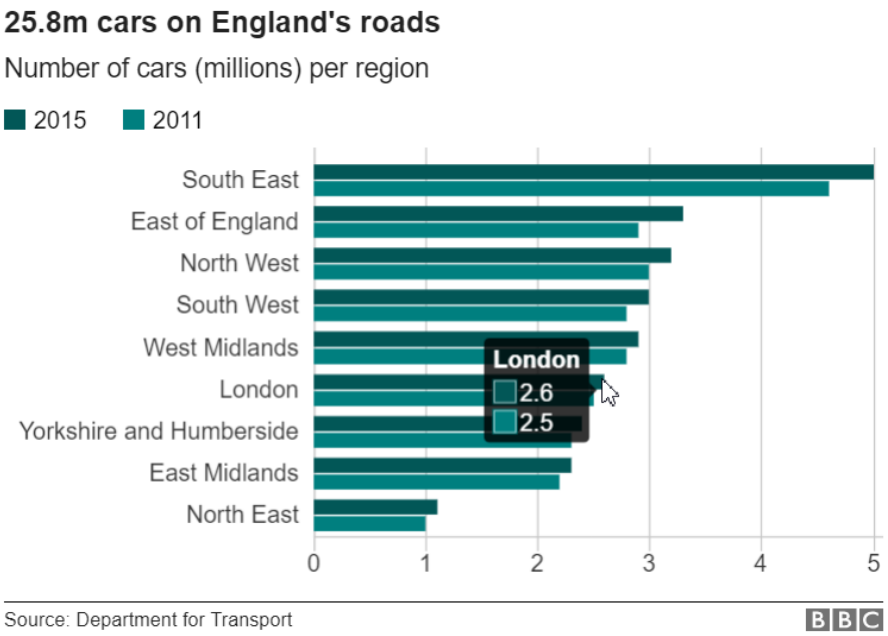
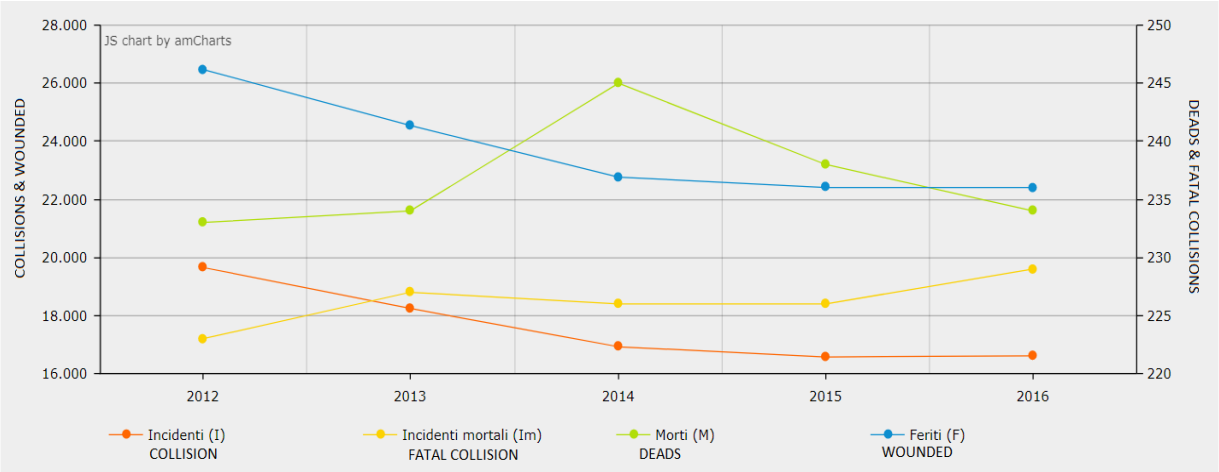


Figure 21: Number of cars (millions) in London city region (Cars in England's roads)

7.1.2 Case study: Rome

As well as previous London case study we collected some useful data about Rome city in order to give some statistical data to improve inferences about car accident and achieve the goal shown in the 3rd point.

- **Total amount of collisions in urban areas**: According to the (Automobile Club Italia (ACI)) we can share some data about car collision in Rome. Unfortunately in this case we has data related to the province of Rome, not just metropolitan city of Rome. However the city circulating vehicle park represents about the 66% of the total circulating vehicle park of the province. The first picture shows the progressive collision statistics along years 2012-2016. The second picture gives a glance of URBAN vs EXTRA-URBAN/HIGHWAYS collision ratio.
- **Mean insurance annual fees for client**: Between 20% and 45% more than the average fee payed in Italy: 660-790 Euros per year
- **Size of the circulating park**: Around 1,780,000 vehicles (Italian Circulating Car Park) with respect to 2,678,000 vehicles of the whole province of Rome.



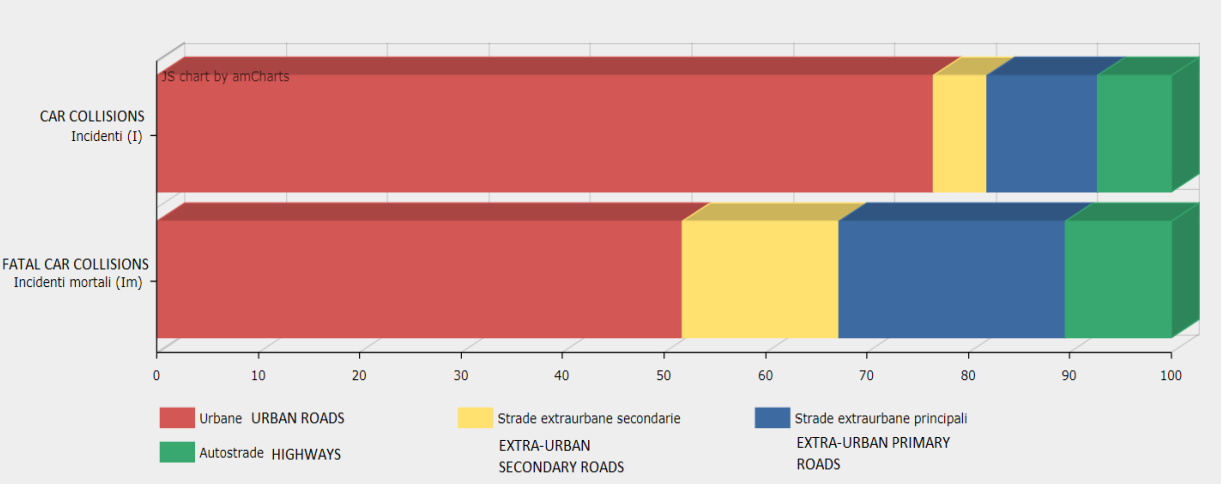


Figure 22: Some graphs regarding car accidents in Rome (Automobile Club Italia (ACI)).

7.1.3 Case study: Tuscany, Italy

Thanks to the same fonts founded for Rome case study we are providing a similar statistics for the Italian region of Tuscany.

- **Total amount of car collision in urban areas and extra-urban area**: According to the (Automobile Club Italia (ACI)) we can share some data about car collision along the last six years in urban areas and extra-urban area of Tuscany, Italy.
- **Mean insurance annual fees for client:** In agreement with the average fee payed in Italy: 550-650 Euros per year
- **Size of the circulating park**: Around 2,378,000 vehicles (Italian Circulating Car Park).

7.2 About the car pooling and car sharing market analysis

In this chapter we present some information about different paradigm of ride-sharing in the three geographical areas considered. The goal of this chapter is to show the actual scenario of ride-sharing and focus on how a rising car-pooling service could compete with transportation in a metropolitan point-to-point environment as well as in a long-term trips. In this way we can address the 2nd goal of the pilot project.

Some definitions and some history

"**Car sharing**" means the rental of a car owned by third parties, generally short-term and in urban contexts. The same car is made available to more drivers who use it for a limited period of time. Traditional car sharing began in the 1990s and early 2000s. The concept is simple: provide short-term personal vehicle transportation for consumers who would not utilize a personal vehicle often enough to justify owning one. Essentially, a company owns fleet of cars that are strategically placed in high-density environments and can be rented by people on an hour, daily or monthly basis. An evolution of "**traditional car sharing**" is "**peer to peer car sharing**": an innovative approach to vehicle sharing in which vehicle owners temporarily rent their personal car to others in their surrounding area. Peer-to-peer car sharing belongs to the larger sharing economy, an economic model based on the notion of collaboration as opposed to ownership. The proliferation of smartphones and social networking sites, especially since 2006, has influenced the development of peer-to-peer car sharing, in fact, at the base of this new economy model there are online platforms provided by organizations, through which people can share their own vehicle.

Peer-to-peer car sharing requires much less initial capital investment than traditional car sharing because no cars need to be purchased. The costs for creating and hosting a web or a mobile app are comparable. Another advantage of the Peer-to-peer model is that it has the potential to be used in less-dense environments, once software, hardware and customer service components are in place. Therefore, a tech-oriented and automated peer-to-peer company provides technology, hardware and software, to install in the owner's car. The installed hardware and software integration allow the driver to unlock the rented car through a RFID/NFC card, a smartphone or a key fob. The purpose of the hardware is to create an easier user experience for both renters and owners. Another strong point of tech-oriented model is the ease to locate the car, both for owner, he will always know where his car is, and for renter, he can easily find the vehicle before starting driving.

A prosumer-based evolution of peer-to-peer car sharing get into the concept of "ride sharing", that refers in general to the activity of sharing car passages, also in order to produce a profit (in this case called "on demand sharing") while the concept of "**car pooling**" presupposes that sharing of the trip does not provide a gain for the driver but only a sharing of costs, or a courtesy transport activity made by usual couriers or privates. The difference from car-sharing is remarkable, in fact in the last case we observe a sharing of a service rather than a commodity, however the market needs may be the same as peer-to-peer car sharing, that means essentially the costs for creating and hosting a web or a mobile app and insurance service costs. Also, the business model is quite similar due to the fact that about 65% of the income belongs to the prosumer (that eventually can re-invest in the service as a consumer), about 20% is for insurance fees and about 15% is for the service maintaining.

For our analyses we choose the top market players. that are essentially two: UBER and BLABLACAR, both currently present in London and Rome and Tuscany area.

The players: UBER & BLABLACAR

The Uber company, founded in San Francisco in 2009, is one of the main society that represent the "sharing economy" concept as its best. The initial investment of 250 K$ yields dividens for 1,25 M$ in just the first year, when the company establish its first business in San Francisco city with with prices that were 50% more expensive than other cab companies. In 2011 the company settle down in New York, Seattle, Chicago, Boston, Washington D.C and Paris. In 2012 Uber reach to lower prices of 35% launching the "dynamic pricing model" algorithm to automatically tune prices based on weekday, level of demand and car ranking. Today the Uber's values is about 65 billions of dollars.

The BlaBlaCar company, was founded in 2012 today from a French pilot website called Covoiturage.fr, born in 2004. The current value is more than 1.5 billions of dollars with 4 main country in Europe and an increasing market all over the word. It counts more than 400000 usual driver in the whole Europe and more than 2 bilions of accounts. The BalBlacar policy is inspired to the sharing economy system so, as opposed to Uber, the driver does not expect any earning but just a mileage reimbursement.

7.2.1 In-urban point to point car pooling: The metropolitan city of London case study

The most popular ride sharing companies in the UK are of course BlaBlaCar, Carpooling, Liftshare. Uber has been recently banned in London. In fact On September 22th, 2017, TfL, which regulates transport in the capital, announced that it would not renew Uber's operating licence. UberEats, the company's food delivery service, is not affected by the dispute. Therefore, the move affects more than 40,000 Uber working drivers and 3.5million customers. So, the scenario favors car-pooling rather than paid car-sharing services. A taxi ride in London costs about 5 Euros per kilometer in the central traffic jam, with a 10-15% surplus for holiday and weekends. The cost can span a lot due to position (central vs peripheral), traffic jam and weekday.

7.2.2 In-urban point to point car pooling : The metropolitan city of Rome: UBER, and other case studies

It is well known that Uber's rise brought several protests in cab companies in different states. Italy was one of them. There were several taxi strikes in Rome to protest the ride sharing app because its drivers do not have to go through the same process of obtaining an expensive taxi license. As a result of protests, Uber was indeed temporarily banned in April 2017. However, the company quickly appealed the ruling and it was swiftly overturned. As a result, Uber remains legal to use in major cities such as Rome and Milan with its luxury and most expensive service "UberBlack" with a regular license, with comparable prices respect to regular cabs. Tourists' seems to prefer "UberBlack" for airport-urban paths and cabs for in-urban trips.

In this scenario two smart minor players are appearing: "Scooterino", a ride-sharing no-profit organization that sends you a moped with an extra helmet that quickly zip you almost anywhere in Rome, and "Welcome Pickups" for airport-city paths. In the meantime other start-up players are getting in the peer-to-peer market : "Enjoy", "ZigZag", "Car2Go", "Share 'n Go" and the public service "carsharing mobilità" more similar to a car-rental. Here is a glancing info-picture.

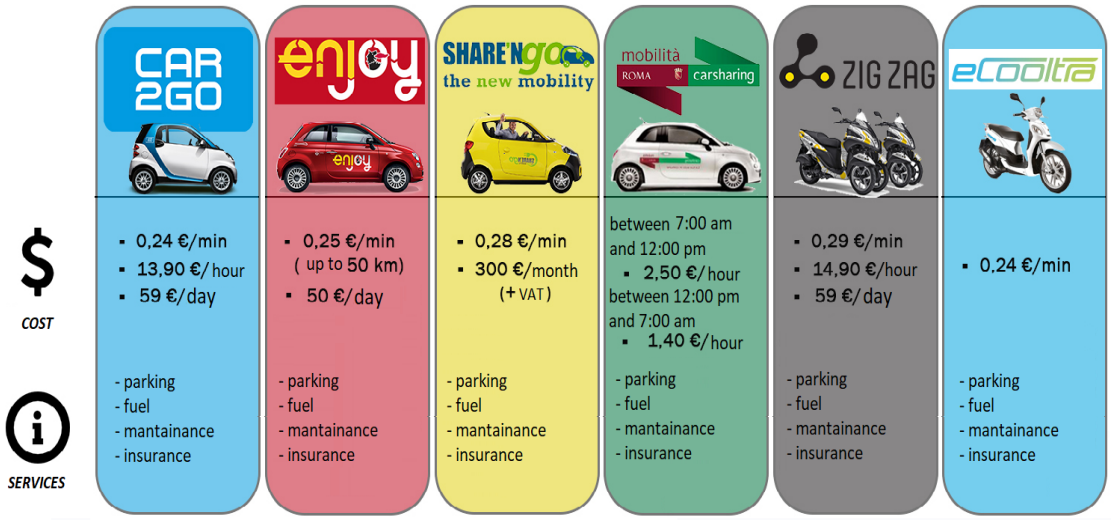


Figure 23: A glimpse of car sharing market (Car sharing Roma).

7.2.3 Long term point to point trip sharing: Tuscany, Italy: BLABLACAR case study

BlaBlaCar popularity in Italy is well known. The company has 2,5 milions of registered users and boasts 1.5 billions of offered kilometers in 5 years of activity in Italy. The service is increase at the rate of 200-300% per year and the main reason why it is still chosen is because it represents the best choice with respect not to cab but to trains. Is a fact that Italian high-speed trains work well for the mainstream routes connecting two or more large cities with few intermediate terminal stops. However minor railway routes connecting the cities to countryside or countryside to countryside are still too much expensive and, first of all, still too much time consuming due to the typical hard Italian geographical topology and the twisting railways. An example: Milan-Massa by train costs 32 Euros ant it takes more than 3 hours and half. A Blablacar shared trip for the same route would costs around 35% of that train cost and it will takes just 2 hours from point to point.

Tuscany is in fact one of the main business market for blablacar and one of the main Italian route with several main actracting cities and a lot of countryside town and villages to connect. A perfect case study to test efficiency of car sharing with respect to trains.

7.3 About the supply developing of the plug-in electric vehicles

Some definitions and some history

A "**Hybrid**" vehicle is intended to be any vehicle that has two main sources of energy that generally works in synergy, but just a one-way main refueling method, usually by fossil propellent. A "**Plug-in**" vehicle is a mono- or bi-fuel vehicle, that at least one of its source of energy is a rechargeable electric battery pack. So, we can tell a "**Hybrid Plug-in**" vehicle from a "**Full Electric**" vehicle by the number of refueling way we have.

In this section, we will investigate several factor that can lead use to an inference about the market growing of the plug-in electric vehicles such as:

- Number of free charging urban point in the analyzed areas related to the total number of circulating vehicles in that areas.
- Number of charging points where to charge vehicles, in the analyzed areas (from 4,5 kW up to 7,0 kW)
- Number of charging station in the city/path.
- Economical eco-incentives pro-capite
- Difference between the cost per kilometer from Plug-in Vehicles and other propellent vehicles (CH4, GPL, Gasoline, Diesel)

The goal of this section is to enrich the time-path information given by the dataset with a further glance of the cost of the technology in order to achieve the 3rd goal of the pilot project.

7.3.1 Case study: London

The UK zap-map is certainly the most complete and user-friendly web platform to locate electric charging point in the country (ZAP MAP); we collect some data regarding the London scenario:

- Total amount of charging point in the area: Around 3500 (21% of total UK charging point)
- Size of the circulating park: 2.6 millions of vehicles (last update 2015 as previously shown according to BBC)
- Vehicles/charging point ratio: Around 750
- Costs of in-house energy charging (from 4,5 kW up to 6,0 kW): Currently the cost for in-house charging is 0.18 Euro per KWh including taxes and distribution, for in-house charging. (see the picture below).
- Economical eco-incentives pro-capital:
    - Purchase grants: 5000 up to 9000 Euros
    - Company car tax: 4% up to 8% less than a diesel
    - Provate installation costs: 75% off of for private installation at house or work
    - Complete exempt from annual road tax.
- Difference between the cost per kilometer from Plug-in Vehicles and other propellent for vehicles (CH4, GPL, Gasoline, Diesel):
    - GASOLINE (MINI-HYBRID):
        - Average Price per liter : 1.34 Euros
        - Urban Average Fuel Efficiency : 17 (21) km/liter
        - Average Cost per 100 kilometers : 7,88 (6,38) Euros
        - Average Maintenance Costs & mechanical parts wearing per 100 kilometers: 2,3 Euros
    - DIESEL (Blue):
        - Average Price per liter : 1.40 Euros
        - Urban Average Fuel Efficiency : 19 (22) km/liter

- Average Cost per 100 kilometers : 7.36 (6,36) Euros
- Average Maintenance Costs & mechanical parts wearing per 100 kilometers: 1,8 Euros

- CH4 :
  - Average Price per kg : --
  - Urban Average Fuel Efficiency : 18 km/kg
  - Average Cost per 100 kilometers: --
  - Average Maintenance Costs & mechanical parts wearing per 100 kilometers: 2,0 Euros
- GPL :
  - Average Price per liter : 0,68 Euros
  - Urban Average Fuel Efficiency : 13 km/liter
  - Average Cost per 100 kilometer : 5,23 Euros
  - Average Maintenance Costs & mechanical parts wearing per 100 kilometers: 2,0 Euros
- FULL ELECTRIC :
  - Average Price per Kwh : 0,18-0,33 Euros (depending to low-fast charging)
  - Urban Average Fuel Efficiency : 55 Km/Kwh
  - Average Cost per 100 kilometers (flat-land) : 0,65 up to Euros
  - Average Maintenance Costs per 100 kilometers : 0.05

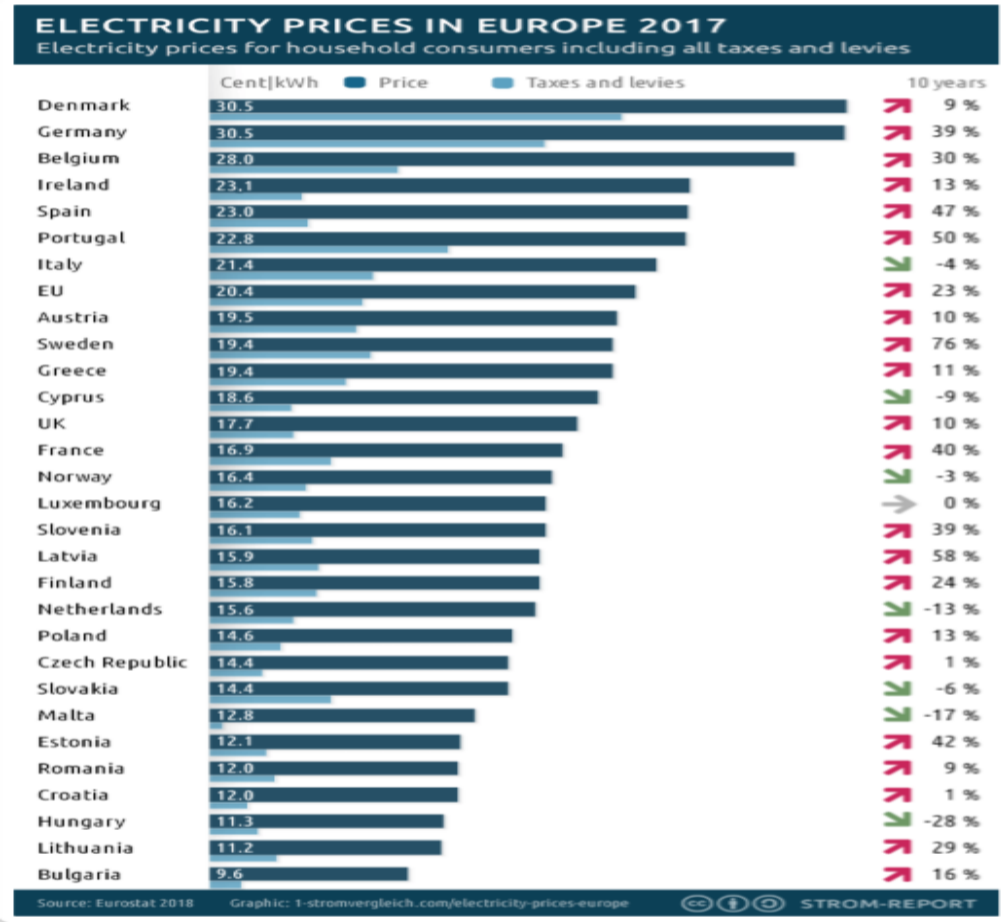


Figure 24: Mean cost of electricity per KW in EU countries. [Eurostat 2018]

7.3.2 Case study: Rome

We also collect some data regarding the Rome scenario:

- **Total amount of charging point on the urban area**:
  - Currently Roma has about 120 charging point in the urban area an about other 12 in the neighboring that is going to increase to 12880 charging point till 2020 according to the "NUOVO PIANO GENERALE DEL TRAFFICO URBANO DI ROMA CAPITALE" (new general urban mobility plan for the metropolitan city of Rome).
  - Some of the charging points will be "free charging point" depending to private dealer policy. We don't know the exact number of free charging point but currently it should be around 5-8 %.
  - According to the plan Roma will have at least a 4:1 ratio between slow- and fast-charging points by 2020.
  - The plan provides a parcellization of the urban area that will lead that 60% of the charging point will be equally distributed according to the urban mobility plan, with a density of 40 point per area of 300 meters of diameter, and the remaining 40% up to providers.
- **Size of the circulating park**:
  - Around 1780000 vehicles (Automobile Club Italia (ACI)) with respect to 2678000 vehicles of the whole province of Rome.
- **Circulating vehicles/charging point ratio**:
  - 130-140 by 2020 in metropolitan city of Rome
- **Costs of in-house energy charging (from 4,5 kW up to 6,0 kW)**:
  - Currently the cost for in-house charging is 0.24 Euro per KWh including taxes and distribution fees that can reach 0.32 Euro in case of high power fast-charging (6 KW). That is less with respect to Germany but is certainly more than UK mean value. (see the picture in the previous section).
- **Economical eco-incentives pro-capite**:
  - Electrical vehicles owners do not pay annual circulation tax for the first 4 years, and just 25% of the total amount starting from the 5th year. The annual circulation tax spans between a mean value of 200 Euro to 600 Euros per year depending on vehicles nominal power.
  - In several zones of Rome parks are free, that means a saving of about 380-960 Euros per year
  - Grants for electrical vehicles purchasing of about 2000-45000 Euros
  - Access to the urban center with no limitation due to fueling type
- **Difference between the cost per kilometer from Plug-in Vehicles and other propellent for vehicles (CH4, GPL, Gasoline, Diesel)**:
  - GASOLINE (MINI-HYBRID):
    - Average Price per liter: 1,75 Euros
    - Urban Average Fuel Efficiency: 17 (21) km/liter
    - Average Cost per 100 kilometers: 10,30 (7,40) Euros
    - Average Maintenance Costs & mechanical parts wearing per 100 kilometers: 2,3 Euros
  - DIESEL (Blue):
    - Average Price per liter: 1,60 Euros
    - Urban Average Fuel Efficiency: 19 (22) km/liter
    - Average Cost per 100 kilometers: 8,50 (7,30) Euros
    - Average Maintenance Costs & mechanical parts wearing per 100 kilometers: 1,8 Euros

- CH4:
  - Average Price per kg: 0,90 Euros
  - Urban Average Fuel Efficiency: 18 km/kg
  - Average Cost per 100 kilometers: 5 Euros
  - Average Maintenance Costs & mechanical parts wearing per 100 kilometers: 2,0 Euros
- GPL:
  - Average Price per liter: 0,70 Euros
  - Urban Average Fuel Efficiency: 13 km/liter
  - Average Cost per 100 kilometers: 5,40 Euros
  - Average Maintenance Costs & mechanical parts wearing per 100 kilometers: 2,0 Euros
- FULL ELECTRIC:
  - Average Price per Kwh : 0,25-0,42 Euros (depending to low-fast charging)
  - Urban Average Fuel Efficiency: 55 Km/Kwh
  - Average Cost per 100 kilometers (flat-land): up to 0,99 Euros
  - Average Maintenance Costs per 100 kilometers: 0.05

7.3.3 Case study: Tuscany, Italy

We also collect some data regarding the Tuscany scenario:

- **Total amount of charging point on the regional area**:
  - Currently Tuscany has about 310 charging point in the urban and extra-urban area, we don't have any other information regarding how this number will rise by 2020. Otherwise Tuscany is certainly one of the main focal area involved in the national plan for charging point installation that will provide 14000 electric charging point by 2022 all over Italy.
  - Some of the charging points will be "free charging point" depending to private dealer policy. We don't know the exact number of free charging point but currently it should be around 5-8 %.
  - Currently Florence is the first Italian city for charging point/circulating vehicle ratio
- **Size of the circulating park**:
  - Around 2.378.000 vehicles (Automobile Club Italia (ACI))
- **Vehicles/charging point ratio**:
  - Currently: 7600, trending to 170-150 by 2022
- **Costs of in-house energy charging (from 4,5 kW up to 6,0 kW)**:
  - Currently the cost for in-house charging is 0.24 Euro per KWh including taxes and distribution fees that can reach 0.32 Euro in case of high power fast-charging (6 KW). That is less with respect to Germany but is certainly more than UK mean value. (see the Fig. 1).
- **Economical eco-incentives pro-capital**:
  - Electrical vehicles owners do not pay annual circulation tax for the first 4 years, and just 25% of the total amount starting from the 5th year. The annual circulation tax spans between a mean value of 200 Euro to 600 Euros per year depending on vehicles nominal power.
  - In several zones of Rome parks are free, that means a saving of about 380-960 Euros per year
  - Bonus for electrical vehicles purchasing of about 2000-45000 Euros
  - Access to the urban center with no limitation due to fueling type
- **Difference between the cost per kilometer from Plug-in Vehicles and other propellent for vehicles (CH4, GPL, Gasoline, Diesel)**:
  - GASOLINE (MINI-HYBRID):

        - Average Price per liter: 1,75 Euros
        - Non Urban Average Fuel Efficiency: 24 (27) km/liter
        - Average Cost per 100 kilometers: 7,3 (6,4) Euros
        - Average Maintenance Costs & mechanical parts wearing per 100 kilometers: 2,3 Euros
    - DIESEL (Blue):
        - Average Price per liter: 1,60 Euros
        - Non Urban Average Fuel Efficiency: 25 (27) km/liter
        - Average Cost per 100 kilometers: 6,5 (6,0) Euros
        - Average Maintenance Costs & mechanical parts wearing per 100 kilometers: 1,8 Euros
    - CH4:
        - Average Price per kg: 0,90 Euros
        - Non Urban Average Fuel Efficiency: 26 km/kg
        - Average Cost per 100 kilometers: 3,5 Euros
        - Average Maintenance Costs & mechanical parts wearing per 100 kilometers: 2,0 Euros
    - GPL:
        - Average Price per liter: 0,70 Euros
        - Non Urban Average Fuel Efficiency: 20 km/liter
        - Average Cost per 100 kilometer: 3,3 Euros
        - Average Maintenance Costs & mechanical parts wearing per 100 kilometers: 2,0 Euros
    - FULL ELECTRIC:
        - Average Price per Kwh : 0,25-0,42 Euros (depending to low-fast charging)
        - Non Urban Average Fuel Efficiency: 48 Km/Kwh
        - Average Cost per 100 kilometers (flat-land): up to 0,875 Euros
        - Average Maintenance Costs per 100 kilometers: 0.05

7.4 Related Open Data

Automobile Club Italia (ACI). URL: http://www.lis.aci.it/en/dati/#/generali/2016/12/058; http://www.lis.aci.it/en/dati/#/generali/2016/12/09

UK Collision Map. URL: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

ZAP MAP – London Charging Points, UK. URL: https://www.zap-map.com/location-search/london-charging-points/; https://www.zap-map.com/live/

7.5 References

Automobile Club Italia (ACI). URL: http://www.lis.aci.it/en/dati/#/generali/2016/12/058; http://www.lis.aci.it/en/dati/#/generali/2016/12/09; http://www.datiopen.it/it/opendata/Parco_veicolare_per_categoria_e_comune_nel_2014

Bain, Y., Yang, C., Zhao, J.L., Liang, L. (2017) Good drivers pay less: A study of usage-based vehicle insurance models. – Transportation Research Part A. Elsevier.

Car sharing Roma: confronto prezzi e condizioni del servizio. URL: http://www.ultimora.news/Car-sharing-Roma-confronto-prezzi-e-condizioni-del-servizio

Cars in England's roads – BBC news. URL: http://www.bbc.com/news/uk-england-35312562

Italian Circulating Car Park, Italian Gov., Italy. URL: http://dati.mit.gov.it/catalog/dataset/parco-circolante-dei-veicoli

London Collision Map. URL: http://www.collisionmap.uk/

London Collision Map: Fatal and Serious Collision during 2016. URL: https://tfl.gov.uk/corporate/safety-and-security/road-safety/london-collision

UK Collision Map. URL: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

World Insurance Report (2017) – Capgemini. URL: https://www.capgemini.com/it-it/service/world-insurance-report-2017/

ZAP MAP – London Charging Points, UK. URL: https://www.zap-map.com/location-search/london-charging-points/; https://www.zap-map.com/live/

8 Market Analysis – the Healthcare Business Case

The Market Analysis started with a clear definition of the specific Healthcare Business Case to be addressed by Track&Know. The following steps were required to outline the Scope:

- Description of the health care service under consideration at Royal Papworth Hospital, as well as the components that will be subject of analysis
- Description of the health service business case
- Questions to be answered with the BD analytics
- Description of the contribution of BDA beyond the current state of the art
- Description of the impact (national, international) and relevance to healthcare business cases in other domains
- Clarification as to what is out of scope

As a next step the current state of the art in health care analytics was examined, as well as available tools and methodologies to answer the specific business questions.

Subsequently, other business domains that might have similar business questions and analytical needs were examined, along with tools and methodologies deployed there, which might serve as a benchmark for the healthcare business case.

Finally, relevant open data sources were sought in order to add value to the analytics of RPH's data.

Several Methodologies to define the scope and extract the market information were deployed in parallel:

- Analysis of written service descriptions provided by RPH
- Interviews of RPH staff
- Desktop research to gain a general understanding of the service at global scale, keywords…, results
- Desktop research (publications, websites, online textbooks) to gain an understanding of: current analytics deployed in similar health services, use of BD in health services, use of BD outside the health domain but for comparable business questions, open data sets available.
- Activation of partner networks to provide expertise and request for information from e.g. NHS Digital, CCG, Societies, HOPE.be (European Hospital and Healthcare Federation)

8.1 Healthcare Service Description

8.1.1 The Medical Condition

Obstructive Sleep Apnoea syndrome (OSAs) is a prevalent chronic sleep- related disease, associated with a varying degree of upper airway collapse during sleep. If not diagnosed and treated OSA can lead to the development of cardiovascular and cerebrovascular disease, diabetes and increased mortality, as well as to an increased risk of Motor Vehicle Accidents (MVAs). Due to daytime sleepiness people with the condition are up to 5 times more likely to be involved in motor vehicle accidents (MVAs) than people without the disorder.

Reported estimates of prevalence vary due to variable study approaches. A systematic review (Myers et al. 2013) estimated a prevalence range of 2 to 14 percent. The incidence and prevalence of OSA have been increasing in most developed countries due to rising rates of obesity (Peppard et al. 2013, Young et al. 2009). In the UK for example the proportion of the population recorded as obese has risen from 15% in 1993 to 27% in 2015. OSA affects an estimated 1.5 million people in the UK, up to 50% are undiagnosed, therefore untreated.

In short, OSA is currently a significant health care burden in the UK, for individuals, for the NHS, and for society as a whole. This burden can be decreased if diagnostic services could be provided in a timely

manner where they are most needed, to allow initiation of treatment. The mainstay of treatment in the UK is a home device, which provides continuous positive airway pressure (CPAP) during sleep time, as recommended by the National Institute for Clinical excellence / NICE in 2008.

With effective treatment the symptoms of OSA can be reversed with improved cognitive function (Dalmases et al. 2015) and reduction in blood pressure in some individuals (Pedrosa et al. 2013). The risk of MVC's has been shown to return to baseline with effective CPAP treatment (George et al. 2001). The impact on reduction of risk of cardiovascular and cerebrovascular morbidity and mortality needs further studies (McEvoy et al. 2016).

8.1.2 The Obstructive Sleep Apnea / OSA Service

OSA services in the UK

The provision of services for the diagnosis and management of the condition is not equitably distributed around the country. In many cases, the service is less well developed in areas where the risk of the condition as calculated by the demographics of the population, is highest (Rejón-Parrilla et al. 2014). As it is deemed necessary for service planning, the relative risk of OSA across the UK has recently been estimated and plotted geographically (BLF OSA Toolkit). Local population datasets were used to calculate relative risk for each of the health areas across the four nations (Clinical Commissioning Groups (CCGs) in England, Health and Social Care Trusts (HSCTs) in Northern Ireland, NHS Health Boards (HBs) in Scotland and Local Health Boards (LHBs) in Wales). In addition to the risks the map (see Figure 25) also shows the locations of known NHS sleep clinics, and colour codes them according to the type of sleep study undertaken there if known. The total number of identified sleep units was 289. There were large differences in the number of available sleep centres per health area, ranging from no sleep centres in 66 health areas to 9 in one large urban area. Some people in areas without sleep centres live close to an urban area and therefore have a short travel distance to access sleep services. Others, specifically in more rural areas or islands, where the population tends to be older and more likely to have OSA, have far greater distances to travel.

The provision of care in the UK has struggled to keep up with the increasing number of patients presenting for diagnosis both through increased prevalence of the disorder but also increased awareness. The NHS publishes median waiting times for 15 categories of diagnostics including, for example, endoscopies and radiological tests. In the 6 months from August 2017 to January 2018 sleep studies had the longest wait of all 15 categories (NHS Statistics).

OSA services at Royal Papworth Hospital / RPH and region

The Respiratory Support and Sleep Centre (RSSC) at RPH is one of the largest in the UK and was the first, and remains one of only two, to be accredited by the European and British Sleep Societies as a mark of the quality of its service. Striving to go beyond state of the art and having observed potentially significant efficiency gaps, RPH intends to use the Track&Know platform to obtain service delivery analytics in an innovative and more informative way. The medical business case will focus on the effective delivery of the diagnostic service, which comprises the main activity of the interaction between the health care system and patients in OSA and where efficiency gains will be measurable (KPI 1: reduction of unnecessary travel, KPI 2: improvement of service response times).

The diagnostic service is managed in a network with the following geographic layout:

Central site RPH	= 1
Outreach sites (clinics)	= 7
GP practices	= 13

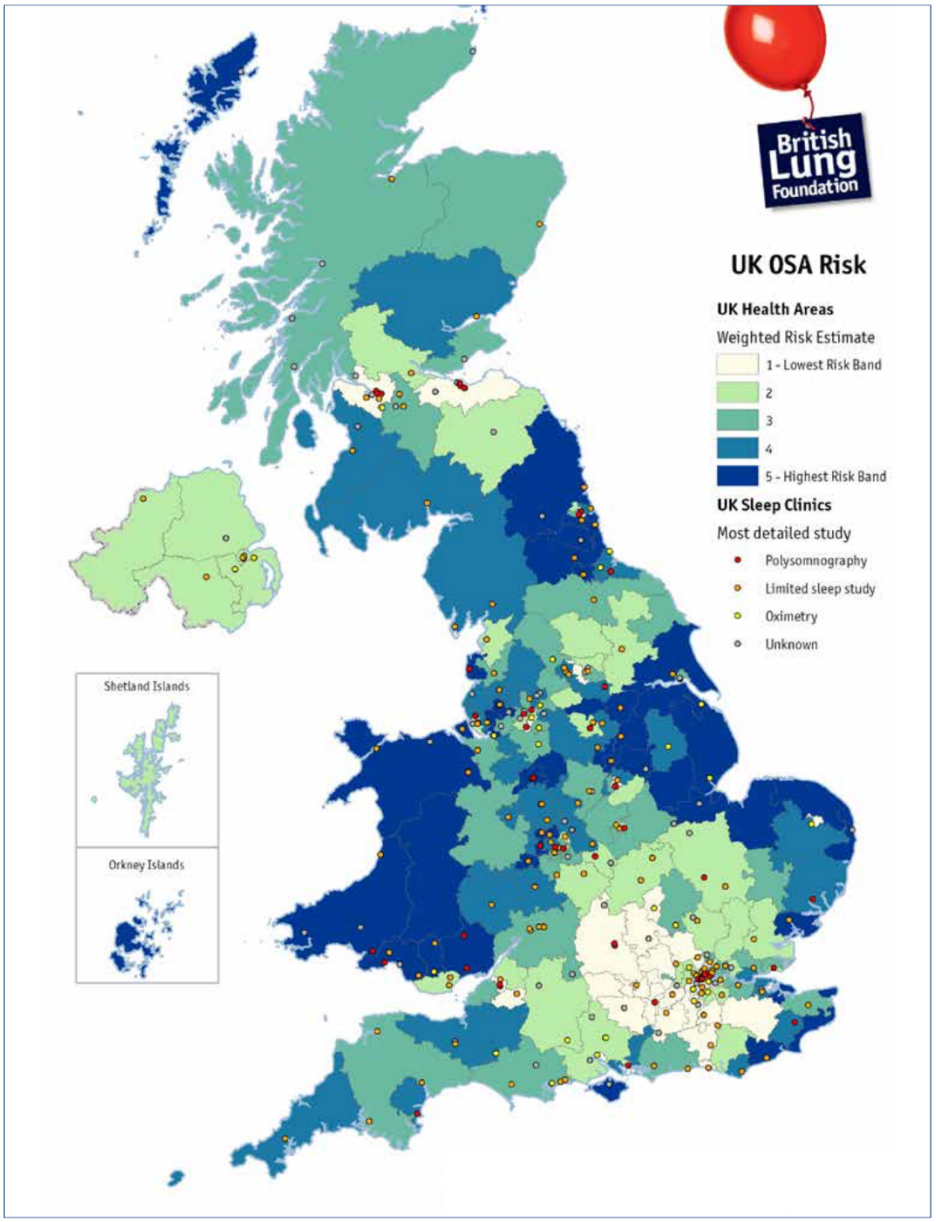


Figure 25: OSA risk map.

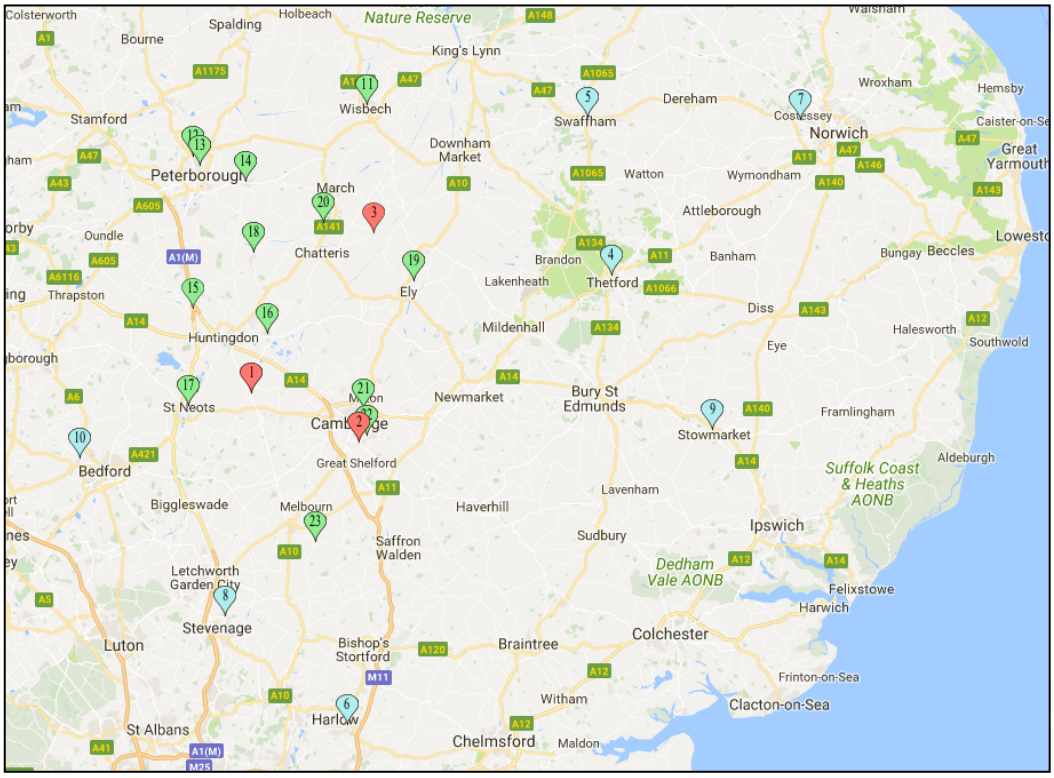


Figure 26: Royal Papworth Hospital RSSC network.

In order to perform the initial diagnostic test, a patient has to travel from home or work to be seen at one of the OSA service sites and has to be prescribed a wearable monitoring device, which he/she takes home, to measure the oxygen saturation during sleep for one night. The device has to be then brought by the patient from home back to one of the sites, the results are downloaded for evaluation by medical staff and a decision is made regarding need for further investigations or need for treatment.

Depending on patient case, at some point later this same diagnostic test has to be repeated, i.e. the patient needs to be prescribed a test with a wearable monitoring device, which he/she picks up, monitors overnight, returns etc. This set of activities may have to be performed in regular intervals in some patients, this is especially true for DVLA (annual test repeats for heavy vehicle drivers and 3 yearly for other drivers).

The OSA service has three different types of diagnostic monitoring devices, two of those types are portable:

Device type 1 /DT1 = 160 (pulse oximeter, small)

Device type 2/DT2 = 6 (respiratory polygraphy, medium size, portable)

The devices are not stored in one place, rather they move around depending on perceived demand. However currently their storage and distribution across different sites is not based on precise analytics, therefore there are always devices that are not in use. At the same time some patients wait for devices to become available in order to pick them up, take them home and have the sleep study.

In order to save too long trips for patients, medical staff from Papworth Hospital are rostered to provide outreach clinic services at an outreach site in regular intervals. That means on certain days staff moves to the outreach clinic sites to see patients. However, this rostering is not necessarily aligned with the regional demand. At the same time, decisions on which patient should drive from his address to which clinic are managed in a reactive manner, depending on where the staff is rostered for appointments as well as on the available device location and the return location instructions for the result reading.

In summary, currently the service-staff-device-patient locations are matched manually and there is no business intelligence in the system to determine the best possible locations to save unnecessary travel and reduce waiting time for device availability and diagnosis.

Technically speaking, the OSA service entails movements of multiple subjects and objects across multiple locations to specific "events". The movements are currently captured by event logging (doctors, appointments, locations, patients, devices) and there is need for visualization of the service.

Track&Know aims to capture these movements within the OSA service by applying visual analytics and big data tools. From September 2018 the VA will be aided by the provision of GPS location signals to track the DT1s.

This will enable the design and validation of a real-time model, which will also allow simulation of effectiveness of variations of the service flow, as well as predictions for need and recommendations for most effective deployment.

The models should be scalable to deploy at national and international scale, after successful validation at regional scale.

8.1.3 The Economic Significance of OSA

This section summarizes cost figures published by the British Lung Foundation in the BLF OSA Toolkit.

Once diagnosed, OSA is relatively easy to treat, and its treatment has been proven to be cost effective. For many years, research has shown that treatment is associated with better outcomes for the individual and for society. And yet, only an estimated 330,000 adults are currently being treated in the UK, out of an OSA population of 1.5 million. As evidenced by the long waiting times mentioned in the previous section, the diagnostic services are the bottleneck.

Treating OSA would save the NHS millions of pounds. The estimated annual savings to the NHS in the UK would be £28 million and 20,000 extra QALYs (quality-adjusted life years), if all people with moderate to severe OSA were to be diagnosed and treated, compared with the current estimated level of treatment of 330,000 adults. These estimates of NHS cost savings are due to reductions in consequential acute events (including stroke, cardiovascular events and road accidents) resulting from treatment with CPAP.

OSA is not just a risk for the individual, but also to society as a whole due to the increased risk of serious accidents that lead to fatalities, lifelong disability, emotional distress and broader societal costs (BTS 2014). One fatal accident is estimated to cost approximately £1.5 million to society (Mackay et al.2010). Sleepiness accounts for around 20% of all road collisions and many of these are likely to be caused by people with undiagnosed OSA. It is estimated that if everyone likely to have moderate to severe OSA in the UK were treated, this could result in 40 000 fewer road accidents each year relative to the current level of treatment. As some of these accidents result in injury or even fatality, the health gains are considerable.

The impact on future savings is even bigger if one takes into account that the prevalence of OSA will rise in the coming years, particularly due to an increasing prevalence of obesity and the increasing age of the UK population.

8.1.4 Questions to be answered with BDA

- Does the historical data set provide evidence for ineffective service flow, e.g. mismatch between demand and device and clinic availability?
- Would a different device management model be more effective to improve device availability and travel distances for pickup/drop-off by patient, as well as actual device transportation distances (e.g. by nurses, couriers)?
- Can the schedule and frequency of outreach clinics be modified to better fit demand and travel distances (both of patients and staff).

- Would different clinic locations provide a more effective service (measured by a higher number of patients seen compared to the optimized number in the current service settings).
- Does the number of new patient referrals correspond to the areas of high need?
- Do patients from affluent areas (by postcode) of one town have better access to services than patients from less affluent areas of the same town?
- By how much can the waiting times for diagnosis be reduced by optimization of device and clinic availability, given the current staff and device numbers?
- What is the minimum staff and device number needed to provide a service without waiting times?
- What are the savings from shortening of travel distance, as well as the environmental impact on CO2 emissions.
- Does waiting time and travel distance shortening impact on patient satisfaction?
- Can Track&Know propose an improved service flow and delivery model, which can be scaled and implemented nationwide?

While some of the above business questions could be answered by applying existing business intelligence methods, it is important to point out that currently there is no model that would allow answering all questions en bloc and testing different models of care via simulation. By visualizing the service flow in context of prevalence and needs, T&K takes operational research in the health domain beyond the state of the art. In addition, the application of big data tools in T&K will enable simulations of service flow, predictions, as well as adjustments of service response based on real-time demand and capabilities.

8.1.5 Out of scope

It is important to understand the limitations of scope and these pertain to the differences between service decisions and medical treatment decisions. While the T&K model will inform the OSA service administrator about effective service flow and will lay the foundation for decision support points regarding service optimization, it is outside the scope of this project to provide analytics for medical decision-making. For example, one research question could be: Can the referral information/patient profile and/or the monitoring output help predict the patient case management in order to make further savings, e.g. a patient with specific profile that predicts that he is unlikely to have a conclusive result from the mobile device and hence should not waste time and travel, the diagnostic test should be done in the sleep lab in the central site.

While T&K will deploy machine learning, it will serve the purpose of operational decision support. Creation of medical decision support would be a future step, i.e. after the T&K platform has been deployed and validated. Although BDAs are used increasingly for clinical pattern recognition and decision support, it is necessary as a first step to build the actual system, which will visualize the service and model different flows in order to gain better understanding. After that, the decision points can be inserted and the algorithms to capture medical decision making can be provided.

8.2 Current State of the Art in Health Service Flow Analytics, Available Tools and Methodologies

Health service managers have a rather low awareness of what BD analytics and VA can offer, and the adoption of operational research solutions in healthcare is in general disappointingly low (Brailsford et al. 2009).

System flow improvement has a crucial role to play in driving up service quality and productivity. The importance of flow is increasingly recognised by practice leaders and policymakers throughout the UK. For example, there have been recent flow improvement programmes in both Scotland (NHS Scotland) and Wales (1000 Lives). The concept of improving flow is also referenced nationally and locally, across

the UK, in strategies for service configuration and for tackling emergency and elective access challenges (Monitor). Where providers have been able to match capacity and demand and enable better flow between departments and organisations, there have been impressive results. However, to date, virtually all attempts to improve flow have focused on single organisations or pathways (Fillingham et al. 2016).

Most of the improvement work relates to in-hospital services and not to the provision of a multi-site service with multiple moving components, such as the OSA service. Our research has not identified any application of visual analytics for service flow improvements. Some common approaches that have been used in the UK and other countries to understand flows across complex organisations or care journeys with many variables and interrelationships are described below.

Simulation and modelling

Simulation and modelling of patient or service user flow can provide insight into where bottlenecks occur in a health care system. They allow service planners to evaluate the benefits and pitfalls of potential improvements before enacting them.

In health care, simulation and modelling approaches have been used to manage bed capacity, schedule staff, manage admission and scheduling procedures, and to test the value or functionality of new initiatives and services before they are implemented. For example, a Swedish hospital has used a simulation model to support discussions about the resources, capacity and work methods that would be required on a maternity ward that was shortly to be built.

Value stream mapping

Value stream mapping (VSM) is an approach that produces a visual map of a system or process. It is often used by multidisciplinary teams to improve processes as part of lean/continuous improvement projects.

Using VSM, a team can produce a visual map of the 'current state', identifying all the steps in a patient or service user's care journey. The team then focuses on the 'future state', which often represents a significant change in the way the system currently operates. Of course, this also means that the team needs to develop an implementation strategy to make the future state a reality.

Using VSM can result in streamlined work processes, reduced costs and increased quality and this method has been included in the NHS guidelines (NHS-III, NHS-SQI).

As an example of applying VSM in practice should be mentioned that in Ireland, researchers used lean principles and the theory of constraints to identify bottlenecks in patient journeys through A&E. For each stage of the patient journey, average times were compared and disproportionate delays were identifed using a significance test. A value stream map and the five focusing steps of the theory of constraints were used to analyse these bottlenecks (Ryan A et al, 2013).

Queuing theory

Queuing theory, or the study of waiting lines, or queues, can help to understand and address mismatches between service demand and capacity. Usually a mathematical model is constructed to help predict queue lengths and waiting times. Historical data are analysed to explore how to provide optimal service while minimising waiting, thus providing an objective method of determining staffing needs during a specific time period. Popular in other industries, queuing theory has also been used in health care, particularly by hospitals wanting to understand waiting times for unscheduled care or the time spent waiting for specific equipment, surgery or laboratory results. It is also applicable to wider systems of care or transitions.

For example, a hospital in England used queuing theory to analyse one year's worth of data to help understand the practical challenges associated with variation in patient demand for services and length of stay Allder et al, 2010). The analysis found that daily bed shortages are mostly influenced by

the timing of arrival and discharge of patients with a short length of stay, and that bed shortages around holiday periods are not due solely to increased demand, but also a reduction in staff and service capacity in and out of hospital around these times.

The adoption of Operational Research / OR solutions in healthcare is, in general, disappointingly low. Although In some organisations a great deal of data is being generated, they are not translated into actionable knowledge or effective organisational responses (Dixon-Woods et al, 2013).

There are a similar set of problems faced in patient transport services, with the scope to balance patient journey times, aggregated route lengths and the times between drop-off (pick-up) and the start (end) of patient appointments (Bowers et al, 2012).

Location and allocation modelling

The geographical placement of health services can influence access, use and equity in health service delivery. Location and allocation modelling enables different options for the placement of clinics, ambulance stations etc. to be evaluated in terms of the average distance or time travelled by patients to access services or the proportion of demand a population that lies within a certain distance or travel time of a service. It also permits the optimal placement of one or more clinics within a region to be determined (Li et al, 2012), although this is one area where it cannot always be assumed that an agreed and explicit objective function exists. The availability of sophisticated GIS systems has enabled the development of location models that incorporate that incorporate time-varying patterns of demand and journey times.

Although notions of services having mutually exclusive geographical catchment areas are not consistent with current policies concerning patient choice, allocation modelling still has useful insights to offer, particularly for non-elective services and when patient journey times are considered to influence strongly choices by patients as to which services to use. We provide in Figure 27 and Figure 28 below two examples for location modeling (ORAHS07).

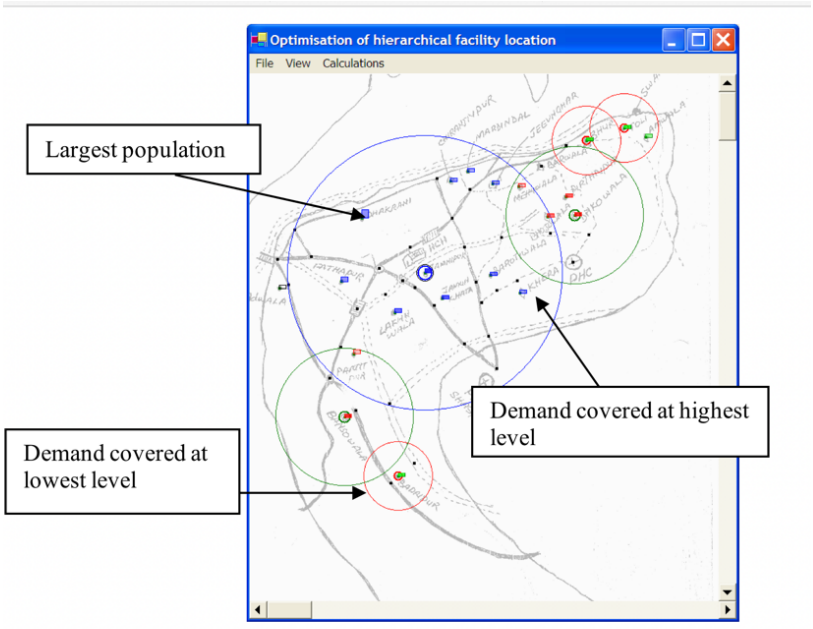


Figure 27: Location modeling facilities vs. demand

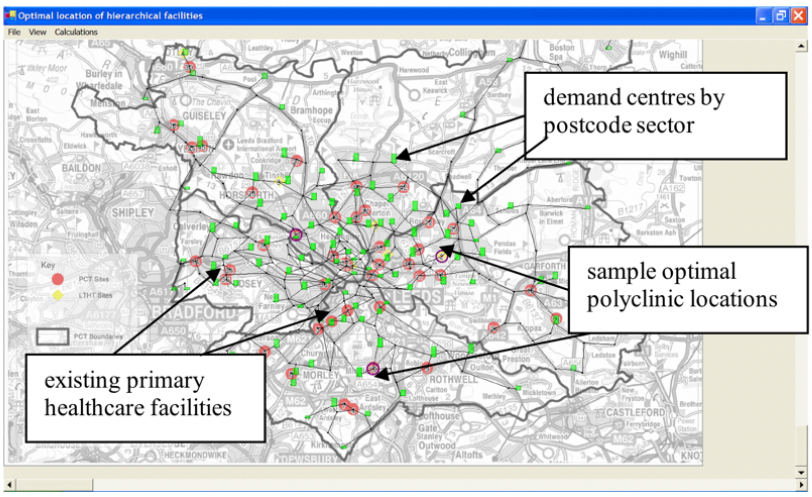


Figure 28: Location modeling optimal location polyclinics to bus routes

Modelling remains however an alien form of evidence to many in the health service and to other academic disciplines and creative work to improve the communication of what modelling is and how it differs from other research methods is needed.

8.3 Other business domains with similar business questions

In the current chapter we are going to provide a relevant example from another business domain with similar business questions, that has applied specific mathematical and geographical models to meet the needs.

Bike Sharing

Bike-sharing, or public bicycle programs, have received increasing attention in recent years with initiatives to increase cycle usage, improve the first mile/last mile connection to other modes of transit, and lessen the environmental impacts of our transport activities. Originally a concept from the revolutionary 1960s, bike-sharing's growth had been slow until the development of better methods of tracking bikes with improved technology. This development gave birth to the rapid expansion of bike-sharing programs throughout Europe and now most other continents during this decade. Today almost 1,000 bike sharing schemes exist and the market is expected to grow by 20% by 2020. See Figure 29 (Roland Berger) for global presence overview.

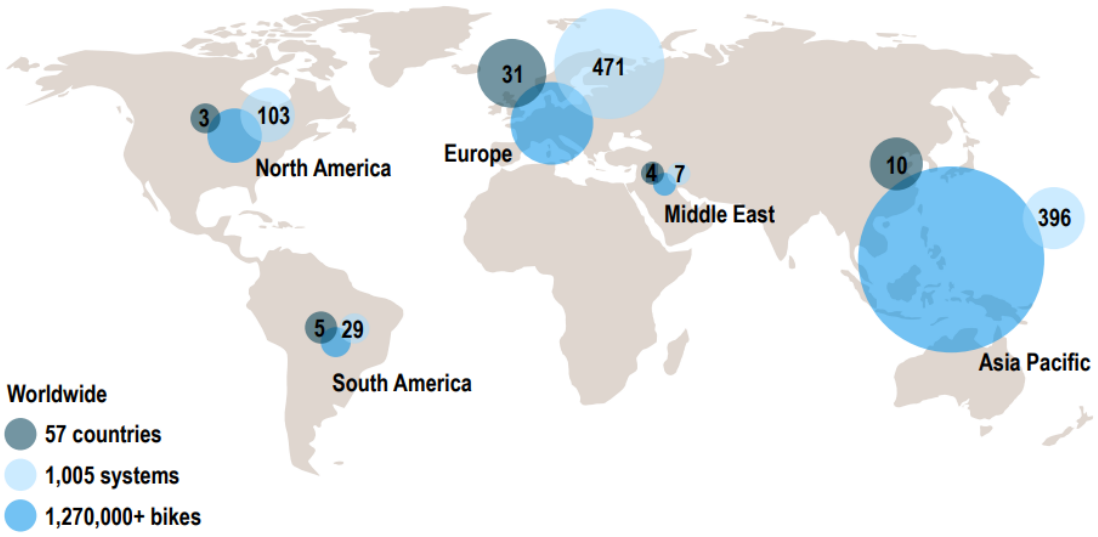


Figure 29 Global presence of bike sharing systems December 2015

Successful examples of Bike Sharing schemes in Europe are Santander Cycles in London, Bicing in Barcelona and Velib in Paris.

Basic Bike Sharing Models

During the evolution of Bike sharing throughout the years two basic sharing models have dominated in the market, trying to cope with specific network developing problems.

a) Free-Floating Bike sharing: This solution refers to Inner- city rentals without any fixed pick –up points within a defined urban area, where bicycles can be picked out and dropped off at any intersection and the transaction can be performed usually by phone or app.

b) Station-based bike sharing: This model includes Inner- city rental of bicycles from specific pick –up points where people can rent and return the bikes to these specified pick-up points and the transaction can normally occur at a user terminal at the station or by app. This model is similar to the envisioned OSA oximetry management model in T&K.

Challenges risen and problem-solving methods on establishing a Bike Sharing Network

In order for the operators to meet with the above key factors in a successful way, and especially with the first problem of "where should they situate the dock stations and how to develop the bike lanes network", they have to use several mathematical and geographical methods and models. The most widely known and used of the mentioned methods are queuing networks (queuing theory) and location and allocation method. In the following paragraphs the two methods are described providing also an overview of the problems they were used to provide solution.

Queuing Theory: The main objective of this theory is to reduce the average time a customer spends in the system, focusing on customer wait time as well as other areas that can be improved. For design and operations of the bike sharing systems, it has become a basic and interesting topic to assess and ensure the quality of service from a user's perspective. In general, the quality of service of a bike sharing system may be evaluated from two basic points: (a) The bike non-empty. Some bikes have been parked at the stations such that any arriving customer can rent a bike from his entering station. (b) The parking non-full. Some parking places (or lockers) become empty and available so that a rider can immediately return his bike at a destination station. Based on the two points, the bike-empty or parking-full stations are called problematic stations, while the probability of problematic stations can be used to measure the quality of service of the bike sharing system. In general, computing the

probability of problematic stations is always very difficult and challenging. On the other hand, it is worthwhile to note that recent interesting research of bike sharing systems is also related to the probability of problematic stations. In this case, the queuing theory should be one of the best approximate methods for understanding dynamic behavior of more general bike sharing systems.

Location and Allocation Method: This method is widely used in bike sharing business in order to determine potential locations for docking stations in the selected area. The main objective is, to locate specific sites for bike-share stations out of several candidate locations based on demand, using a geographic information system (GIS), more specifically by making use of a location-allocation analysis. The entire process from inputting data to performing a multi-criteria analysis can be summed up in 4 generalized steps:

1. *Find facilities*: meaning to locate potential sites for bike-share stations.
2. *Determine demand:* locate demand for the service.
3. *Define barriers:* define any point/line/polygon barriers.
4. *Calculate result:* find the best locations.

The outcome is usually a map with the most appropriate locations for siting bike sharing docks. This analysis could be much more detailed depending on the result we target to e.g. general population, tourists, working group, nighttime population, daytime population.

8.4 Related Open Data

The following open data sets have been identified as sources of information relevant to the medical pilot.

- http://geoportal.statistics.gov.uk. The Open Geography portal from the Office for National Statistics (ONS) provides free and open access to the definitive source of geographic products, web applications, story maps, services and APIs. All content is available under the Open Government Licence v3.0, except where otherwise stated.
- https://www.ons.gov.uk/search?q=health&size=25&page=3. Provides file rft−−−mid−2012 Health−geography−population−estimates.zip. Health geography populations estimates. Population by gender and age (years, not categories) for 2012 for each Clinical Commissioning Group / CCG.
- https://ons.maps.arcgis.com/home/item.html?id=726532de7e62432dbc0d443c22ad810f NHS Postcode Directory UK Full (May 2018). This contains a table 'CCG names and codes' and the dataset possibly allows to get the number of addresses in each CCG. Coordinates are not specified (see also property constraints on 'Ordnance'and 'Royal Mail' data. Hence synthetic disaggregation will be required. Caution: the documentation mentions 195 CCG's instead of 211 CCG's, requires further clarification.
- https://borders.ukdataservice.ac.uk/bds.html; https://borders.ukdataservice.ac.uk/easy_download_data.html?data=England_ccg_2013. Shapefile for English NHS CCG's. shp and csv files specifying the 211 CCG's for England. Identifiers in the shp are compatible with those in rft−−−mid−2012−health−geography−population−estimates.zip
- https://census.ukdataservice.ac.uk/use−data/guides/boundary−data. UK Data Service: Census Support Guides (manuals) about the use of "digitised boundary datasets". These contain "Ordnance" data, clarification is required which of those are publicly available.
- http://www.nomisweb.co.uk/census/2011. Census 2011 data tables.
- http://www.nomisweb.co.uk/census/2011/origin_destination. Origin−destination data (also known as flow data) will include the travel−to−work and migration patterns of individuals,

cross−tabulated by variables of interest (for example occupation). Uses "local authority districts", requires alignment with the codes used in the CCG tables.

- http://geoportal.statistics.gov.uk/datasets?q=ONS%20UPRN%20Directory%20%28May%202018%29&sort=name. The ONS UPRN Directory (May 2018) User Guide: TTWA (Travel to work areas) is of interest (self contained zones for work commuting) see ONSUD_User_Guide_May_2018.pdf sec:17. This guide/manual could provide the details about how to align CCG identifiers with identifiers used in the census/2011/origin_destination files.
- https://www.ordnancesurvey.co.uk/business−and−government/products/addressbase.html. Supplies an address database but it is not open/free (could have been useful for address disaggregation).
- http://results.openaddresses.io/. Street address databases for most countries, unfortunately excluding UK, but could be used for the expansion of the validation beyond the UK.
- http://geoportal.statistics.gov.uk/datasets?q=LSOA_CCG_LAD_LU&sort=name. A lookup file between 2011 Lower Layer Super Output Areas (LSOA), to clinical commissioning groups (CCG) and local authority districts (LAD) in England, as at 1 April 2018. (File size − 8MB). May be useful code/id conversion table; there are multiple versions: see http://geoportal.statistics.gov.uk/datasets/lower−layer−super−output−area−2011−to−clinical−commissioning−group−to−local−authority−district−april−2018−lookup−in−england
- http://geoportal.statistics.gov.uk/datasets/rural−urban−classification−2011−of−ccgs−including−population−in−england. Rural Urban Classification (2011) of CCGs including population in England. CCG's are classified according and for each CCG the population is specified for multiple rural and urban type parts separately (may be useful for predictions). However, no geographical data are provided to describe the parts. file=Rural_Urban_Classification_2011_of_CCGs_including_population_in_England.csv.
- https://www.england.nhs.uk/statistics/statistical-work-areas/diagnostics-waiting-times-and-activity/monthly-diagnostics-waiting-times-and-activity/monthly-diagnostics-data-2017-18/. Provides number of diagnostic tests and waiting times. https://calculator.blf.org.uk/?_ga=2.54404523.268454387.1528899844-830150035.1524770270. Provides statistics about the propensity vs. population density. However, manual search by health area needed to extract numbers.
- http://ec.europa.eu/eurostat. Population health data useful for potential expansion of simulation to EU level.

The open data list above may be expanded further as required.

8.5 References

Allder S, Silvester K, Walley P. (2010) Understanding the current state of patient flow in a hospital. Clinical Medicine, 10(5):441–4. www.ncbi.nlm.nih.gov/pubmed/21117373

BLF OSA Toolkit – A Toolkit for commissioning and planning local NHS services in the UK (2015) British Lung Foundation, UK. URL: https://www.blf.org.uk/sites/default/files/OSA_Toolkit_2015_BLF_0.pdf

Bowers J, Lyons B, Mould G (2012) Developing a resource allocation model for the Scottish patient transport service, Operations Research for Health Care 1:84-94

Brailsford SC, Harper PR, Patel B, Pitt M (2009) An analysis of the academic literature on simulation and modelling in health care, Journal of Simulation, 3:130-140

BTS 2014 – British Thoracic Society: Position statement on driving and obstructive sleep apnoea (OSA)

Dalmases M et al. Chest. Effect of CPAP on Cognition, Brain Function, and Structure among Elderly Patients with OSA: A Randomized Pilot Study. Chest 2015;148: 1214-1223

Dixon-Woods M, Baker R, Charles K, et al. Culture and behaviour in the English National Health Service: Overview of lessons from a large multimethod study. BMJ Quality & Safety 2013

Fillingham D. , Jones B, Pereira P. (2016) The challenge and potential of whole system flow, The Health Foundation, URL: https://www.health.org.uk/sites/health/files/ChallengeAndPotentialOfWholeSystemFlow.pdf

George GF. Reduction in motor vehicle collisions following treatment of sleep apnoea with nasal CPAP. Thorax 2001;56:508–12

Li X, Zhao Z, Zhu X, Wyatt T (2011) Covering models and optimization techniques for emergency response facility location and planning: A review, Mathematical Methods of Operations Research, 74:281-310

Mackay, T. 2010: OSA working towards the development of minimal standards for referral, investigation and treatment in Scotland

McEvoy RD et al, CPAP for Prevention of Cardiovascular events in Obstructive Sleep Apnea. N Engl J Med 2016; 375:919-931

Monitor. Improving patient flow: Evidence to help local decision-makers. London: Monitor; September 2015. URL: www.gov.uk/government/uploads/system/uploads/attachment_data/file/499229/Operational_ productivity_A.pdf

Myers KA, Mrkobrada M, Simel DL. Does this patient have obstructive sleep apnea? The Rational Clinical Examination systematic review (Structured abstract). JAMA. 2013;310(7):731-41. PMID: DARE-12013049250

NHS-III - NHS Institute for Innovation and Improvement. The Handbook of Quality and Service Improvement Tools. NHS Institute for Innovation and Improvement, 2010. www.nhsiq.nhs.uk/media/2760650/the_handbook_of_ quality_and_service_improvement_tools_2010.pdf

NHS Scotland Quality Improvement Hub. Whole System Patient Flow Improvement Programme. Available from: www.qihub.scot.nhs.uk/quality-and-efficiency/whole-system-patient-flow.aspx

NHS – SQI - NHS Scotland Quality Improvement Hub. Value Stream Mapping. Available from: www.qihub.scot.nhs.uk/ knowledge-centre/quality-improvement-tools/value-stream-mapping.aspx

NHS Statistics - https://www.england.nhs.uk/statistics/statistical-work-areas/diagnostics-waiting-times-and-activity/monthly-diagnostics-waiting-times-and-activity/monthly-diagnostics-data-2017-18/

NICE. 2008. Continuous positive airway pressure for the treatment of obstructive sleep apnoea/hypopnoea syndrome. NICE Technology Appraisal guidance 139. London

ORAHS07 - https://eprints.soton.ac.uk/64208/1/HKSmith_ORAHS07_submit.doc

Pedrosa PD et al. Effects of OSA Treatment on BP in patients with Resistant Hypertension. 2013; 144: 1487–1494

Peppard PE, Young T, Barnet JH, et al. Increased prevalence of sleep-disordered breathing in adults. Am J Epidemiol. 2013 May 1;177(9):1006-14. Epub: 2013/04/17. PMID: 23589584

Rejón-Parrilla JC, Garau M and Sussex J Obstructive Sleep Apnoea Health Economics Report Office of Health Economics September 2014. https://www.blf.org.uk/support-for-you/obstructive-sleep-apnoea-osa/health-care-professionals/health-economics-report

Roland Berger. https://www.rolandberger.com/publications/publication_pdf/roland_berger_bike_sharing_4_0.pdf

Quan-Lin Li, Rui-Na, "A Mean-Field Matrix-Analytic Method for Bike Sharing Systems under Markovian Environment" Fan School of Economics and Management Sciences. China, February 2018

Quan-Lin Li, Chang Chen, Rui-Na Fan, Liang Xu and Jing-Yu Ma, "Queueing Analysis of a Large-Scale Bike Sharing System through Mean-Field Theory" School of Economics and Management Sciences Yanshan University, China, December 2016

Carter Xin, Locating Potential Bike-Share Stations in Vancouver, http://blogs.ubc.ca/vanbikesharestations/

Santander Cycles: https://tfl.gov.uk/modes/cycling/santander-cycles

Ryan A, et al. STEPS: Lean thinking, theory of constraints and identifying bottlenecks in an emergency department. The Irish Medical Journal, 2013;106(4):105–7

Young T, Palta M, Dempsey J, et al. Burden of sleep apnea: rationale, design, and major findings of the Wisconsin Sleep Cohort study. WMJ. 2009 Aug;108(5):246-9. Epub: 2009/09/12. PMID: 19743755

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/613532/obes-phys-acti-diet-eng-2017-rep.pdf

1000 Lives. The Patient Flow programme. Available from: www.1000livesplus.wales.nhs.uk/flow (accessed 27 September 2016)

9 Market Analysis – the Transport Business Case

The rapid progress of information and telecommunication technologies have resulted major leaps in transportation data collection and management practices. Connected and automated vehicle systems exploit these technologies resulting the potential to enhance fleet operations, transform transportation system management, improve data collection methods and modify the transportation data exploitation. For example, in a request for information (RFI) about the Connected Vehicle Pilot Deployment Program issued by the US Department of Transportation (USDOT), "data" was the most frequent word that appeared summed across all responses[3] (Figure 30).

Figure 30 'data' as the most frequently used word in connected car RFIs.

9.1 Big Data in the Transport Industry

Big data is a term used for very large data sets that have more varied and complex structure. These characteristics usually correlate with additional difficulties in storing, analyzing and applying further procedures or extracting results. Big data analytics is the term used to describe the process of researching massive amounts of complex data in order to reveal hidden patterns or identify secret correlation. Big data sets generally have some or all of following features:

- Digitally generated
- Passively produced
- Automatically collected
- Geographically or temporally trackable
- Continuously analyzed

Thanks to on-board devices, sensors, and wireless connectivity, big data have been expanding into the automotive and transportation sectors. For example, the automotive industry is using big data to improve operational efficiency in designing, building, and servicing vehicles. Transport operators are also increasingly connecting data from vehicles and people's behavior to the data about the environment in which the vehicle is operating (weather, traffic, hazardous situations, etc.). Furthermore, the transport sector's increasing ability to track the location of mobile devices has enabled both the monitoring of traffic to save time and reduce congestion.

For example, the trucking industry is using telematics and electronic on-board recorders (EOBRs) to collect data and communicate in real time to improve safety and operational performance. The future of big data in trucking will be cross-referencing real-time driver data with data on weather, parking availability and traffic delays to deliver information to the driver as quickly as possible. Many of these applications are already providing encouraging and potentially lucrative results. Many transportation

3 *"Connected vs. Automated Vehicles as Generators of Useful Data"*, Center for Automotive Research, Michigan Department of Transportation, September 2014

agencies see big data and its applications as an opportunity to improve the management and operation of transportation systems, increase the accuracy of prediction, enable informed decision making, and optimize transportation services.

However, data collections are growing too fast, or are becoming too complex for existing information technology platforms to handle them. The challenges include capture, curation, storage, security, search, sharing, transfer, analysis and visualization. Track & Know has identified these needs and is developing specialized applications (Toolboxes), tailor-made to the specific needs of the Transport industries. In the remaining section we are focusing on Fleet Management domain which is on of the three Track and Know Demonstrators, the other two being: car insurance; and, healthcare telematics.

9.2 Fleet management systems

A modern Fleet Management system requires continuous real time monitoring and updates from the vehicles in order to achieve high efficiency and quick responses to the fleet owner and to the fleet operator. Vehicle tmonitoring uses Technologies such as mobile communications, GPS (Global Positioning System) and GIS (Geographical Information Systems), combined with information systems that are storing collected data and provide the applications to end users (customers). Such integrated systems help to track the position of the vehicle while it is operating to deliver the products[4].

The key areas are of a Fleet Management System are:

- Acquisition of data: Vehicles are equipped with an embedded device that is sending data regarding the current state of vehicle or any information other information collected during the vehicle operation. The geographic position is automatically captured from the embedded GPS receivers within the vehicle. Exact location can also be determined with the help of mobile phone operators mobile network positioning (Cell information)
- Communications: In order to receive and transmit information between the vehicle and the fleet operators data center wireless mobile technologies need to be employed, for example GPRS, 3G etc.
- Information management systems: These are system that are responsible for the management of data received. Information is stored and kept for a long time period (1-3 years), providing records of positioning and vehicle information.
- Fleet Portal and applications. In turn, information is compiled by the application server and user interfaces that reflect the needs of the customer are used to deliver the Fleet Management Application. Fleet management application provide combination of various functions of operational control. The fleet applications to generate information that will help fleet operators to take informed decisions related to the proper planning of vehicles. Therefore, fleet application require a certain flexibility in order to adapt to different types of vehicle fleets. Also the fleet applications may need to integrate with a Enterprise Resource Planning applications already being used in the customers enabling new ap-plications such as: purchasing, cost structuring, assets control, employee productivity etc. Additionally fleet applications may generate many reports, in an interactive form where the users can use whatever they think is important.
- Integration with enterprise systems, can improve fleet data to include information such as:
    - Tyres: Registration of tyres, shipping, movement and receiving of re-treads, evaluation of the tyre retread performance, number of retreads, statistics.

4 Paul Buijs, J.C. (“Hans”) Wortmann. 2014. Joint operational decision-making in collaborative transportation networks: the role of IT. Supply Chain Management: An International Journal 19:2,200-210.

- Vehicles: Odometers and hour meters, insurance, documentation, traffic tickets, accidents and other events, vehicle age, Truck registration and truck spare parts etc.
- Fuel and lubricants: Controlling total fleet fuel costs can help streamlining and optimizing the fuel supply. There is also potential of checking the average consumption and to identify of abnormalities.
- Fleet production: Integration with highway "no-stop" toll type systems to reduce waiting time and with customers for informing them regarding their shipment. Record of travels, routes, driver, production by vehicle and customers.
- Preventive and corrective maintenance: Collect service orders which help in the integration with the storeroom, purchasing and parts control. Plans for predefined maintenance activities could also be carried out.
- Employees: registration of drivers, records, productivity, operators, salesmen, mechanics, traffic tickets, accidents, Control of documentation etc;
- Costs: Collect approximation of indirect costs and other costs per unit as they occur. Thus, maintaining a record of total cost of fleet ownership.

9.3 Driver behaviour and autonomous vehicles

The ‘driverless car’ concept has received a great degree of attention in recent years. In the UK, the government has pledged the testing of driverless cars since 2013. The future of autonomous vehicles is much favoured by the automotive industry and related stories are currently quoted in the media on an almost daily basis, promoting the plan to initiate the sale of ‘self-driving cars’ by 2020.

Still, from the research perspective, this domain has seen gradual development activities and the impact of such vehicles in the context of road traffic management research, is not yet well developed (Excell, 2013). The technology to design and implement such cars is already advanced and perhaps readily available, still, the challenge for researchers to ensure that the drivers of such vehicles are able to comprehend the capabilities and limitations of the systems in place, is not yet well developed.

Until now human factors research on vehicles automation mainly focused on the interaction with Automatic Cruise Control (ACC) systems. This research revealed that an increase in automation could reduce drivers’ awareness and decrease driver performance, especially during risky conditions (Endsley & Kaber, 1999). In recent years, a number of projects in this area (conducted mainly in Europe) have attempted to progress beyond the ACC, adding lane keeping assistance, for example, and transferring the degree of automation from function specific automation (Level 1) to combined function automation (Level 2) and also limited self-driving (Level 3). Projects that which have considered driver behavior implications in automated vehicles include CityMobil (see Merat & Jamson, 2009; Toffetti et al., 2009), InteractIVe (Hesse et al., 2011) and HAVEit (Happee et al., 2008). In Level 3 automation there is very limited understanding of drivers’ behaviour and performance. In these settings the driver needs to remain ‘in-the-loop’ and maintain his awareness in order to resume control to his driving, if required. Still, research on understanding the human factors of how drivers are involved in the ‘‘occasional control’’ of the vehicle and what constitutes ‘‘comfortable transition time’’ is currently very limited.

Track & Know will demonstrate driver behavior in traditional professional vehicle scenarios. Further to this, the consortium recognizes the evolving autonomous vehicle domain and the impact that operator behavior is having to the success of new commercial vehicles and monitoring solutions.

9.4 Vehicle control and driver assistance

The term ‘Vehicle Control’ refers to the set of tasks required to driving and manoeuvring a vehicle. Advanced electronic systems provide varying levels of assistance to drivers in carrying out these tasks.

Driver-assistance for vehicle control aims to reduce undesirable consequences that driver-only vehicle control can produce. For example road traffic accidents or wasted fuel can be reduced by vehicle control systems, since human factors such as different performance and competence levels vary between different drivers or in the same driver over time.

The increasing availability of information and control system technologies both within and external to vehicles provides the potential to compensate for inconsistent driver performance and decision making. So, the main users of Vehicle Control are the drivers of professional vehicles who use the equipped vehicles in their daily assignment. Driver-assistance functions are addressing the needs of:

- drivers of individual vehicles, providing driver assistance in critical situations;
- fleet operators, allowing for reduced wear and tear on vehicles and improved overall operating efficiency of their fleets;
- freight forwarders, improving reliability of service and reduced risks of transported goods;
- transport infrastructure operators, improving emergency responses, traffic efficiency, driver and road

Track & Know has recognized the gap for fleet management solutions to provide support for driver assistance and vehicle control feedback. Although this option is not within the scope of the project demonstrations, still it is a case that the project would be willing to consider during the exploitation phase.

9.5 Data collection and Open Data sources

Fleet management solutions depend on transportation data collection to support fleet operations and decision-making processes. Data collection starts with the logging and aggregation of vehicle generated data (see section 9.5.1). Vehicle data are stored in fleet systems databases where operational and financial KPIs are monitored. Ideally external data sources such as the rapidly developing intelligent transportation system's technologies and new open data sources, would be correlated to deliver meaningful value-added services.

Track & Know has recognized this potential and will try to include access to open data sources via its Big Data Platform, making these sources horizontally available to all demonstrators in a common and efficient sharing method.

Key data principles for open data sources to be considered in the Track & Know Platform are summarised below[5]:

- VALUABLE - Data are a core business asset that has value and are managed accordingly.
- AVAILABLE - Access to data is critical to performing duties and functions, data must be open and usable for diverse applications and open to all.
- RELIABLE - Data quality is acceptable and meets the needs for which the data are intended.
- AUTHORIZED - Data are trustworthy and safeguarded from unauthorized access, whether malicious,
- fraudulent, or erroneous.
- CLEAR - Data dictionaries are developed and metadata established to maximize consistency and transparency of data across systems.
- EFFICIENT - Data are collected once and used many times for many purposes.
- ACCOUNTABLE - Timely, relevant, high quality data are essential to maximize the utility of data for decision making

5 Proposed by the American Association of State Highway and Transportation Officials (AASHTO)

In our days, the significant increase in the number of personal mobile devices and amount of crows generated information has created new dynamics of the transportation open data sources. The advancement of connected and automated vehicle technologies, as well as, the inclusion of various actors from the private, business, and public sectors) is creating key transformations with respect to big data in transportation.

9.5.1 Vehicle Data

Moving vehicles are increasingly used as data sources for numerous purposes. In a connected vehicle environment, the vehicle subsystem (VS) provides the sensory, processing, storage, and communications functions necessary to support efficient, safe, and convenient travel. These functions reside in various types of vehicles, including automobiles, and commercial, emergency, construction, maintenance, and transit vehicles. Advanced sensors, processors, enhanced driver interfaces, and other on-board units (OBU) are able to record and deliver the data through wireless networks.

The vehicle data may include basic vehicle measures, vehicle safety data, environmental probe data, vehicle diagnostics data, and vehicle emissions data. Specific data elements from connected vehicles include, but are not limited to:

- Vehicle type and characteristics (length, width, bumper height)
- Time stamp
- Speed and heading
- Vehicle acceleration and yaw rate
- Turn signal status
- Brake status
- Stability control status
- Driving wheel angle
- Vehicle steering
- Tire Pressure
- Traction control state
- Wiper status and run rate
- Exterior lights
- GPS status and vehicle position (longitude, latitude, elevation)
- Obstacle direction
- Obstacle distance
- Road friction
- Current and average fuel consumption
- Vehicle emissions data - measured emissions of specific vehicles comprised of exhaust pollutants including hydrocarbons, carbon monoxide, and nitrogen oxides
- Air temperature and pressure
- Weather information such as rainfall rate and solar radiation data
- Electronic Stability Control

In the case of Track & Know Fleet Management Pilot, the vehicle data provided to Toolbox developers are described in D6.1 “Pilot Specifications”.

9.5.2 Infrastructure Data

The infrastructure subsystem is a critical component of the connected vehicle environment. The subsystem is managed by the infrastructure operator and could provide data collected by roadside equipment (RSE) and traditional ITS equipment distributed along the roadways. Example equipment include traffic detectors, environmental sensors, traffic signals, highway advisory radios, dynamic

message signs, Closed Circuit Television (CCTV) cameras and video image processing systems, grade crossing warning systems, and ramp metering systems.

The infrastructure data may be available to Fleet Operators via open ITS data feeds, aiming to provide and exchange data related to road characteristics and conditions, intersection status, field equipment status etc. Specific data elements include, but are not limited to:

- Roadway characteristics
- Friction coefficient
- Road geometry and markings
- Road conditions
- Surface temperature
- Subsurface temperature
- Moisture
- Icing
- Treatment status
- Road surface weather conditions
- Air temperature
- Wind speed
- Precipitation
- Visibility
- Intersection status
- Current operational status Signal phase and timing
- Intersection geometry
- Approaching vehicle information (position, velocity, acceleration, and turning status)
- Field equipment status
- Dynamic message signs
- Variable speed limit signs
- Dynamic lane signs or control devices
- Ramp meters
- Parking information
- Location of parking facilities
- Spaces available

In the case of the Fleet management Pilot, Track & Know will not consider ITS feeds. However, in the case of data correlation big data services, open ITS data may be treated in a uniform way via the Track & Know Big Data Platform.

9.6 Transport Open Data Sources

Location intelligence is a methodology for transforming location data into business outcomes. Location data can be anything from addresses, latitude and longitude coordinates, or existing points, lines, or polygons. More and more business operators and government bodies are turning to open data and Location Intelligence to optimize current services while also preparing more sustainable solutions in light of anticipated burst of user generated content, information aggregation technologies and big data platforms.

The status of open transport data sources in Europe should be considered at both national and European level. At national level European countries such as UK[6], Sweden, Finland[7] and France have

[6] https://www.transportapi.com/
[7] http://wiki.itsfactory.fi/index.php/Tampere_Public_Transport_SIRI_Interface_%28Realtime%29

released significant amounts of open transport data. On the other hand, countries such as Germany, Belgium, Austria, Switzerland, Poland and Portugal are not members of the Open Government Partnership[8] (OGP) and appear to lack national level commitment to open data. Still, cities such as Berlin[9] and Vienna[10] have opened transport data locally. At EU level the Intelligent Transport Systems (ITS) Directive[11] has addressed the release of open transport data across the Union, but with limited outcomes so far.

In this ecosystem, Track & Know acknowledges that correlation of open data sources with operational databases have the potential to deliver innovative new services. We single-out below examples of interesting open data sources.

Open Street Map - OpenStreetMap[12] is built by a community of mappers that contribute and maintain data about roads, trails, cafés, railway stations, and much more, all over the world. OpenStreetMap emphasizes local knowledge. Contributors use aerial imagery, GPS devices, and low-tech field maps to verify that OSM is accurate and up to date. OpenStreetMap's community is diverse, passionate, and growing every day. Contributors include enthusiast mappers, GIS professionals, engineers running the OSM servers, humanitarians mapping disaster-affected areas, and many more. OpenStreetMap is open data: it is free to use for any purpose as long as OpenStreetMap and its contributors are credited. In the case of alterations or built-up upon the data in certain ways, the result may only be distributed under the same license.

EC Urban Data Platform - The European Commission's Urban Data Platform[13] is an open source tool facilitating data sharing and comparative research across the European Union. The platform's interactive data visualizations provide users with ample information on a range of topics that can be explored at both local and regional levels. There is a section specifically on Transport and accessibility.

French National Address Base - French National Address Base[14], is an open source database initiated by the General Secretariat for Modernization of Public Action (SGMAP), as a collaborative project. The database uses available open data from civil services and state agencies across France, including IGN, La Poste, Etalab, and OpenStreetMap France. It also invites citizens to contribute more accurate location data in an effort to improve emergency response times, facilitate more efficient public/private partnerships, and allow for more spatial analysis of under-utilized areas.

Foursquare Developers - The Foursquare API[15] provides location based experiences with diverse information about venues, users, photos, and check-ins. The API supports real time access to places, Snap-to-Place that assigns users to specific locations, and Geo-tag. Additionally, Foursquare allows developers to build audience segments for analysis and measurement. JSON is the preferred response format.

Global Open Data Index - The Global Open Data Index (GODI)[16] is the annual global benchmark for publication of open government data, run by the Open Knowledge Network. Our crowdsourced survey measures the openness of government data according to the Open Definition. By having a tool that is run by civil society, GODI creates valuable insights for government's data publishers to understand where they have data gaps. It also shows how to make data more useable and eventually more impactful. GODI therefore provides important feedback that governments are usually lacking.

[8] https://www.opengovpartnership.org/theme/open-data
[9] https://www.programmableweb.com/api/berlin-open-data
[10] https://www.europeandataportal.eu/data/en/dataset/stadt-wien_ffentlichesverkehrsnetzliniennetzwien
[11] https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32010L0040
[12] https://www.openstreetmap.org
[13] http://urban.jrc.ec.europa.eu
[14] https://adresse.data.gouv.fr/api
[15] https://developer.foursquare.com/
[16] https://index.okfn.org

9.7 Fleet Management Market estimations

Fleet management is an ambiguous term used in reference to a wide range of solutions for different vehicle-related applications. As we saw in the previous sections, a fleet management solution is basically a vehicle-based system that incorporates data logging, satellite positioning and data communication to a backoffice application.

Commercial vehicle fleets play an essential role in the European economy. According to European Automobile Manufacturers Association (ACEA), there were over 54 million commercial vehicles in use in Europe in 2015. The 13.04 million medium and heavy trucks accounted for more than 75 percent of all inland transports, forming a € 250 billion industry. Approximately 0.9 million buses and coaches stood for 9.3 percent of all passenger kilometres. The greater part of the 30.0 million light commercial vehicles (LCV) in Europe was used by mobile workers and for activities such as distribution of goods and parcels. It is also interesting to note that there are an estimated 10.6 million passenger cars owned by companies and governments. The tables below provide further analysis on the these statistics and the specific numbers calculated per considered country (source is The 2017 edition of ACEA's 'Vehicles in use' Report[17]).

	2011	2012	2013	2014	2015	%change 15/14
Austria	336,322	346,397	355,214	365,686	375,163	2.6
Belgium	613,343	627,692	640,253	656,691	678,801	3.4
Croatia	110,938	114,930	119,411	121,935	127,395	4.5
Czech Republic	495,639	502,164	504,388	509,378	515,263	1.2
Denmark	427,484	414,725	401,874	398,074	395,645	-0.6
Estonia	49,698	54,139	57,414	61,233	66,297	8.3
Finland	361,499	299,088	301,012	304,255	307,706	1.1
France	5,867,000	5,896,000	5,915,000	5,965,000	5,995,177	0.5
Germany	2,085,258	2,141,457	2,196,265	2,274,261	2,374,822	4.4
Greece	818,818	822,492	825,956	830,935	836,685	0.7
Hungary	354,203	355,042	361,706	373,162	389,980	4.5
Ireland	291,241	281,122	287,587	286,294	299,609	4.7
Italy	3,861,167	3,853,329	3,831,774	3,844,429	3,874,452	0.8
Latvia	39,762	43,119	46,477	50,531	52,612	4.1
Lithuania	97,214	98,459	95,854	44,856	46,342	3.3
Luxembourg	24,800	26,089	27,046	27,635	28,521	3.2
Netherlands	922,000	906,000	890,000	885,000	901,026	1.8
Poland	2,237,729	2,303,433	2,334,415	2,399,323	2,447,764	2.0
Portugal	1,206,000	1,170,000	1,137,000	1,118,000	1,110,000	-0.7
Romania	516,071	555,141	591,978	637,750	670,119	5.1
Slovakia	208,877	215,404	222,464	227,395	235,519	3.6
Slovenia	57,455	61,065	64,751	68,132	71,971	5.6
Spain	4,696,898	4,636,062	4,550,076	4,508,276	4,520,616	0.3
Sweden	467,533	477,094	486,052	501,661	516,168	2.9
United Kingdom	3,614,664	3,631,595	3,706,351	3,842,017	4,007,331	4.3
EUROPEAN UNION	**29,761,613**	**29,832,038**	**29,950,318**	**30,301,909**	**30,844,984**	**1.8**
Norway	468,075	472,799	478,225	480,846	487,674	1.4
Switzerland	300,254	313,541	323,235	333,803	344,853	3.3
EFTA	**768,329**	**786,340**	**801,460**	**814,649**	**832,527**	**2.2**
Russia	3,622,592	3,691,099	3,779,540	3,788,247	3,916,555	3.4
Turkey	3,000,539	3,190,725	3,354,898	3,489,743	3,704,512	6.2
Ukraine	1,392,416	1,401,770	1,409,805	1,423,115	1,428,513	0.4
EUROPE	***38,545,489***	***38,901,972***	***39,296,021***	***39,817,663***	***40,727,091***	***2.3***

Figure 31 Light commercial vehicles operating in Europe.

[17] http://www.acea.be/statistics/article/vehicles-in-use-europe-2017

	2011	2012	2013	2014	2015	%change 15/14
Austria	71,076	70,138	69,538	69,229	68,860	-0.5
Belgium	149,444	147,545	145,694	144,370	143,697	-0.5
Croatia	34,648	37,564	39,925	44,506	45,757	2.8
Czech Republic	187,161	183,704	189,939	192,165	196,816	2.4
Denmark	43,577	42,461	41,654	41,424	41,457	0.1
Estonia	33,562	33,906	34,766	35,389	35,455	0.2
Finland	96,864	96,714	96,733	95,176	95,233	0.1
France	564,000	555,000	547,000	554,000	567,000	2.3
Germany	894,462	889,520	890,410	892,695	902,718	1.1
Greece	231,959	232,065	232,334	232,692	233,159	0.2
Hungary	88,334	86,723	86,780	87,488	86,831	-0.8
Ireland	29,725	28,097	30,262	31,084	30,932	-0.5
Italy	992,173	968,846	936,675	922,824	918,258	-0.5
Latvia	33,748	36,017	38,285	37,414	32,908	-12.0
Lithuania	81,879	83,431	84,866	48,222	50,089	3.9
Luxembourg	11,498	11,462	11,456	11,331	11,384	0.5
Netherlands	159,000	155,000	153,000	149,383	149,588	0.1
Poland	841,112	874,572	908,069	941,293	980,201	4.1
Portugal	129,500	125,000	121,400	119,000	119,000	0.0
Romania	166,964	179,409	194,974	197,382	218,728	10.8
Slovakia	91,914	92,513	93,413	93,109	94,611	1.6
Slovenia	27,806	28,946	30,165	31,068	32,445	4.4
Spain	559,853	535,624	520,098	517,268	526,559	1.8
Sweden	80,739	79,727	79,130	79,544	80,046	0.6
United Kingdom	563,872	557,128	568,993	569,921	581,645	2.1
EUROPEAN UNION	**6,164,871**	**6,131,113**	**6,145,559**	**6,137,977**	**6,243,377**	**1.7**
Norway	101,736	101,335	100,898	100,602	100,095	-0.5
Switzerland	60,241	60,335	59,950	60,602	60,076	-0.9
EFTA	**161,977**	**161,670**	**160,848**	**161,204**	**160,171**	**-0.6**
Russia	4,308,314	4,269,514	4,184,944	4,283,455	4,107,344	-4.1
Turkey	728,458	751,650	755,950	773,728	804,319	4.0
Ukraine	1,885,932	1,844,549	1,816,272	1,735,572	1,733,506	-0.1
EUROPE	***13,249,553***	***13,158,496***	***13,063,573***	***13,091,936***	***13,048,717***	***-0.3***

Figure 32 Medium and heavy commercial vehicles.

	2011	2012	2013	2014	2015	%change 15/14
Austria	417,000	426,081	434,331	444,500	453,702	2.1
Belgium	778,745	791,033	801,722	817,089	838,424	2.6
Croatia	145,586	152,494	159,336	166,441	173,152	4.0
Czech Republic	702,499	705,147	714,043	721,432	732,045	1.5
Denmark	480,154	466,000	452,174	448,267	445,934	-0.5
Estonia	87,416	92,356	96,675	101,240	106,539	5.2
Finland	470,145	407,814	409,928	411,877	415,394	0.9
France	6,517,000	6,538,000	6,550,000	6,608,000	6,652,177	0.7
Germany	3,055,708	3,107,000	3,163,469	3,244,457	3,355,885	3.4
Greece	1,075,445	1,079,290	1,083,064	1,088,498	1,094,851	0.6
Hungary	459,309	458,474	465,466	478,034	494,065	3.4
Ireland	337,926	326,280	335,732	333,202	348,627	4.6
Italy	4,953,778	4,921,712	4,867,000	4,865,167	4,890,701	0.5
Latvia	73,510	79,136	84,762	87,945	85,520	-2.8
Lithuania	193,595	196,175	194,420	100,626	103,578	2.9
Luxembourg	37,934	39,254	40,230	40,725	41,683	2.4
Netherlands	1,092,000	1,072,000	1,053,000	1,044,485	1,059,999	1.5
Poland	3,177,221	3,277,863	3,345,086	3,446,673	3,537,809	2.6
Portugal	1,351,000	1,310,100	1,273,200	1,251,500	1,243,700	-0.6
Romania	701,726	753,539	806,343	855,187	909,970	6.4
Slovakia	300,791	307,917	315,877	320,504	330,130	3.0
Slovenia	85,261	90,011	94,916	99,200	104,416	5.3
Spain	5,319,109	5,232,813	5,130,066	5,085,343	5,107,427	0.4
Sweden	562,219	571,024	579,168	595,197	610,328	2.5
United Kingdom	4,269,641	4,279,078	4,364,718	4,500,576	4,677,162	3.9
EUROPEAN UNION	**36,644,718**	**36,680,592**	**36,814,726**	**37,156,165**	**37,813,218**	**1.8**
Norway	569,811	574,134	579,123	581,448	587,769	1.1
Switzerland	377,275	390,668	399,322	410,118	420,613	2.6
EFTA	**947,086**	**964,802**	**978,445**	**991,566**	**1,008,382**	**1.7**
Russia	7,930,906	7,960,613	7,964,484	8,071,702	8,023,899	-0.6
Turkey	3,948,903	4,178,324	4,330,733	4,474,671	4,725,887	5.6
Ukraine	3,278,348	3,246,319	3,226,077	3,158,687	3,162,019	0.1
EUROPE	***52,749,962***	***53,030,650***	***53,314,465***	***53,852,791***	***54,733,405***	***1.6***

Figure 33 Total commercial vehicles (incl. buses).

Most major analysts agree that the European fleet management market has entered a growth period that will last for several years to come. The number of fleet management systems in active use is forecasted to grow at an estimated annual growth rate of 16.4 percent from 6.6 million units at the end of 2016 to 14.1 million by 2021. Moreover, the penetration rate in the total population of non-privately owned commercial vehicles and cars is estimated to increase from 15.6 percent in 2016 to 31.6 percent in 2021 (source is "Fleet Management in Europe", Berg Insight, 2017).

9.8 Big Data in the Fleet Management sector

Big data is rapidly developing to address many facets of Fleet Management systems - driver performance, tracking data, working times and maintenance history. Therefore, new features are continuously added to fleet management solutions adding more value to end users. Big data is the real gold in the fleet management sector. TomTom Telematics, for instance, processes 25 billion new data points every quarter for its fleet management clients[18]. It can be of immense value to use the large amounts of data generated by a fleet and analyze this data to help the fleet owner to better understand its drivers and assets, which allows it to make improvements that maximize performance. Fleet owners can learn new things about their daily operations by analysing everything from delivery routes to waiting times. Big Data can detect driving behaviours and let employees know how they can change their driving style to save money and potentially even save lives. Big data can also be used to detect problems before they happen, allowing vehicles to be serviced when they actually need maintenance in order to minimize cost.

This specific market opportunity is a main focus of Track & Know Project, which through the development of its Toolboxes and a Big Data Platform that specifically consider the transport big data market requirements, is aiming to change the way Big Data will be used for value generation.

[18] http://iotnowtransport.com/2018/06/05/67533-new-features-business-models-propel-fleet-management-market/

10 The Track-and-Know Online Observatory

The Track-and-Know Online Observatory objective is to provide easily accessible, comprehensive and understandable information on the current state-of-the-art methods related to the project, as well as open access (free of charge online access for any user) to all resulting scientific publications of the project.

As far as the easy access to the provided information is concerned, an online webpage will be created. Thus, the online observatory will be implemented as an integral part of the project webpage. The latter, not only will help in the identification and monitoring of the current state-of-the-art throughout the project methods, but also it will bridge knowledge fragmentation across these methods. Moreover, the online observatory content will be dynamic. That is, the online observatory will be continuously updated as new content is becoming available.

The online version will follow the structure of the document, each chapter and subchapter, although it will have a slightly different layout owing to the fact that it will be developed to make it mobile-friendly, just as the rest of the project website will be. The bone structure of the online observatory will be as presented in Figure 34.

The first version of the webpage of the online observatory will be temporarily published online in M6 with the corresponding report (D1.1). Also, a revised version of the online observatory with updated content will be available online at the end of the project (RD1.1)

Track & Know online observatory

- **Track & Know overview**
  - **Scope and Approach**
  - **The T&K 'big picture'**
  - **Challenges and Objectives**
- **Big Data state-of-the-art**
  - **Big Data Platforms and Infrastructure**
    - **Industry-related Benchmarks**
  - **Big Data Management and Processing**
    - **Big Data Storage and Indexing**
    - **Big Data Processing**
  - **Big Data Analytics and Visualization**
    - **Knowledge Discovery in Big Data**
    - **Complex Network Analysis in Big Data**
    - **Complex Event Recognition in Big Data**
    - **Visual Analytics in Big Data**
- **Ethics, Regulations and Standards Aspects**
  - **GDPR compliance**
  - **Related Standards**
- **Market Analysis**
  - **Market Analysis – the Insurance Business Case**
    - **Related Open Data**
  - **Market Analysis – the Healthcare Business Case**
    - **Related Open Data**
  - **Market Analysis – the Transport Business Case**
    - **Related Open Data**
- **Scientific News & Events**
- **Contact**

Figure 34: Structure of the online observatory.

11 Conclusions

In this deliverable, we presented a thorough state-of-the-art survey of the scientific topics as well as the business case of interest for Track & Know H2020 project. In particular, we first presented the overall Track & Know concept, according to the BDVA reference model. Then, we discussed in detail state-of-the-art on big data platforms and architectures, big data management and processing techniques, big data analytics methods, and big data visualization and visual analytics methods and tools, which are highly relevant to the project. After a discussion, on various aspects on ethics (including personal data protection), compliance to regulations and standards, etc., the report included a market analysis on the three industries that will be addressed in Trach & Know project: car insurance, healthcare services, and fleet management, respectively. We concluded the report by outlining the structure and functionality of the online observatory, to be implemented as an integral part of the project website.