Mining and Analysis of Enriched Trajectories - Application to Museum Visitor Trajectories

By

Alexandros Kontarinis

A thesis submitted to "CY Cergy Paris Université"
Department of Computer Science
for the degree of

Doctor of Philosophy

13 December 2021

Committee :

Reviewers:	José Macedo	Associate Professor	Federal University of Ceará
	Christian Sallaberry	Associate Professor	Université de Pau et des pays de l'Adour
Directors:	Dimitris Kotzinos	Full Professor	CY Cergy Paris Université
	Karine Zeitouni	Full Professor	Université de Versailles St-Quentin-en-Yvelines
Supervisors:	Claudia Marinica	Assistant Professor	Université de Nantes
	Dan Vodislav	Full Professor	CY Cergy Paris Université
Examiners:	Cyril Ray	Assistant Professor	École Navale de Brest
	Mounira Harzallah	Associate Professor	Université de Nantes
President:	José Macedo	Associate Professor	Federal University of Ceará

Typeset in LaTeX 2_{ε}.

To the pursuit of truth.

Abstract/Résumé

Abstract

Over the last decades significant progress has been achieved with respect to the mining and analysis of trajectory data. This Thesis is concerned with the problem of, given as input a set of indoor trajectories and additional contextual data describing those trajectories, how to structure those data and analyze them properly, in order to derive valuable insight about the movement phenomena under examination. The research fields most related to it are *Semantic Trajectory Data Modeling* and *Semantic Trajectory Data Mining*. Moreover, we address particular issues stemming from the application domain of museums, or more generally from human mobility in indoor environments. Thus, our modeling and analysis proposals are inspired by museum visit trajectories, but not limited to them.

The main contributions of this Thesis can be summarized as follows:

(i) An overview and a classification of trajectory data mining tasks according to the state-of-the-practice

(ii) A survey of the *semantic trajectory modeling* literature, the *trajectory pattern mining* literature, and the (non-trajectory) *sequential pattern mining* literature for multidimensional or temporally annotated data.

(iii) One of the first-ever studies on how museums and their visitors can simultaneously benefit from the implementation of *museum visitor trajectory analytics*.

(iv) A *conceptual data model* called Semantic Indoor Trajectory Model (SITM) that aims at representing semantic trajectories of moving objects in indoor environments, allowing a rich representation of movement and supporting advanced types of movement analysis.

(v) An implementation of SITM for representing museum visits as semantic indoor trajectories, and an experimental case study analyzing real visitor trajectories from the Louvre Museum in Paris.

(vi) A trajectory pattern mining approach, extending state-of-the-art algorithms and combining semantics, time, and topology.

Keywords: semantic trajectories, indoor trajectories, trajectory data modeling, trajectory pattern mining, sequential pattern mining, human mobility analytics, museum visitor studies

Résumé

Au cours des dernières décennies, des progrès importants ont été réalisés en ce qui concerne l'exploration et l'analyse de données de trajectoires. Cette Thèse s'intéresse au problème de la fouille de données à partir d'un ensemble de trajectoires en milieu intérieur ainsi que des données contextuelles liées, afin de révéler des structures cachées et des connaissances sur le phénomènes de mobilité étudié. Les domaines de recherche relatifs à ce problème sont la modélisation des données de trajectoires sémantiques et l'exploration de données de ces trajectoires. Par ailleurs, nous intégrons certains travaux issus du domaine d'application cible, à savoir la mobilité dans les musées, ou plus généralement de la mobilité humaine dans des environnements à l'intérieur.

Ainsi, nos propositions de modélisation et d'analyse s'inspirent des trajectoires des visites des musées, mais ne s'y limitent pas. En particulier, nous nous concentrons sur des méthodes d'extraction de motifs séquentiels afin d'extraire des motifs intéressants de trajectoires sémantiques en intérieur.

Les principaux apports de cette thèse peuvent être résumés comme suit:

(i) Une classification des tâches d'exploration de données de trajectoires.

(ii) Une étude de la littérature sur la modélisation de trajectoires sémantiques, sur la fouille de motifs de trajectoires, ainsi que sur la fouille de motifs séquentiels multidimensionnels et/ou temporellement annotées (pouvant servir dans la fouille de trajectoires sémantiques).

(iii) L'une des toutes premières études sur la façon dont les musées et leurs visiteurs peuvent simultanément bénéficier de la mise en œuvre de l'analyse des trajectoires des visiteurs des musées.

(iv) Un modèle de données conceptuel appelé Semantic Indoor Trajectory Model (SITM) permettant de représenter des trajectoires sémantiques d'objets mobiles dans des environnements intérieurs, qui capture la richesse des trajectoires et sert des types d'analyse avancés de mobilité.

(v) Une implémentation de SITM pour représenter les visites de musées comme des trajectoires sémantiques en interieur et une étude de cas sur l'analyse de trajectoires réelles des visiteurs du musée du Louvre à Paris.

(vi) Une approche d'exploration de motifs de trajectoires, étendant des algorithmes de l'état de l'art et combinant sémantique, temps et topologie.

Mots-clés: modélisation de trajectoires, fouilles de trajectoires, enrichissement d'information, analyse de trajectoires, mobilité humaine

Acknowledgments

The following dissertation benefited from the insights and direction of several people.

First and foremost, I would like to express my gratitude towards my PhD supervisors, Prof. Kotzinos, Prof. Zeitouni, Prof. Marinica, and Prof. Vodislav for offering me the opportunity to conduct my doctoral studies on a fascinating subject, and for their continuous guidance, patience, and exemplary work ethic which is truly inspiring. Also, I would like to thank Mme. Anne Krebs for doing everything within her powers to support my research efforts and remove any potential obstacles, but also all the Louvre professionals that I have had the pleasure to work with on occasion.

Equally importantly, I would like to thank the members of the committee Prof. Macedo, Prof. Sallaberry, Prof. Harzallah, and Prof. Ray for accepting to dedicate their precious time to read and review my research work. Their feedback was highly enlightening and motivating.

Moreover, a special note of thanks goes to my friend Dr. Charles Robinson for offering some of his precious time to proofread the manuscript. Of course, I would like to also thank all of the colleagues that I have had the pleasure to meet and work alongside, at the ETIS lab of CY Paris Cergy University and at the DAVID lab of the University of Versailles-St-Quentin-en-Yvelines, simply too numerous to mention by name.

This Thesis was supported by the Paris Seine Graduate School Humanities, Creation, Heritage, Investissement d'Avenir ANR-17-EURE-0021 – Foundation for Cultural Heritage Science. Thus, I would like to thank the Fondation des Sciences du Patrimoine and particularly Prof. Anne-Julie Etter and Mr. Emmanuel Poirault who are doing an exemplary work supporting the PhD students throughout their Theses.

Lastly, I would like to wholeheartedly thank my family.

In addition to the above personal notes of thanks, I would like to express my gratitude to France, the nation of great thinkers.

Contents

Abstract/Résumé **v**

Acknowledgments **vii**

1 Introduction **1**
- 1.1 Research Context . . . 2
  - 1.1.1 Trajectory Data Modeling and Representation . . . 2
  - 1.1.2 Trajectory Data Mining and Analysis . . . 4
  - 1.1.3 Museum Visitor Trajectory Modeling and Analysis . . . 6
- 1.2 Problem Statement and Contributions . . . 7
- 1.3 Thesis Organization . . . 8

2 State of the Art in Trajectory Data Modeling **11**
- 2.1 Introduction . . . 12
- 2.2 Trajectory Data Modeling and Representation . . . 12
  - 2.2.1 Trajectory Definition and Fundamental Trajectory Types . . . 12
    - 2.2.1.1 Outdoor vs Indoor . . . 14
    - 2.2.1.2 Geometric vs Symbolic . . . 15
    - 2.2.1.3 Sequential vs Temporally Annotated . . . 16
    - 2.2.1.4 Semantic Trajectories . . . 16
  - 2.2.2 State-of-the-art Semantic Trajectory Models . . . 18
    - 2.2.2.1 The First Conceptual Semantic Trajectory Model . . . 18
    - 2.2.2.2 Early Models Extending "Stop-Move" Semantics . . . 19
    - 2.2.2.3 A Conceptual Framework For Movement Analysis . . . 21
    - 2.2.2.4 Annotation-based Model Used in SeMiTri Framework 24
    - 2.2.2.5 Annotation-based Model Refined . . . 27
    - 2.2.2.6 CONSTAnT: Conceptual Model of Semantic Trajectories . . . 30
    - 2.2.2.7 Symbolic Trajectory Model used in SECONDO DBMS 33
    - 2.2.2.8 MASTER: Multiple Aspect Trajectory Representation 34
    - 2.2.2.9 STriDE: Semantic Trajectories in Dynamic Environments . . . 37
  - 2.2.3 Applied Trajectory Models: what is used in practice? . . . 39
    - 2.2.3.1 Semantic trajectory models . . . 39
    - 2.2.3.2 Indoor trajectory models . . . 50
- 2.3 Indoor Space Modeling and Representation . . . 54

2.3.1 Fundamental Space Representations 54
2.3.2 Indoor Space Models and the IndoorGML Standard 56
2.3.3 IndoorGML Reception 58
2.4 Chapter Conclusions 62

3 SITM: A New Model for Semantic Indoor Trajectories 63
3.1 Introduction 64
3.2 The Semantic Indoor Trajectory Modeling Problem 64
3.2.1 A Close Look at Current Modeling Limitations 64
3.2.2 A Broad Perspective of Current Modeling Limitations 72
3.2.3 Problem Statement 73
3.3 Semantic Indoor Trajectory Model (SITM) 73
3.3.1 The Indoor Space Model 74
3.3.1.1 Abiding by the IndoorGML standard 74
3.3.1.2 Indoor Space Modeling: Definitions, Aspects, Details 75
3.3.2 The Semantic Indoor Trajectory Model 81
3.4 Chapter Conclusion 87

4 State of the Art in Trajectory Data Mining and Analysis 89
4.1 Introduction 90
4.2 Trajectory Data Mining Landscape 90
4.2.1 Trajectory Data Management & Preprocessing 93
4.3 Sequential Pattern Mining and Trajectory Pattern Mining 94
4.3.1 Trajectory Pattern Mining: History and State-of-the-Art . . . 94
4.3.2 Sequential Pattern Mining 103
4.3.3 The Standard Problem and Algorithms 104
4.3.4 Sequential Pattern Mining and Multiple Dimensions 105
4.3.5 Sequential Pattern Mining and the Temporal Dimension 108
4.4 Chapter Conclusions 115

5 A Novel Sequential Pattern Mining Approach for Trajectory Pattern Mining 117
5.1 Introduction 118
5.1.1 The MultiDimensional Temporally Annotated Sequential Pattern Mining (MD-TAS-PM) Problem 118
5.1.1.1 Temporal Information as Interval Duration Annotations 120
5.1.1.2 Semantic Information as Multiple Dimensions 122
5.1.1.3 Topological Information as Accessibility Constraints . 123
5.1.2 Combining Time, Semantics, and Topology: The Case for a Novel Semantic Indoor Trajectory Pattern Mining Algorithm . 125
5.1.2.1 Background Definitions 126
5.1.2.2 Compatibility with SITM and Trajectory Data Preprocessing 128
5.1.2.3 SITPE: A MD-TAS Pattern Mining Algorithm 131
5.2 Conclusion 138

6 Trajectory Model Validation and the Louvre Case Study 141
6.1 Introduction 142
6.2 The Museum Application Domain 142
6.3 The Louvre Case Study 145
6.3.1 The Louvre Setting 146
6.3.2 Trajectory Dataset and Model 148
6.3.2.1 Trajectory Dataset Description & Data Quality Issues 148
6.3.2.2 Smartphone Application Usage Dataset Description . 151
6.3.2.3 Trajectory Data Model Instantiation 156
6.4 Louvre Visitor Experimental Results 165
6.4.1 Trajectory Data Preprocessing 165
6.4.2 Preliminary Statistical Analysis 167
6.4.3 Visualization & Standard Sequential Pattern Mining 168
6.4.4 SITM-based State-Of-The-Art Sequential Pattern Mining . . . 171
6.5 Chapter Conclusion 178

7 Conclusion and Perspectives 181

Glossary 191

References 193
References 193

List of Figures

2.1 A trajectory represents the abstraction of a continuous part of some real-world movement. 12
2.2 A trajectory differs conceptually from a mere collection of detection records. 13
2.3 A geometric outdoor trajectory (left) and its symbolic counterpart (right), where coordinate points have been grouped at the city level. . 15
2.4 A semantic trajectory can be either geometric (left) or symbolic (right). Here, it is enriched with information about the transportation means. 17
2.5 A "Venn-like" diagram reflecting the components of the spatiotemporal context (gray) of movement data (white), according to the modeling framework of [14, 15]. 22
2.6 The class diagram of the CONSTAnT conceptual model taken from [25]. 31
2.7 The conceptual data model of the multiple aspect modeling approach of [133]. 37
2.8 Structured (left) or ad-hoc (right) representation of a hierarchical space. 38
2.9 The trajectory generator of [142] considers mainly "stop-move" semantics targeting outdoor environments. 41
2.10 An illustration of the sub-stop concept proposed in [17, 18]. 44
2.11 Example of a trajectory defined with respect to space sectors according to [66]. 47
2.12 The concepts and relations of the datAcron ontology for semantic trajectories as proposed by Vouros et al. [177] 48
2.13 Extension of the conceptual model of semantic trajectories of [163] as proposed by Mousavi et al. [135] 48
2.14 A basic taxonomy of indoor space models according to [182] 56
2.15 A detailed taxonomy of indoor space models according to [182] 56

3.1 A 2D multiple floor hierarchical indoor space representation 75
3.2 Common types of binary topological relationships between surfaces. . 78
3.3 Structured (left) or ad-hoc (right) representation of a hierarchical space. 80

4.1 The five most typical visitor behavior patterns found in [90]: 1) "directly go to robot area", 2) "go around and stay at robot area", 3) go around backwards"' 4)"visit every place", 5) "stay for long time". . . . 95

4.2 An occurrence of a trivial T-pattern $(x_0, y_0) \to (x_1, y_1)$ in an input trajectory instance, given a neighborhood function N and a temporal threshold τ, taken from [71]. 97
4.3 Example move concept hierarchy from [24] which assumes that the database instances correspond to the lowest level of granularity g=n. . 99
4.4 Alternative approaches of considering time in T-PM as illustrated in [92]: implicitly in the regions R_i (left) or explicitly in the items $[s_I, d_i]$ (right). 100
4.5 A fine-grained semantic trajectory pattern example according to the SPLITTER algorithm [83]. 102
4.6 The merging of grid cells is used to transition from geometric trajectories to semantic RoIs in the approach of [34]. 103
4.7 An example of semantic trajectories (left) and regional patterns (right) according to [41]. 104
4.8 Example of a hierarchical tree over a “Location” dimension and over a “Product” dimension, taken from [147]. 106
4.9 Example of a tree of generated multidimensional items taken from [147].108
4.10 Example of the database transformation in the proposal of Plantevit et al. [147]. 109
4.11 Example of the output of a standard S-PM process run over the transformed database taken from [147]. 109
4.12 The $MiSTA$ algorithm as presented in [69]. 112
4.13 An example of searching for frequent TASs of length 3 ($a[\xrightarrow{a_1} b[\xrightarrow{a_2} c$) in the corresponding 2D annotation space, taken from [70]. 114

5.1 The TP algorithm proposed by Devitt et al. in [50]. 125
5.2 Data dimension hierarchies corresponding to a Louvre visit dataset. . 138

6.1 Visitors often use The Daru staircase surrounding the Winged Victory of Samothrace to rest (left). The Salon Denon has a triple function: it hosts a museum shop, it serves as a major junction point, and it houses many large-scale neoclassical paintings (right). 145
6.2 The symbolic spatial regions used in past analysis studies of the Louvre's visitor movement data [191–193, 195]. Area coverage has since improved considerably. 147
6.3 Visitor tracking data acquisition infrastructure used in the Louvre. . . 147
6.4 Thematic zones of the Louvre Museum's five floors, hand-annotated with accessibility relations within or across floors. White zones are not covered by the Bluetooth beacons but are included in our graph-based representation of indoor space. 149
6.5 Visual representation of a real Louvre visit trajectory, structured as a sequence of detection intervals inside zones (red). Since “Begin” detection timestamps (blue) do not coincide with their respective previous “End” timestamps (orange), there are detection gaps. 150
6.6 Bar chart illustrating the trajectory dataset's distribution of visit length.151

6.7 Part of a real Louvre visit showcasing a detection overlap example. . . 152
6.8 While in the map tab of the *My Visit to the Louvre* app, the visitor can find out more about individual exhibits, in this case Michelangelo's famous Slave sculptures. 154
6.9 Temporal distribution of the smartphone application's launch events. . 154
6.10 The GPS coordinate positions where the Louvre visitors launched the app. 155
6.11 Based on the chain topology of zones (denoted by alphanumeric IDs), a visitor's presence in the blue zone can be inferred, even when undetected. 157
6.12 Taking into account non-observable areas in the modeling of space can help obtain trajectories that are more faithful to the actual real-world movement. 158
6.13 Indicative Regions of Interest belonging to two different thematic zones. 160
6.14 $T_{vis0058,16:00:00,16:45:00} = (trace_{vis0058,16:00:00,16:45:00}, \emptyset)$ is a subtrajectory example composed of 6 presence intervals in several rooms of the Louvre's Denon wing. Green rooms house Italian Renaissance paintings. Cyan rooms house Ancient Greek sculptures. 163
6.15 A simple 2-level domain-specific instantiation of CIDOC concepts maps the RoIs to exhibits providing structure to the interpretation of indoor space. 165
6.16 UI action logs (right) can in principle enrich trajectory data (left). . . 166
6.17 Choropleth map of the Louvre's zones (-1 floor missing from the data). 168
6.18 The main hall of the Louvre Museum is its most frequented area. . . . 169
6.19 Space-time graphs [14] of the trajectories of two real Louvre visitors: single-visit visitor #4645703 (left) and returning visitor #4549164 (right). 169
6.20 Most frequent zone transitions happen in the southern part of the Louvre. 171
6.21 Multimedia content consumption from the smartphone application can theoretically provide the semantic enrichment of the trajectories of the Louvre museum visitors. 173
6.22 Normal distribution of the Louvre visitors' duration of stay in each zone, under two different interpretations of the detection gaps. 176
6.23 The four frequent Louvre TAS patterns of length 3, for $supp_{min}$=5% and τ=117sec. 177
6.24 The four frequent Louvre TAS patterns of length 3, for $supp_{min}$=5% and τ=117sec. 178

List of Tables

4.1 State of the practice classification of trajectory data mining tasks. . . 92

6.1 Our classification of visitor trajectory data mining and analysis goals in the museum domain [100] 143
6.2 A single log record created by the *map_popup* event of Figure 6.8. . . . 153
6.3 Information about the zones encountered in the trajectory of Figure 6.13. 160
6.4 The preprocessed trajectory dataset in tabular form: each row corresponds to a single zone detection. Detections of zero duration are discarded. 167
6.5 Ten most frequent Louvre zone co-occurences and zone transitions. . . 170
6.6 A single log record created by the *poi_audio* event of Figure 6.21. . . . 173

7.1 Glossary of special terms and abbreviations used in the present manuscript.191

1
Introduction

Contents

1.1 Research Context 2
1.1.1 Trajectory Data Modeling and Representation 2
1.1.2 Trajectory Data Mining and Analysis 4
1.1.3 Museum Visitor Trajectory Modeling and Analysis 6
1.2 Problem Statement and Contributions 7
1.3 Thesis Organization 8

1.1 Research Context

This introductory chapter sets the research context of the Thesis and the application-oriented motivation behind it. The main thematic axes most related to our work are *Trajectory Data Modeling and Representation*, *Trajectory Data Mining and Analysis*, and *Museum Visitor Mobility Studies*. The Thesis was conducted in partnership with the Louvre Museum in Paris, which provided us with a unique real-world case study.

1.1.1 Trajectory Data Modeling and Representation

In the data analysis domain, movements consist of spatiotemporal records out of which individual trajectories can be formed. When the moving objects are people, with the exception of applications requiring very precise tracking of different body parts, trajectories are typically represented as moving points. Then, a trajectory essentially consists of a sequence of timestamped locations. But even though for human mobility data, the moving object is considered to exist in a specific location, oftentimes it is preferable to capture only the area containing that location. In such cases, a trajectory essentially consists of a sequence of timestamped spatial regions.

A considerable amount of research work has dealt with modeling and analyzing people's trajectories in a variety of application domains, ranging from urban transportation [80, 113] and road-network vehicle movement [10], to animal migration [126] and air pollution studies [26, 27].

Yet most trajectory-based human mobility research works focus solely on outdoor trajectories, driven by the fact that Geographic Information Science (GIS) has traditionally only supported outdoor spatial information. For indoor environments there are considerable differences, mainly due to interior architectural components constraining (or otherwise affecting) the way people move. For example, an indoor trajectory model has to consider multiple room entrances, floor changes, specific building entrances/exits, sensor coverage gaps or detection area overlaps (due to obstacles, sensor malpositioning, etc.), varying spatial granularity in the data, and numerous other challenges.

This not only affects the analysis' goals, but oftentimes also the quality of the movement data, and as a result the methods required to achieve those goals. For example, without proper installation planning, detection beacons may produce low-quality tracking data riddled with uncertainty issues (e.g. long periods of non-detection), whereas without proper maintenance planning, they may start logging completely erroneous data due to battery depletion [159].

Another difference is that indoor trajectory analytics may gain from avoiding cumbersome calculations over geometric representations of space and objects within it, that are typical of outdoor settings. The reason is that an indoor space is typically clearly divided into subspaces. Therefore, operations such as intersection, containment, and proximity can be simplified in order to prioritize the non-geometric aspects of movement [81] (e.g. topological properties), instead of the metric aspects which typically focus on Euclidean distances from potential targets. These are only some of the reasons why indoor environments deserve special attention when analyzing human mobility data.

Moreover, in order to reason about movement in information-rich domains, a trajectory model must also account for multiple types of contextual and semantic information. As identified by Peuquet in [143] and further explored by Andrienko et al. in [14, 15], there are three fundamental sets pertinent to movement, representing the *where* (set of locations), *when* (set of instants or intervals), and *what* (set of objects) of spatiotemporal data. This is true across applications, as well as across application domains. Distinguishing among semantics of time, semantics of places, and semantics of moving objects, in addition to the semantics of the movement itself, can empower a synergistic interplay between them.

In practice however, the most crucial pieces of semantic information are derived either from the moving object's environment or from external data sources. Such information can then be used to add a meaningful dimension to the large volumes of purely spatiotemporal location feeds being amassed nowadays. These are often referred to as *raw Big Trajectory Data*. Unfortunately, semantic trajectory models have - to a large extent - targeted outdoor settings where the semantics are quite different than those found in indoor applications. Therefore, we need to come up with new ways of integrating contextual information.

For example, the trajectory literature has typically used ad-hoc speed thresholds, temporal thresholds, or spatiotemporal thresholds, to derive the stops and moves from raw trajectories, in an attempt to enrich them with the correponding semantics. But as explained in [139], *the characterization of a stop can imply no movement at all, slow speed, movement within constrained area, or proximity to some POI*[1]*, to name a few.* More generally, this means that the same approach that works in outdoor settings might not be nearly as relevant in an indoor environment. For instance, the moving object's location information is directly linked to a finite number of predetermined well-separated spatial regions, and thus movement is more naturally discretized without necessarily involving stops and moves. Also, the crowdedness of a room may force a stop or even determine the whole route, whereas something similar can only happen outdoors if the movement takes place over a network. In addition, an object moving indoors generally does not travel such long distances nor does it reach such high speeds, compared to objects moving outdoors, which is why an indoor stop may hold a different significance.

Stops and moves constitute only a small example of the intricacies of trajectory semantics that merit further study. By targeting outdoor settings, the related literature has also emphasized the enrichment of Global Positioning System (GPS) data, the identification of transportation means, and other modeling concepts and issues that are, either not interesting or simply not transferable in an indoor setting. On the other hand, the adoption of some modeling approaches, such as the segmentation of trajectories into episodes, or the use of semantic annotations at different levels of spatial granularity, seems to be equally promising for indoor environments. In this Thesis, some of these more broadly applicable ideas are combined with new modeling concepts to derive a novel semantic trajectory model for indoor spaces.

[1] Point of Interest.

1.1.2 Trajectory Data Mining and Analysis

During the last fifteen years, there has been a gradual increase in semantic trajectory research activity, not only with respect to the design of trajectory models and representation formats, but also with respect to the mining methods used to derive knowledge from them. Location-based applications have been a major driving force for this progress. As will be detailed in chapter 2, trajectory data mining already includes a broad range of problems and methods, like characterization of moving objects (e.g. profiling), discovery of relationships between them (e.g. social relationships), discovery and characterization of interesting places or regions, recognition of social events, prediction and recommendation, among others.

However, trajectory data mining research has not yet fully caught up with the increased inclusion of semantic information in the representation of trajectories. One of the reasons is that - similar to modeling works - trajectory data mining works have mostly focused on outdoor long-distance trajectory applications, like tourist travels or urban mobility. As a result the semantics utilized in the mining tended to reflect those domains. Another reason is that trajectories had in the past almost become synonymous to GPS trajectories, and therefore almost any abstraction from coordinate positional data was considered as "semantic enrichment". But lately, there is increased interest for trajectory-based research over a plethora of more specialized domains, like maritime trajectories [42], team sports trajectories [174], mobile crowdsensing [59], indoor-outdoor trajectory integration [136], museum visit trajectories [37], and various others. Therefore, there lies considerable potential for research breakthroughs in the field of semantic indoor trajectory data mining and analysis, and this is precisely the focus of this Thesis.

In particular, the Thesis explores Trajectory Pattern Mining (T-PM), which in simple terms is the task of finding all trajectory parts that occur frequently enough in a given set of trajectories. Albeit in this case, the mining approach needs to be adapted to indoor environments and semantic information. Our methodology of choice for doing so is to extend and combine general-scope pattern mining algorithms, especially those looking for sequential patterns.

To illustrate the importance of coming up with new T-PM methods, let us consider a set of shopping mall client trajectories and assume that, as managers of this shopping mall, we wish to provide clients with a comfortable and enjoyable experience. At the same time we want to find out which stores are more attractive and why, so that we may increase our revenue. Hence, we decide to run a standard Sequential Pattern Mining (S-PM) algorithm (e.g. GSP [165]), in order to find out typical mobility behaviors in our shopping mall. Indeed, we learn that for instance a very frequent pattern is:

$$Yves\_Rocher \rightarrow H\&M \rightarrow Promod \rightarrow MacDonald's.$$

Sequential patterns like the above one, consist of chains of real-world *symbolic locations*, or as in this case *spatial regions* representing the presence of a client inside of them. Such patterns are indicative of the shopping activity of clients, but do not offer conclusive insight about it. To understand why this is the case, let us assume that the previous pattern is somehow enhanced with temporal information becoming:

$$\overset{15min}{Yves\_Rocher} \rightarrow \overset{1min}{H\&M} \rightarrow \overset{5min}{Promod} \rightarrow \overset{30min}{MacDonald's}$$

where each time annotation indicates the approximate duration of stay in the corresponding shop. Why is this new representation more insightful than the previous one? By taking time into account, we are now in a position to understand our clients' actual shopping behavior much better. The amount of time consistently spent at $Yves\_Rocher$ indicates a clear interest in shopping there, whereas similarly, the low duration value at $H\&M$ indicates clients are only passing by it. As for $Promod$ they do stop to take a quick look at it or maybe they tend to pre-order and they stop to pick-up their products. The pattern also reveals that the typical shopping mall experience involves $MacDonald's$ clients who prefer to spend time there after shopping in the aforementioned shops. Judging from the time spent there, they eat on spot rather than take something to eat on the go. Evidently, adding time into the patterns can expand our understanding of practically any type of movement phenomena.

Now, we take a hypothetical look at the floor plans of our shopping mall, and notice that due to a peculiar architectural design of the building, a client actually has to pass through $H\&M$ to arrive at $Promod$. This is a piece of information that we would like our analysis to be aware of, as it makes us confident about discarding the corresponding item $\overset{1min}{H\&M}$ from the sequence as unimportant. Or perhaps to the contrary, we notice that $H\&M$ is located on a different floor level than $Yves\_Rocher$ and $Promod$ and is not directly accessible from either one, again making us confident about discarding it, this time as a sensor misfire. In either case, by taking the building's topology into account, our analysis now outputs the pattern:

$$\overset{15min}{Yves\_Rocher} \rightarrow \overset{5min}{Promod} \rightarrow \overset{30min}{MacDonald's}$$

which better captures the true essence of this particular client mobility behavior.

Now, let us take one step further and assume that we have developed and launched a smartphone application offering Location-Based Services (LBS) to our clients, in order to facilitate their experience in our shopping mall. This application provides near real-time semantic information about the client's activity, enabling us to enrich the purely spatiotemporal client trajectories with additional non-spatial non-temporal information. In turn, our mining efforts shall result in yet more detailed trajectory patterns. For example, now the most frequent pattern is:

$$\overset{15min}{\{Yves\_Rocher, disc\_buy, reg\_cust\}} \rightarrow \overset{5min}{\{Promod, prod\_replacement, cas\_cust\}} \rightarrow \overset{30min}{\{MacDonald's, reg\_buy, new\_cust\}}$$

where the two extra data dimensions represent *shopping actions* and *shopper type*.

Provided that we can indeed obtain or estimate their values, the extra semantic dimensions enable a much more meaningful interpretation of client movement: we learn that regular customers (*"reg_cust"*) like to visit $Yves\_Rocher$ specifically to take advantage of its discounts (*"disc_buy"*), whereas casual customers (*"cas_cust"*) tend to visit $Promod$ to return or replace products (*"prod_replacement"*), and $MacDonald's$ attracts mostly new customers (*"new_cust"*) instead of returning ones. Of course, other semantic data dimensions and other types of contextual information can also be used, such as demographics, satisfaction level, product wishlists, appointments, fidelity points, public happenings, etc.

The above example goes to show the types of insight that can be extracted from the discovery of qualitatively richer mobility patterns, when combining spatiotemporal data, semantic information, and the intricacies of indoor space. From now on, the focus will shift to a related indoor application domain, namely museum visits. However, as illustrated by the shopping mall example, the proposed solutions will be equally useful to any application domain of indoor mobility and especially human indoor mobility.

1.1.3 Museum Visitor Trajectory Modeling and Analysis

Let us now put ourselves in the position of the manager, not of a shopping mall like before, but this time of a museum. As is well known, the role of a museum mainly lies in collecting, storing, preserving and exhibiting natural and man-made objects [169], but museums also emphasize the visitor experience [62], especially since the expectations of the museum-visitor interaction have changed for both sides [130, 131]. Actually, it has long been of paramount importance for museums to *know* their visitors, meaning to study and understand their motivations, expectations, engagement, and satisfaction.

In this regard, they have traditionally relied on visitor study methods based on human observation, survey questionnaires, and interviews. Throughout the years, these methods have been enhanced with digital information resources and technologies, while still retaining the human element at their core. For example, video recording has been used to help a human observer detect visiting behaviors that he or she might have missed in person, and digital questionnaires and handheld devices have addressed the inefficiencies of hand-written note taking. This has allowed museum visitor studies to become less intrusive for the visitors.

Nevertheless, these studies still lack precision and generally rely upon human-error prone ways to derive the necessary data. Moreover, since they are labor intensive and resource demanding, they have limited scaling potential[2], i.e. they can only support studies of limited duration and visitor sample size. In turn this makes the analysis vulnerable to biases (e.g. seasonal bias, demographics bias), more dependable on the ability of the visitors to accurately report their visiting experience, and in general less likely to reach confident conclusions.

Despite all these limitations, traditional visitor study methods still to this day constitute the primary way for studying and profiling the visitors. But this is starting to change as museums become aware of the implications of the virtual and digital dimensions of visiting exhibitions. For example, the Louvre Museum has identified that only 5% of its visitors are purely "physical" (i.e. with no relationship to the museum's website) as opposed to 27% being purely "virtual" (i.e. with no physical relationship to the museum) [61]. What is more, these figures correspond to the

[2]See [194] for a historic comparison of museum visitor movement data collection techniques including in terms of sample size.

period before the COVID-19 pandemic, which made museums like the Louvre rely even more on social media and digital means of communication to keep the active participation of the audience alive [44].

Among others, one of the most effective ways for museums to understand their visitors is to study their *movement* in the exhibition space. For a long time, this has been attempted in person by following visitors around as their visit progresses [4]. But as of late, the advent of diverse wireless indoor positioning technologies has contributed in LBS (e.g. way-finding, contextualized content delivery) becoming almost standard museum multimedia guide functionality. These LBSs grant museums access to an unprecedented wealth of visitor movement data which, despite privacy restrictions [137], can reveal many aspects of the visitors' behavior and experience if properly analysed.

As a general rule, since museums engage primarily in understanding their visitors, they are more interested in descriptive rather than predictive mobility analytics. The specific analysis goals vary considerably as described in greater detail in chapter 6, but can be grouped into three main categories: improving the visitor experience, aiding the managerial decision making, and effectively managing the visitor crowd, as identified by the author in [100]. For all three areas of improvement, the *indoor context* and the *semantic aspect* of movement remain key modeling elements.

Finally, studying human mobility behavior through semantic indoor trajectory analysis is of great interest not only to museums, but also to sectors such as healthcare, universities, retail, and airports. In all of these application domains, similar opportunities to collect vast amounts of individual trajectory data exist [87]. This is why our modeling and analysis proposals are inspired by, but not exclusive to the museum domain.

1.2 Problem Statement and Contributions

A concise description of the problem that this Thesis attempts to study and help solve follows below.

Given a raw tracking dataset of individuals moving in an indoor environment, and given any type of semantic information related either to those moving objects or to their movement, we wish to derive knowledge that will help us understand their movement in depth. Our goal is to do so, first by structuring the input data in a properly designed semantic trajectory format, and then by applying proper data mining methods over those trajectories.

As will be explained in chapter 2, there is a great plethora of data mining methods to consider for deriving knowledge from trajectory data, however not all are relevant for descriptive semantic indoor trajectory analytics. This is why the focus will be onSequential Pattern Mining (S-PM) methods in particular.

The previous problem statement is straightforward, but encompasses many interesting subproblems including:

- How does the indoor environment affect a trajectory?
- How to best represent the indoor space when modeling trajectories?

- How to model an indoor trajectory that is semantically rich?
- How to mine frequent patterns from semantic indoor trajectories?
- How to treat spatial hierarchies and semantic hierarchies in the mining process?
- What are the practical challenges related to the quality of real-world indoor trajectory data?
- What can be done about them at the modeling level, at the data pre-processing level, or at the analysis level?
- What are the typical visiting behaviors in the world's most frequented museum?

In order to help address the aforementioned problem, and all of the modeling and analysis issues stemming from it, this Thesis contributes the following elements:

1. A survey of semantic trajectory modeling literature and a survey of trajectory pattern mining literature.
2. An overview of trajectory data mining literature accompanied by a proposed classification of the related tasks according to the state-of-the-practice.
3. A study on the benefits of trajectory data analytics research for museums [100].
4. A new model for spatiotemporal indoor trajectories enriched with semantic annotations, called Semantic Indoor Trajectory Model ($SITM$) [102].
5. A validation of $SITM$ by its instantiation in the case of the Louvre Museum in Paris, and the application of standard and state-of-the-art mining algorithms over real-world Louvre visit trajectories expressed in $SITM$ form.
6. A formalization of the problem of mining semantic indoor trajectory patterns from input trajectories, and a novel T-PM approach to solve it, along with a corresponding proposed algorithm called Semantic Indoor Trajectory Pattern Extractor (SITPE).
7. A unique analysis study of the Louvre visitors' mobility patterns, constituting one of the rare cases of real-world museum visitor trajectory data-based computational studies.

1.3 Thesis Organization

The rest of this Thesis is divided as follows:

- **Chapter 2** presents an extended overview of the related work in the research fields of Semantic Trajectory Modeling, Trajectory Data Mining, T-PM, and non-trajectory S-PM.
- **Chapter 3** proposes a Semantic Indoor Trajectory Model.

- **Chapter 5** introduces the Semantic Indoor Trajectory Pattern Mining (SIT-PM) problem and proposes an algorithmic method to solve it.

- **Chapter 6** introduces a compelling real-world case study for validating our trajectory model, involving visitor trajectory data from the Louvre Museum.

The Thesis is complemented by examples illustrating how the proposed trajectory model serves its purpose in practice and how the proposed mining method will enable us to extract richer mobility patterns. Finally, it includes a discussion on trajectory data-based museum visitor studies, and identifies important goals related to this very promising new application domain of trajectory data mining research.

A more detailed summary of the Thesis follows.

Chapter 2 describes the state-of-the-art works for targeting the aforementioned problem statement. It contains a survey of Semantic Trajectory Models and an overview of the field of Computational Trajectory Analytics, both receiving considerable attention currently in the research world. It also encompasses a classification of trajectory data mining tasks, before delving deeper into the work being done in T-PM in particular. Our focus is given to Semantic Trajectories and Indoor Trajectories, both independently and in combination, and a complete survey-level overview is provided. The second part of the chapter has a narrower scope dealing with a specific type of trajectory data, namely museum visits. It looks into how the movement of museum visitors has so far been represented in trajectory form, and proceeds into detailing the importance of semantics and the indoor context in the modeling effort.

Chapter 3 describes the proposed conceptual data model, called Semantic Indoor Trajectory Model ($SITM$), which is illustrated for the museum domain, but is actually applicable across all indoor environments. The chapter starts by making a concrete case for why this model was developed, and then proceeds to define a standards-based representation of the indoor space itself. This indoor space representation can be considered as part of the trajectory model itself, in the sense that the spatial dimension of the trajectory data refers to it. Then, the relevant trajectory concept definitions are provided. Our extensive study of semantic trajectory models in chapter 2 is of paramount importance here, because it enables us to identify concepts necessary for representing semantic trajectories in the indoor context while avoiding others that are suited mostly to outdoor environments. It also provides a bird's eye view of the related literature which helps us avoid confusion around the terminology used to describe similar concepts (like episodes, segments, and subtrajectories, or annotations and labels, or places, PoIs, and RoIs, etc). Finally, $SITM$ fits the space representation with the trajectory concepts and the semantic mechanisms cohesively into a single model.

Chapter 5 is dedicated to the problem of extracting frequent patterns from a set of input trajectories. First, it deals with how to apply methods reviewed in chapter 2 over trajectory data. The goal is to be able to find not only sequential patterns, but patterns containing either temporal or semantic information. Then, it extends and combines such methods in order to mine qualitatively richer patterns than those which the state-of-the-art currently allows. More specifically, the pseudo-code of the algorithm is given, its implementation process is illustrated, and it is discussed how it differs or improves upon previous algorithms. To the author's knowledge, it is the first algorithm to take into account multiple hierarchical spatiosemantic dimensions, temporal annotations, and topological constraints, resulting in a qualitatively richer description of mobility behaviors.

Chapter 6 presents a real-world case study from the Louvre museum, in which a historic dataset of around 5,000 visits is structured into trajectories and then, based on the proposed trajectory model is subjected to gradual analysis, revealing interesting visiting behavioral patterns. Also, analysis work is provided which is not related to T-PM but nevertheless provides an exclusive view at an intriguing case study, not only of the most frequented museum but also of the largest indoor beacon installation at the time globally. This includes issues and conclusions on trajectory data quality, trajectory matching, trajectory data pre-processing, and trajectory semantics. Apart from validating our trajectory model proposed in chapter 3 and deriving insight on the visitors, and despite the practical constraints preventing us from validating the proposed T-PM methodology, it is - to our knowledge - the first attempt at extracting temporally-aware trajectory patterns of museum visits, and more generally the fist attempt at analyzing museum visitor trajectories at such scale. Thus it served to the Louvre as an evaluation of the visitor behavior analysis potential of permanent beacon installations, and provides various contributions in the field of Museum Visitor Studies especially in terms of methodology.

Chapter 7 offers our conclusions and perspectives about all of the issues addressed in this Thesis. This is a very important chapter as the main ways to drive forwards the research field of T-PM is identified, based on the extensive experience with one of the few real-world case studies in the museum domain. Related to this, a personal view of the current state of Trajectory Data Mining more generally is provided, along with concluding remarks for what should be tackled next in this field, especially in terms of modeling and mining.

2

State of the Art in Trajectory Data Modeling

lscape

CONTENTS

2.1 INTRODUCTION 12
2.2 TRAJECTORY DATA MODELING AND REPRESENTATION 12
2.2.1 Trajectory Definition and Fundamental Trajectory Types 12
2.2.1.1 Outdoor vs Indoor 14
2.2.1.2 Geometric vs Symbolic 15
2.2.1.3 Sequential vs Temporally Annotated 16
2.2.1.4 Semantic Trajectories 16
2.2.2 State-of-the-art Semantic Trajectory Models 18
2.2.2.1 The First Conceptual Semantic Trajectory Model 18
2.2.2.2 Early Models Extending "Stop-Move" Semantics 19
2.2.2.3 A Conceptual Framework For Movement Analysis 21
2.2.2.4 Annotation-based Model Used in SeMiTri Framework 24
2.2.2.5 Annotation-based Model Refined 27
2.2.2.6 CONSTAnT: Conceptual Model of Semantic Trajectories . . 30
2.2.2.7 Symbolic Trajectory Model used in SECONDO DBMS . . . 33
2.2.2.8 MASTER: Multiple Aspect Trajectory Representation 34
2.2.2.9 STriDE: Semantic Trajectories in Dynamic Environments . . 37
2.2.3 Applied Trajectory Models: what is used in practice? 39
2.2.3.1 Semantic trajectory models 39
2.2.3.2 Indoor trajectory models 50
2.3 INDOOR SPACE MODELING AND REPRESENTATION 54
2.3.1 Fundamental Space Representations 54
2.3.2 Indoor Space Models and the IndoorGML Standard 56
2.3.3 IndoorGML Reception 58
2.4 CHAPTER CONCLUSIONS 62

2.1 Introduction

In this chapter, the background works in the trajectory modeling research domain are presented, especially those relating to the author's own corresponding contributions. In addition, an exhaustive overview of the semantic trajectory modeling literature is provided and the limitations of state-of-the-art conceptual models are identified, before proposing ways to overcome them in the case of movement in indoor environments in chapter 3. An overview of the trajectory data mining landscape is also provided, and it is identified why state-of-the-art methods and algorithms are lacking for finding patterns in temporally annotated datasets, before proposing ways to overcome them as far as trajectory data analysis is concerned in chapter 5.

2.2 Trajectory Data Modeling and Representation

2.2.1 Trajectory Definition and Fundamental Trajectory Types

When the term *trajectory* is encountered in everyday discussions, it is usually interpreted as a specific instance of movement of somebody or something. In the scientific literature, its meaning is actually not that different, although it varies according to the particular field of interest. In Computer Science, the tendency is to use the term *moving object trajectory*, as opposed to more specific terms such as *gene expression trajectories* (in Biology), *particle trajectories* or *quantum trajectories* (in Physics), *molecular trajectories* (in Chemistry), which all narrow down the type of the moving object. In the same spirit for example, the term *trajectory data management* is also often referred to as *moving objects databases* [28, 75].

Thus, it is best to start by explaining how the related terminology is understood in this Thesis, before moving on to the discussion of modeling a trajectory. This ¿Thesis is concerned with moving object trajectories, which can be defined as follows:

Definition 2.2.1 (moving object trajectory)
A moving object trajectory is a spatiotemporal trail generated over time by a person, or an animal, or an inanimate object.

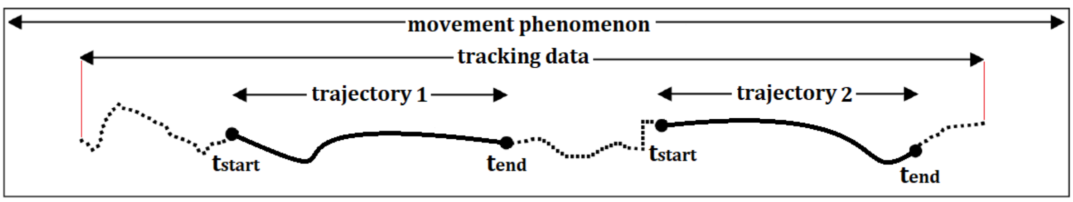


FIGURE 2.1: A trajectory represents the abstraction of a continuous part of some real-world movement.

As shown in Figure 2.1, a moving object trajectory does not necessarily coincide with the entire movement of that object, nor does it have to last for the entire observation period during which tracking data have been generated. Instead, it might as well represent any meaningful and well defined part of that movement, spanning over any window of the entire observation lifetime. In fact, a trajectory does not even presuppose the existence of tracking data, as long as it is possible to somehow guess or estimate the location of the moving object for the duration of a gap, as illustrated in Figure 2.2. This goes to show that despite the large observation gap in the middle of the movement, in principle it is still possible to derive a close approximation to the two actual trajectories in Figure 2.2.

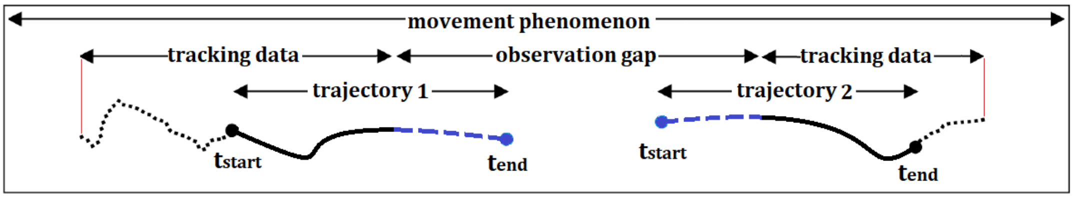


FIGURE 2.2: A trajectory differs conceptually from a mere collection of detection records.

Until the 1990s, temporal and spatial databases were completely separate areas of database research [103], and only gradually did conceptual data models start to combine the spatial and temporal aspects, thanks to works such as [181] and [140]. By similarly restricting - for now - our attention to the spatial and temporal dimensions separately, a few basic trajectory types can be distinguished.

Given Definition 2.2.1, one of the naturally arising fundamental modeling questions at the lowest levels of abstraction is whether the moving object is described as a *moving point* or as a *moving region*. Of course, this depends on the nature of the moving object and that of the movement phenomenon itself. For example, applications modeling geophysical processes such as storms or cyclones often require a certain amount of realism that can not be attained by a moving point representation, and thus spatial variability is introduced into the moving object representation (e.g. random field representation [138]). For human mobility data in particular, the distinction remains pertinent because some human activities may be described in enough detail through the evolution of each person's unique representative location (e.g. navigation), whereas others would require the evolution of different body part locations (e.g. gait analysis).

For the rest of this chapter and Thesis, the focus will be only on the first type of human mobility data representation, namely moving points, especially given that it fits our own application domain of interest which concerns museum visitor trajectories. Finally, oftentimes when human movement data are unavailable (e.g. in migration studies [161]) the first type of trajectories might degenerate into consisting of only two timestamped positions: the origin and the destination. However, this Thesis focuses on the non-trivial case where intermediate positions do exist, since its interest lies in studying human mobility at the individual level.

2.2.1.1 Outdoor vs Indoor

Arguably, outdoor spatial informatics have received much more attention than indoor spatial informatics, and by consequence, the same can be said about trajectory data research. The distinction between outdoor trajectories and indoor trajectories is rather straightforward: depending on the type of environment in which the related movement phenomenon takes place, a trajectory is either of the former type or of the latter type. But this distinction is not a mere formality, as it has less obvious implications on how trajectories can and should be thought of and represented.

Many differences between the two types of environments with respect to the computational analysis of mobility data have already been identified during the last decade. According to [87], the two major ones are, first that indoor spaces are composed of entities unique to them (e.g. rooms, hallways) which constrain movement, and secondly that unlike GPS, indoor positioning technologies (e.g. proximity-based ones) can not report velocities or accurate locations, resulting in uncertain tracking data and an increased need for symbolic models (e.g. graph-based). With respect to positioning technologies, [72] stresses that indoor environments are lacking a universally used one, like GPS or Galileo serving outdoor applications. [180] goes a step further reporting more specifically on the problems related with the various indoor localization technologies: position accuracy affected by limitations in signal transmittance, absence of precise and synchronous clocks between transmitters and receivers, non-line of sight, signal attenuation and interference due to environmental factors, time consuming calibration phase, demanding infrastructure maintenance (hardware and software), multipath propagation effects, energy inefficiency, cost, etc. These factors are hardly as relevant to outdoor applications. [72] detects some other differences when studying indoor - as opposed to outdoor - movement: the existence of multi-level routes, the different function of landmarks, the focus on pedestrian (rather than vehicular) traffic, the more “regular” geometries of rooms, the layered and often complex structure of buildings. With respect to the latter, [6] mentions the hierarchical structure of an indoor environment and the different levels of granularity as a major difference, and also claims that small-scale indoor spaces encompass more specific properties and hold more interactions between moving objects (e.g. humans) in comparison to large-scale open spaces. [156] considers the concept of indoor geography to be heavily context-based, due to the fact that the modeling of an indoor structure is strongly intertwined with the associated application field of the building. This is a remark also made in [6] where the design of a comprehensive indoor spatial data model is said to require the integration of data from diverse sources and the user’s context.

To conclude by adding our personal view on the outdoor-indoor trajectory distinction, the main differences between the two types can be summarized as follows:

- the nature of the location information (discrete spatial entities indoors, precise points outdoors);
- the levels of positional data uncertainty (higher indoors);
- the existence of movement constraints (indoors as opposed to only sometimes

outdoors);

- the floor changes (indoors);
- the ability to more readily capture the semantics of space (indoors);
- the changes in speed (greater outdoors).

2.2.1.2 Geometric vs Symbolic

Another clear distinction is the one between geometric and symbolic trajectories [5], which is based upon how the moving object's location information is modeled. Trajectories of the former type are typically characterized by a coordinate system-based representation of the location information, whereas trajectories of the latter type use symbols to represent it, in other words discrete spatial entities such as particular points or areas. Those entities can either be predefined or they can be found by processing the tracking data themselves, as explained in section 4.2.

Geometric trajectories do not necessarily comprise coordinate values, as long as they are represented on the basis of geometric primitives such points, lines, areas, and volumes. But in practice, these basic shapes are almost always measured within a given coordinate system, and therefore a geometric trajectory is typically composed of a sequence of timestamped coordinate tuples, e.g. timestamped pairs (x, y, t) in the 2-dimensional case, or timestamped triples (x, y, z, t) in the 3-dimensional case. This type of representation enables accurate location and distance information [6] and is prevalent in outdoor applications.

Symbolic trajectories on the other hand are based on a modeling of space which treats entire spatial entities as first-class citizens, instead of relying on precise - but sometimes meaningless - coordinate values. As a result, a symbolic trajectory is typically composed of a sequence of named places. This of course emphasizes the topological properties of space, facilitates the representation of dynamically changing environments, and in general is characterised by attributes that will be further elaborated in section 2.3.2. In addition, even though symbolic trajectories are not necessarily semantic, they do favor a semantic modeling of the environment where movement takes place.

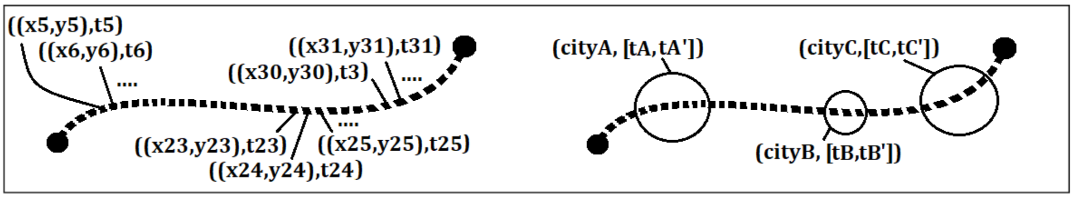


FIGURE 2.3: A geometric outdoor trajectory (left) and its symbolic counterpart (right), where coordinate points have been grouped at the city level.

2.2.1.3 Sequential vs Temporally Annotated

With respect to the temporal aspect of trajectories, there exists an implicit distinction having to do with whether the temporal information is kept in some form, or is discarded in favor of the mere order of location records. We can refer to the first case as a *temporally annotated trajectory* and to the second case as a *sequential trajectory*. Interestingly, whereas trajectory modeling research has always adopted a temporally annotated type of trajectory, which is a sensible approach given that a considerable amount of explanatory power lies within the temporal information of any movement, a considerable amount of application-driven research works actually use simple sequential trajectories, as will be shown in our literature review in the rest of this chapter. This can no longer be attributed to a lack of timestamped data because, as explained in chapter 1, even for indoor environments there is now a broad range of wireless indoor positioning technologies generating large volumes of timestamped detection logs. Instead, it can be partly attributed to some application cases being sufficiently served by ordered positional information alone. But mainly to the lack of a cohesive strategy for addressing the temporal dimension in trajectory mining and analysis, without having to retreat to time series analysis methods that would take the focus away from individual trajectories. Put more simply, it may have sometimes been more convenient to disregard time completely, than find out how to properly account for it in the analysis of trajectories, when order information is "good enough" with respect to the analysis requirements.

Finally, there exists a third intermediate category, or to be more precise, a subcategory of temporally annotated trajectories, which is *interval-based trajectories*. These are trajectories encompassing temporal information, not in absolute time moment form, but rather in the form of time durations. Each interval may correspond either to a specific position of the moving object or to the transition period between two such consecutive locations. This type of trajectories has been used in various works (e.g. [31]) but has very seldom been explicitly studied in trajectory data research [183].

2.2.1.4 Semantic Trajectories

Recently, apart from the usual[1] issues related to their large volumes or fast process rates, Big Trajectory Data have also started being studied with respect to issues of data heterogeneity [152]. This does not simply constitute an attempt at finding methods to combine multiple trajectory datasets, or even to fuse trajectory datasets together with relevant non-trajectory datasets. Instead, it is a more elaborate effort to enrich the way in which trajectories are being modeled, processed, and analysed, with information that allows them to become more meaningful, and to serve as more comprehensive abstractions of real-world mobility phenomena.

More specifically, the inclusion of contextual elements such as properties of space, moving object profiles, personal preferences, events, or even entire ontologies, leads

[1] In the sense that they are nowadays encountered with all types of data.

to the formation of *semantic trajectories*, which precisely embed the additional contextual information to the main geometric or symbolic nature of the data. Whereas these additional data dimensions are indeed contextual, related to the environment of the movement and to the agents or objects that participate in it, the semantic scope of trajectories is only really limited by the particular application domain or use case requirements. Different semantic facets may be useful for different types of scenarios and analyses, and so there is no reliable way to determine a priori which types of semantic information should be taken into account and which not. For example, in section 1.1.2 an indoors example was examined of how the addition of semantic aspects to trajectories can be envisaged, to enable new methods of trajectory analysis, qualitatively superior in comparison to what would be possible solely based on their spatiotemporal aspects.

Let us now more formally define a *semantic trajectory* as follows:

Definition 2.2.2 (semantic trajectory)
A semantic trajectory is a moving object trajectory, whose spatiotemporal trail is enriched with information relating to the context, the environment, and/or the domain of movement, resulting in a more meaningful representation of it.

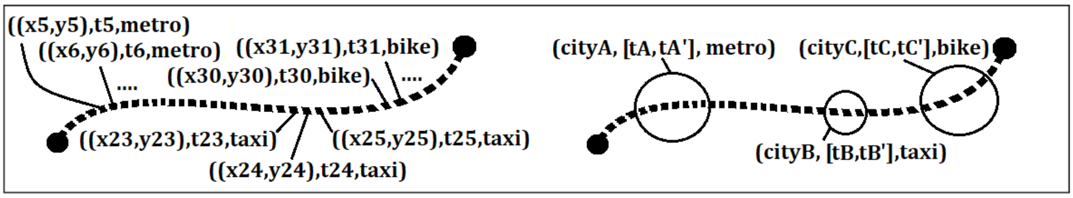


Figure 2.4: A semantic trajectory can be either geometric (left) or symbolic (right). Here, it is enriched with information about the transportation means.

Finally, it should be noted that semantic trajectories can be either geometric or symbolic, but the latter is typically more convenient, as exemplified in Figure 2.4. This can be more easily appreciated, if one notices that in the trajectory research literature, geometric trajectories have been almost exclusively outdoor trajectories, whereas symbolic ones have been used in both outdoor and indoor settings. The link between symbolic trajectories and indoor trajectories clearly lies in the existence of well-defined spatial areas in almost any building, which can be fittingly represented as symbols. Hence, these spatial entities often also have clearly distinguishable semantics (e.g. the function of a space, the temperature of a room). However, the connection between symbolic trajectories and outdoor trajectories is less evident, and has its roots in efforts (over the last fifteen years) to transition from purely spatiotemporal outdoor trajectories, composed of GPS coordinate pairs, to semantic trajectories. These efforts lead to the realization that, single coordinate points only rarely hold significant semantic content so as to justify treating them as the atomic unit of spatial information, or to use the words of [187], that *annotating each GPS point may result in information overload*. Consequently, concepts like the *semantic point* [25] are often not very useful. Instead, the preferred approach by far has been to group coordinate points into symbolic regions, and only then proceed into adding semantic information to the symbolic trajectories. Therefore, whereas indoor symbolic trajectories were conceived intuitively, in the world of outdoor GIS, symbolic trajectories have only really served as a step towards semantic trajectories.

This Thesis is particularly interested in exploring how semantic trajectories can serve human mobility analytics in indoor spaces, especially in the museum domain which will be the focus of chapter 6.

2.2.2 State-of-the-art Semantic Trajectory Models

Over the last ten to fifteen years, considerable progress has been made in the field of semantic trajectory data modeling. Here, this progress is overviewed in loose chronological order, starting from trajectory models encompassing a basic level of semantics and moving on to ones with more complex semantics. Particular attention is paid to the modeling ideas that withstood the test of time, in order to gain the necessary understanding for developing a new conceptual model in chapter 3, but also for informative and educational purposes. The limitations of these models are identified here but will be addressed in detail in chapter 3.

2.2.2.1 The First Conceptual Semantic Trajectory Model

In [163], the first formal definition of a semantic trajectory at the conceptual level was proposed by Spaccapietra et al. in an effort to structure movement data into identifiable and countable units. A moving object is said to potentially produce many trajectories during its lifespan, some of which are meaningful and can be semantically segmented. Thus, the authors define a *trajectory* as the user defined record of spatiotemporal evolvement of the position of the moving object, during a given time interval (part of its lifespan), as indicated by the mapping $[t_{begin}, t_{end}] \rightarrow space$, and in order to achieve a certain goal.

On top of this definition, the authors propose a characterization of trajectories with semantic annotation properties (e.g. the name of the moving object's location) and spatial and/or thematic integrity constraints (e.g. movement restrictions based on the type of moving object). It should be clarified that these are to be understood as generic annotations, i.e. as text attachments or unstructured content tags, because the authors do not specify any particular format. With respect to the content of the annotations however, the authors do differentiate between semantic properties that hold a constant or a time-varying value, either throughout the whole trajectory or during the corresponding trajectory component that they characterize.

Their modeling approach is illustrated by a hypothetical example of the study of the behavior of white storks during their migration period. The type of semantics in relation to this scenario includes, with respect to stops, the type of stop (nightly or longer), the food availability, the bird activities during the stop (feeding, resting), the bird attributes (weight, percentage of fat, body temperature, health condition), and with respect to moves, the bird's flying altitude, the topography, the weather conditions, any potential obstacles (natural or artificial), the flock that the bird is part of, etc.

The authors also propose the semantic segmentation of a trajectory into *stops* and *moves*, that they believe to be inherent semantic facets of any trajectory. Thus, they view a trajectory as a sequence of time sub-intervals where the moving object's position alternates between changing and remaining fixed. More specifically, the two concepts are purposefully defined generically, so as to be refined by the modeler according to the needs of the application at hand (e.g. assigned specific geometries):

Definition 2.2.3 (trajectory stop)
A stop is defined as a non-empty (temporally disjoint with any other) trajectory part $[t_{beginstopx}, t_{endstopx}]$ *where the traveling object does not move as far as the application view is concerned.*

In the context of the previous white stork migration example, this concept would describe where and why a bird makes a stopover during its migration.

Definition 2.2.4 (trajectory move)
A move is defined as a non-empty (temporally disjoint with any other) trajectory part $[t_{beginmovex}, t_{endmovex}]$ *delimited by stops and/or the temporal extremities of the trajectory itself* t_{begin}, t_{end}.

Again in the context of the previous white stork migration example, a move would be the representation of a bird going from one place to the next, or in the context of a human mobility example, the transition between two different cities.

Finally, the proposed model is accompanied by two materialization methods. The first is a design pattern which is imported by a particular application database schema and personalized accordingly. The alternative is the use of two dedicated *TrajectoryType* and *TrajectoryListType* data types which hide the handling of the trajectory and its components behind a set of basic methods that a particular application will be using.

Although [163] was a pioneering effort with respect to trajectory modeling, its main limitations are, first that stops and moves add only a first level of semantic content to the trajectory which is not enough for certain modern day applications, and secondly that it is more pertinent to outdoor trajectories, as indicated by the aforementioned example.

2.2.2.2 Early Models Extending "Stop-Move" Semantics

In [11], Alvares et al. identify that if trajectories simply consist of data in the form (tid, x, y, t) where tid is the moving object identifier, x and y are the spatial coordinates, and t is a timestamp, then trajectory analysis becomes computationally expensive and complex from the user's perspective, and lacks semantic information at the representation and data manipulation levels. This is why they adopt the trajectory model of [163], whose main semantic facet is the segmentation of a trajectory into stops and moves. More specifically, the authors define the concepts of a *sample trajectory*, a *candidate stop*, an *application*, and a *stop*, as follows:

Definition 2.2.5 (sample trajectory)

A sample trajectory T is a list $< (x_0, y_0, t_0), (x_1, y_1, t_1), ..., (x_N, y_N, t_N) >$ of space-time points, where x_i, y_i, $t_i \in \mathbb{R}$, $i = 0, ..., N$, $t_0 < t_1 < ... < t_N$.

Definition 2.2.6 (candidate stop)

A candidate stop C is a tuple (R_C, Δ_C) whose geometry R_C is a topologically closed polygon in $\mathbb{R}^2$ and whose minimum time duration Δ_C is a strictly positive real number.

Definition 2.2.7 (application)

An application $\mathcal{A}$ is a finite set of candidate stops $\{C_1 = (R_{C_1}, \Delta_{C_1}), ..., C_N = (R_{C_N}, \Delta_{C_N})\}$ with mutually non-overlapping geometries $R_{C_1}, ..., R_{C_N}$.

Definition 2.2.8 (stop)

A stop of trajectory T with respect to application $\mathcal{A}$ is a tuple (R_{C_k}, t_i, t_{i+l}) for which $\exists$ a subtrajectory $< (x_i, y_i, t_i), (x_{i+1}, y_{i+1}, t_{i+1}), ..., (x_{i+l}, y_{i+l}, t_{i+l}) >$ of T and a candidate stop $(R_{C_k}, \Delta_{C_k}) \in \mathcal{A}$, such that $\forall j \in [i, i+l] : (x_j, y_j) \in R_{C_k}$ and $|t_{i+l} - t_i| \geq \Delta_{C_k}$.

Finally, the authors illustrate their framework with an example of a 2-month dataset of tourist visits in the city of Paris, where candidate stops consist of feature types such as *hotel*, *touristic place*, *shopping area*, *train station*, *airport*, etc. or more specific ones such as *Ibis Hotel*, *Eiffel Tower*, etc. Timewise a discretization at multiple levels of granularity is proposed, resulting in values such as *morning*, *afternoon*, *evening*, *rush hours*, *[07:00-09:00]*, *[17:00-19:00]*, *weekdays* , *weekend*, etc. They also implemented it as part of the Weka data mining toolkit, under the name of Weka-STPM (Semantic Trajectory Preprocessing Module) [12].

The advantage of the model proposed by [11] is that, by adopting the stop-move representation of [163], it enables non-spatial queries and data mining tasks related to behaviors such as arriving at the airport, going from hotels to touristic places, going from touristic places to shopping areas, etc. On the other hand, the main limitation is again that the trajectory representation is geared towards outdoor geometric trajectories, and as a result proposes a semantic enrichment based on geographic information alone.

In [23], Bogorny et al. also adopt the conceptual model of trajectories of Spaccapietra et al. [163] along with its semantic interpretation of a trajectory as a sequence of alternating stop and move segments. Stops are here associated with important visited places. However, the authors extend the model with some fundamental data mining concepts in the form of new classes, attributes, and methods, in order to enable trajectory mining functionality i.e. the extracting of frequent patterns, sequential patterns, and association rules. More importantly, the authors define a *semantic trajectory* and the pattern mining notion of *support* as follows:

Definition 2.2.9 (semantic trajectory)

A semantic trajectory is a finite sequence $< I_1, I_2, ..., I_n >$ where each item I_k is a stop or a move, and has a spatial dimension (a spatial feature name and type) and a temporal dimension (a generalized time description).

Definition 2.2.10 (support)

The support $s(X)$ of a set of items $X = \{x_1, x_2, ..., x_n\}$ with respect to a set of trajectories T is the fraction of trajectories in T that contain X

Definition 2.2.11 (trajectory frequent pattern & trajectory sequential pattern)

If X is a set and $s(X) \geq minSup$ then X is a trajectory frequent pattern with respect to T. If X is a time-ordered sequence and $s(X) \geq minSup$ then X is a trajectory sequential pattern with respect to T.

For example, a trajectory frequent pattern might be
$\{ReligiousPlace_{[weekend]}, Restaurant_{[weekend]}\})$
whereas a trajectory sequential pattern might be
$\{Work_{[morning]}, ShoppingCenter_{[afternoon]}, Gym_{[afternoon]}\})$.
The authors also define a *trajectory association rule*:

Definition 2.2.12 (trajectory association rule)

A trajectory association rule with respect to T, $minSup$, and $minConf$ is an implication of the form $X \implies Y$, where X and Y are two disjoint sets of items, if $s(X \cup Y) \geq minSup$ and $c \geq minConf$ where $s(X \cup Y)$ is the support of the rule $X \implies Y$, and the confidence c is defined as $\frac{s(X \cup Y)}{s(X)}$.

At the level of actual data storage, the authors represent a *stop*, a *move*, and the three types of trajectory patterns, as database relations. Finally, their work is closely related to [24], where Bogorny et al. propose a trajectory data mining query language called *ST-DMQL*, and implement it as an extension of spatial SQL. The main modeling limitation in both works, is that a semantic trajectory is considered to *aggregate the geographic information that is necessary for the analysis of the trajectory*, but can not represent other types of semantics not related to the geographic space. For instance, a semantic trajectory example given in [24] is

$Airport$ [08:00–08:30] $\rightarrow$ $IbisHotel$ [09:00–12:00] $\rightarrow$ $EiffelTower$ [13:00–15:00] $\rightarrow$ $LouvreMuseum$ [16:00–18:00],

and another example given in [24] is

$Home$ [8PM–7AM] $\rightarrow$ $Work$ [8AM–1PM] $\rightarrow$ $ShoppingCenter$ [2PM–5PM] $\rightarrow$ Gym [6PM–7PM].

However, these trajectories are practically almost symbolic, because the name of a place alone is not informative enough to justify viewing the trajectory from a semantic perspective, unless more information about those places enters the model.

2.2.2.3 A Conceptual Framework For Movement Analysis

In [14, 15], Andrienko et al. propose a generic conceptual modeling framework aiming to connect the analysis of *movement* data with the *spatiotemporal context* of movement. The two concepts are illustrated in Figure 2.5 and defined as follows:

Definition 2.2.13 (movement)

A movement is the change of the spatial position(s) of one or more moving objects over time.

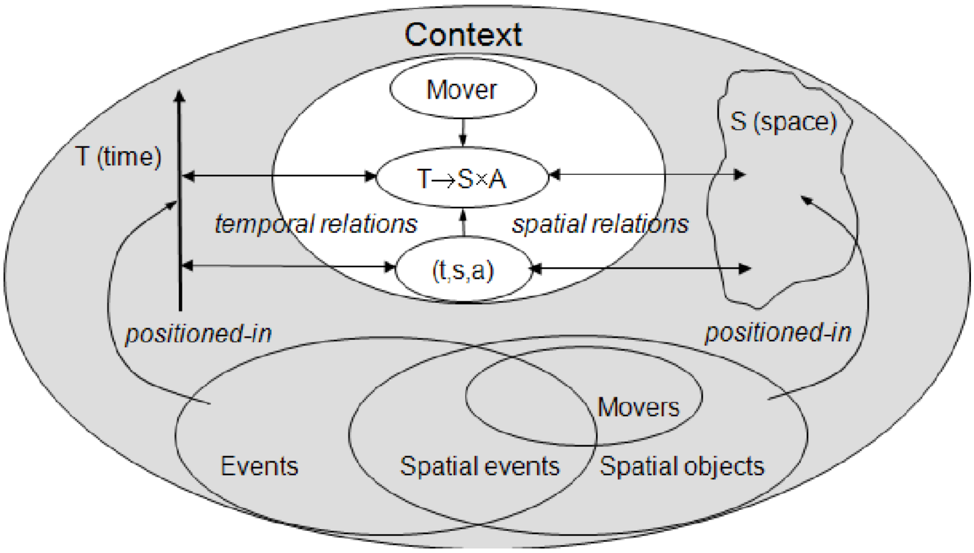


FIGURE 2.5: A “Venn-like” diagram reflecting the components of the spatiotemporal context (gray) of movement data (white), according to the modeling framework of [14, 15].

Definition 2.2.14 (spatiotemporal context / environment)
The spatiotemporal context of movement is the physical space and time (along with their properties) where it takes place, together with the objects and events that co-exist in that space and time. It is fundamentally composed of the set of locations S (representing space), the set of time units T (representing time instants or intervals), and the set of objects O (representing physical and abstract entities)

Locations in specific may be assigned arbitrary geometric shapes (e.g. points, lines, areas, volumes), whereas objects are categorized based on their spatial and temporal properties into:

- *spatial objects* (having a particular position in space at any moment of their existence);
- *temporal objects / events* (having a particular position in time, i.e. a limited time of existence with respect to the time period of observation);
- *spatial events* (having a particular position in space and time);
- *moving objects / movers* (having a changing spatial position over time);
- *spatiotemporal objects* (i.e. spatial events and movers).

Additionally, the authors introduce the notion of a *movement event*:

Definition 2.2.15 (movement event)
A movement event (t, s) (where $t \in T$ and $s \in S$) is an elementary or composite - i.e. consisting of consecutive lower level ones - spatial event involved in a movement.

Hence, under the modeling framework of Spaccapietra et al., movement essentially becomes a collection of so-called spatial events, represented by the mapping $\tau : T \rightarrow S$ for a single mover or $\mu : OxT \rightarrow S$ for multiple movers. Moreover, the authors define the concept of *presence dynamics* and *spatial configuration* as follows:

Definition 2.2.16 (presence dynamics)
Presence dynamics is a dynamic (time-dependent) attribute of a location which characterizes it in terms of the objects that are present in it ($T \rightarrow P(O)$)

Definition 2.2.17 (spatial configuration)
Spatial configuration is an attribute of a time unit which characterizes it in terms of the objects existing in it and in terms of their spatial positions ($O \rightarrow S$).

Elements of the three sets S, T, O comprising the spatiotemporal context of movement may have properties represented as attribute values, which in turn can involve other elements of S, T, O. If its values are not purely spatiotemporal, then an attribute is said to be a *thematic attribute*. Thematic attributes of objects and locations are further distinguished into *static* ones and *dynamic* ones, depending on whether they change over time or not. Dynamic thematic attributes of movers may be derived from trajectories (e.g. speed, direction) and dynamic thematic attributes of locations may be derived from presence dynamics (e.g. counts of the objects, statistics of the objects' attributes or time spent in the locations).

Furthermore, each movement event may be linked to one or more elements of the spatiotemporal context via *relations*. The framework models the occurrences of such relations as spatial events. *Temporal relations* are categorized into *binary topological*, *ordering*, and *distance* relations, while *spatial relations* are categorized into *binary topological*, *directional* and *distance* relations. Topological and ordering relations can be modeled as predicates $PxQ \rightarrow \{true, false\}$. Distance and directional relations can be modeled either quantitatively by numeric functions $PxQ \rightarrow [0, \infty]$ or qualitatively by application-specific predicates (e.g. "near", "north"). The aforementioned types of relations can also be used to build more complex ones such as *density* relations (e.g. clustering, dispersion), *arrangement* relations (e.g. temporal sequence, spatial alignment), and *spatiotemporal* relations (i.e. movement patterns such as approaching, entering, following, concentrating). As a result, under the framework of Andrienko et al. the trajectories of movers are linked with locations via spatial relations and linked with times via temporal relations.

The difference between [14] and [15] lies only in the use cases used to illustrate the conceptual framework: in [14] the indicative trajectory consists of a sequence of geographically referenced photo taking events by a single Flickr user, whereas in [15] it consists of a sparse sequence of positions of wildlife animals. It is worth noting that, even though they pertain to completely different application domains, both use cases concern outdoor trajectories, which indicates the proclivity of the model towards outdoor settings. Despite this, the representation of the movement phenomenon is broken down into its most basic components in a very detailed manner, and all are interconnected thanks to the concepts of spatial events and relations. This makes the framework of Andrienko et al. one of the most exhaustive ones, in terms of linking all of the fundamental constituting elements of a trajectory together, and in principle applicable to indoor trajectories.

2.2.2.4 Annotation-based Model Used in SeMiTri Framework

In [190], Yan et al. propose a modeling and computing platform for inferring semantic abstractions from raw GPS data. Their goal is to take a step towards bridging the gap between semantic trajectory modeling and conventional data mining and machine learning, by forming semantic trajectories from low-level GPS and other real-life mobility feeds such as cellular location data. The proposed platform is based upon a so-called hybrid trajectory model, which adopts its semantic concepts mainly from the works of Yan et al. [189] and Spaccapietra et al. [163]), and is comprised of three submodels:

- The *data model* encapsulates the low-level representation of trajectories as derived from the characteristics of raw mobility data, based on temporal (e.g. hourly/daily/monthly) or spatial (e.g. geo-fenced) trajectory division points.
- The *conceptual model* encapsulates a mid-level semantic abstraction of trajectories as series of non-overlapping *episodes* of stops and moves.
- The *semantic model* encapsulates the "spatio-semantic" behavior of trajectories via semantic annotations of their episodes or of themselves as a whole.

In [187], Yan et al. present a closely related application-independent framework, called *SeMiTri*, for the semantic enrichment of raw GPS trajectories in the form of annotations. In [188], Yan et al. further extend the trajectory model of SeMiTri. Although the main goal of SeMiTri is to support multilevel trajectory abstractions, it also explores the annotation algorithms that take into account, not only the spatial and temporal properties of the raw data stream or its derived features (e.g. velocity, acceleration), but also contextual geographic and application data (e.g. road types, transport networks).

This is evident in the "shopping" and "traffic" semantics in the following indicative semantic trajectory example of [190]:

$home \overset{traffic}{\rightarrow} office \overset{traffic}{\rightarrow} market \overset{traffic}{\rightarrow} home,$

which also reveals the model's focus on outdoor trajectories. The same trajectory example is refined in [188] by adding the transportation means semantics:

$home \overset{road(bus)}{\rightarrow} office \overset{train(metro)}{\rightarrow} market \overset{pathway(walk)}{\rightarrow} home.$

Likewise, the semantic trajectory example given in [187]:

$(home, -9am, -) \rightarrow (road, 9am - 10am, on - bus) \rightarrow (office, 10am - 5pm, work)$
$\rightarrow (road, 5pm - 5.30pm, on - metro) \rightarrow (market, 5.30pm - 6pm, shopping) \rightarrow$
$(road, 6pm - 6.20pm, on - foot) \rightarrow (home, 6.20pm-, -)$

also includes transportation means semantics, whereas the similar semantic trajectory example offered in [188]:

$(Begin, home, -9am, -) \rightarrow (move, road, 9am - 10am, on - bus) \rightarrow$
$(stop, office, 10am - 5pm, work) \rightarrow (move, road, 5pm - 5.30pm, on - metro) \rightarrow$

$(stop, market, 5.30pm - 6pm, shopping) \rightarrow (move, road, 6pm - 6.20pm, walking) \rightarrow (End, home, 6.20pm-, -)$,

contains explicit annotations of each episode's basic type ("stop" or "move").

Furthermore, event examples such as "Harry just reached office", "Sally is shopping in the Owings Mills mall", and "Dave is stuck in traffic" reveal how the authors envisage trajectory semantics.

As made apparent by all of the above examples, the semantic content captured by SeMiTri constitutes an upgrade from previous related works reviewed, where it was limited to a very thin layer of "stop-move" and "named place" semantics. More precisely, the works of Yan et al. adopt a "stop-move" segmentation into episodes inspired by Spaccapietra et al. in [163] but then each episode is potentially enhanced with additional semantics, integrated from third party geographic sources such as geographic databases or social networks. Let us now look more closely at the formal concepts used to make this possible.

In [188, 190] a *GPS feed* and a *spatiotemporal trajectory* are defined as follows:

Definition 2.2.18 (GPS feed)
A GPS feed is a raw sequence $\mathcal{G} = \{p_1, p_2, ..., p_m\}$ of spatiotemporal points $p_i = (x_i, y_i, t_i)$ of a moving object.

Definition 2.2.19 (spatiotemporal trajectory)
A spatiotemporal trajectory T_{spa} is a cleaned subsequence of a GPS feed $\mathcal{G}$ that represents a meaningful unit of movement and corresponds to a given time interval $[t_{begin}, t_{end}]$ during which it does not contain any significant spatial or temporal gap.

In [187], a *raw trajectory* is defined as follows:

Definition 2.2.20 (raw trajectory)
A raw trajectory is a sequence $T = \{Q_1, ..., Q_m\}$ of spatiotemporal points $Q_i = (x, y, t)$ each described as a $(longitude, latitude, timestamp)$ triple.

The conceptual model as presented in [187] and [188] enables the low-level enrichment of the location feed with semantic places:

Definition 2.2.21 (semantic place)
A semantic place $sp_i \in P = P_{region} \cup P_{line} \cup P_{point}$ is a meaningful geographic object having an extent that is either a region ($sp_i \in P_{region} = \{r_1, r_2, ..., r_{n_1}\}$ where each r_i represents a park, an administrative region, a residential land use cell, or any other possible Region Of Interest [ROI]), or a line ($sp_i \in P_{line} = \{l_1, l_2, ..., l_{n_2}\}$ where each l_i represents a jogging path, a highway, a road, or any other possible Line Of Interest [LOI]), or a point ($sp_i \in P_{point} = \{p_1, p_2, ..., p_{n_3}\}$ where each p_i represents a bar, a restaurant, a shopping mall, or any other possible Point Of Interest [POI]).

In [187], an *episode* is defined as follows:

Definition 2.2.22 (episode)
An episode is a maximal trajectory subsequence such that all its spatiotemporal positions comply with a given predicate that bears on the spatiotemporal positions

and/or their annotations (e.g. based on speed, transportation means, time periods, traveled zones)

Whereas, in [188, 190] an *episode* is similarly defined as follows:

Definition 2.2.23 (episode)
An episode is defined as an abstraction of a subsequence of a spatiotemporal trajectory's points $\{p_1^{e_i}, ..., p_k^{e_i}\}$ *that are highly correlated with respect to some identifiable spatiotemporal feature (e.g. velocity, acceleration, orientation, density, time interval).*

Either way, the conceptual model enables the automated structuring of trajectories into episodes based on different algorithms and techniques. Hence, in [187], a *structured semantic trajectory* is defined as follows:

Definition 2.2.24 (structured semantic trajectory)
A structured semantic trajectory is a sequence of episodes $SST = \{ep_1, ep_2, ..., ep_m\}$*, where each episode corresponds to a subsequence of the original trajectory* $ep_i = (sp, time_{in}, time_{out}, A)$*,* $sp \in P$ *is a link to a semantic place,* $time_{in}$ *and* $time_{out}$ *are respectively the time that the moving object enters and exits sp, and* A *is a set of other annotations associated to the whole episode (e.g. the activity of a stop, the transportation means of a move).*

Similarly, in [188, 190] a *structured trajectory* is defined as follows:

Definition 2.2.25 (structured trajectory)
A structured trajectory is a sequence of episodes $T_{str} = \{e_1, e_2, ..., e_m\}$ *where each episode is denoted* $e_i = (time_{from}, time_{to}, da, rep)$*,* $time_{from}$ *is the instant of the first point of the episode,* $time_{to}$ *is the instant of the last point of the episode, da is the episode's defining annotation (i.e. the spatiotemporal characteristics shared by all the spatiotemporal points of the episode), and rep is the episode's spatiotemporal or spatial representation (i.e. a sequence of the episode's points or a spatial abstraction of those such as its two extremity points, or its center point, or its bounding rectangle).*

In [188, 190], a *semantic trajectory* is defined as follows:

Definition 2.2.26 (semantic trajectory)
A semantic trajectory is a structured trajectory enhanced with semantic annotations of its episodes $T_{sem} = \{se_1, se_2, ..., se_m\}$ *where each semantic episode is denoted* $se_i = (da, sp_i, t_{in}^{sp_i}, t_{out}^{sp_i}, tagList)$*, the semantic position* sp_i *is a real-world geographic object (e.g. a building, a road segment, an administrative region) or one of its characteristics (e.g. its type) and represents the episode's location at the semantic level,* $t_{in}^{sp_i}$ *and* $t_{out}^{sp_i}$ *are the incoming and outgoing timestamps for the trajectory entering and leaving* sp_i *respectively (approximated by the episode's* $time_{from}$ *and* $time_{to}$ *values), and tagList is a list of additional semantic episode annotations (e.g. activity performed, transportation mode).*

In [187], a *semantic trajectory* is defined as follows:

Definition 2.2.27 (semantic trajectory)
$ST = \{Q'_1, Q'_2, ..., Q'_m\}$ *as a raw trajectory whose spatiotemporal positions* $Q'_i = (x, y, t, A)$ *have been complemented with associated annotations* A *that are links to semantic places that the moving object visited.*

From the above definitions, it can be noticed that [188, 190] use the term "semantic trajectory" differently in comparison to [187]. However, despite such terminology differences, and despite the more recent experimental results of [188], all three works share an almost identical conceptual trajectory model and the same two ways to semantically enhance a spatiotemporal/raw trajectory: either at the level of individual spatiotemporal points or at a higher level of grouped spatiotemporal points called *episodes*. The difference between the two is precisely the one already highlighted in section 2.2.1.4 and illustrated in Figure 2.4.

2.2.2.5 Annotation-based Model Refined

In [162], Spaccapietra et al. define *movement* as the timestamped changes in the spatial position of a moving object. They also define a *movement track* as the sequence of spatiotemporal positions that contains the evolving position of a moving object, while *raw data* are defined as the captured data that represent such a movement track, typically in the form of $(instant, point)$ pairs. Temporal gaps in the movement track greater than the sampling rate of its raw data, are said to be either accidental and called *holes*, or intentional and called *semantic gaps*. A similar concept called *invisibility duration* has also been proposed by Teng et al. in [168] to address uncertainty in location information.

Let us proceed with the main conceptual modeling definitions proposed by [162].

A *raw trajectory* is defined as follows:

Definition 2.2.28 (raw trajectory)
A raw trajectory is a tuple (trajectoryID, movingObjectID, track: LISTOF position(instant, point)).

An *application data repository* is defined as follows:

Definition 2.2.29 (application data repository)
An application data repository is the external sources (e.g. databases, GIS, web pages) that offer information (e.g. goal, activity, or transportation means of that part) that can be attached to parts of the movement track.

An *annotation* is defined as follows:

Definition 2.2.30 (annotation)
An annotation is any (captured or inferred) additional data that enrich the knowledge about a trajectory or any part of a trajectory: an attribute value, a link to an object, or even a complex value composed of both attribute values and links to objects.

Next, with an *episode* defined verbatim from [187][2], a *trajectory* is defined as the segments of an object's movement that are of interest for a given application, whereas a *trajectory interpretation* is defined as a list of episodes constituting a trajectory. Thus, in the modeling approach proposed by Spaccapitera et al., a *semantic trajectory* is understood as a trajectory enhanced with annotations and/or one or several alternative interpretations, and is more formally defined as follows:

Definition 2.2.31 (semantic trajectory)
A semantic trajectory is a tuple (trajectoryID, objectID, trajAnnotations, track: LISTOF position (t, p, posAnnotations), semanticGaps: LISTOF gap (t_1, t_2), interpretations: SETOF interpretation (interpretationID, episodes: LISTOF episode (t'_1, t'_2, type, episodeAnnotations)))

In the above semantic trajectory definition, *trajAnnotations* is the (possibly empty) set of annotations associated to the trajectory as a whole (e.g. goal, cause, duration, length), *track* is the (temporally ordered according to ascending t) list of spatiotemporal positions of the moving object, p is a spatial element (assumed to be a 2D/3D-coordinate point), *posAnnotations* is a (possibly empty) set of annotations associated to a spatiotemporal position (the first being *Begin* and the last being *End*), *semanticGaps* is a (possibly empty) list of semantic gaps in the trajectory, *episodes* is the list of episodes reflecting a particular interpretation of the trajectory, *type* is the type of the episode (e.g. "stop", "move", "playing", "eating", "resting"), and *episodeAnnotations* is the (possibly empty) set of annotations associated to the episode.

Interestingly, there can be multiple alternative interpretations of the same trajectory, and therefore multiple corresponding segmentations of it into episodes. This is actually one of the main differences between the semantic trajectory model proposed by [139, 162] presented here, and the one proposed by [187, 188, 190] previously reviewed. In other words, if we take any semantic trajectory and find two different (no matter how similar) meaningful ways to divide it into episodes, according to Spaccapietra et al. we are still left with one semantic trajectory, whereas according to Yan et al. we end up with two different semantic trajectories. The repercussions of this modeling viewpoint choice have never been considered to the best of our knowledge.

Apart from the mild refinements in the core trajectory modeling part, the difference between [162] and the works of Yan et al. lie in the modeling of trajectory behaviors. Spaccapietra et al. define a *trajectory behavior/pattern* by its Boolean predicate $p(T)$, which - based on some logic formalism - indicates if a given trajectory T complies with the corresponding behavior or not. More specifically, a *trajectory behavior predicate* is said to denote some distinguishable spatial (e.g. "CrossAreaA", "Co-location", "Concentration"), temporal (e.g. within a certain time interval, or starting before a certain time), spatiotemporal (e.g. "Convergence", "Expansion", "Progression", "Flock", "Sequence", "TemporalCross"), or semantic (e.g. "Shopping" defined as the total duration of stops in places of type "Shop" being greater than a certain percentage of the total duration of stops) characteristic.

[2]As a maximal subsequence of a semantic trajectory, such that all its spatiotemporal positions comply with a given predicate, bearing on the spatiotemporal characteristics of the positions.

Behaviors can also be defined as any combination of such characteristics. For example, a "TouristGroup" can be defined as any "Flock" for which most stops are in places of type "Museum", or "Monument", or "SouvenirShop", or "Restaurant".

Finally, behaviors are classified into:

- *global behaviors*, whose predicates constrain the whole trajectory;
- *local behaviors*, whose predicates constrain only a part of the trajectory;
- *simple behaviors*, whose predicates consist in a set of conditions connected by the regular Boolean operators AND, OR, NOT;
- *complex behaviors*, whose predicates consist in a set of conditions connected by the Boolean operators and by at least one sequence operator;
- *individual behaviors*, whose predicates $p(T)$ need to be checked independently against any single trajectory T (e.g. "Tourist" behavior);
- *collective behaviors*, whose predicates $p(S)$ need to be checked simultaneously against a non-empty set of trajectories S of different moving objects (e.g. "Flock" behavior).

Clearly, the notion of a complex behavior proposed by Spaccapietra et al. largely coincides with that of a *sequential pattern*, that is used more broadly and will be examined in detail in section 4.3. Also, in between the last two types, we find behaviors that characterize a single trajectory but with respect to a group of trajectories (e.g. "Leadership" behavior). Finally, the authors refer to a priori known groups of trajectories as *groups* and to a posteriori identified groups of trajectories as *cohorts*.

In their survey work of [139], Parent et al. retain the model proposed by Spaccapietra et al. in [162], but refine some of the definitions and add a few new ones.

Definition 2.2.32 (raw trajectory)
A raw trajectory is a tuple (trajectoryID, movingObjectID, trace: LISTOF position(instant, point, δ)) where δ represents a (possibly empty) list of additional raw data (e.g. speed, direction).

In the above definition of a *raw trajectory*, an improvement over the model of [162] is the inclusion of the element δ, which serves to capture additional attribute-like data.

Definition 2.2.33 ("sound" raw trajectory)
A "sound" raw trajectory is a clean (i.e. noiseless), accurate (i.e. map-matched), and compressed (i.e. compact) trajectory.

The authors also overview common approaches for deriving "sound" raw trajectories from raw data, which is essentially a trajectory data preprocessing layer.

Definition 2.2.34 (semantic trajectory)
A semantic trajectory is a tuple (trajectoryID, movingObjectID, trajectoryAnnotations, trace: LISTOF position (instant, point, δ, positionAnnotations), semanticGaps: LISTOF gap (t_1, t_2), segmentations: SETOF segmentation (segmentationID, episodes: LISTOF episode (t_3, t_4, definingAnnotation, episodeAnnotations))).

In the above definition of a *semantic trajectory*, at least one of *trajectoryAnnotations* or *segmentations* must be non-empty, while *semanticGaps* are optional. When compared to Definition 2.2.27, various differences in terminology (e.g. "track" becomes "trace", "interpretation" becomes "segmentation", "type" becomes "definingAnnotation", "application data repository" becomes "contextual data repository") can be noticed, but more important is the addition of attribute data δ. Other than that, the concept of a semantic trajectory remains the same in the two works.

With respect to the concept of a *trajectory behavior/pattern*, Parent et al. retain its definition from [162] as a set of distinguishing characteristics that identifies a particular bearing of a moving object (or of a set of moving objects), and claim that a trajectory behavior may be used as a criterion for segmenting a trajectory into episodes i.e. homogeneous meaningful segments. Lastly, they distinguish between an *individual trajectory behavior* defined as a trajectory behavior whose predicate $p(T)$ bears on a single trajectory T (e.g. "Tourist"), and a *collective trajectory behavior* defined as a trajectory behavior whose predicate $p(S)$ bears on a non-empty set of trajectories S (e.g. "Flock").

In conclusion, [162] and [139] define meaningful abstractions in an attempt to form a semantic counterpart to raw GPS records. Compared to the models in [187, 188, 190], the major difference in their proposal is that a semantic trajectory is not annotated either at the level of points or at the level of episodes, but potentially at both levels at the same time. Also the semantic gaps is a novel feature.

2.2.2.6 CONSTAnT: Conceptual Model of Semantic Trajectories

In [25], Bogorny et al. present a conceptual semantic trajectory data model named CONSTAnT. It is inspired by the semantic trajectory model proposed by Parent et al. in [139] and complements it with an integrated conceptual schema, shown in Figure 2.6. Thus, it tries to organize the semantic information into concepts and relations, instead of using annotations like the semantic trajectory model of Yan et al. [187].

First, the authors define a *point* p as a tuple (x, y, t) where x, y are spatial coordinates and t is the timestamp in which the point was collected. Given this, a *trajectory* and a *subtrajectory* are defined as follows:

Definition 2.2.35 (trajectory)
A trajectory T is an ordered list of points $< p_1, p_2, ..., p_n >$ where $p_i = (x_i, y_i, t_i)$ and $t_1 < t_2 < ... < t_n$.

Definition 2.2.36 (subtrajectory)
A subtrajectory s of a trajectory T is an ordered list of points $< p_k, p_{k+1}, ..., p_{k+l} >$ where $p_i \subset T$, $k \geq 1$, $k + l \leq n$.

Contextual information (e.g. moving object temperature, place temperature, weather, transportation means, movement objective, activity) can enrich a trajectory into becoming a *semantic trajectory*, which is defined as follows:

Definition 2.2.37 (semantic trajectory)
A semantic trajectory T defined is a tuple (tid, oid, S, g, d), where tid and oid

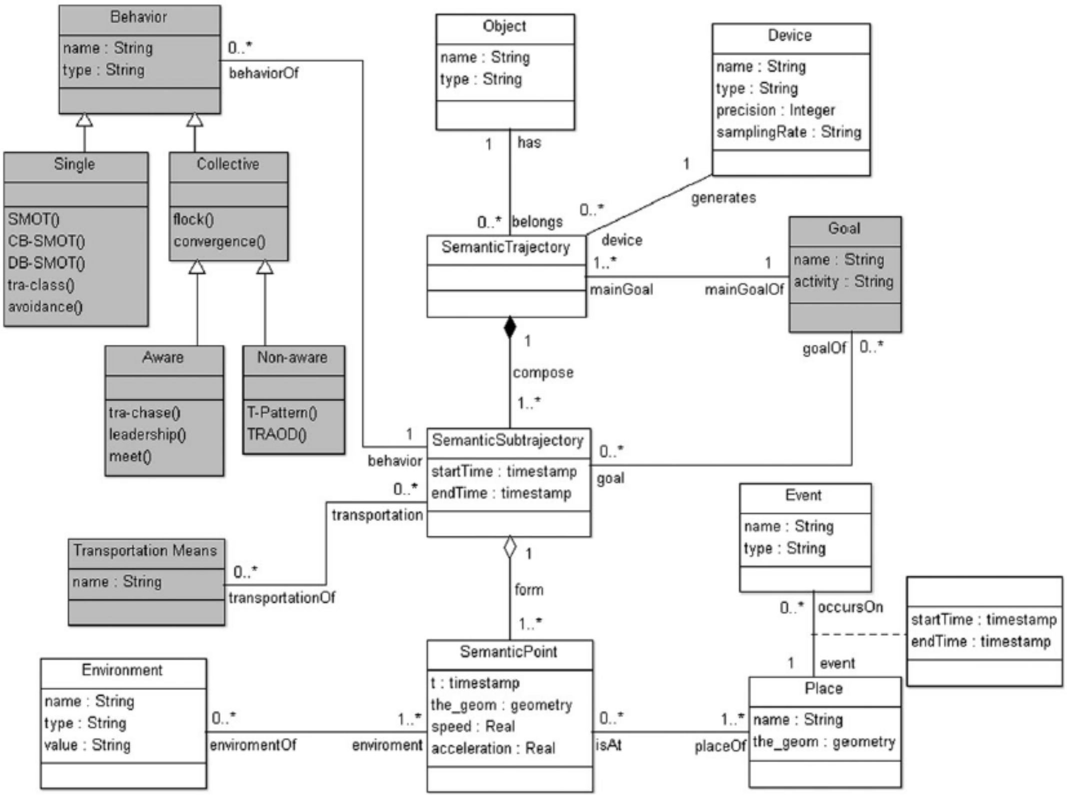


FIGURE 2.6: The class diagram of the CONSTAnT conceptual model taken from [25].

are the trajectory and moving object identifiers respectively, S is a non-empty list of semantic subtrajectories, g is the (required) general goal of the trajectory (the reason/objective of the movement), and d is the device that generated the trajectory (i.e. the apparatus that collected the sequence of points).

In the above definition, it is worth noticing that g is required and S must contain at least one semantic subtrajectory, which means that a semantic trajectory must have exactly one goal and at least one meaningful part.

Next, a *semantic subtrajectory* is defined in a rather different way:

Definition 2.2.38 (semantic subtrajectory)
A semantic subtrajectory $s \subset T$ is a tuple (tid, sid, P, G, M, B, startTime, endTime), where P is a list of consecutive semantic points, G is a set of goals of the subtrajectory, M is a set of transportation means, B is a set of behaviors, and at least one of G, M, B must be non-empty.

Thus, a semantic subtrajectory must correspond at least to one goal or one transportation means or one behavior, and potentially to multiple ones. This bears the question of how Bogorny et al. define the *goal* of a trajectory:

Definition 2.2.39 (goal)

A goal g is a triple $(id, name, activity)$*, where name is the objective that the moving object wants to achieve (e.g. "going to the gym", "work", "eat", "having fun"), and activity is what the object is going to do (e.g. "going to work", "reading a paper in the office", "teaching", "visit a city").*

As it is obvious from its definition, a goal can be related to an activity, or sometimes even be an activity itself (e.g. "jogging", "eat something", "watch a movie").

Finally, a *semantic point* is defined as follows:

Definition 2.2.40 (semantic point)

A semantic point $p \subset s$ *is a tuple (pid, x, y, t, V, L, sid), where x, y are geographic coordinates collected at time t, V is a set of environments related to where the point was collected, L is a set of places where the point is located, and at least one of* V*, L must be non-empty.*

In relation to the above definition, an *environment* $v \in V$ is defined as follows:

Definition 2.2.41 (environment)

An environment is a triple $(name, type, value)$*, where name is the name of an environmental attribute (e.g. air temperature, moving object temperature, humidity, pollution degree), type is whether the attribute refers to external or internal to the object information, and value is the value of the attribute.*

Definition 2.2.42 (event)

An event e is defined as a tuple (l, name, type, startTime, endTime), where l is a place, name is the name of the event, and type is the kind of the event.

Two examples of events given by the authors are, a musical show on May 13th, 2012 from 8 to 11 p.m. at Main Square in the city of Madrid, and a football match at Bernabéu Stadium on May 14th, 2012 from 4 to 6 p.m..

The model also adopts from [139] the definition of a *trajectory behavior*, and each semantic subtrajectory, as well as the whole semantic trajectory, may be related to one or several behaviors during their lifetime.

Finally, the model also adopts from [139] the distinction between individual - here called *single* - and *collective* behaviors. However, it extends the latter to also include patterns defined by multiple trajectories of the same moving object. It also further classifies collective behaviors into:

- *aware behaviors*, in which at least one trajectory is affected by one or more other trajectories (e.g. "chasing" behavior).
- *non-aware trajectories*, in which when trajectories behave similarly by coincidence and not intention (e.g. "highway driving").

Just like the annotation-based conceptual models proposed by Yan et al. and Parent et al., CONSTAnT primarily attempts to broaden the scope of semantic information included in the trajectory model, beyond just stops and moves. In place of episodes, it proposes the notion of semantic subtrajectories. Unlike those works however, it does not adopt a generic annotation-based approach to representation;

instead it tries to walk the fine line of specifying semantics, enough so that it better reflects the meaning of a movement, but less than a domain-specific ontology would do. For example, CONSTAnT's concepts have stricter semantic requirements in comparison to the aforementioned models: a semantic trajectory requires a goal, a semantic subtrajectory requires either a goal, a behavior, or a transportation means, an event requires an event type, a goal requires a related activity, etc. The downside is of course that there can be applications where these restrictions render the model unfit. For example, a museum exhibition can be visited without any overarching goal other than the (tautological) general goal of seeing the exhibit(s), but still be comprised of parts having their own goals such as avoiding congested rooms or resting.

2.2.2.7 Symbolic Trajectory Model used in SECONDO DBMS

In [77], Güting et al. propose a generic trajectory model to capture a wide range of meanings in an originally geometric trajectory. Mentioned examples include, understanding that a tourist is visiting the Louvre or having dinner at a restaurant (instead of just being at some geographic coordinates in France), understanding that a car was in a traffic jam during a certain period, understanding an animal's migration behavior, understanding that a bird is flying in a swarm, understanding whether a person moving around is walking, going by bicycle or using a bus, and others. Hence, the focus is once again on outdoor trajectories. However, unlike previous trajectory modeling works using the adjective "semantic", Güting et al. prefer to characterize their modeled trajectories as "symbolic". In our view, this adjective would be more fitting for earlier trajectory models whose inclusion of semantic information was absolutely minimal (e.g. meaningful places).

More specifically, they define a *symbolic trajectory* as follows:

Definition 2.2.43 (symbolic trajectory)
A symbolic trajectory is a temporally ordered sequence of pairs $< (i_1, l_1), ..., (i_n, l_n) >$ *where* i_j *is a time interval (disjoint from others) and* l_j *is a time dependent label i.e. a short character string.*

For example, a semantic trajectory might be:
<([8:30-8:45], walk), ([8:45-9:13], train), ([9:13-9:19], walk)>

The authors claim that, although symbolic trajectories can be used on their own, if they are combined with geometric trajectories then their labels essentially become the annotations of the geometric trajectory. Accordingly, they distinguish three kinds of *trajectory annotations* reflecting:

- semantics obtained from data mining (e.g. transportation modes, activities);
- relations to the spatiotemporal environment (e.g. temperature, weather, states, districts, cell towers);
- properties derived directly from the raw geometric trajectory (e.g. direction, speed, acceleration, altitude).

Unlike previous works reviewed in this section, [77] focuses more on the database querying aspects rather than the conceptual modeling ones, and actually also proposes a language for pattern matching and rewriting of symbolic trajectories.

A label is thus also modeled as an abstract data type called *moving label* or *mlabel*, and integrated into the data type framework of the SECONDO prototype DBMS for moving objects [75], whose generic operations it inherits. Similarly, *moving places* or *mplaces* are defined in the framework. Also, a *pattern* is defined as a sequence containing the same type of pairs that make up a symbolic trajectory (called *units*) and wildcards that match any sequence of units, as well as regular expressions over such elements. The corresponding pattern language proposed has been demoed by Valdés et al. for the case of Aircraft Traffic Control data composed of timestamped aircraft position and altitude recordings in [173]. Finally, in [172], Valdés et al. extend the flexibility and expressiveness of the pattern language by changing its semantics, making it fit for analyzing large datasets having any number of time-dependent attributes of different types.

2.2.2.8 MASTER: Multiple Aspect Trajectory Representation

In [67], Ferrero et al. envision a *multiple aspect* representation of trajectories, motivated by the fact that, even though trajectories are gradually being represented as more complex data types bearing several dimensions, there has never been any concentrated effort to tackle multiple aspect trajectory data analysis and mining. By *aspect* the authors actually refer to the main data dimension by which the trajectory is being interpreted, be it a spatial / temporal dimension (e.g. "raw data" aspect) or a semantic dimension (e.g. "stop" aspects, "transportation means" aspect, "activities" aspect, "weather conditions" aspect), or perhaps a combination of them. They argue that such a viewpoint allows analytical queries to concern multiple aspects and dimensions at once, and as a result help answer questions such as: "Which transportation means do individuals use when it is raining?", "Do groups of friends visit specific places only by bus and with good weather?", "How do weather conditions affect traffic jams?", "Which is the transportation pattern at a beach town on a rainy weekend and a sunny weekend?".

Under this type of modeling, an aspect may encompass multiple data dimensions: space, time, semantics, or any of their combinations. Let us relate these notions to some query examples given in [67]: "the name and location of the stops with duration above 1 hour" refers to one aspect (stops) and three dimensions (stop location, stop duration, stop name), whereas "the average speed of the moving object when traveling by car when it is raining" refers to three different aspects (raw data, transportation means, weather conditions).

According to Ferrero et al., a typical semantic trajectory example in the related literature would be:

Hotel[8:00,10:00] → Mall[13:00,19:00] → Restaurant[20:00,21:30].

But based on their proposed model, the above representation only encodes the trajectory's stops aspect, which encompasses a semantic dimension (the type of place) and a temporal dimension (the duration of stay).

The same trajectory may also be represented using a transportation means aspect:
On foot[8:00,10:00] $\overset{by\_car}{\rightarrow}$ On foot[13:00,19:00] $\overset{by\_bus}{\rightarrow}$ On foot[20:00,21:30].

Or alternatively, it can be represented by incorporating a weather conditions aspect:
Rainy[8:00,13:00] → Sunny[13:00,19:00].

In general, the main idea presented in [67] is that trajectories have several aspects to be considered in their analysis. Then, this serves as motivation for Mello et al. to propose MASTER [133], a conceptual semantic trajectory model, converted into a logical RDF Schema and implemented using a middleware which stores RDF data into multiple NoSQL databases. MASTER focuses on the heterogeneity of the semantic information of trajectories by reintroducing the notion of an aspect and introducing that of its type:

Definition 2.2.44 (aspect)
An aspect $asp = (desc, SAT)$ is a real-world fact relevant to the trajectory data analysis, where:

- *desc is the aspect description;*
- *$SAT = \{sat\}$ is a set of aspect types that the aspect may hold;*
- *$sat_i = (asp_{type\_k}, ATV_k)$, $sat_i \in SAT$;*
- *an aspect type asp_{type_k};*
- *a non-empty set $ATV_k = \{a_1 : v_1, a_2 : v_2, ..., a_n : v_n\}$ of attribute-value pairs so that each pair $(ai : vi) \in ATV_k$ is an instantiation of a property a_i of asp_{type_k} with an atomic or multivalued value v_i.*

A trajectory may have numerous aspects, complex contextual data dimensions of heterogeneous form such as numbers, ranges, text, geometries (e.g. the shape of a hurricane at a specific time instant), complex objects, etc. Of course, the specific aspects to be included in the representation depend on the application. To illustrate this, the authors claim that a tourism example might include, visited POIS along with their categories, prices, and reviews, transportation means, social status of the tourist, weather conditions, general mood or opinions about the town or local services, etc. A smart city example may include the aspects of home and work, working hours, transportation mode, social status, weather condition, etc. A bird migration example may include regions where birds fly, rest, or eat, relationships with other species, the temperature, the types of vegetation, etc.

Definition 2.2.45 (aspect type)
An aspect type $asp_{type} = (desc, ATT, asp_{supertype})$ is a category of aspects that is composed of:

- *a description desc of the aspects,*
- *a set of attributes $ATT = \{a_1, a_2, ..., a_k\}$ that hold their properties,*

- *a (possibly empty) supertype aspect* $asp_{supertype}$

Aspect types act as the metadata of aspects, akin to a semantic taxonomy. For example, a "weather condition" aspect type may contain attributes like "temperature", "wind speed", "climate".

For example, a "train" aspect belongs to a "transportation mode" aspect type, and a "rain" aspect belongs to a "weather condition" aspect type. An aspect type has a set of attributes and it may also be a subtype of a more general aspect type, allowing the modeling of an aspect type $subtypeOf$ hierarchy (e.g. POI, accommodation, hotel).

As another mentioned example, an "Il Campanario Resort" aspect belongs to a "hotel" aspect type which has the attributes "geographic coordinates", "address", "stars", "types of rooms", and "facilities". Hence, this aspect will have corresponding attribute-values such as "geographic coordinates":-27.439771,-48.500802, "address":Buzios Ave., Florianopolis, "stars":5, "types of rooms":{suite, suite junior}, "facilities":{gym, swimming pool, restaurant, bar, beach service}.

Similarly, a "happy" aspect belongs to a "mood" aspect type which has the attributes "emoticon" and "intensity". Hence, it has the attribute-values: "emoticon": :-D, "intensity": high.

Moreover, an aspect may hold several real-world meanings. An indicative example given in [133] is the "Sao Paulo" aspect which might mean the "town", the "state", the "soccer team" or even the "holy Sao Paulo". Thus, the notion of *semantic meaning* is defined to give the context of the aspect:

Definition 2.2.46 (semantic meaning)
A Semantic Meaning $SM = (asp, asp_{type})$ *is an association between an aspect* asp *and an aspect type* asp_{type}*, so that the latter belongs to former's aspect types.*

To highlight the general scope of the semantic information that might be of interest to trajectories, the authors mention, data about the place (e.g. temperature, air pollution, noise, luminosity), about the object that is moving around or inside this place (e.g. the heart rate, the emotional status, blood pressure, sleeping stages).

Given all of the above concepts, a *multiple aspect trajectory* is defined as follows:

Definition 2.2.47 (multiple aspect trajectory)
A multiple aspect trajectory is a tuple $mat = (P, S\_LTA, mo, desc)$ *where:*

- $P = <p_1, p_2, ..., p_n>$ *is a sequence of points* $p_i = (x_i, y_i, t_i, S\_VA)$ *each consisting of timestamped* (x, y) *coordinate points and a (possibly empty) set* $S\_VA = \{SM_{va}\}$ *of volatile aspects (i.e. changing semantic meanings),*

- $S\_LTA = \{SM_{lta}\}$ *is a set of long term aspects (i.e. not changing semantic meanings),*

- $mo = (motype, desc, S\_PA)$ *is the moving object consisting of a description* $desc$*, a set of (possibly empty) permanent aspects* S_PA*, being* $S\_PA = \{SMpa\}$*, a set of semantic meanings, and a type* $motype$ *that categorizes it,*

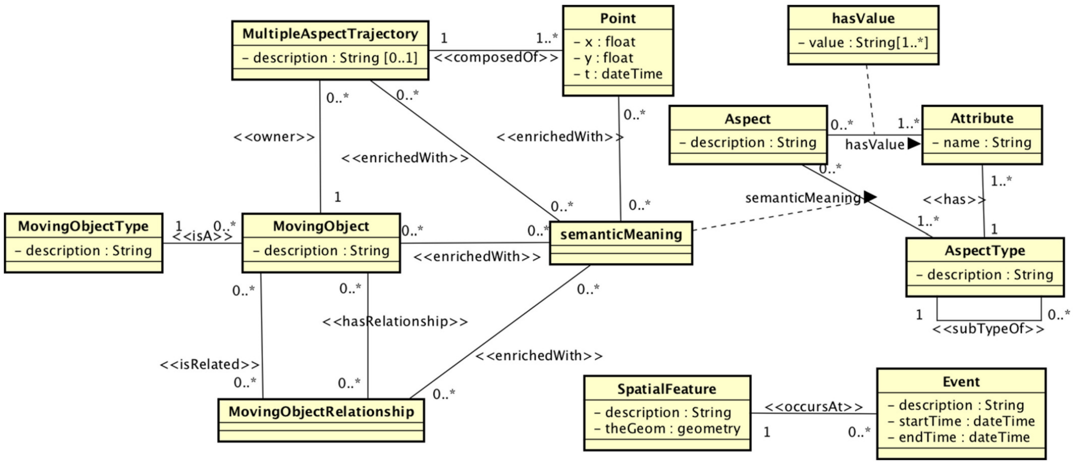


FIGURE 2.7: The conceptual data model of the multiple aspect modeling approach of [133].

- *desc is an aspect description.*

Noticeably, *volatile aspects* vary during a multiple aspect trajectory and are thus associated with each trajectory point (e.g. stops, heart rate), whereas *long term aspects* remain the same and are thus associated with the entire trajectory, and *permanent aspects* remain unchanged even across different trajectories since they are associated with the moving object and not with the movement.

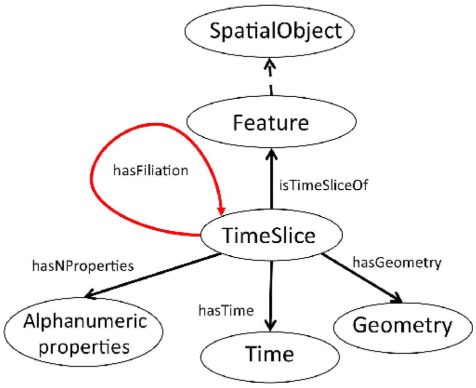
Subtrajectories are not defined in the model as the authors opt to leave any type of granularity representation to the analysis process (e.g. as a segmentation step). However, MASTER does define a *spatial feature* as a relevant PoI not spatially related to any trajectory point (e.g. a church between two PoIs), and an *event* as a happening at a spatial feature (thus again not related to the trajectory itself but perhaps relevant to the analytical query) for a valid period.

Finally, Mello et al. focus on the concept of a *moving object relationship* that is actually neglected by most other related modeling works. They define it as follows:

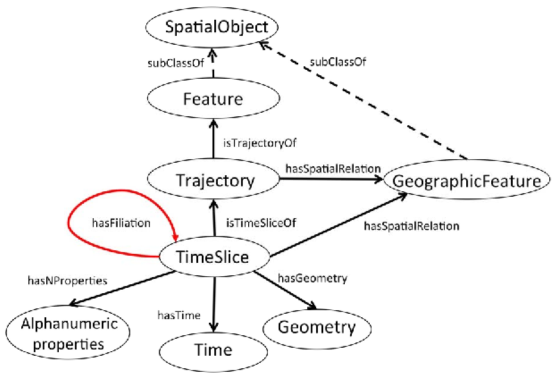
Definition 2.2.48 (moving object relationship)
A Moving Object Relationship $mor = (mo1, mo2, S\_RA)$ *is a relevant association between two moving objects* $mo1$ *and* $mo2$ *that holds a (possibly empty) set of relationship aspects* S_RA*, where* $S\_RA = \{SM_{ra}\}$ *is a set of semantic meanings.*

2.2.2.9 STriDE: Semantic Trajectories in Dynamic Environments

In [45], Cruz et al. propose one of the very few trajectory models oriented both towards indoor environments and towards the semantic aspects of movement. They introduce it in the form of a geographic ontology-based conceptual trajectory model called *Semantic Trajectories in Dynamic Environments* (*STriDE*), focusing on the representation of moving objects in dynamically changing built environments. By dynamic, the authors refer to environments that include moving objects and that may change in shape, size, or attributes (e.g. an entry gate becomes closed, a simple room becomes a meeting room). STriDE actually extends the Continuum model proposed by Harbelot et al. [79], which represents dynamic entities using *ephemeral timeslices*.



The Continuum model proposed in [79].



The extension of the Continuum model proposed in [45].

FIGURE 2.8: Structured (left) or ad-hoc (right) representation of a hierarchical space.

A timeslice is composed of an object identity, a set of object (alphanumeric) properties, a geometric spatial representation, and a valid period. Any time there is a change either in the identity, or in the geometry, or in the properties, a new timeslice is created. Thus, *filiation relationships* between consecutive timeslices associated with the same entity, are used to represent the entity's spatial or semantic evolution while preserving its identity. More generally, filiation relationships represent the succession between different representations of the same object at different instants of time, thereby being also useful for representing divisions or mergers of entities. Moreover, ontological taxonomies enable the association of the same timeslice to concepts at different levels of spatiosemantic granularity.

In STriDE in particular, a moving object is an instance of the class *Feature* and a semantic trajectory is defined as follows:

Definition 2.2.49 (semantic trajectory)
A semantic trajectory is a set of timeslices having a starting and an ending spatiotemporal point.

Finally, instead of using land parcels as in [79], Cruz et al. use building elements, represented by 2D floor plans with an additional third dimension. Then, semantic trajectories are linked to building element trajectories, using spatiotemporal relationships between instances of $TimeSlice$ objects and spatiotemporal integrity constrains. Finally, the temporal domain is represented - as suggested in [61] - as a linear structure composed of a set of strictly ordered $TemporalPoint$ objects, any pair of which defines a time interval.

2.2.3 Applied Trajectory Models: what is used in practice?

The purpose of designing a trajectory model is naturally to use it towards tackling some type of trajectory data mining and analysis task. Therefore, it is worth finding out to what degree formal trajectory models presented in section 2.2.2 have affected (or not) these works. Do trajectory mining and analysis methods already adopt semantic trajectory representations as envisioned more than a decade ago? If that is the case, to what extent do they use their offered features? If not, do they instead resort to simpler ad-hoc trajectory models? These are the types of questions that this section aims to address.

2.2.3.1 Semantic trajectory models

The previous section reviewed how trajectory data mining and analysis research works have handled the modeling of spatiotemporal trajectories. Within the scope of this Thesis however, the interest is in finding out how they deal with the modeling of semantic trajectories, and especially how they represent their semantic aspects.

SPLITTER method. In [83], Huang et al. propose a method called *SPLITTER* for retrieving a set of spatially coarse trajectory patterns and then progressively breaking down each pattern using a so-called weighted snippet shift algorithm. Their goal is to find interesting patterns in semantic trajectories which are defined as follows:

Definition 2.2.50 (semantic trajectory)
A semantic trajectory is a sequence $< (p_1, t_1), (p_2, t_2), ..., (p_l, t_l) >$ *of timestamped places, where each place is described by a spatial location as well as a semantic label (e.g., office, park).*

Hence, the authors only consider semantic trajectory aspects by virtue of a 2-level hierarchy comprised of location names and location types, with no support for additional data dimensions or deeper spatiosemantic hierarchies.

Hermoupolis simulator. In [142], Pelekis et al. propose a pattern-aware synthetic network-constrained trajectory generator, which produces semantically annotated moving object trajectories along with respective artificial "GPS-like" records. Their aim is to enable the validation of existing mobility data management and mining techniques, given that there exist no openly accessible semantic trajectory datasets to include both the raw GPS records and the corresponding semantic annotations.

The simulated movement patterns follow, either certain mobility profiles given as input (*Hermoupolis* simulator), or alternatively mobility profiles discovered in (typically small) real semantic trajectory datasets (*Hermoupolis* $^{by-example}$ simulator). Examples of such patterns include "from home to work and back to home", "from home to a mall for shopping, then to a restaurant for dining, and back to home", etc. Moreover, the authors represent a road network as follows:

Definition 2.2.51 (road network)
A road network N is a graph $G(V, E)$ *consisting of, a set of vertices* $V =$

$v_1, ..., v_n$ each corresponding to a geographical location (x, y), and a set of edges $E = \{e_{i,j} = (v_i, v_j)|v_i, v_j \in V, i \neq j\}$ each belonging to one of a small number of categories (i.e. road types) and being associated with a maximum speed and capacity at a specific time instance.

What is worth noticing in the above definition is the assignment of different types and attributes to the edges of the network, a useful way to semantically enhance network-based trajectories. Another type of information source used is the set of Points of Interest (PoIs), where a point of interest $poi \in PoI$ in the simulation's region is defined as follows:

Definition 2.2.52 (point of interest)
A point of interest is a tuple $< poi-id, poi-loc, poitags, poi-cat >$ corresponding to a vertex of the network, where $poi-id$ is its unique identifier, $poi-loc \in V$ is its location corresponding to a vertex, $poi-tags$ is a set of tags describing its utility (e.g. "caf'e del mar", "Greek tavern Parthenon"), and $poi-cat$ is its category (e.g. "caf'e", "restaurant").

Moreover, a (raw) *trajectory* is defined as follows:

Definition 2.2.53 (trajectory)
A trajectory τ is a tuple <o-id, traj-id, T>, where o-id is the identifier of the moving object, traj-id is the identifier of its specific trajectory, and T is a 3D polyline consisting of a sequence of $|T|$ pairs (p_i, t_i), $0 \leq i \leq |T|-1$, where in turn each p_i is a 2D point (x_i, y_i) laying over a vertex / edge of the road network, and t_i is its corresponding timestamp.

With regards to the polyline, the authors assume linear interpolation between consecutive pairs (p_i, t_i) and (p_{i+1}, t_{i+1}). A raw trajectory can be partitioned into a sequence of *raw sub-trajectories* defined as follows:

Definition 2.2.54 (subtrajectory)
A subtrajectory τ' (of trajectory τ) - valid in the interval $[t_i, t_j]$, $t_0 \leq t_i < t_j \leq t_{|T|-1}$ - is a tuple <o-id, traj-id, subtraj-id,T'>, where T' is the portion of T between timestamps t_i and t_j.

Then, the authors propose their own versions of an episode and of a semantic trajectory. In place of the former, they define a *LifeStep* as follows:

Definition 2.2.55 (LifeStep)
A LifeStep ls is a tuple <ls-id, ls-flag, MBB, tags, T-link>, where ls-id is the LifeStep's identifier, ls-flag is a flag taking the values "Move" or "Stop", MBB is a tuple $< MBR, [t_{start}, t_{end}] >$ corresponding to the 3D approximation of τ', MBR is the 2D Minimum Bounding Rectangle enclosing the spatial projection of τ' in a 2D plane, $[t_{start}, t_{end}]$ is the interval of the temporal projection of τ' in a 1D timeline, tags is a set of keywords describing the corresponding activities and semantic annotations related to this portion of movement (e.g. category of PoI for stops, type of road for moves), and T-link is a link to τ'.

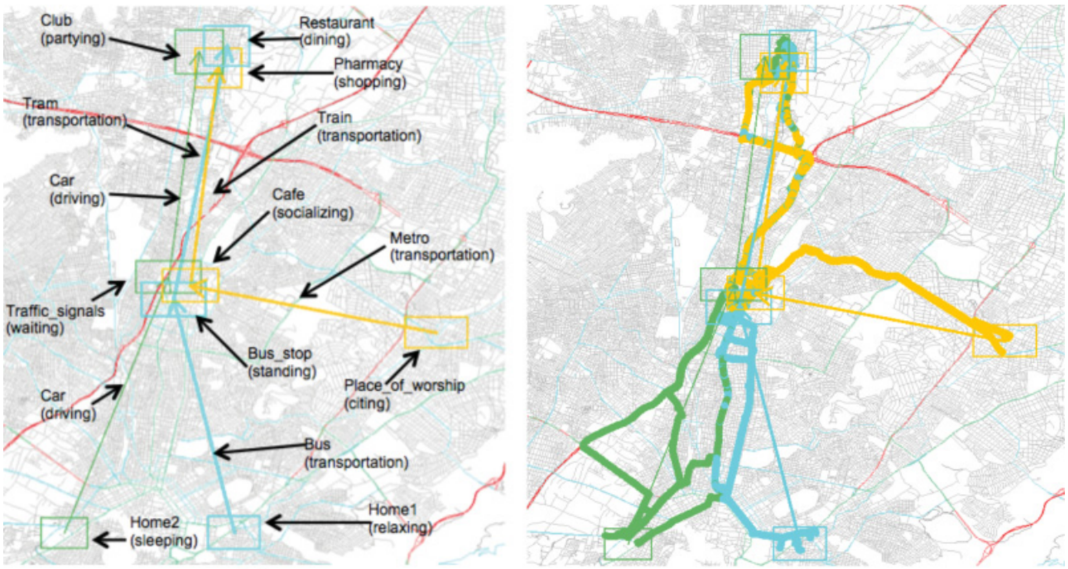


FIGURE 2.9: The trajectory generator of [142] considers mainly "stop-move" semantics targeting outdoor environments.

Similarly for the latter, they define a *mobility TimeLine* as follows:

Definition 2.2.56 (mobility TimeLine)
A mobility TimeLine mtl is a (raw) trajectory τ of a moving object valid in G, represented as a tuple <o-id, mtl-id,TLS>, where o-id is the moving object identifier, mtl-id is its mobility timeline identifier, and TLS is a sequence of successive $LifeSteps$ *of trajectory* τ $(ls_i.t_{end} = ls_{i+1}.t_{start})$.

Finally, the authors define a *Generalized LifeStep (GLS)* as follows:

Definition 2.2.57 (Generalized LifeStep)
A Generalized LifeStep is a tuple <gls-id, gls-flag, stop-params, move-params>, where:

- *gls-id is the GLS identifier*
- *gls-flag indicates whether the GLS corresponds to a Stop or a Move*
- *stop-params / move-params are optional parameters of Stop / Move LifeSteps*
- *stop-params is a tuple <MBB,* σ^2_{range}, σ^2_{dur}, *poi-cat> where:*
  - *MBB is a spatiotemporal region wherein all simulated Stop LifeStep reside*
  - σ^2_{range} *is the variance of the spatial range of the simulated Stop LifeSteps*
  - *poi-cat is the PoI category that simulated Stop LifeSteps should belong to*
- *move-params is a tuple <speedmax, move-tags> where:*

- *speedmax is the maximum allowed speed of this movement*
- *move-tags is a set of annotations attached to the simulated Move LifeStep*

As can be told from the parameters of their trajectory data generator, Pelekis et al. mainly consider in their semantic trajectory (i.e. LifeSteps) modeling approach the "stop-move" semantics of movement, as well as some move attributes, but not much else.

Stop Activity Inference. In [18], Beber et al. propose a method to integrate GPS human trajectory data with social media data and census data, aiming at multiple activity recognition in certain Points of Interest (PoIs) and the discovery of all individuals participating in each activity. Based on whether people who are together performing the same activity are connected (either directly or by sharing a common connection) or not, they infer whether that activity was performed in a group or individually. This extends the authors' proposed activity recognition algorithm from [17], which is based on a matching process that takes into account the similarity between a trajectory of georeferenced (associated with Foursquare) tweets and a knowledge base (extracted from Twitter and enriched with statistics) that describes the activities that can be performed at any given PoI.

In both works, Beber et al. use the same trajectory model. They define a *raw trajectory* as a temporally ordered sequence of spatial positions which do not present explicit semantics, and simplify the semantic trajectory definition of [25] as follows:

Definition 2.2.58 (semantic trajectory)
A semantic trajectory is a sequence of stops $S =< s_0, s_1, ..., s_n >$*, where the i-th stop is a tuple* $s_i = (x_i, y_i, startTime_i, endTime_i, poi_i)$ *composed of:*

- x_i *and* y_i *are the spatial coordinates of the stop at poi*
- $startTime_i$ *until* $endTime_i$ *denote when the stop takes place*
- *the tuple* $poi = (type, x, y, ot, ct)$ *is a Point of Interest (PoI), where in turn, type is the type of the PoI (e.g. restaurant), x and y are its spatial coordinates, and ot and ct are its opening and closing hours respectively.*

Also, the authors define an *activity* as follows:

Definition 2.2.59 (activity)
An activity is a tuple $a = (act.startTime, act.endTime, label, P)$*, where* $act.startTime$ *and* $act.endTime$ *are the starting and ending times of the activity* $label$*, and* P *is the set of moving objects performing this activity.*

Then, they define an *activity trajectory* as follows:

Definition 2.2.60 (activity trajectory)
An activity trajectory is a sequence $T =< t_0, t_1, ..., t_n >$ *of tuples* $t_i = (s_i, A_i)$*, where* s_i *is a stop, and* $A_i = \{a_0, a_1, ..., a_n\}$ *is the set of activities performed at* s_i*.*

In [17] in particular, the authors define a *georeferenced tweet* as a tuple (*text*, *time*, *day_week*, *POI*, *act*) where *act* is the activity extracted from the tweet text *text* shared at time *time* of the day of the week *day_week* at the PoI *POI*. This goes to show how social media are affecting the modeling of semantic trajectories, because oftentimes in practice the models follow data availability.

At the same time, motivated by the task of identifying group activities based on similarity measures, the authors in fact extend the model of [25], with respect to the semantics of places (by introducing the notion of *PoI profiles*) and in [18] also with respect to collective activities (by introducing the notions of *encounters* and *relationship degrees* between the moving objects).

In both [17, 18], in order to know which activities can happen at each PoI type, the authors introduce the notion of a *PoI type profile* defined as follows:

Definition 2.2.61 (PoI type profile)
A PoI type profile is a tuple $pro = (POItype, act, meanTime, sdTime, meanDuration, sdDuration, frequency)$, *where:*

- $meanTime$ *is the mean time of the observed occurrences of the activity act at the PoI type POItype*
- $sdTime$ *is the standard deviation of that time*
- $meanDuration$ *is the mean duration of act at POItype*
- $sdDuration$ *is the standard deviation of that duration*
- $frequency$ *is the frequency of act at POItype relative to the total number of activity occurrences observed at PoIs of the type POItype*

In other words, PoI profiles consist of an activity knowledge base representing the distribution of activity time and duration.

Moreover, to distinguish between different possible activities at the same PoI, they introduce the notion of a *sub-stop* defined as follows:

Definition 2.2.62 (sub-stop)
A sub-stop is a tuple $sub = (s, t.startTime, t.endTime, x, y)$, *where sub takes place inside the stop s from* $t.startTime$ *until* $t.endTime$, *and* x, y *are the coordinates of the centroid of the sub-stop.*

Actually, modeling sub-stops is simply another convention aimed at capturing trajectories at a finer spatiotemporal level of granularity.

Since they are targeting the task of activity recognition, the authors look at the similarity between PoI type profiles and trajectory sub-stops, and accordingly choose the most similar activity. More specifically, they define two similarity metrics (a time-based one and a frequency-based one). To this end, they also consider the relative frequencies of the activities.

Finally, they also give a threshold-based definition of an *encounter* and a *group activity*:

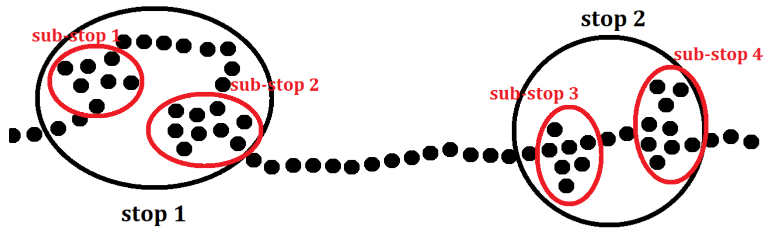


FIGURE 2.10: An illustration of the sub-stop concept proposed in [17, 18].

Definition 2.2.63 (encounter)
An encounter is two or more moving objects meeting at the same place, at the same time, and for a minimum amount of time.

Definition 2.2.64 (group activity)
A group activity is an activity performed by a group of people that have a certain relationship degree and that are at the same place, at the same time.

Unlike other types of trajectory mining and analysis, Beber et al. propose a plethora of trajectory-related concepts, despite the fact that their trajectory modeling is not their primary concern. This is justified by the fact that the problem of activity recognition is inherently semantic in nature, and therefore lends itself perfectly to the use of rich trajectory semantics.

Ranking semantic data sources. In [111], Leme et al. propose a method to rank datasets according to their suitability for enriching mobility data, based on the purpose of movement, to the extent that this can be induced by the sequence of places visited i.e. the stops. For example, the sequence [hotel, stadium, restaurant, hotel] in the city of Rio de Janeiro suggests that this could be a tourist trajectory. They describe a *raw trajectory* as a track collected by a mobile device representing the geometric facets of movement data and define it as follows:

Definition 2.2.65 (raw trajectory)
A raw trajectory is a sequence $\rho_0 = (p_1, p_2, ..., p_n)$ of spatiotemporal points such that the timestamp of p_1 is earlier than the timestamp of p_{i+1}.

Then, they define a *segmented trajectory* as follows:

Definition 2.2.66 (segmented trajectory)
A segmented trajectory of the raw trajectory ρ_0 is a sequence $\sigma_0 = (g_1, g_2, ..., g_n)$, where each segment g_i is a fragment (continuous subsequence) of ρ_0 in which a given property holds.

The authors adopt the "stop-move" segmentation strategy from [188] by merging the positions of corresponding consecutive Foursquare check-in tweets into move or stop segments, based on a threshold value of the time interval between them. This segmentation approach is motivated by the typical sparseness of trajectory data extracted from social media (which prohibits speed-based criteria for example). Then, the authors label each segment with taxonomic classifications of the place visited at the end of the segment.

Moreover, Leme et al. define a *contextual resource* as follows:

Definition 2.2.67 (contextual resource)
A contextual resource r of a dataset d is a pair $r^d = (r, d)$ where $r \in d$.

They also define the notion of *contextual information* as follows:

Definition 2.2.68 (contextual information)
A contextual information c of a segment g of the segmented trajectory σ_0 is a set of contextual resources $c = \{r_1^{d_1}, ..., r_n^{d_n}\}$ enriching g.

The above concepts are used to define a *semantic trajectory*:

Definition 2.2.69 (semantic trajectory)
A segmented trajectory σ_0 is a sequence $\tau_0 = (< g_1, c_1 >, ..., < g_n, c_n)$, where $< g_i, c_i >$ is a pair indicating that segment g_i is enriched with contextual information c_i.

In other words, according to [111] a semantic trajectory enriches a segmented trajectory with contextual information retrieved from external datasets.

For mobility data captured from social media in particular, they similarly define a *labeled trajectory*:

Definition 2.2.70 (labeled trajectory)
A labeled trajectory is a sequence $\lambda_0 = (< g_1, l_1 >, ..., < g_n, l_n >)$ where $< g_i, l_i >$ is a pair indicating that segment g_i is enriched with a set l_i of labels.

Finally, by defining Σ as a set of segmented trajectories, Λ as a set of labeled trajectories of the trajectories in Σ, T as a set of semantic trajectories of the trajectories in Σ, Δ as a set of datasets (available on the Web) of the contextual resources of the trajectories in T, and P as an assessment function of the likelihood that a dataset d_i contains enrichments for a segmented trajectory $\sigma_0 \in \Sigma$ with respect to the labeled trajectory $\lambda_0 \in \Lambda$, the authors define the problem of finding a ranking function $rank : \Lambda \rightarrow \cup_{n=1}^{\infty} \Delta^n$ such that $rank(\lambda_0) = [d_1, ..., d_n] \implies P(\lambda_0, d_i) > P(\lambda_0, d_{i+1})$, for $i = 1, ..., n-1$.

They proceed into solving this as a supervised multi-class classification problem, assessing the computed ranking function by the Mean Average Precision (MAP) of the rankings of a set of trajectories. The enrichment process is implemented through a combination of a matching process that computes the similarity of the visited places to the entities contained in each dataset, based on their geometric distance and the Levenshtein distance of their names, and of a manual matching decision.

DART project. Within the scope of the DART project [66], Fernandez et al. explore how machine learning methods (in particular Hidden Markov Models, kernel-based distance metric clustering methods, and non-linear regression models) as well as agent-based modeling, can help Air Traffic Management systems make single aircraft trajectory predictions based on weather data and other contextual information.

The goal is to consider the demand-capacity balance in order to prevent hotspots of excessive demand of airspace use, by the surrounding traffic.

More specifically, instead of a uniform grid of $3 + k$ dimensions, where k is the number of additional enrichment parameters (e.g. local weather), the waypoints of the filed flight plans of each specific flight are used as reference points for the Hidden Markov Model states. Each of these is matched to the closest point of the medoid of the cluster that each flight is assigned to during the first phase (using the properly defined similarity metric).

T is a set of trajectories that must be executed over the airspace in a period of p time instants (e.g. hours), and S is a set of sector comprising the airspace of interest. Time is divided in intervals Δt equal to the duration of a measured occupancy period reflecting the demand, which in turn is considered to be the number of trajectories co-occurring over of a period p in the same sector.

Within the DART project, a *trajectory* is viewed as the time-evolution of the position of the aircraft's center of mass and other state variables. From a sequence of timed positions in airspace, it is then transformed into a series of sectors that each flight crosses, together with the entry and exit time for each one:

Definition 2.2.71 (flight trajectory)
A flight trajectory is a sequence $T = \{(s_1, entryTime_1, exitTime_1),$ $(s2, entryTime_2, exitTime_2),$ *...,* $(s_m, entryTime_m, exitTime_m)\}$ *where* $s_i \in S$*,* $i = 1, ...m$ *are the sectors that the flight crosses, and* $entryTime_i, exitTime_i$ *are the flight's corresponding entry and exit times to those sectors.*

Thus, Fernandez et al. adopt a grid-based representation of the airspace of interest and ignore all flight parts outside of it. For example, the trajectory shown in Figure 2.11 would be encoded as: $T = \{(s_{25}, t_1, t_2), (s_{26}, t_2, t_3), ..., (s_8, t_{11}, t_{12})\}$

Finally, trajectory semantics are not modeled anew but instead are adopted, thanks to the use of the SemT-OPTICS clustering algorithm that Pelekis et al. propose in [142] for discovering typical mobility profiles. Hence, the authors essentially follow the tag-based semantic representation of a *LifeStep* which as already seen in [142] is a network-constrained equivalent of a semantic trajectory, instead of proposing their own semantic modeling of trajectories.

STOM: Semantic Trajectory Ontology Model. In [135], Mousavi et al. introduce a new ontology-based approach so as to extract different types of human activity from GPS data. Their proposed conceptual model addresses the interpretation of movement patterns by extending the model of Spaccapietra et al. [163]. In specific, a *raw trajectory* may be divided into *semantic subtrajectories*, each composed of *stops* (the spatial part related to a stop interval) and *moves* (the maximal subsequence between two consecutive stops).

Moreover, the authors model the following domain concepts and relations:

- *semantic features* of stops: frequency, average duration, land use type, POI category type, start time

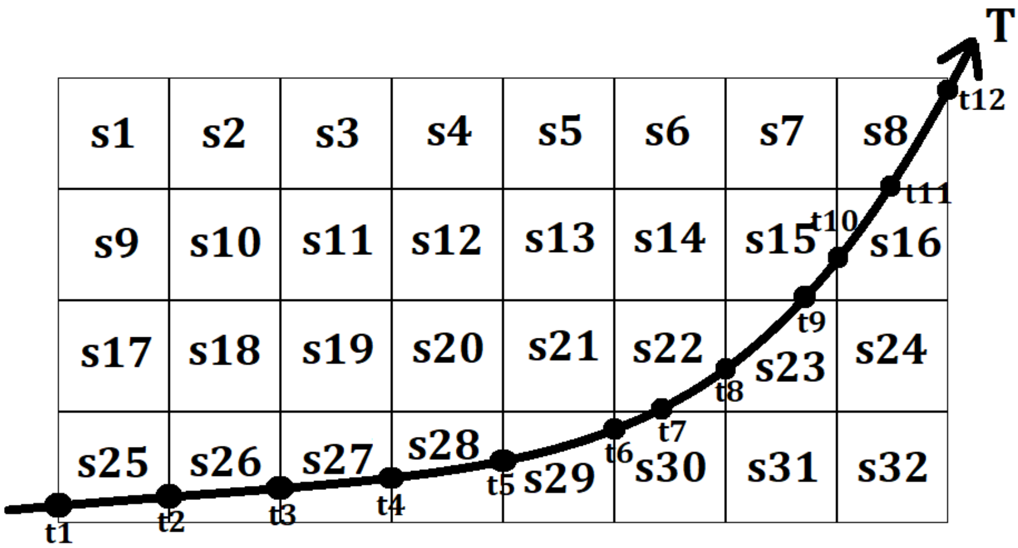


FIGURE 2.11: Example of a trajectory defined with respect to space sectors according to [66].

- *semantic place*: environmental description/information of stop locations such as land use type, PoI type
- *activity*: a movement objective belonging to one of four possible activity types.
- *activity type*: the category of activity: recreation (e.g. go to the theater), profession (e.g. working), shopping (e.g. buying food), or other (e.g. relaxing at home, cultural activities").

According to the authors, the activity coincides with the objective of the movement, while its type is associated to places where they are typically performed, but also to the time of day and the duration spent in the place. They capture this by defining it as follows:

Definition 2.2.72 (activity type)
An activity type is a function $AT = f(P, L, S_f, T_b, S_d)$ *of several features:*

- P *is the POI type that is around the stop*
- L *is the land use type where the stop has occurred*
- S_f *is the frequency of the stop in a week*
- T_b *is the start time of the stop in the place*
- S_d *is the average duration of the stop in a week*

Based on the above concepts, Mousavi et al. propose an ontology model called *STOM*, which consists of a *spatial ontology* (generic concepts about the geometric component of a trajectory), a *temporal ontology* (uses the *OwlTime* ontology

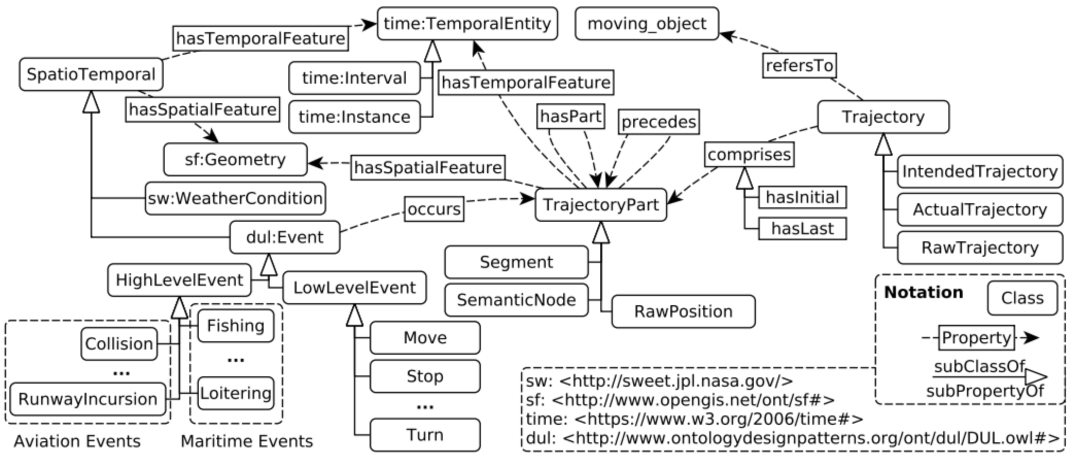


FIGURE 2.12: The concepts and relations of the datAcron ontology for semantic trajectories as proposed by Vouros et al. [177]

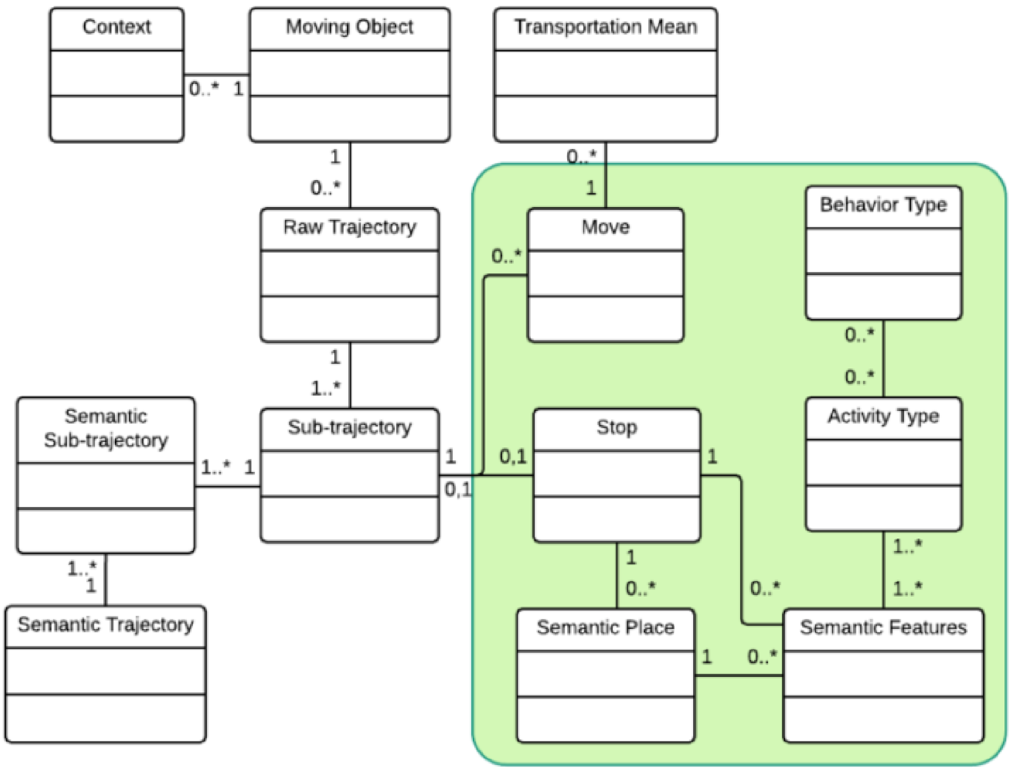


FIGURE 2.13: Extension of the conceptual model of semantic trajectories of [163] as proposed by Mousavi et al. [135]

to integrate time concepts and rules), a *geographic ontology* (describes a stop place with: land use types, road networks, POIs layer), and a *thematic ontology* (gathers a wide range of application-dependent concepts). Integrating these ontologies together (by setting up rules between them) provides the semantic description of application-relevant trajectories within each specific domain.

Geo-tagged photo trajectory mining. The line of works consisting of [21, 31–34] attempts to mine sequential trajectory patterns from geo-tagged photos.

A trajectory is defined as a sequence of geographic coordinates with time information:

Definition 2.2.73 (trajectory)

A trajectory is a sequence $T = < (x_1, y_1, t_1), (x_2, y_2, t_2), ..., (x_n, y_n, t_n) >$*, where* x_i *and* y_i*,* $1 \leq i \leq n$*, are attached geographical coordinates of a geo-tagged entity, and* t_i *is the corresponding timestamp.*

Out of these works, only [33] and [34] consider semantic trajectories and define them as follows:

Definition 2.2.74 (semantic trajectory)

A semantic trajectory is a sequence $SemT = < (SemA_0, t_0), ..., (SemA_n, t_n) >$*, where each semantic element* $SemA_i = (e_i, V_i)$ *contains a set of basic semantics* e_i *and a set of additional semantic annotations* V_i *pertaining to a RoI, and ti,* $1 \leq i \leq n$ *is the corresponding timestamp.*

In essence, in the above two works, Cai et al. represent trajectories as time ordered sequences of basic semantic spatial information with additional aspatial semantic information. To be precise, each sequence item is a spatial region annotated with contextual place type semantics, giving rise to what is called a *semantic RoI.* These RoIs are extracted through means of a grid-based mining method that considers such contextual semantics. A *multi-dimensional semantic RoI* is then defined as a semantic RoI with additional semantic features such as temporal information and weather condition. Noticeably then, this semantic trajectory modeling approach is yet another example of focusing on outdoor applications, and limiting the semantic scope of the analysis to place semantics.

Finally, two indicative trajectory examples used in [34] are the following:
$((Beach[clear], t1), (Park[rain], t2), Home[rain], t3))$
$((Beach[clear], t1'), Lake[clear], t2'), (Park[rain], t3'), (Home[rain], t4'))$

Another work which attempts to mine sequential trajectory patterns for semantic trajectories is [41]. However, as in earlier works the semantics here only concern the types of places visited by the moving object. More specifically, a semantic trajectory is defined as a sequence of pairs of PoIs together with a timestamp: $T = < (p_1, t_1), ..., (p_{l(T)}, t_{l((T)}) >$ where $l(T)$ is the length of the trajectory T. And each PoI p corresponds to a semantic category c: γ

In [186], the authors provide the outline of a moving objects database system, aimed at integrating multiple movement data models (e.g. road network models, region-based outdoor models, indoor models) paying attention to the support of semantics and multiple descriptive attributes. A data type called *mpoint* is defined for representing spatiotemporal trajectories having m attributes $A_1, ..., A_m$. Thus, a multi-attribute trajectory consists of a sequence of time-stamped locations and a set of attributes characterizing diverse aspects. The system is intended to also include a preprocessing tool for the detection and reparation of GPS data errors, a supervised-learning classifier for handling natural language queries, and a prediction model for indicating the 3D R-tree's leaf where nodes are stored. In [185], the authors investigate queries which return trajectories whose multiple attributes contain the expected values and whose locations are within a range from a query trajectory during the whole overlap period. This line of work goes to show that semantic trajectories are gradually starting to be supported at the lowest system levels.

2.2.3.2 Indoor trajectory models

Let us now take a closer look at how application-oriented works consider modeling trajectories, when the movement environment is specifically indoors. Then, by comparing with the findings of the previous section, where practically all of the works treated outdoor mobility data, the reader can be better positioned to appreciate the particular ways in which an indoor environment may affect computational mobility data analysis, so that in turn a novel modeling approach and a novel mining approach can be proposed in the next two chapters.

In [60], Elmamooz et al. propose an architecture to manage museum visitor mobility data. It consists mainly of a so-called *museum data management* component that continuously processes data incoming from and feeding other (sub)components, namely *mobility sensors* (e.g. cameras, WiFi trackers), *mobile museum guides* (i.e. a multimedia application), a *curator decision support* application (visual data analysis), and a *museum graph editor* (i.e. a modeling tool for planning the exhibitions, game tasks, room layouts, etc.). In the end, semantic trajectories are produced and fed to various trajectory mining tasks to produce mobility models, such as frequent mobility patterns and trajectory clusters.

In particular, based on the overview and classification of location models proposed in [19], the authors in [60] opt for a graph model of the museum, that represents all assets and objects (e.g. exhibits, visitors, tasks, areas, sensor locations, routes) as nodes, and all their relationships (mainly expressing containment and connectedness) as edges. This graph model is considered by the authors as the context information of the trajectories. They adopt the property graph approach for their model which allows the inclusion of multiple attributes to the nodes and edges (e.g. unique ID, name, type, subtype, description, topics, temporal availability). In the example provided, they use four different types of nodes: "PoI" (e.g. an exhibit, a service), "location" (e.g. a room), "passage" (e.g. a door), and "activity" (e.g. a task), as well as three different types of edges: "connected", "inside", and "assigned" (expressing the assignment of a task offered to the visitors to an exhibit). The authors identify the fact that the museum environment is dynamic and public, as the main challenges with respect to managing the data related to the museum, and indicate some ways to meet those challenges.

The actual visitor movements coming from the mobile sensors and the actual usage data coming from the mobile museum guide, are eventually separately modeled as so-called *mobility models*. The authors pinpoint to the adoption of existing preprocessing techniques (e.g. stay point detection, map-matching), primary route extraction, trajectory segmentation techniques, and trajectory mining tasks (e.g. trajectory clustering, stay point clustering), for producing more than one such mobility models to be used for obtaining knowledge serving the applications.

In [60], the authors focus specifically on data coming from proximity sensors attached to visitors and exhibits, however the proposed graph model of the museum can be expected to effectively support other types of raw movement data as well. In addition, it is flexible and scalable, as it is easy to add or remove nodes and their properties and relationships. It does not however account for a description of the museum at multiple spatial granularities.

Moreover, the authors partly overstress the dynamic nature of the museum environment. While indeed different parts of the graph model need to be updated with variable frequency, that frequency is almost always larger than the lifecycle of an individual trajectory. For example, a new room is not added to the museum very often, nor does an existing room change its availability so often, so as to affect evolving individual trajectories. On the other hand, such topological changes might for sure affect the aggregate movement over larger periods of time (e.g. from month to month). Thus, it is useful that the authors identify graph features that can assist in dealing with the dynamic nature of the environment as well as with data privacy concerns (due to the public nature of museums), but they do not propose any specific implementation of those features.

More importantly however, the authors do not propose any trajectory model. Even though the graph model of the museum and the consideration of stay points suggest what the main trajectory modeling direction would probably be, a lot of modeling aspects are left unspecified.

In [86], the authors propose a graph model for indoor space with the aim to improve the indoor tracking accuracy from a data management perspective. They define *base graphs* as the labeled multi-graphs resulting from applying the *Poincaré duality* mapping to a floor plan. A base graph thus incorporates the basic connectivity and accessibility information of an indoor space. More specifically, the *connectivity base graph* is defined as the triple $G_{conn} = (V, E_d, \Sigma_{door})$, where V is the set of vertices, E_d is the set of undirected edges $E_d = \{(\{v_i, v_j\}, k)|v_i, v_j \in V, k \in \Sigma_{door}\}$, Σ_{door} is a set of edge labels that represent connections (k values distinguish edges connecting the same vertices). Similarly, the *accessibility base graph* is defined as the triple $G_{acc} = (V, E, \Sigma_{door}, l_e)$, where V is the set of the vertices, E is the set of directed edges $E = \{v_i, v_j|v_i, v_j \in V, v_i \neq v_j\}$, where l_e is a function that maps edges to subsets of the set of doors $l_e : E \rightarrow \Sigma_{door}$. A base graph is enriched with geometric information as well, thanks to a mapping $BuildingPartitions : V \rightarrow Polygons$ of vertices to polygons, and to a mapping $Doors : \Sigma door \rightarrow Line\_Segments$ of edges to line segments.

The base graphs are first derived along with relevant mappings that represent the topology of the indoor space at different levels. Then, assuming RFID readers with disjoint activation ranges that are embedded in known positions of the indoor space, a deployment graph can be constructed that represents their deployment. Other deployment graphs may represent other indoor positioning technologies. For the specific case of RFID, the authors propose using RFID readers, either as *partitioning readers* (dividing the indoor space into cells such as a reader deployed by the single door of a room) or as *presence readers* (sensing the presence of RFID tags within their detection range). The resulting spatial cells correspond to the deployment graph's vertices, while the partitioning readers correspond to its edges (since by definition they signify a transition from one cell to another).

The raw movement data consist of RFID reader detections collected continuously with a certain sampling frequency. They are in the form $< readerID, tagID, t >$, where $readerID$ is the reader identifier, $tagID$ is the detected tag identifier, and t is the detection time. Time intervals during which the object is not observed by any reader are defined as *vacant time intervals*. An *off-line trajectory* is defined as a sequence of observation records of the form $< readerID, tagID, t^{\vdash}, t^{\dashv} >$, where $tagID$ is always the same (denoting a specific moving object) and $t^{\vdash}$ and $t^{\dashv}$ are the first and last time points of continuous detection of $tagID$ by $readerID$.

Such a trajectory is refined with the help of the deployment graph, by obtaining the graph element corresponding to the $readerID$ of each observation in the raw trajectory, as well as the spatial cells (in geometric form) that the object could have possibly been in during vacant time intervals. The latter are filtered using the graph topology and maximum speed constraints (based on the intersection of polygons and ellipses and having assumed circular detection regions). This so called *refinement* process is essentially a mapping, from a sequence of reader detections, to a sequence of graph elements, a more intuitive trajectory form which corresponds directly to the indoor environment.

The trajectory model proposed in [86] is primarily symbolic. Geometrical information is only used to infer the potential location of the moving object when the latter is not being detected. This showcases the difference with the trajectory models typically used in outdoor environments.

As mentioned in the future work section of [86], the model does not account for overlapping activation ranges of sensors, and has not yet been tested with multiple deployment graphs representing several positioning technologies used in parallel.

The trajectory model does not account for semantic information, apart from some trivial semantics of the indoor spaces (e.g. awareness that a certain spatial cell is a room or that a certain transition corresponds to traversing a door).

In [114], the authors analyse episodic movement data that may contain uncertainties with respect to temporal continuity, spatial accuracy, and moving object coverage. They use aggregations to tackle such uncertainties through means of visual exploration. In specific, they use a previously trained Spatial Bayesian Network to denote the conditional probabilities of visits to discrete locations and visualize them on 3D thematic maps. The Spatial Bayesian Network is used as an intermediate data structure holding only the required data instead of complete trajectories.

In the work of [114], even in their rawest form, the trajectory data are not geometric coordinate data, but instead log entries of the form: [$timestamp$], [$sensorID$], [$sha256(MAC)$], [$signalstrength$], coming from 15 Bluetooth beacons. This leads the authors to implicitly define a trajectory as a sequence of discrete visited locations which are truly regions (each region corresponding to a part of a football stadium in this specific application). This serves as an indicative example of why symbolic trajectories are generally a better fit for pedestrian mobility analysis in indoor or combined indoor/outdoor environments.

Some elements of the workflow used are of course not always relevant. For example, deciding on the number and location of the sensors to be deployed based on the application requirements, is not always possible, in case the tracking infrastructure is pre-deployed.

In [150], the authors propose an algorithmic (bootstrapping) approach for mitigating error biases and other types of noise and inaccuracies of the positioning estimates of indoor trajectories, as well as for infering an approximation of the underlying route network from the trajectory data themselves. Their algorithm produces in particular the following output: a reconstructed route network (reflecting the actual route segments taken by the moving objects and useful in detecting implicit route segments), deviation maps (taken by comparing representative trajectories - which result from aggregating all subtrajectories corresponding to a particular segment of the inferred (or given) route network into an average or median trajectory - with the location estimations of route segments), outlier-free trajectories (either by mapping each trajectory to the representatives of its segments or by checking if its geometric distance from the representative trajectory surpasses a given threshold), detected outliers in movement patterns (again by comparing the position traces collected to the representative trajectory), self-healing analysis (thanks to trajectory cleaning and outlier detection getting better as more trajectories are being collected over time).

These deliverables are produced by a processing pipeline consisting of three main modules. Firstly, the "route network reconstruction" module - as its name suggests - analyzes the noisy collected trajectories and finds the intersection points and major segments of the underlying route network, based on the "majority angle" of the trajectories in each cell of a grid covering the area over which the trajectories extend. Secondly, the "trajectory segmentation" module performs a type of "map matching" by iteratively subdividing the input trajectories into subtrajectories that connect intersection points of the extracted route network, effectively aligning each trajectory with the edges (corridors) of the route network. It takes special care for cases where a subtrajectory misses intermediate interesection points or where yet undected corridors exist between the subtrajectory's endpoints. It also filters the input data, because if two consecutive elements of the resulting subdivision of a trajectory are assigned to non-adjacent corridors, they are marked as an overly noisy part of the trajectory, which is then excluded from consideration in computing representative trajectories. Thirdly, the "computation of representative trajectories" module aggregates the subtrajectories assigned to each corridor of the route network, either by replacing them with their average trajectory (better suited for regular layouts) or with their median trajectory (better suited for curved trajectories and outlier movement detection, because it always results in a traversable representative). The proposed approach was evaluated on trajectories of hospital employees collected from smartphones based on Wi-Fi signal strength measurements.

Even though the authors do not explicitly offer a definition or model of trajectories in their work, it is apparent that they treat them as sequences of points represented by coordinates in a 2D grid cell space, where each cell has a given size σ. Over this grid representation of indoor space, they implicitly interpolate the moving object's positions between two consecutive detection points of a trajectory (trivially assuming straight lines). Thus, thanks to the grid providing a rasterization of the trajectories, they are also able to extract the (quantized) angles between two consecutive straight segments of a trajectory. This is an example of a work which does not focus on indoor trajectory modeling at all but rather implicitly adopts a simple indoor trajectory representation.

2.3 Indoor Space Modeling and Representation

This section examines how indoor environments are represented for trajectory analysis purposes. Naturally, whenever trajectory data are processed towards some analytical goal, then the spatial aspect of the environment within which the movement phenomena that produced those data took place has to be modeled either implicitly or explicitly. This may range from a trivial representation to a very elaborate model. The same applies for indoor environments in specific, despite the argument posed in section 2.2 that by far most trajectory data-based research works still target outdoor environments. Actually, this makes it even more important to consider how outdoor space is typically modeled so as to help determine a proper indoor representation.

2.3.1 Fundamental Space Representations

In order to represent movement phenomena in terms of trajectories, first a formal spatial model is needed to provide an abstraction of their physical environment. Thus, every trajectory model proposed in the literature, either explicitly or more usually implicitly, uses a certain model of location and therefore space. Similarly, an indoor trajectory model always entails some understanding of the related indoor space, because at its core lies the representation of indoor location information. At the same time, it is important to realize that modeling indoor spaces is not an effort undertaken exclusively for trajectory research purposes: moving object tracking, navigation (including routing and guidance), ubiquitous computing, ambient assisted living, crowd management, location-based systems, smart cities, spatial analysis, and many other families of applications have all been motivating factors for modeling indoor space, whether the data are structured in trajectory form or not.

Having stated that, this Thesis is interested in modeling indoor space for individual trajectory analytics purposes. In this regard, a fundamental distinction exists between quantitative and qualitative spatial representation approaches. The former are preferable when precise spatial information is important, while the latter when it is unnecessary or unavailable [39].

A qualitative spatial representation formalism, coupled with qualitative relations between spatial objects and qualitative reasoning about spatial knowledge, constitutes what is known as Qualitative Spatial Reasoning (QSR) [153]. Two of the most widespread qualitative spatial calculi are RCC (Region Connection Calculus) [43] and n-intersection [55].

RCC theory in particular, considers spatial regions as its primary spatial primitive and the reflexive and symmetric *is connected to* dyadic relation as its primitive relation [47]. Based on it, various constraint languages have been defined. For example, RCC-8 defines eight JEPD (Joint Exhaustive and Pairwise Disjoint) relations: *is disconnected from*, *is externally connected with*, *partially overlaps*, *equals*, *is a tangential proper part of* and its inverse, *is a non-tangential proper part of* and its inverse.

Alternatively, n-intersection theory is based on point-set topological theory and considers a spatial region as a 2D point set x embedded in $\mathcal{R}^2$, related to its interior, its boundary, and its exterior [56]. In particular, the 4-intersection formalism ignores the exterior, and based on the intersection combinations of the interiors and boundaries of two regions, results in eight binary topological relations: *disjoint*, *touch* (*meet*), *overlap*, *contains*, *insideOf*, *covers*, *coveredBy*, *equal* [57], equivalent to those of RCC-8.

From a more applied perspective, indoor space models can be broadly categorised into *semantic* and *spatial* models [182]. The former represent the different types of fixed (e.g. walls, rooms, doors, windows, floors, landmarks) and mobile (e.g. people, furniture, equipment) entities, their properties, and the relationships between them. The latter consist of *topological* models and *geometrical* models. The first are concerned with the various connectivity properties of space, and typically represent the building's structural aspect as a *primal/topographic* space and its connectivity aspect as a *dual/path* space. As appreciated by [182], the two types can be combined into *hybrid* models in which case "geometry adds quantification of distance and angle to the connectivity and accessibility provided by topology". It is also worth noting that despite being a review of indoor space models in general, [182] focuses solely on graph-based models, which goes to confirm the common intuition of how fitting a graph is as the basic representation component of an indoor space.

Looking more closely, most indoor spatial data models can be classified into geometric ones and symbolic ones [5]. The former focus on representing the geometry of indoor features using primitives such as points, lines, areas, and volumes. The latter focus on representing the ontological aspects of spatial units and the topological relationships between them, maintaining a more abstract view of indoor space [6]. Symbolic indoor space models in particular, are typically either set-based or graph-based (when capturing topological information). Hybrid models represent both symbolic concepts and geometric properties. Geometric and symbolic indoor space models largely correspond to the aforementioned quantitative and qualitative approaches of representing space in general, but focus on the conceptual data structures that hold the spatial information rather than on its mathematical formalism.

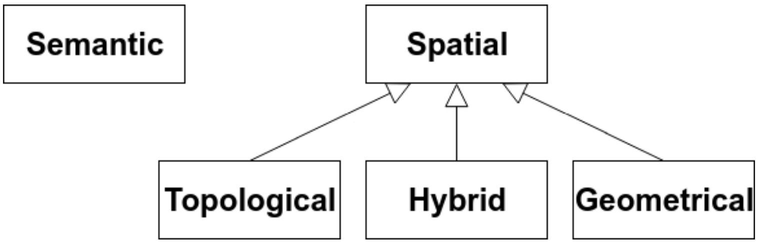


FIGURE 2.14: A basic taxonomy of indoor space models according to [182]

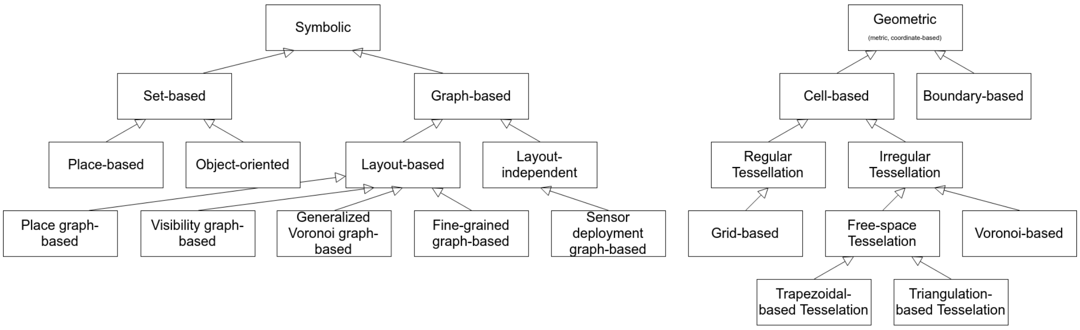


FIGURE 2.15: A detailed taxonomy of indoor space models according to [182]

2.3.2 Indoor Space Models and the IndoorGML Standard

A long line of research works on indoor space modeling ([29], [82], [107], [166], [20], etc.) has culminated into the development of IndoorGML [108, 109][3], an OGC standard aimed at representing and allowing the exchange of geoinformation that is required to build and operate indoor navigation systems. It does so by providing encoding features for indoor spatial information, modularly structured as a core data model (an application schema of GML3.2.1 [2]) and thematic extension data models. The core module considers an indoor space as a *cellular space* i.e. a set of non-overlapping cells that represent the smallest organizational/structural units of space: $S = \{c_1, c_2, ..., c_n\}$, $c_i \cap c_j = \emptyset$. Since IndoorGML is not concerned with the architectural components themselves (e.g. roofs, ceilings, walls, doors), cells represent the spaces that those components define (e.g. rooms, corridors, stairs) and where objects may actually be located in and navigate. Thus, they may share boundaries with each other but are assumed not to overlap, and as a result, an object may be at most in one cell at any given point in time.

Technically, IndoorGML describes a hybrid indoor space model since it captures the topological information of cells as well as an optional quantitative description of their spatial characteristics. The cell space and the topological relationships among

[3]For an overview of how IndoorGML was inspired by previous hybrid indoor space models in an attempt to meet the same requirements for a more general application scope, the reader is advised to refer to [91] as well as the standard's original requirements in [108].

its objects are represented by one or more *Node-Relation Graphs (NRGs).* A NRG is a graph whose nodes represent indoor space cells and whose edges represent topological relationships between any two such cells. Consequently, a NRG's nodes also correspond to potential locations of a navigating object, which is why they are also referred to as *states*. The node representing the actual cell that the object belongs to at a given point in time is called the *active state*. Similarly, edges are also called *transitions* whenever they reflect the movement of a navigating object from one cell to another. Essentially, a NRG simplifies complex spatial relationships based on graph theory concepts [107] and the Poincaré duality in particular: a cell (e.g. room) becomes a node and a cell boundary (e.g. a thin wall) becomes an edge. More precisely, an Adjacency NRG is directly derived from the *Poincaré duality* which maps k-dimensional objects in an N-dimensional primal (here a 2D/3D physical) space to (N-k)-dimensional objects in a dual space. If cell boundary semantics (e.g. doors, walls, ramps) and/or boundary constraints (e.g. door width) are also taken into account, then a connectivity and/or an accessibility NRG may in turn be derived from the Adjacency NRG. *Connectivity* suggests that there exists an opening in the common boundary of two cells. *Accessibility* additionally suggests that the opening can be crossed by the moving object.

The basic principle of IndoorGML is that any meaningful definition of space can be used to decompose it into cells. Thus, each NRG is treated as a separate graph layer. E.g. the same indoor space may be interpreted as a topographic space composed of rooms, corridors, stairs, etc., or as another topographic space defined at a different level of granularity and composed of floors, inner courts, etc., or as a WiFi/RFID sensor space composed of sensor coverage cells, or even yet as a security space composed of private offices, check-in areas, boarding areas, crew areas, etc. Moreover, there is a modeling choice to be made between a *thin/paper wall* model where walls and doors are considered to be boundaries in the primal space and thus get mapped to edges of the NRG, and a *thick wall* model where walls and doors are considered to be cells of a certain thickness in the primal space and thus get mapped to nodes of the NRG. The same distinction is made for doors as well leading to a *thin door* model nad a *thick door* model. In addition, depending on whether its nodes and edges contain any geometric properties or not, a NRG can either be a *Geometric NRG* or a *Logical NRG*. In the case of the former, IndoorGML does not define its own geometric description, but instead offers two ways of representing the spatial characteristics (i.e. form, extension, and position) of the 2D or 3D Euclidean spaces corresponding to cells: either using the data model of ISO 19197 (e.g. *c.geom* is "GM_Solid" or "GM_Point", *e.geom* is a "GM_Surface", "GM_Curve") or using simple external references to objects defined in other datasets (*c.xlink* is a $1:1$ or $n:1$ mapping from cells to objects). This mechanism of optional external references can also be used to represent indoor features bearing external domain specific semantics. Finally,

Perhaps more importantly, IndoorGML's Multi Layer Space Model (MLSM) can be used to represent spatial hierarchies. The combination of multiple layers into a single multi-layered graph is assisted by *joint edges* which connect nodes belonging to different layers. To this end, in [91] the authors define an IndoorGML hierarchical graph as a direct adaptation of the hierarchical graph definition of [167]. While the intra-layer edges represent either adjacency, connectivity, or accessibility relations between non-overlapping cells, the aforementioned joint edges are inter-layer relationships denoting the potential locations where a physical object might actually reside. Therefore, given that a physical object may be in only one active state at each layer at any given point in time, joint edges practically express all the valid active state combinations (called *overall* states) and are derived by pairwise cell intersection. Equivalently, a joint edge represents any of the eight binary topological relationships derived by the n-intersection model [55][4], except for *disjoint* and *meet*, because a physical object can not simultaneously coexist in two cells that are completely disjoint or simply touch each other. Instead, the intersection of the interior of their corresponding cells must be non-empty.

Finally, with regards to topology, IndoorGML also provides the concept of *anchor nodes* as a special node type connecting indoors to outdoors and potentially providing the necessary information for converting coordinates from a relative indoor Coordinate Reference System (CRS) to a different outdoor CRS (e.g.rotation origin point, rotation angles, rescaling factor, translation vector). Anchor nodes may actually belong to external datasets instead of being defined within IndoorGML.

2.3.3 IndoorGML Reception

In this section, review works related to IndoorGML are briefly discussed in order to gain a deeper understanding about how to meet our indoor space modeling needs, and thereby our indoor trajectory modeling needs. Studying the standard's technical specification is of course necessary, but alone does not suffice to gain a deep understanding of its function, advantages, and limitations. This is why related research works that have made use of IndoorGML or at least commented on it are also reviewed. It is important to state here that IndoorGML leaves many modeling details unspecified, and rightfully so since it is essentially a framework providing a modeling basis and only some guidelines for the rest.

In [155], Ryoo et al. study how the IndoorGML standard may be used in complement to CityGML [1] which is another OGC standard. In particular, they compare IndoorGML to CityGML Level of Detail 4 (the finest out of five levels), which describes indoor spatial objects (e.g. windows, doors, floors, furniture), as opposed to describing indoor spaces like IndoorGML. The authors identify a few requirements that CityGML fails to meet in comparison to IndoorGML: the cellular representation of space is unclear (only "Room" may represent a cell), the topological representation of space is incomplete, multiple interpretations of space are difficult to derive, references to external objects are not supported. The authors also detail the differences in modeling space closures, stairs and staircases, walls, nested rooms, cell decomposition criteria, door surface orientation, wall texture and materials. They illustrate their observations, with two use case studies: three buildings hosting shopping malls in the

[4]Resulting from the 9-intersection of two objects i.e. the emptiness or non-emptiness of the intersections of their three topologically distinct parts: interior, boundary, and exterior.

Lotte World Tower complex, and a two-level subway station in Seoul.

Similarly in [91], Kang et al. briefly explain how IndoorGML differs from CityGML and from the Industrial Foundation Classes (IFC) standard: although they all target built environments (3D buildings and indoor space), the latter are based upon feature modeling (e.g. walls, doors, slabs, windows, spaces), whereas IndoorGML is based upon space (e.g. room) modeling. Hence, IndoorGML reflects important indoor space properties such as topological ones. The authors also find IndoorGML to be complementary to the IndoorLocationGML standard [115, 184, 199] which emphasizes more the modeling of moving object locations. Furthermore, by comparing to outdoor space, they extract what they consider to be the typical characteristics of indoor space and the requirements of indoor spatial data models in general:

- non-Euclidean distance definitions considering architectural components, topological constraints, and verticality
- support of indoor space structures that are complex in terms of geometry
- network connectivity and multiple interpretations
- support for indoor cell-awareness by defining an identifiable unit of space clearly bounded by a closed geometry
- an indoor position corrective function (akin to an outdoor map-matching function)
- a mechanism of integration with the outdoor space (either through physical data transformation between standard data models, or through external references to them)

Then, the authors proceed to describe the basic concepts of IndoorGML that meet the above-specified requirements, namely the cellular space model, the optional cell geometry, the cell topology encoded as a NRG, the cell semantics encoded as a classonomy, the MLSM, and the standard's modular structure. Finally, they propose their own ways of applying IndoorGML, in particular with relation to cell granularity, subspacing determination, indoor distance computation, and indoor context-awareness services implementation.

In [52], Diakite et al. study in detail how IndoorGML's cell subdivision may be applied to support automatic subspacing, especially for fine-grained furnished 3D spaces. Using the notion of *static occupancy*, they describe objects physically occupying a fixed location in the indoor environment. Even though a cell corresponding to a room is typically classified as a "navigable-general" space, more often than not it contains a "non-navigable" part. This is why the authors claim that finer than room cells need to be considered when applying the IndoorGML standard for navigation purposes. As for automatically choosing how a cell should be subdivided, one approach suggested is to use purely *geometric criteria* (e.g. length, surface area, volume) and another is to use primarily *topological criteria* (e.g. contact with other cells) potentially enriched with geometric and/or *semantic criteria* (e.g. contact with wall cells, navigation constraints).

Moreover, since [52] is primarily concerned with the navigability of the spatial cells' interior, it addresses the question of exactly which position within the spatial cell a NRG node represents. With regards to this, they identify that the common practice in indoor navigation is to represent the centroid of the cell, but that it risks residing outside a non-convex cell. Thus, they propose to rely on a per-case combination of geometric, topological, and semantic information (e.g. specific types of furnishing elements). Finally, related to the previous issue, the authors identify a potentially problematic case: if the geometry of an edge of the NRG is interpreted as the actual path followed by the moving object when that edge is traversed, then if the positions of the two nodes (that the edge connects) are not on a navigable (e.g. there are furnishing elements, or walls between them) then the representation is wrong. This issue of edges crossing through non-navigable elements is actually closely related to the issue of defining a proper way to measure indoor distance that has been the focus of past works such as [91, 116, 123]

In [184], Xiong et al. introduce the user requirements of a GML-based Chinese national standard (under development at the time) that aims to provide an exchange format for location information in indoor routing and navigation applications. They claim that an indoor location model must support the representation of both an *absolute location* as well as a *relative location*. The former is defined as a point in geographic space measured with respect to the origin of a standard coordinate system. The latter is defined as a position measured or described with respect to another location. Having identified as a shortcoming IndoorGML's (also CityGML's) lack of indoor location information representation, the authors propose a first version of the class diagram upon which their IndoorLocationGML standard was based.

Without going into the details of their classonomy[5], the main idea is rather simple: the are two ways to define a location, either in absolute terms, or in relation to some well-defined reference location. In either case, its definition may be "specific" (i.e. geometric) or "rough" (i.e. semantic). The authors illustrate the proposed location model with a use case in which it complements a sensor model and a 3D building model that structures the building into four parts: portal, container, surface, obstacle.

Later on in [199], Zhu et al. revise IndoorLocationGML and formalize it as a standard. Thus, they present the formal definitions for the following notions:

- *indoor location*: a location of an object in enclosed indoor spaces
- *indoor absolute location*: a unitary structured description and identification of an object in an indoor space, only relevant to the spatial reference system in which it is defined
- *indoor relative location*: a structured description and identification of an object by a spatial relationship between that object and other references
- *indoor spatial reference system*: a spatial coordinate (geographic or local Cartesian) reference system of the indoor space, associated with an indoor reference

[5]Containing classes such as *Life Circle*, *Indoor Location*, *Indoor Absolute Location*, *Indoor Relative Location* , *Geometrical Absolute Location*, *Geometrical Relative Location*, *Semantic Absolute Location*, *Semantic Relative Location*, *Spatial Cell Location*, etc.

and an indoor target

- *indoor spatial relationship*: a relationship between any two or more target objects in indoor space (e.g. directional, distance, order, topological)
- *multi-dimensional location*: the information used to describe both indoor absolute and relative locations from the perspective of space, time, and semantics

In [115], Liu et al. study the interaction between the two indoor modeling standards of IndoorGML and IndoorLocationGML. They first compare their area of focus, which according to them is the following: IndoorLocationGML focuses on the definition and description of indoor locations and location-based applications, whereas IndoorGML focuses on the definition of navigation networks and representation of different space subdivisions, and pathfinding services. Then, they proceed to describe each one's main structure and most important classes. Omitting the technical details, they argue mainly that IndoorGML offers no definition of a PoI, which can instead be represented by instances of IndoorLocationGML's *IndoorAbsoluteLocation* and *IndoorRelativeLocation* classes. The authors then describe how the two standards can be used in a complementary way for indoor navigation, by enriching IndoorGML's nodes and edges with semantics encoded in IndoorLocationGML. They detect as two associations as the key to integrating the two standards. First, the association between *IndoorLocation* and *CellSpace*, according to which a *CellSpace* may contain none or many *IndoorLocation* and an *IndoorLocation* is always contained within one *CellSpace*. Secondly, the association between *IndoorLocation* and *RouteNode*, according to which a *RouteNode* may link to none or many *IndoorLocation* and an *IndoorLocation* may link to none or one *RouteNode*. As for the practical side of the integration, the authors state two alternative approaches: either the use of unified data documents according to a joint UML model, or the development of a parser to acquire data of the two different types.

However, while indeed IndoorLocationGML is better equipped to represent the location of a PoI, an IndoorGML cell is still enough to represent a PoI, because it can be arbitrarily small and contained within bigger cells thanks to the MLSM. Besides, as Kang et al. note in [91]: "...while cells have spatial extents in most cases, there are also cases where no spatial extent is necessarily required except a point".

Finally, as of late a future version 2.0 of the IndoorGML standard was announced by Diakite et al. [51] which will tackle several limitations of the previous version 1.1, including some of the issues discussed above[6]. More specifically, IndoorGML 2.0 will introduce:

- a renaming of classes: nodes and edges will no longer be called *states* and *transitions* to avoid confusion in non-navigation applications
- a simplification of the navigation module: new definitions of "TransferSpace" and "GeneralSpace" classes, deletion of "RouteSegment" and "RouteNode" classes

[6]The indoor space model of the proposed trajectory model [101] presented in chapter 3 does not consider these future changes as it was already published before [51].

- exclusion of the "thin door" model concept (can nonetheless still be handled with the help of the "NavigableBoundary" class)
- introduction of the "level" attribute in the "CellSpace" class
- redefinition of geometry: it will no longer consist of a separate class but of the new attributes "CellSpaceGeom", "CellBoundaryGeom", "Geometry", and "externalReference"
- clarification of the notion of a *layer*: a combination of primal space features, dual space features, or both via the introduction of the "ThematicLayer" class which is an aggregation of "PrimalSpaceLayer" and "DualSpaceLayer" instances and has the "semantic" and "Theme" attributes to designate the type of layer (e.g. "TOPOGRAPHIC", "SENSOR", "LOGICAL", "TAGS", "UNKNOWN", "LEGAL")

Finally, two changes are perhaps the most interesting. First, IndoorGML 2.0 will introduce a "PoI" attribute to the "CellSpace" class to allow for cell subdivisions at a suitable granularity level [95], whereas the proposed model dedicates an entire graph layer (as detailed in section 3.3.1) to do it. Secondly, the lack of a strategy to partition space will be addressed on the basis of the ability of the objects within those spaces to move or to be moved. Hence, the indoor space will be divided into three main subspaces: *object spaces (O-Spaces)* representing the spaces physically occupied by static and semi-mobile objects, *functional spaces (F-Spaces)* representing the spaces related to the function of an O-Space, and the *remaining free spaces (R-Spaces)* representing the spaces that are freely available for agents' navigation.

2.4 Chapter Conclusions

In this chapter, the fundamental as well as the state-of-the-art research works in the trajectory data modeling research domain were described. The focus was on semantic trajectories and indoor trajectories. As argued in depth throughout the chapter, modeling semantic trajectories has so far neglected the intricacies of indoor environments, whereas in terms of data mining even well-established "outdoor-inspired" semantic concepts have not been sufficiently utilized yet. Chapter 3 will build upon this awareness of the related works' limitations to provide a conceptual model for semantic indoor trajectory data at multiple levels of spatiosemantic granularity.

3

SITM: A New Model for Semantic Indoor Trajectories

Contents

3.1 Introduction 64
3.2 The Semantic Indoor Trajectory Modeling Problem 64
3.2.1 A Close Look at Current Modeling Limitations 64
3.2.2 A Broad Perspective of Current Modeling Limitations 72
3.2.3 Problem Statement 73
3.3 Semantic Indoor Trajectory Model (SITM) 73
3.3.1 The Indoor Space Model 74
3.3.1.1 Abiding by the IndoorGML standard 74
3.3.1.2 Indoor Space Modeling: Definitions, Aspects, Details 75
3.3.2 The Semantic Indoor Trajectory Model 81
3.4 Chapter Conclusion 87

3.1 Introduction

Having described among other things the state-of-the-art in Trajectory Modeling in section 2.2, the focus now shifts to its limitations and a new conceptual data model of trajectories is introduced, named *Semantic Indoor Trajectory Model (SITM)*.

In specific, *SITM* is a model of spatiotemporal indoor trajectories enriched with semantic annotations. It accounts for semantic and indoor space information and supports the design and implementation of context-aware mobility data mining and statistical analytics methods. It does so by means of a standardized (primarily symbolic) indoor space modeling framework, instead of modeling space on a 2D coordinate reference system, as is typically the case with formal trajectory models. Moreover, it integrates semantic annotations at different levels of spatiosemantic granularity, in order to allow a detailed description of movement.

More generally, motivated by what is perceived as a lack in indoor trajectory research, this Thesis combines in *SITM* different aspects of state-of-the-art semantic outdoor trajectory models, along with a semantically-enabled hierarchical symbolic representation of the indoor space which abides by OGC's IndoorGML standard [109]. Finally, this chapter also drives the discussion on modeling issues that have so far been overlooked in the literature.

3.2 The Semantic Indoor Trajectory Modeling Problem

Let us now proceed into a detailed identification of the shortcomings in the state-of-the-art semantic trajectory models. Those will then be used to formalize the proposed modeling problem.

3.2.1 A Close Look at Current Modeling Limitations

In [163], Spaccapietra et al. identify the need to model movement data from the application's perspective by structuring them into meaningful trajectories. At the time, this constituted an important breakthrough as it allows the model to avoid unnecessary computational complexity issues, i.e. repeating low-level geometric calculations for finding the characteristics of and the relations between points. It also enables the representation of semantic aspects of trajectories such as their goal-oriented nature, although it is not clarified how a trajectory goal may be represented, or calculated / extracted. For example, in the bird migration application example used to illustrate the trajectory model's usage, a bird's goal may be to "search for food".

Even though they only segment a trajectory into stops and moves, they suggest that this be done at an application-specific level. Combined with the fact that semantic annotations may be used to represent any type of semantic information, either related to the trajectory itself or to the moving object, this results in a relatively flexible and generalizable trajectory model.

In comparison, Alvares et al. [11] adopt the model in [163] but refine its definition of stops based on some temporal threshold value of stay, in order to propose a particular stop-move detection algorithm. Both works are actually targeting outdoor environments, where stops may be identified by intersections with background geometric information (i.e. places of interest). This is also made evident by referring to *trajectories* and *moving objects* as *travels* and *traveling objects* respectively. For indoor applications however, a distinction between stops and moves is often not that significant. Instead, other types of movement semantics indeed are more important such as with respect to the functional interpretation of the environment. Moreover, a symbolic (rather than geometric) trajectory representation is better suited to capture the indoor space semantics.

The model of Bogorny et al. [23] succeeds in representing the trajectory patterns at multiple granularity levels thanks to the methods already proposed previously in [24]. These methods discretize the temporal dimension either in several predefined granularities (e.g. month) or in user defined intervals (e.g. 14:00-18:00), and the spatiosemantic dimension either in two predefined granularities (namely *feature instance* and *feature type*) or in user defined intermediate granularity levels (namely the class *GenericSpatialFeatureType*). Thus, the model proposed by the authors is flexible with respect to the representation of space and time and suffices to perform certain pattern mining tasks over trajectory data.

However, it adopts the limitations of the model of Spaccapietra et al. [163] due to its being oriented towards outdoor settings. Most importantly, the semantic aspects of the mining process are exhausted in the names and types of the geographic locations of interest and do not support associating trajectory patterns with other types of contextual information (e.g. the type of the traveling object). Topological or geographical relationships between the various interesting places that constitute the items of the patterns are also not possible. Instead, the standardized spatial features adopted by the model of [163] may be used to define the geometry of stops and moves, but this information can only be leveraged for processing the results of the mining algorithm, not in the main mining process.

The conceptual framework proposed by Andrienko et al. in [14, 15] exhaustively describes and categorizes the types of information that may be represented by movement data. In this sense, it offers a deeper theoretical foundation than all other trajectory data models reviewed in the previous chapter. By effectively breaking movement down to its most essential elements: *space*, *time*, and *objects*, it enables the sound definition of *events* and relations between objects and events. This in turn allows the definition of higher level concepts (e.g. trajectories, spatiotemporal context of movement) in a way that captures the interactions between them.

Typically, in the literature the term *event* is loosely used to describe external phenomena that may affect the movement. In [14, 15] however, an event is precisely defined as any object having a particular position in time and hence a lifecycle, while a spatial event is simply defined as a spatially positioned event. As a result, any trajectory or any part of a trajectory is actually a spatial event (consisting of a sequence of elementary spatial events).

Thanks to this representation, relations between events in the movement data (e.g. visiting a particular place) and events in the context data (e.g. closing down of particular subway station due to a strike) may themselves be defined as spatial events, without resorting to ad-hoc modeling approaches. Relations between moving objects can also be similarly described. Even individual occurrences of such relations are spatial events. Spatial events can be even used to represent activities of objects but this idea is not explored.

The aforementioned representation sets the ground for the extraction of interesting events from trajectories, and the subsequent use of those events as independent objects. In other words, the framework proposed by Andrienko et al. offers a way of using movement data analysis results for performing further analysis.

More broadly, different types of spatial events may be used to reflect different interpretations of movement from the application viewpoint. In this sense, the framework constitutes a generalization of the conceptual model of Spaccapietra et al. [163]. However, the authors do not offer any direction on how to proceed with this type of semantic modeling.

Related to this, dynamic thematic attributes are defined to represent any attribute available in the movement data or *"any other existing or conceivable thing"*, including semantic pieces of information. This effectively turns a thematic attribute into the equivalent of a semantic annotation in the model of [163]. But apart from proposing this conceptual mechanism, the authors do not address at all the issue of how to implement it.

From the examples used to illustrate the usefulness of the framework, it is apparent that the model of Andrienko et al. primarily targets outdoor environments where movement data originally consist of geometric coordinates. Characteristically, it is stated in [14] that *"positions in movement data are most often specified by coordinates"*, and in [15] that *"most often, movement data have the form (object identifier, time reference, spatial coordinates, attribute values). Movement data available in other forms can be transformed to this form"*. Hence, while the framework does not prohibit the use of other types, it is oriented towards coordinate positional data.

The conceptual model of semantic trajectories proposed by Yan et al. in [187] and the very similar one proposed in [188, 190] are more flexible than the model in [163], because they support heterogeneous semantics, thanks to the concept of *episodes* functioning as a generalization of stops and moves.

In addition, their model allows the semantics of movement to be captured even at the level of spatiotemporal positions via unrestricted annotations of any type. However, annotating individual positions is not an efficient way to capture even slightly abstract semantics. It is instead more useful to model high-level trajectory semantics with equally high-level annotations, each encompassing multiple position records.

Apart from enabling different interpretations of the same movement, another upside of an episodic interpretation of trajectories is that it enables trajectory data compression, because individual GPS records may be replaced by episodes. For example, as stated by Yan et al. in [190], an episode can be efficiently stored as a tuple

$$e_i = (time_{from}, time_{to}, bounding_{rectangle}, center).$$

Storing such episodic intervals is much less memory costly than storing all timestamped positions. In addition, unrestricted episode annotations even allow multiple semantic episodes

$$se_i = (da, sp_i, t_{in}^{sp_i}, t_{out}^{sp_i}, tagList)$$

to refer to the same semantic place sp_i, which again is less costly than storing repeating the same semantic place tags. The downside is of course that all of the semantic modeling effort falls upon the shoulders of the model adopter, because the model offers no predefined (domain-specific or generic movement-related) semantic specifications or restrictions. For example, it is not specified whether one or more trajectory goals are required, optional, or not supported at all.

Moreover, while in [187] it is left unspecified, in [188, 190] the episodes of a semantic trajectory are specifically assumed to be non-overlapping, without any reasoning in support of this assumption: only mutually exclusive episode predicates are considered (i.e. $[P_1] + [P_2] + ... + [P_n] = 1$ where $[P] = 1$ if $P = true$ and $[P] = 0$ if $[P] = false$). Perhaps this can be attributed to a particular focus on identifying stops and moves, which are normally non-overlapping.

In the more general case however, two meaningful trajectory subsequences - each satisfying a different predicate - may share a common part. For example, if an episode reflects a specific movement goal and multiple movement goals are modeled in parallel, then it could be the case that two episodes overlap. A trajectory model should ideally support segmentation strategies that allow for such overlaps to occur. The implication of this on the results of the extraction of sequential movement patterns is also discussed in chapter 6.

The conceptual semantic trajectory model proposed by Spaccapietra and Parent in [162] and refined in [139], adopts from the model of Yan et al. in [187] its main structure as a sequence of (potentially annotated) timestamped coordinate positions or episodes. However, it extends it in the following ways:

- Firstly, it more precisely defines and groups trajectory behaviors, thus enabling new types of analyses.
- Secondly, it accounts for purposefully missing data in the trajectory's trace (thanks to *semanticGaps*). Distinguishing between accidental and deliberate (hence meaningful) missing movement information can be useful for dealing with data quality and uncertainty issues[1].
- Thirdly, it accounts for the representation of additional types of raw data (e.g. speed) besides coordinate positions (thanks to δ).
- Lastly, it encodes multiple different episode segmentations within the same copy of the trajectory (thanks to *segmentations*).

By allowing multiple alternative interpretations of the same trajectory and implementing them via unrestricted semantic annotations, the model of Spaccapietra and Parent significantly generalizes the stop-move interpretation of [163]. Also, it represents the semantics of movement at three different levels of granularity (i.e. points,

[1]For example, when a moving object being tracked regularly skips detections (e.g. active tracking such as badge scanning systems), or when a moving object often enters areas not covered by the tracking hardware.

episodes, and whole trajectories). However, even though an episode may group together any number of positions, this does not mean that it can correspond to any level of granularity desired, because episodes are defined according to [187] as *maximal* subsequences whose spatio-temporal positions all comply with a given predicate.

In addition, the distinction between *individual* and *collective* trajectory behaviors constitutes a first step towards modeling phenomena that involve multiple simultaneously moving objects (e.g. visitor group interactions). In particular, the fact that a trajectory behavior can be defined with respect to another trajectory behavior (e.g. "Leadership" requiring the presence of "Flock") can enable a hierarchy of trajectory behaviors, with "macro-behaviors" building upon "micro-behaviors".

The main technical limitation of the conceptual model of [139] is that the spatial dimension of the trajectories is purely geometrical: a spatiotemporal position is actually a timestamped point which consists of latitude and longitude coordinates, and a *trace* (called *track* in [162]) is simply a list of such positions. On the contrary, an indoor trajectory model should primarily support symbolic trajectories, where location information is encoded as a symbol, instead of a coordinate tuple. Therefore, even though it would be possible to derive symbolic spatial entities after a first step of (light) semantic enrichment, it is preferable to first symbolically model indoor space at the application level, and then define trajectories directly over this space model, effectively bypassing the coordinate representation altogether.

This limitation also concerns the models of [162, 187, 188, 190], all oriented towards handling GPS data collected in a coordinate form, and in contrast to raw indoor movement tracks which are often collected in a symbolic form (e.g. for compression purposes). Hence, these models simply can not support certain types of movement data (e.g. proximity sensor readings) since they either do not readily offer any higher-level construct to function as a spatiotemporal unit, other than the timestamped coordinate position, or if they do, they still require coordinate data to calculate its values.

Another modeling aspect of the works of [139, 162, 187, 188, 190], that may prove to be limiting under certain advanced analysis scenarios, is that the trajectory model allows for the addition of semantic information strictly with respect to the movement itself (via the mechanism of annotations of the trajectory itself and/or of its parts i.e. positions and episodes). However, as identified by [143] and described in the conceptual framework of [14, 15], there are three fundamental sets pertinent to movement, representing the "where" (set of locations), "when" (set of instants or intervals), and "what" (set of objects) of spatiotemporal data. This is true across applications.

For some annotation-based models such as the one of [139], the distinction between semantics of *time*, semantics of *places*, semantics of *moving objects*, and semantics of *movement* itself, can be accounted for by using the annotations as references to external application-specific objects. To put it more simply, the trajectory model itself does not distinguish between the different types of semantics but can be extended to do so.

Ideally however, semantic information should be captured in a way that empowers the trajectory model not only to distinguish between these four entities, but also to capture their interplay. In this respect, it would be useful to complement annotations with other mechanisms, in an effort to capture the interactions of the fundamental types of trajectory semantics. For example, if the semantics of indoor spaces remain the same throughout the lifetime of a trajectory, then it is probably not worth capturing them as yet another annotation value set.

Finally, as in the models of [139, 162, 187, 188, 190], also in [25] trajectories are composed of 2D-coordinate timestamped points $p = (x, y, t)$, due to the model being oriented towards handling GPS data. In indoor environments however, symbolic positions are preferred and the raw movement track is not always provided in coordinate or even geometric form. The model proposes an optional classification of semantic points into geometrically described places but only as part of the enrichment process. As previously explained, the opposite approach is actually preferable for indoor trajectories: symbolic places should be the de facto position representation, and only optionally mapped to their corresponding geometries and coordinate systems. Also inspired from outdoor long-distance trajectories, is the importance given by the model to the transportation means: a semantic subtrajectory can be defined purely based on its transportation means. On the contrary, transportation means in indoor environments are typically restricted to a static distinction of the moving objects (e.g. people with or without walking disabilities) and therefore do not constitute an important part of the model.

The conceptual model of Bogorny et al. [25] is inspired by the model of Parent et al. [139], but specifies more precisely the supported types of trajectory semantics, by organizing the contextual and semantic information into concepts and relations between concepts. The proposed conceptual schema is still flexible enough to allow the user to select the types of information with which to enrich a trajectory, as well as the granularity level of the enrichment (i.e. trajectory, subtrajectory, or point levels).

The semantic trajectory definition proposed by Bogorny et al. is stricter than the one in [139]:

- Firstly, semantic annotations are required to be either goals, or transportation means, or behaviors, or place names, or environmental information, as opposed to any type in [139].

- Secondly, a semantic trajectory needs to have exactly one goal and at least one semantic subtrajectory, as opposed to any number and any type of semantic annotations and meaningful segmentations in [139].

- Thirdly, some basic application-independent attributes of the contextual objects are already predefined (e.g. point-level “speed” and “acceleration”), whereas in [139] they are supported but not predefined.

Although, some of the additional modeling concepts are both interesting and useful (e.g. distinction between “aware” and “non aware” behaviors) and some modeling choices are sound (e.g. separate classes representing objects, events, places, and environments), the model of Bogny et al. also makes a few unnecessarily restrictive assumptions.

For example, the dependence of a semantic trajectory upon the existence of its semantic subtrajectories, and in turn the dependence of a semantic subtrajectory upon the existence of its semantic points, reflects a bottom-up structuring of the trajectory semantics, consistent with applications trying to enrich GPS datasets. It is not however necessarily preferable for other trajectory data mining tasks or for indoor trajectories where all spatial entities are already predefined. For example, it can be argued that a trajectory having an explicit goal should qualify as being semantic, even if it does not include any semantic subtrajectory at all.

In [77], Güting et al. study annotated trajectory databases. They consider a symbolic trajectory to be a sequence of pairs, called *units* and consisting of a time interval and a label. An important quality of the model they propose is that it provides a consistent framework for querying both symbolic and geometric trajectories. An advantage in comparison to other models is that, its proposed abstract data types are integrated into an existing framework of data types for moving objects (introduced in [76]). As a result, the trajectory model inherits that framework's generic operations.

The modeling framework of Güting et al. is a time interval-based one which fits perfectly the intuition of symbolic indoor trajectories. However, it is oriented towards lower-level modeling elements, offering a pattern matching language and query operations, since it aims to improve storage space and response time in comparison to raw geometric trajectory representations. Thus, it is not of primary interest to this Thesis.

The semantic trajectory model used by Beber et al. in [17, 18] once again showcases that, for outdoor trajectory mining and analysis applications the focus is typically on stop episodes. The authors opt to ignore most of the semantics represented in the similar model of Bogorny et al. [25]: environments, events, goals, transportation means, behaviors (in the more general sense than mere activities). In this way, they avoid some of its restrictive requirements such as the whole trajectory always having a general goal. Instead their semantic trajectory model only represents stops corresponding to particular PoIs, each mapped to a specific 2D-coordinate pair.

At the same time however, Beber et al. do extend the model of [25], first with respect to the semantics of places with the notion of *PoI profiles*, and also with respect to collective activities with the notions of *encounters* and moving object *relationship degrees*. Other works typically ignore group activities or simply assume that groups are known in advance.

This goes to show that it is difficult to strive a fine balance when introducing semantics to a general scope conceptual trajectory model, because each application task may put the emphasis on different semantic aspects of the movement. As a general rule, indoor trajectory applications are more often and more deeply interested in space semantics than their outdoor counterparts.

Finally, the definition of *sub-stops* as additional stops corresponding to a particular stop and its PoI, is an unsophisticated - yet effective - way to account for the representation of trajectories at multiple levels of spatiotemporal granularity. It allows the identification of multiple finer activities happening during a single stop, in an otherwise limited (in terms of expressiveness) model.

In [111], Leme et al. are interested only in finding the most relevant datasets (through classification algorithms) for the semantic enrichment of trajectories, based on the types of places visited. As a result, similarly to the works of [17, 18], the expressiveness offered by existing conceptual models of semantic trajectories is not needed. Instead, the authors rely on a simple stop-move heuristic and a simple categorization of the "check-in" places used for labeling them within the trajectories.

While Ferrero et al. do not specifically define a semantic trajectory, their work of [67] outlines how to do it in a way that enables the discovery of new and more complex types of patterns. Examples envisaged include the discovery of patterns of moving object relationships, group awareness, individual-group movement influence. Indeed, it is the first work in our knowledge to point to this direction, although it targets the broader trajectory data mining landscape (including trajectory segmentation, feature extraction, similarity measures, etc.), whereas this Thesis is specifically focused on trajectory pattern mining.

Despite the fact that the resulting MASTER model introduced in [133] targets only geometric trajectory data, it is actually compatible with the model proposed in this chapter, because it focuses on the modeling of the relationships between moving objects (e.g. friendship, professional, family) and a handful of other concepts (e.g. *events*) that are left largely unspecified in the approach of this Thesis. At the same time, MASTER does not consider data at multiple levels of granularity and would therefore most fittingly be implemented in combination with or even parallel to a multigranular trajectory model.

The model proposed by Cruz in [45] is arguably an interval-based one, since a transitory structure, namely a timeslice, constitutes its most basic component. Unfortunately, the model does not scale well when representing data of fine granularity. The latter seems to be partly due to the fact that a relationship (*meets*) and a property (*hasFiliation*) are needed to specify the temporal relationship between even two consecutive timeslices. It would instead be simpler to structure timeslices in a fixed manner per identity value, and order them (either all together or per moving object) according to their starting time.

Timeslices are better suited for trajectories of changing objects, but this Thesis is only interested in trajectories of well established individual moving objects whose identity does not change over time. Thus, representing the parent-child relationships between timeslices seems to be a cumbersome representation not yielding any practical returns. Another limitation of the model is that it does not account for movement uncertainty. A separate ontology could be developed to address this, but this idea is neither pursued nor suggested.

As with many semantic trajectory modeling efforts, especially those tackling the problem of location-based service user activity inference, Mousavi et al. [135] introduce a trajectory model focused on outdoor trips. This again leads to the logical representation of a trajectory as a set of stops and moves. Given their focus on activity recognition from GPS data, the authors succeed in offering a balanced level of semantic specification, so as to make their formalism useful for any given application case related to the aforementioned task.

The ontology model is wisely composed of four separate ontologies: time, place, stops, and activity. This actually resembles a refinement of the proposal in [189] which consists of three different ontologies, namely a *Geometric Trajectory Ontology* for the basic concepts of the spatiotemporal aspect of a trajectory (e.g. temporal points, areas, lines), a *Geography Ontology* for describing domain-specific natural and artificial features of interest, and an *Application Domain Ontology* for higher level concepts related to specific domains.

Here however, the temporal ontology is concerned with time in qualitative terms (e.g. morning, evening). More generally, capturing the temporal dimension in such discretized semantic form can be useful but does not have to fully replace a metric approach. In addition, the temporal ontology does not include crucial temporal features of stops (e.g. average duration, begin time). It would be interesting for the trajectory model to offer some interconnectedness between the two, perhaps via a common representation as intervals which would also allow their superimposition.

3.2.2 A Broad Perspective of Current Modeling Limitations

In the earlier semantic trajectory modeling literature, semantics were largely exhausted in the names and types of the geographic places of interest related to the moving object's physical stops. Whereas other types of contextual information, or relations between places, were rarely taken into account.

In general, the relevant context information is acquired from three different types of sources: users, sensors, and inference systems [158]. As covered in detail in section 2.2.1.4, efforts have since been undertaken to integrate movement ontologies, linked open data, information extracted from social network platforms, or complementary case-specific datasets, with spatiotemporal trajectory data. Even the basic concept of (trajectory) episodes can be viewed as a generalization from stop-move segments to more diverse and heterogeneous semantics.

However, as more and more types of semantics become relevant given the increasing interest in context-aware location based services and applications (e.g. context-aware museum guidance [105]), and as Big Trajectory Data are characterized more and more by their variety and not just volume, semantically rich models become all the more appreciated. Unfortunately, targeted trajectory semantics have so far largely concerned outdoor contexts, as often made evident by the terminology (e.g. *traveling objects* [163]) and definitions used.

On the contrary, a model for semantic trajectories in indoor environments needs to at least consider the building's topology and space semantics. The interior of buildings is typically divided into clearly delimited spatial entities such as rooms, halls, corridors, floors. This physical segmentation already holds a considerable amount of semantic information that a typical outdoors-inspired trajectory model does not capture.

Furthermore, for most semantic trajectory models, the sole spatial primitive is a 2D coordinate position relative to the GPS's or to the specific application's coordinate reference system. In contrast, raw indoor movement tracks are often collected in symbolic form, either due to indoor positioning technologies being better suited for compartmentalized tracking (e.g. proximity sensor readings), or due to pressing data compression needs. The latter is particularly important in the context of Big

Trajectory Data, because the indoor topology can be used to reduce the massive storage needs, in a way much similar to how a road network [149] can compress a vehicle trajectory dataset.

In any case, a model representing movement in symbolic terms has the advantage of being independent from the positioning technology used for collecting the movement data [151]. At the same time, knowing in advance the spatial entities that a moving object may find itself in (e.g. a list of rooms) makes encoding them as symbols conceptually and computationally more practical. Therefore, symbolic and hybrid indoor space models become an attractive building block for modeling indoor trajectories.

3.2.3 Problem Statement

With the above points in mind and within the scope of a broad range of real-world applications (e.g. airports, exhibition spaces, museums, etc.), the modeling problem can be reduced into finding a trajectory data representation which consistently:

1. captures the meaningful aspects of indoor movement;
2. captures the intricate effects of the indoor environment upon movement;
3. remains application domain-independent.

More particularly, there is a need to design a trajectory model which supports spatiotemporal types of analysis and semantics-based types of analysis, at multiple levels of spatiosemantic granularity, for multiple moving objects, and at the same time account for trajectory data quality and uncertainty issues.

3.3 Semantic Indoor Trajectory Model (SITM)

In this section, a new model is defined for semantic trajectories in indoor environments, named Semantic Indoor Trajectory Model ($SITM$), aimed at supporting:

- all types of indoor settings;
- different types of semantics;
- both human and inanimate moving objects;
- mining and analysis applications using statistical and reasoning approaches (applied both at the individual and collective trajectory level).

$SITM$ consists of, a semantically enriched representation of indoor space, and a semantically enriched sequence representing an individual moving object's spatiotemporal presence.

The proposed indoor space representation is a layered multigraph. Its nodes symbolically represent indoor spatial regions, and its edges represent topological relationship information between those regions. Static semantic information about the regions is represented through node classes and attributes as well the grouping of nodes and edges into layers. The proposed representation is compatible with OGC's IndoorGML standard and can be viewed as an extension of it. It is described in section 3.3.1.

The proposed representation of an individual moving object's trajectory is a couple consisting of, a trace of consecutive presence intervals inside the indoor regions represented by the graph's nodes, and a set of semantic annotations describing the trajectory in its entirety. It uses the aforementioned indoor space representation and is described in section 3.3.2.

3.3.1 The Indoor Space Model

As argued in section 2.3, any type of trajectory model has to somehow represent location information, and consequently the spatial environment where movement takes place. For indoor spaces in particular, symbolic representations are generally preferred for reasons already discussed in chapter 2. Set-based ones are of course simpler than graph-based ones, but lack the means to represent space connectivity, a feature that is too important to ignore as far as trajectories go.

3.3.1.1 Abiding by the IndoorGML standard

As explained in section 2.3.2, the IndoorGML standard [109] serves to represent indoor space as a set of cells. Further information can of course be represented thanks to the graph structure (topological information), the ISO19197 spatial features and the external references (optional geometric information), as well as the classification of cells and the Multi-Layered Space Model (semantic information).

Hence, IndoorGML offers a flexible indoor space model that combines topological, geometric, and semantic information. In particular thanks to the graph-based representation of the spatial entities, it offers a computationally efficient way to not only model indoor navigation and routing problems, but also implement innovative indoor trajectory pattern mining methods.

In specific, based on the modeling framework provided by IndoorGML and its Multi-Layered Space Model (MLSM), a 2D multiple floor (i.e. 2.5D) indoor space is defined as follows:

Definition 3.3.1 (2D multiple floor indoor space)

A 2D multiple floor indoor space is represented as a layered multigraph $G = (V, E)$ with $m + 1$ layers where

- $V = \bigcup_{i=0}^{m} V_i$
- $E = E^{top} \cup \bigcup_{i=0}^{m} E_i^{acc}$
- *Each layer constitutes an accessibility Node-Relation Graph (NRG) $G_i = (V_i, E_i^{acc})$, $0 \leq i \leq m$, where E^{top} represents binary topological relationships between two cells of different layers*

- *The graph G is composed of $m+1$ different layers of nodes and edges, each comprising a NRG G_i and corresponding to a different decomposition of the indoor space.*

On the one hand, node $v \in V_i$ represents a cell belonging to the i-th layer and an edge $e \in E_i^{acc} \subseteq V_i \times V_i$ represents the accessibility between two cells of the i-th layer. On the other hand, a joint edge $e' \in E^{top} \subseteq V_i \times V_j$ represents a binary topological relationship between two cells of different layers ($i \neq j$).

Figure 3.1 illustrates an example of such an indoor space graph representation consisting of five hierarchical layers: *Region of Interest*, *Room*, *Floor*, *Building*, *Building Complex*, which will be explained in greater detail in the following paragraphs.

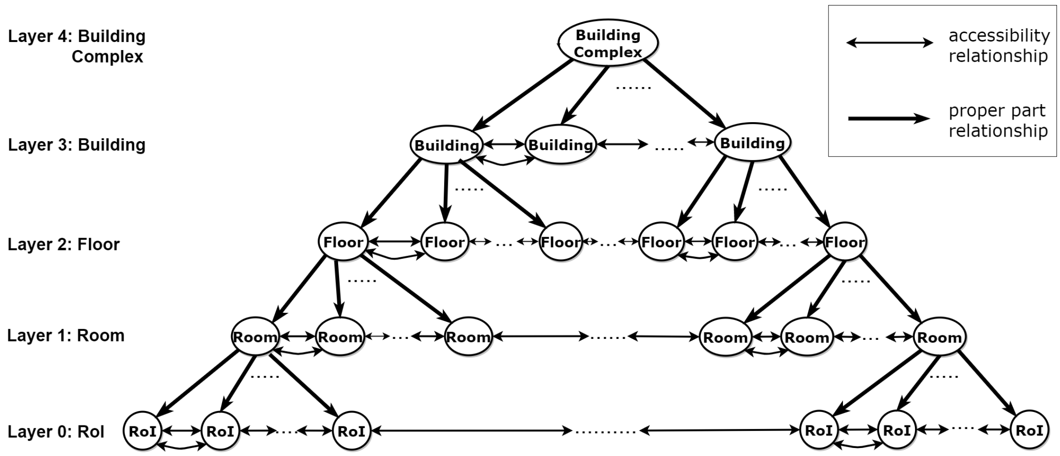


FIGURE 3.1: A 2D multiple floor hierarchical indoor space representation

3.3.1.2 Indoor Space Modeling: Definitions, Aspects, Details

Despite the IndoorGML standard offering a core framework for modeling indoor space in a meaningful way, it is still necessary to clarify, restrict, or adapt certain parts of it. More generally, there exist important modeling details or issues that have never been addressed before in the literature. Thus, along with presenting the indoor space model, this section also examines various modeling choices.

First, let us define a layer hierarchy as $k+1$ ordered layers G_i ($0 \leq i \leq k$, $k \geq 2$) of G, only consecutively connected by joint edges. The hierarchical structure of an indoor space is often perceived based on its function/usage and is thus not necessarily the same as its architectural structure. On this matter, Diakite et al. [52] propose a categorization of specific criteria to automate the subspacing procedure:

- geometry-driven criteria (e.g. split if some cell dimension surpasses a certain value),
- topology-driven criteria (e.g. split depending on which cells a cell is connected to before/after the split),

- semantics-driven criteria (e.g. split depending on what type of cells a cell is connected to),
- navigation-driven criteria (e.g. split if a cell has both walkable and non-walkable parts).

Despite the fact that [52] focuses on furnished 3D indoor spaces, this categorization can also be used as a guideline for coming up with 2D space partitioning criteria, despite the implementation differing considerably from case to case. However, a specific splitting strategy is missing. Unfortunately, hierarchical subspacing based on loose guidelines requires a considerable amount of modeling effort, because there are many possible ways to structure the multi-layered graph G, some more complicated than others.

The straightforward modeling approach, followed here, is to define specific levels of spatiosemantic granularity with respect to the main architectural elements of indoor environments, and devote a separate layer to each one. Then, additional layers reflecting purely semantic interpretations (e.g. usage) of space can also be added according to application needs. These may or may not be part of the initial spatial hierarchy. Semantic layers can actually relate to a basic architectural hierarchy, or be completely independent of them, even form their own hierarchies (e.g. an ontology or a taxonomy) in parallel to the topographic one.

The alternative approach would be to define fewer layers of *mixed* (i.e. non-hierarchical) spatial granularity in order to satisfy a specific application's requirements. This would lead to a more compact ad-hoc space model where for example rooms and artworks may be represented in the same layer, whereas buildings and floors in a different one. In the limit case where an application is only relevant to a specific level of granularity (e.g. room level), it suffices to model indoor space with a single NRG.

However, since the model's goal is to support trajectory modeling for a variety of application cases, a hierarchical approach is preferred. Nevertheless, the multigraph G does not have to be "entirely" hierarchical, meaning that not all of its m layers need to belong to the same hierarchy.

More specifically, the space model requires a certain 3-layer core hierarchy to account for the fact that virtually any indoor environment consists of: a *Building* layer, a *Floor* layer, and a *Room* layer.

Therefore, G must include 3 layers representing static hierarchical levels of spatiosemantic granularity. Other layers are of course optional and may also integrate with this core layer hierarchy, in which case $k > 2$.

As mentioned, layer hierarchies comprise either topographic layers, or semantic layers, or both. It is evident that the proposed 3-layer core hierarchy is basically a topographic one. The *Building* and *Floor* layers are spatially defined, since the architectural structure alone is mostly enough to determine which space constitutes a building and which space constitutes a floor. The *Room* layer is also predicated spatially, but in a looser way since it may on occasion contain cells whose boundaries are not necessarily physical (e.g. functionally independent subspaces of a big hall or of a great room). Hence, it may actually contain any type of room-level navigable spatial cell: rooms, chambers, halls, lobbies, cellars, terraces, corridors, hallways, big staircases, etc.

Furthermore, the 3-layer core hierarchy makes the model generalizable to different tracking technologies and infrastructures, an important enabler for the fusion of heterogeneous Big Trajectory Data. In addition, it simplifies the indoor space model as it follows common intuition about the building blocks of indoor space.

This is especially important when more than one hierarchical interpretations are needed, for example when the usage of the spaces has to be added to the model. The reason is that each usage layer (e.g. sensor layer) might need to be related to multiple layers of the architectural structure (e.g. a sensor might cover a part of a room while another sensor might cover a whole room plus a part of the next room). Specifically, for N architectural structure layers and M usage layers, there might be MxN different groups of inter-layer connections, without accounting for the possibility of having relationships amongst different usage layers.

In addition to the core hierarchy, two optional layers are proposed for typical cases, as shown in Figure 3.1: a *Building Complex* root layer and a *Region of Interest (RoI)* leaf layer. The *Building Complex* layer is defined to represent the indoor space of a site comprised of multiple buildings, such as a hospital spanning multiple attached wings or a university campus spanning multiple independent edifices. The *RoI* layer is defined to represent navigable sub-room level spatial cells of application-specific interest, such as *"you-are-here"* map installations in a shopping mall or individual exhibit displays in a museum.

The *Building Complex* and *RoI* layers are only relevant per case, and can be properly integrated into the core layer hierarchy: *Building Complex* → *Building* → *Floor* → *Room* → *RoI*. In such case, a *Floor* object describes a single building's floor level and not that floor level in general (e.g. $FloorA1 \neq FloorB1$ for two different buildings $BuildingA$ and $BuildingB$).

Concerning G's edges, intra-layer edges and inter-layer edges are always of different type, and therefore G can be considered as an edge-coloured multigraph which can be mapped to a multilayer network [97]. Identifying representation equivalencies like the aforementioned one is important, as it allows trajectory data modelers and analysts to tap into methodologies already developed in the respective research fields (e.g. Graph Theory).

With regards to the interpretation of the graph's joint edges, the IndoorGML standard defines them as representing the *contains*, *within*, *overlaps*, *crosses*, *equals*, and *intersects* topological relationships, illustrated in Figure 3.2. However, it is crucial to specify restrictions imposed over these relationship types depending on the goals of the model and its desired functionality.

More specifically, in the proposed implementation of the indoor space hierarchy, the *overlaps*, *crosses*, *intersects*, and *within* relationships are excluded from the set of possible joint edge interpretations, as far as hierarchies go. This leads to a simplified top down hierarchical model which offers some advantages for analytical operations as will be explained later in this section. However, all types of *intersects* relationships are still allowed for relating "in parallel" the core 3-layer (or the extended 5-layer) structural hierarchy to one or more semantic layers.

The *equals* relationship could also be excluded to prohibit node repetition and instead enforce a proper hierarchy, in other words to ensure that going from one layer to the next, every node will reflect a strictly finer spatial entity than its parent node.

Related to this, since the indoor space model adopts IndoorGML's implicit assumption that each node belongs to a single layer ($\bigcap_{i=0}^{m} V_i = \emptyset$), it follows that if a node is relevant to multiple layers, then essentially it has to be replicated in each one. Effectively, this means that all of its consecutive copies have to be connected via *equals* joint edges. This is only needed in exceptional use cases where repeating the same node in different layers both makes sense and serves a purpose (e.g. particularly large rooms, single floor buildings).

Finally, Figure 3.2 illustrates common topological relationships between surfaces which may sometimes be mixed due to different naming conventions.

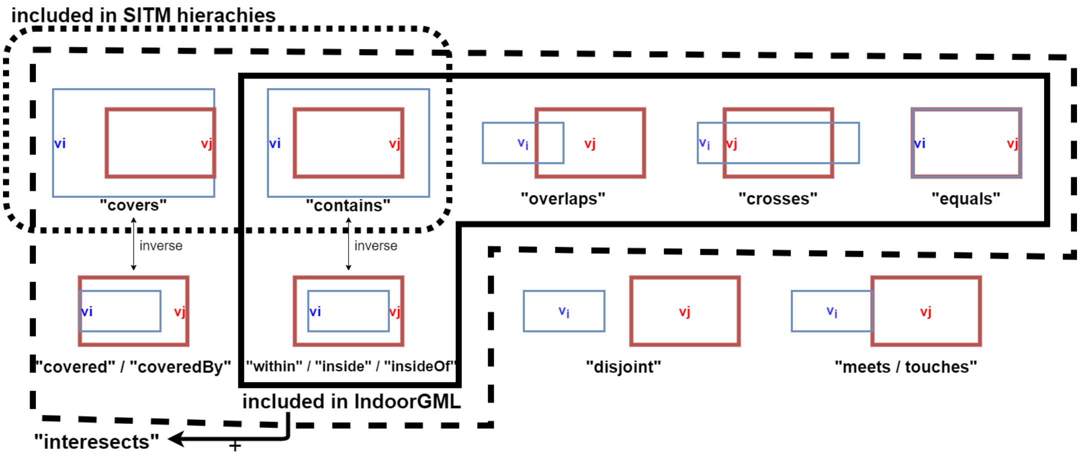


FIGURE 3.2: Common types of binary topological relationships between surfaces.

With regards to intra-layer edges, according to the IndoorGML standard they may represent *adjacency*, *connectivity*, or *accessibility* relationships. Of course, accessibility presupposes connectivity which in turn presupposes adjacency, as argued in section 2.3.2.

Adjacency and connectivity can be said to be less useful for most types of trajectory analysis, because for any moving object to be able to transition between states accessibility is the true requirement. Therefore, the meaning of the intra-layer edges is restricted solely to accessibility between spatial cells.

Modeling adjacency and connectivity can still be useful when changes in the indoor environment itself need to be taken into consideration. In the museum domain for example, it may help with route optimization for emergency response planning, or with interpreting certain mobility behaviors. For instance, a regular visitor getting lost might be explained by an old accessibility relationship being "downgraded" to a connectivity one (e.g. a passage or a door temporarily closing down for renovation reasons) without the visitor being informed about it.

Moreover, given that cells in our core model represent the physical reality of planar space (instead of some form of conceptual space) and that same-layer cells do not overlap at all, an intra-layer edge $e \in E_i^{acc}$ actually presupposes the *meets* topological relationship between its two cells, because the corresponding spaces need to share a common surface for the moving object to be able to physically transition between the two.

Another important indoor space modeling decision has to do with whether G is directed or not. Although IndoorGML does not explicitly assume either case, it considers undirected edges in all of its examples. As far as intra-layer edges go, *adjacency* and *connectivity* can be thought of as being symmetric relationships. However, the proposed model only considers the *accessibility* relationship which is not symmetric since indoor movement is often only unidirectionally possible due to technical, safety or other limitations.

In the right part of Figure 3.3 for example, *Room*4 (*Salle des États*) houses the *Mona Lisa* and accommodates a vast number of visitors on a daily basis. To facilitate their flow, entering it from *Room*2 is often prohibited by the museum personnel, whereas exiting it that way is allowed. Therefore, accessibility NRGs are directed as indicated by the downwards pointing arrowheads in Figure 3.3.

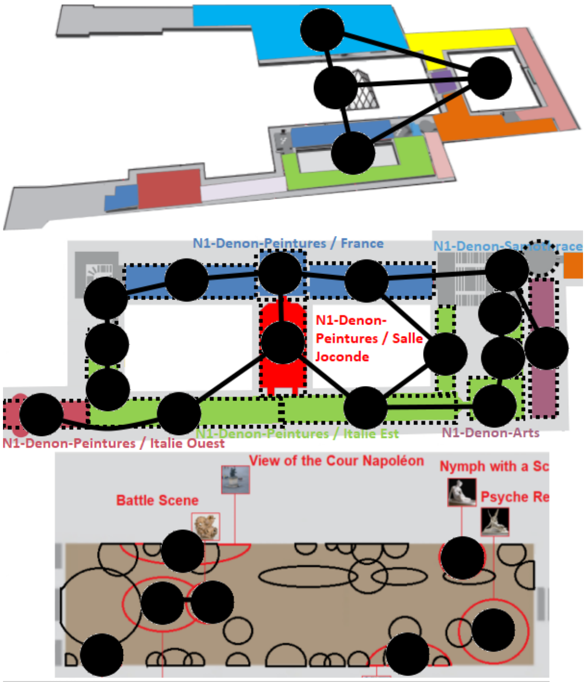
As far as joint edges go, while *overlaps* and *equals* relationships can be thought of as symmetric, *contains* relationships can not. Therefore, again directed joint edges are assumed (as seen in Figure 3.1). If a modeler was only interested in capturing intersection non-emptiness instead of the specific nature of their relationships, then undirected joint edges would suffice, but the hierarchical representation of space would have to be dropped. In conclusion, the entire multigraph G is directed.

Having explained the specific nature of joint edges, it can now be appreciated how a static predefined layer hierarchy enables a structured reasoning about the trajectories at multiple levels of granularity, as opposed to local ad-hoc node subdivisioning schemes.

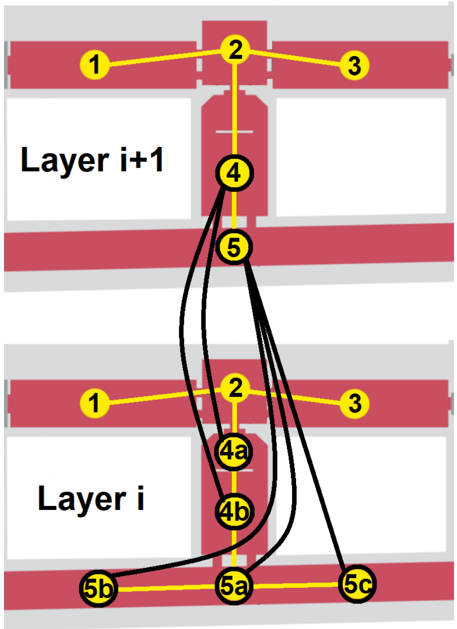
First, by only allowing *proper part* types of relationships, a direct inference of a moving object's location at all levels above the detection/observation data level becomes possible. This in turn allows developing reasoning mechanisms to cope with missing or uncertain location information. For example, in the frequent cases of double detections in one layer (e.g. room layer), such conflicts may be resolved by considering presence at the above layer (e.g. floor) instead of risking an estimation. It also enables the identification of certain types of movement patterns at the room level for instance, and at the same time of other types of patterns at the floor level, all from the same trajectory dataset.

Moreover, multiple layer hierarchies may be defined *in parallel* to each other via parallel joint edges that can additionally include the *equals* and *overlap* relationships. In that case, thanks to the transitivity of parthood (isomorphic to set inclusion) in classical mereology, each layer hierarchy only needs to connect to other layers or layer hierarchies at the lowest possible level. This is because an equals or overlap relationship between two nodes means that an overlap relationship also holds between any two of their predecessors in their respective hierarchies.

Related to this, the graph representation assumes that the indoor area designated



Indoor space represented at different granularity levels by different NRGs.



A 2-level hierarchical graph of the Denon wing's central part.

FIGURE 3.3: Structured (left) or ad-hoc (right) representation of a hierarchical space.

by each node is fully covered by the areas represented by its child nodes [2].

As a reminder from section 2.3.1, IndoorGML distinguishes between a Geometric NRG and a Logical NRG, depending on whether its nodes and edges have geometric properties or not. However, a logical NRG does not address at all how the space being symbolised is being represented, as remarked by [184]: *"an explicit indoor location information is outside the scope of CityGML and IndoorGML"*. Also, according to [52] IndoorGML *"does not seem to provide restriction concerning the placement of such node, but for the targeted applications, such constraints may be necessary. It is implicitly assumed that a node representing a space should be at least contained in that space"*.

Intuitively one may assume that a node in a NRG represents the centroid of the corresponding spatial cell. However, in the proposed indoor space model a different convention is adopted: the location of the moving object is considered to be the whole spatial cell corresponding to the node. This straightforward interpretation implicitly mentioned in [155] can be very useful in applications involving uncertainty in the positional data.

Finally, *SITM* follows an *entrance/exit* node convention: only entrances may

[2]The alternative is IndoorGML's *open-world*-like assumption that *"the union of all cells is a subset of the given indoor space"* ($\bigcup c_i \subseteq U$) [91], under which, if the moving object is present in a given node, then it should certainly also be present in one of its child nodes.

generate moving objects and only exits may consume moving objects. All other nodes are assumed to have equal input and output flows at the end of each day, which can serve for correcting tracking errors in the data. IndoorGML uses the concept of *anchor* nodes, to bidirectionally connect indoors with outdoors, and to contain information for transforming between the respective coordinate reference systems. Entrance and exit nodes can thus be viewed as specializations of anchor nodes, and are meaningful to have even when the outdoor environment is not modeled.

3.3.2 The Semantic Indoor Trajectory Model

Automatically collected raw movement data typically consist of spatiotemporal records, out of which individual trajectories can be extracted. Depending on the application and on the type of moving object, only the evolution of its representative location may be relevant (e.g. museum visit analysis) or perhaps also its shape and parts' movements (e.g. skeletal tracking for sports performance analysis).

In the former case, a trajectory is typically represented as a sequence of timestamped spatial points. Due to a building's clearly separated spaces however, regions (instead of points) are considered to be the model's primary primitive spatial entities, in the spirit of Qualitative Spatial Representation [43] and IndoorGML's cellular space [109], and according to the indoor space model proposed in section 3.3.1.

Following is the presentation of the proposed model of semantic trajectories taking place in an indoor environment. Louvre visits will be used as trajectory examples, since this fits the particular application case, studied in greater detail in chapter 6. Let us first start by providing the formal definition of a semantic trajectory.

Definition 3.3.2 (semantic trajectory)
A semantic trajectory is defined as the couple of its spatiotemporal trace and the set A_{traj} of its semantic annotations:

$$T_{ID_{mo},t_{start},t_{end}} = (trace_{ID_{mo},t_{start},t_{end}}, A_{traj})$$

where ID_{mo} is the identifier of the moving object, t_{start} and t_{end} are the trajectory's starting and ending timestamps, $trace_{ID_{mo},t_{start},t_{end}}$ is a semantic trajectory trace representing the spatiotemporal aspect of the trajectory as a sequence of timestamped semantically annotated presence periods/intervals, and A_{traj} is a set of semantic annotations describing the trajectory in its entirety.

Assuming that no moving object can be in two different places at the same time, its identifier along with the two limit-case timestamps, can be used to identify each of its trajectories $T_{ID_{mo},t_{start},t_{end}}$. The first element $trace_{ID_{mo},t_{start},t_{end}}$ of such a trajectory will be more thoroughly described in the following definition.

The trajectory's second element A_{traj} is a non-empty set of semantic annotations $a_{traj} \in A_{traj}$ characterizing the trajectory in its entirety. Trajectory annotations are not confined within specific types of information, but would typically be chosen to represent an activity, a behavior, or a goal showcased by the complete trajectory.

These terms are often ambiguously used in trajectory literature. Risking oversimplification, the distinction between these three types of trajectory semantics can be summarized as follows:

- *activities* concern targeted/conscious actions performed by a moving object; for example:
  a_{traj} = "*visit_temporary_exhibition*";
- *behaviors* concern less intentional actions or reactions; for example:
  a_{traj} = "*follow_'Masterpieces'_guided_tour*"
- *goals* concern the motivations which affect the movement; for example:
  a_{traj} = "*visit_Mona_Lisa*"

The first two types describe the *actuality* of movement, whereas the third one instead describes the *potentiality* of movement. For example, many trajectories in the Louvre Museum are greatly affected by the visitor's intention to see the *Mona Lisa*, irrespective of whether this goal will eventually be accomplished or disrupted due to overcrowding.

Naturally, a trajectory may well be characterized by multiple types of semantics, e.g.: A_{traj} = {*behaviors* : ["*follow_'Masterpieces'_guided_tour*"],
goals : ["*visit_Mona_Lisa*"]}.

Semantic annotations also have the advantage of allowing the modeler to specify them to the extent desired or needed for a particular application, rather than embark on an ambitious attempt to define all useful movement concepts and relationships of movement in general. For example, they can be used as attribute types in a relational model as suggested in [77], or even matched to the concepts of an ontological model which imposes its own semantic rules and restrictions over what can be represented and what not.

However, they mainly just act as simple labels (i.e. textual annotations) to keep the trajectory model as flexible as possible.

Definition 3.3.3 (semantic trajectory trace)

Let us consider a 2D multiple floor indoor space represented (as detailed in section 3.3.1) by a layered multigraph $G = (V, E)$, $V = \bigcup_{i=0}^{m} V_i$, $E = E^{top} \cup \bigcup_{i=0}^{m} E_i^{acc}$.

A semantic trajectory trace is defined as:

$$trace_{ID_{mo},t_{start},t_{end}} = (e_k, v_k, t_k^{start}, t_k^{end}, A_k)_{k\in[1,n]}$$

where v_k *is the state where the moving object* ID_{mo} *finds itself from* t_k^{start} *until* t_k^{end}, $e_k = (v_{k-1}, v_k) \in \bigcup_{i=0}^{m} E_i^{acc}$ *is the transition that led the moving object from state* v_{k-1} *to state* v_k *(i.e. which boundary was crossed; for example which door, staircase, or elevator was used to get there), and* A_k *is a potentially empty set of semantic annotations describing that specific stay.*

As an example of a semantic trajectory and its corresponding semantic trajectory trace, let us consider a visitor's 2-hour morning visit to the Louvre:

$T_{vis0042,11:30:00,13:30:00} = ($
$\quad trace_{vis0042,11:30:00,13:30:00},$
$\quad \{goals : [\text{“}visit\_temporary\_exhibition\text{”}]\}$
$)$

$trace_{vis0042,11:30:00,13:30:00} = \{$
$\quad$ (*entrance01*, *“PH”*, 11:30:00, 11:32:30, $\emptyset$),
$\quad$ (*ticketcontrol02*, *“TE”*, 11:32:30, 13:00:00, $\{mo : [\text{“}home\_ticket''] \}$),
$\quad$ (*ticketcontrol02*, *“PH”*, 13:00:00, 13:02:00, $\emptyset$),
$\quad$ (*opening02*, *“MS”*, 13:04:00, 13:28:30, $\{activity : [\text{“}shopping''] \}$),
$\quad$ (*opening01*, *“IPH”*, 13:28:30, 13:30:00, $\emptyset$)
$\}$

The above semantic trajectory represents the movement of visitor *vis0042* in the Louvre in order to visit the temporary exhibition. The visitor moves through 4 different zones, first appearing in and passing twice from the *Pyramid Hall (“PE”)*, then going to the *Temporary Exhibition (“TE”)* zone where he/she stays for a long time, and then going back again in the *Pyramid Hall (“PH”)*, before entering the *Museum Shop (“MS”)* area and exiting from the *Inverse Pyramid Hall (“IPH”)*.

Unsurprisingly, while in the *Museum Shop* the visitor did some shopping. It is not the trajectory model's goal to decide how to obtain such semantic aspects from the spatiotemporal context (they may even be simply provided to the analyst) but rather how to represent them in a way that facilitates their extraction and subsequent usage for analysis purposes. As pointed out in [34], semantic annotations “could be arbitrary combinations of the initial multiple semantics”, meaning that they can even be partly provided and partly extracted as part of the analysis. For example, as will be mentioned in section 4.2, a lot of trajectory data mining works have focused on the task of enriching geometric trajectories with place semantics, either as part of the preprocessing process or even as the main means of knowledge discovery.

To accommodate for trajectory *holes* and *semantic gaps* [162] as well as detection data uncertainty issues in general, the spatiotemporal trace is allowed to contain temporal gaps where the presence of the moving object is unknown. This is the case in the above example when the visitor disappeared for a couple of minutes before entering the *Museum Shop*. Allowing for such gaps enables the design of analysis mechanisms [198] treating the uncertainty that is especially prevalent in raw form Big Trajectory Data.

Next, a semantic subtrajectory is defined for all practical purposes as a semantic trajectory - similar to how a mathematical subsequence is itself a sequence - but necessarily referable to some other main semantic trajectory.

Definition 3.3.4 (semantic subtrajectory)
Given a semantic trajectory

$$T_{ID_{mo},t_{start},t_{end}} = (trace_{ID_{mo},t_{start},t_{end}}, A_{traj})$$

a semantic subtrajectory of it is defined as:

$$T'_{ID_{mo},t'_{start},t'_{end}} = (trace'_{ID_{mo},t'_{start},t'_{end}}, A'_{traj})$$

iff $trace'$ *is a proper subsequence of* $trace$*:*
$t_{start} \leq t'_{start} < t'_{end} < t_{end}$ *or* $t_{start} < t'_{start} < t'_{end} \leq t_{end}$.

A subtrajectory's set of semantic annotations A'_{traj} may or may not be the same as that of its main trajectory A_{traj}, contrary for example to [25] where these are (by design) enriched with different types of semantic information.

Let us now consider another visitor's semantic trajectory:
$T_{vis0043,13:00:00,13:33:00} = ($
$trace_{vis0043,13:00:00,13:33:00}$,
$\{goals : [\text{“}visit\_temporary\_exhibition\text{”}]\}$
)
$trace_{vis0043,13:00:00,13:33:00} = \{$
(*entrance01*, *“PH”*, 13:00:00, 13:02:00, $\emptyset$),
(*opening02*, *“MS”*, 13:04:00, 13:28:30, $\{activity : [\text{“}shopping''] \}$),
(*opening01*, *“IPH”*, 13:28:30, 13:33:00, $\emptyset$)
}

In comparison to the previous trajectory example, this one concerns a more casual type of Louvre visitor who is simply shopping in its stores (where a ticket is not required). According to a strict interpretation of Definition 3.3.4, the semantic trajectory of visitor *vis0043* is not a subtrajectory of the semantic trajectory of visitor *vis0042*, because even though they share the same 3-zone pattern of visit “PH” $\rightarrow$ “MS” $\rightarrow$ “IPH”, and the exact same semantics, *vis0043* arrives at *“PH”* via a different edge and stays in *“IPH”* a little longer than *vis0042*.

Therefore, for practical reasons and depending on the application case, the *proper subsequence* requirement needs to be mathematically relaxed according to a fitting interpretation of trace similarity. This can be either with respect to spatiotemporal similarity or semantic similarity or both.

For instance, if we ignore the traversed edges and allow a temporal deviation of 5 minutes in the timestamps of each presence interval, then $T_{vis0043,13:00:00,13:33:00}$ is indeed a semantic subtrajectory of $T_{vis0042,11:30:00,13:30:00}$, because their last timestamps differ only by 3 minutes (less than the aforementioned threshold value of 5 minutes).

Finally, as made evident by the last example, a semantic subtrajectory may concern a different moving object than its main semantic trajectory. This is because in the general case, the proposed model is concerned with studying movement patterns irrespective of who performed them.

Next, an episode of a semantic trajectory is defined as any particularly meaningful part of it.

Definition 3.3.5 (episode)
Given a semantic trajectory

$$T_{ID_{mo},t_{start},t_{end}} = (trace_{ID_{mo},t_{start},t_{end}}, A_{traj})$$

an episode of it is defined as:

$$T'_{ID_{mo},t'_{start},t'_{end}} = (trace'_{ID_{mo},t'_{start},t'_{end}}, A'_{traj})$$

iff

(1) $T'_{ID_{mo},t'_{start},t'_{end}}$ *is a semantic subtrajectory of* $T_{ID_{mo},t_{start},t_{end}}$

(2) $A'_{traj} \neq A_{traj}$

(3) $T'_{ID_{mo},t'_{start},t'_{end}}$ *satisfies a domain-dependent and user-defined spatiotemporal and/or semantic predicate* $P_{ep} : T'_{ID_{mo},t'_{start},t'_{end}} \rightarrow \{true, false\}$

Moreover, an *episodic segmentation* of a semantic trajectory is simply any subset of its episodes that covers it time-wise. Contrary to typical literature practice, we allow an episodic segmentation to contain episodes that overlap in time, since the exact same movement part may have multiple meanings depending on the broader context or on the scale at which it is examined.

For example in the museum domain more precisely, it may be useful to enrich a visit trajectory with PoI theme instances corresponding to where the moving object stopped (e.g. the artwork type), with contextual event instances affecting the movement (e.g. a fire alarm activation that took place at a specific time), with the current status of the moving object (e.g. audio description being played), etc. However, using state of the art annotation-based semantic trajectory models would already require three separate segmentations of the same trajectory trace. One segmentation would include *spatiotemporally*-defined "visit" episodes with annotation values such as "Painting_X", a second segmentation would include *temporally*-defined "event" episodes with annotation values such as "Fire_Alarm" or a link to a "Fire_Alarm" object, and a third segmentation would include *semantically*-defined "listening" episodes with annotation values such as "Sculpture_Y_Audio".

Instead, it is much more convenient to allow overlapping episodes, since for instance

Let us consider an enriched version of visitor vis_{0042}'s previous semantic trajectory example and a non-overlapping activity-based episodic segmentation of it:

$episode_{seg} = \{$

$episode_1$ (arrival):

$T_{vis0042,11:30:00,11:32:30} = ($

$trace_{vis0042,11:30:00,11:32:30},$

$\{activities : [\text{“}arrive\_Louvre\text{”}]\}$

$)$

$trace_{vis0042,11:30:00,11:32:30} = \{$

(*entrance01*, *"PH"*, 11:30:00, 11:32:30,$\emptyset$)

$\}$

$episode_2$ (temporary exhibition visit):

$T_{vis0042,11:32:30,13:00:00} = ($

$trace_{vis0042,11:32:30,13:00:00}$,
$\{activities : [\text{“}visit\_temporary\_exhibition\text{”}],$
$goals : [\text{“}visit\_Salvator\_Mundi\text{”}]\}$
)
$trace_{vis0042,11:32:30,13:00:00} = \{$
(*ticketcontrol02*, *“TE”*, 11:32:30, 13:00:00, {mo:[“home_ticket”]})
}

episode$_3$ (departure):
$T_{vis0042,13:00:00,13:30:00} = ($
$trace_{vis0042,13:00:00,13:30:00}$,
$\{activities : [\text{“}shopping\text{”}, \text{“}exit\_Louvre\text{”}]\}$
)
$trace_{vis0042,13:00:00,13:30:00} = \{$
(*ticketcontrol02*, *“PH”*, 13:00:00, 13:02:00, $\emptyset$),
(*opening02*, *“MS”*, 13:04:00, 13:28:30, $\{activity : [\text{“}shopping''] \}$),
(*opening01*, *“IPH”*, 13:28:30, 13:30:00, $\emptyset$)
}
}

Interestingly, even though it describes temporally continuous movement phenomena, $SITM$ is still an event-based model: only a change of the spatial cell that the moving object is located in, or a change of the semantic information regarding the moving object's presence in that cell, requires a seperate tuple. Hence, each tuple's *begin* and *end* timestamps denote the natural time interval that corresponds to the moving object's physical presence given stable semantics. One may argue that it then becomes difficult to correctly split and annotate the trajectory because what if the intervals become overloaded with different many types of semantics. Whereas this would be true in a stops and moves model, as warned by [67], here it is countered by the multi-granular nature of our representation: not all spatial abstraction levels need to correlate to all semantic semantic abstraction levels. For example, if the spatial granularity of a museum visitor's trace is at the RoI layer, then perhaps we need to take into consideration annotations describing which particular artwork caught that visitor's attention. If on the other hand our data's spatial granularity stops at the room layer, then perhaps we prefer to skip the lowest-level of semantic information that we possess and only use the artwork theme of the rooms or the room congestion level.

Although, we took advantage of spatial semantics in our model, the proper layer equivalence between spatial information and movement semantics is ultimately application dependent. Therefore, it should be up to the analysis method (or to a specific model extension) to decide whether it wants to derive higher level semantics from lower level semantics, attempt to extract lower level semantics from higher level ones and/or from the spatiotemporal data, etc. Finally, such a representation also suits most raw indoor mobility datasets, which typically consist of individual sensor detections.

3.4 Chapter Conclusion

In this chapter, we presented a new conceptual model of trajectories, which accounts for semantic and indoor space information. More specifically, it formally defines a new model for spatiotemporal indoor trajectories enriched with semantic annotations, called Semantic Indoor Trajectory Model ($SITM$). The proposed model makes selective use of the IndoorGML standardized indoor space modeling framework, and integrates semantic annotations at different levels of granularity in order to allow a detailed description of the movements, thus enabling more interesting types of analysis.

Many of the concepts used in our proposed model (e.g. episodes, semantic annotations) were adopted from state-of-the-art conceptual models, as presented in section 2.2.2 and modified to fit cohesively with each other and with the new ideas (e.g. multilayered graph indoor space representation, selective topological relation types, overlapping episodes). In chapter 6 we will illustrate $SITM$ with respect to the modeling of visitor trajectories in the Louvre Museum.

4

State of the Art in Trajectory Data Mining and Analysis

lscape

Contents

4.1 Introduction 90
4.2 Trajectory Data Mining Landscape 90
4.2.1 Trajectory Data Management & Preprocessing 93
4.3 Sequential Pattern Mining and Trajectory Pattern Mining 94
4.3.1 Trajectory Pattern Mining: History and State-of-the-Art 94
4.3.2 Sequential Pattern Mining 103
4.3.3 The Standard Problem and Algorithms 104
4.3.4 Sequential Pattern Mining and Multiple Dimensions 105
4.3.5 Sequential Pattern Mining and the Temporal Dimension 108
4.4 Chapter Conclusions 115

4.1 Introduction

This chapter offers a research overview and categorization of computational processing and analysis tasks related to trajectory data. It does not constitute a systematic literature review, nor does it aim to cover completely the whole spectrum of corresponding methods in the bibliography for tackling these problems. Our aim is to provide help in orientating the researchers and professionals who deal with mobility datasets, and to assist them in identifying which types of analyses fit which data and fields of expertise, and inversely which types of mobility data and analytical methods are most suitable for solving a particular movement analysis problem. This original overview also serves as an introduction to Trajectory Pattern Mining (T-PM) which is the main focus area for the rest of this chapter and for which background work will be examined.

4.2 Trajectory Data Mining Landscape

Data processing, mining, and analysis have been extensively studied for at least three decades already [144], but have typically relied on a data instance independence assumption. Naturally, this is not the case with spatiotemporal movement data, where data instances are integrally interrelated, and even the temporal data dimension alone requires special consideration. Moreover, the apparition of semantic trajectory data throughout the last decade - however they may be represented - makes it even more important in the data mining world to build new specialization around trajectory data mining tasks.

In section 2.2 the focus was on the state-of-the-art on modeling trajectories, but structuring mobility data into a trajectory form is only the first step towards extracting knowledge from them. Towards this end, this section is primarily interested in T-PM, yet it begins by presenting an integrated view on a broad range of trajectory data mining tasks.

Oftentimes the terms *trajectory data mining* and *trajectory pattern mining* are incorrectly used interchangeably, when in fact the former encompasses numerous tasks including - but not limited to - frequent pattern extraction from an input trajectory dataset. The motivation for proposing this problem overview is twofold. First, a delay in acquiring the application case datasets (i.e. the Louvre museum visitor datasets) lead the author to first study and consider the broader trajectory data mining landscape, before narrowing down on a set of potential target problem(s) based on his personal literature review and examination of the available data. Thus, it was experienced first-hand how real-world data availability and quality can dictate the pursual of certain trajectory data mining tasks, whereas others clearly become unfeasible despite being equally interesting from a theoretical point of view. Secondly, similar efforts have so far been undertaken in survey works such as [198], [65], [132], [16][1], and [179]. There also exist survey works which only cover narrower segments

[1] In this survey Atluri et al. actually review all types of spatiotemporal data and consider trajectories as a particular class of data instances.

of the trajectory data-related research spectrum, such as taxi trajectory data mining [36], semantic trajectory data management [85], road transportation systems analysis [129], maritime trajectory management and mining [42, 177], context-aware movement analysis [158] and others. Unfortunately, the attempted trajectory data mining taxonomies proposed by such surveys have a limited utility lifespan, given the pace of advances in the related research field. Hence, a concise up-to-date task taxonomy which emphasizes awareness about the interrelatedness of said tasks, can serve the trajectory research community.

Existing trajectory data mining surveys follow largely different categorization criteria, which almost always fall within either a methodological point of view or an application point of view. Indeed, there is a plethora of useful criteria that one may come up with: the *type of information* sought, the analysis' *level of abstraction*, the *nature of the analytical task* (e.g. descriptive-predictive dichotomy), the *methods/techniques* being applied, the *execution order* (within the broader knowledge discovery process), the *dependence* or *exclusivity* or perhaps even *orthogonality with respect to other tasks*, the *domain(s) of application*, etc.

Whereas each criterion has its own merit, multiple tables would be needed to reflect all of them properly. Therefore, in an attempt to maximize its practicality, the proposed categorization in Table 4.1 is based on a mixture of these criteria that best reflects the current state-of-the art practices. It follows a pragmatic perspective considering the essential nature of the tasks themselves: if a group of tasks is distinctly defined by their *analytic goals* then that group constitutes a task category. The same is true even if the tasks don't share the same analysis goals but are almost always tackled by the same *mining methods*.

Also, *popularity* plays a role in determining whether a task is classified separately from "similar" ones. For example, trajectory-based recommendation could comprise a promising application-inspired task category on its own, but it has not yet gained enough traction so as to justify considering it separately from the broader trajectory similarity and clustering task category, given that they share the very same methods.

Similarly, trajectory segmentation and hotspot extraction sometimes constitute by themselves the end goal of an application, but most of the times are simply used as an input data wrangling step towards enabling further trajectory data mining and analysis actions. As a result, we assign them to the trajectory data processing task category. In relation to this, some surveys such as [179] distinguish between *first-tier* methods and *second-tier* methods, with the former being essentially trajectory clustering and classification and the latter being all of the other types of analysis that run on the already clustered or classified trajectories. Indeed, similar requirement chains do exist between trajectory data mining tasks, where one task depends on the output of another, and are worth studying further, especially as the trajectory data pipeline keeps evolving due to the continuous progress in the applied tracking and storage technologies. However, task dependencies do not qualify as a primary criterion for the "reference guide" type of taxonomy envisaged here and presented in Table 4.1.

Category	Analysis/Mining Task	Typical Application & Application Domain
Trajectory Data Preprocessing	trajectory data cleaning	filtering out erroneous detection records (e.g. impossible/invalid locations, timestamps, transitions) based on spatiotemporal constraints, ground-truth data, etc.
	trajectory calibration & uncertainty handling	resolving conflicting observational data (e.g. double detections), outlier/anomaly detection and handling, fusion or homogenization of trajectory data captured from different sources, handling probabilistic movement data
	trajectory completion/refinement	resolving missing observational data (e.g. detection gaps), extracting representative trajectories by synthesizing fragmented or sparse trajectory data, multiple trajectory datasets integration, trajectory-based entity resolution of unidentified moving objects
	trajectory enrichment	semantic and/or contextual information integration to raw trajectories, adding fine-grained movement information to coarse trajectory data
	map matching	correcting road network-constrained vehicle trajectories, indoor navigation, vehicle navigation, urban resource allocation
	trajectory segmentation	segmentation based on predetermined spatiotemporal thresholds, transportation means extraction, stop-move identification, activity discovery, etc.
	trajectory compression and/or simplification	trajectory sampling (e.g. error-bounded), episode extraction
Trajectory Data Management	trajectory data query processing	trajectory data query language & operator design, location-based queries, range queries, nearest-neighbour queries, spatiotextual queries, similarity queries
	trajectory storage & retrieval	in-memory trajectory storage, trajectory data indexing structures, trajectory datatypes (e.g. moving point/region), distributed data pipelines
	trajectory modeling concept implementation	semantic trajectory data management, trajectory database management systems design, trajectory model integration with existing databases, low-level trajectory conceptualization
Trajectory Similarity and Clustering	clustering (sub)trajectories (of the same or different moving objects)	mobility behavior profiling
	clustering moving objects	moving object profiling, moving object group discovery
	designing trajectory similarity metrics	identifying spatiotemporal and non-spatiotemporal similarities of movements, designing semantic similarity metrics, exploring the correlations between spatiotemporal and semantic types of trajectory similarity
Trajectory Pattern Mining	collective/group pattern mining	urban development (e.g. transportation systems design), group interaction discovery, fleet management, team sports tactical analysis
	sequential pattern mining	discovering typical mobility behaviors, discovering unexpected patterns, place/trip recommendation, crime prevention, metaphorical trajectory analysis (e.g. healthcare trajectories, information trajectories)
	periodic pattern mining	discovering seasonal trends, discovering recurring micro-behaviors, mobility behavior change detection (e.g. abnormal behaviors)
Trajectory Classification	(sub)trajectory classification	predefined behavior recognition, transportation means classification
	moving object relation/role identification	determining the leader of a moving group, distinguishing “personnel” (e.g. museum guards) from “clients” (e.g. visitors) for indoor trajectories
Trajectory Prediction	location prediction	destination prediction, next-location prediction, traffic estimation, emergency response planning
	(sub)trajectory prediction	network route prediction (based on trajectories of the specific moving object or of all moving objects), turn choice prediction, air flight or marine vessel destination prediction
	trajectory attribute prediction	speed prediction, orientation prediction
Trajectory Visual Analytics	trajectory visualization & interactive exploration	interactive mobility pattern extraction, suspicious activity detection, inclusion of domain expert intuition (e.g. visual trajectory prediction), visual report generation for decision support
	multigranular spatio-temporal analytical query answering	activity impact assessment, public transit and traffic control planning
	multi-source visual information fusion/integration	situation overview and real-time monitoring (e.g. city operation control centres, air-traffic & maritime operation safety)
Mobility Behavior Discovery	RoI extraction	hotspot extraction (i.e. areas of high levels of moving object co-occurrence), geographic place discovery, semantic place discovery, event detection
	discovering movement purpose	trajectory-based activity / intent / motivation / goal extraction
	behavior identification	matching trajectory patterns to predefined mobility behaviors, verifying or disproving mobility behavior hypotheses (e.g. typical museum visiting styles)
Miscellaneous	privacy protection/preservation	time distortion anonymization of mobility data, trajectory-based geo-indistinguishability, differential privacy over trajectory data
	routing	shortest/fastest path discovery, accessibility network extraction, personalized route recommendation
	artificial trajectory generation	trajectory data availability for testing analysis and mining methods, enabling experiments in sensitive domains of strong privacy restrictions
	Big Trajectory Data support	data volume scalability, handling heterogeneous data stemming from different detection infrastructures and collection methodologies, handling online trajectory data streams

Table 4.1: State of the practice classification of trajectory data mining tasks.

4.2.1 Trajectory Data Management & Preprocessing

Trajectory data management and trajectory data preprocessing can be viewed separately than main trajectory mining tasks, despite the fact that they overlap in terms of goals and methods.

Trajectory data management often involves algorithmic and data structure efforts to optimize two major types of queries, namely nearest neighbors queries and spatiotemporal range queries. This often includes finding efficient ways to store and retrieve the trajectory data. More recently, particular effort is being put towards addressing Big Data issues as well as semantic data issues. More specifically, recent works have targeted distributed data pipelines including storage and computation engines [54, 112], combined online and offline analysis support [63], interactive visual exploration [38], semantic extensions of database systems [73], data textualization techniques and corresponding semantic querying mechanisms [8], and others.

Whereas in trajectory data management research the main tasks are generally well established, despite some variations imposed by modern computation needs mentioned in the previous paragraph, trajectory data preprocessing encompasses a lot of different tasks and even an even greater number of corresponding techniques, often ad-hoc ones. What they all have in common is that they deal with trajectory data at its rawest form (e.g. tracking data logs). But what constitutes *raw* trajectory data can actually vary considerably from case to case. Depending on the application, trajectory data may be extracted in a form far from the desired one, and hence a lot of preprocessing is required, or to the contrary, it can already resemble closely its final analysis form, in which case preprocessing may even consist of a single straightforward transformation task.

Whatever the case may be, due to the fact that trajectory data are in general more complex than conventional and spatial data [24], their preprocessing requires careful consideration, no matter how simple or complicated. This is why although preprocessing tasks are typically considered to correspond to the preparation of input trajectory data for the main mining step(s), in practice they already result in a great level of knowledge extraction, as they require a deep understanding of the trajectory dataset in terms of its structure and interpretation, as well as strong familiarity with the application domain and even particular use case. This is especially true when the data provider fails to provide the analyst with proper documentation or metadata, resulting in unclear preprocessing requirements even before the beginning of the analysis.

Furthermore, a task like the semantic enrichment of input trajectories, or their segmentation into meaningful parts, can sometimes constitute the mining goal in and of itself, which is why it can be claimed that trajectory data preprocessing is actually an “umbrella term” rather than a clearly defined group of tasks.

4.3 Sequential Pattern Mining and Trajectory Pattern Mining

One of the most interesting family of trajectory data mining tasks is T-PM. As its name suggests it consists in finding the most frequently occurring mobility patterns within a trajectory dataset. But despite the task's simplicity, there are actually many different ways in which the problem can be formulated and tackled.

4.3.1 Trajectory Pattern Mining: History and State-of-the-Art

Trajectory data pattern mining research has its roots in the early 2000s. Related works at the time attempted to adapt classical data mining techniques so as to discover different types of trajectory patterns. Dimitrijevic et al. [53] generalize the Apriori algorithmic framework to discover frequent segments in the trajectories of multiple ice hockey players. They propose a two-phase algorithm, which first transforms geometric trajectories into symbolic sets of trajectory segments, and then through the use of a similarity function compares all segments to the candidate pattern to derive its support. Time is taken into account in the form of constraints that each segment must satisfy (e.g. a certain time interval must exist between the start of any two adjacent segments).

First attempts at mining trajectories as sequences. In [157], Sclarrof et al. first derive the trajectories through sampling and piecewise-linear curve approximation, and then follow a similar approach in transforming the trajectories into sequences of symbolic segments, the so-called *motion signatures* of the trajectories, over which a Longest Common SubSequence (LCSS)-based similarity measure is applied, thereby abstracting time from the mining process.

In [170], Tsoukatos et al. propose a Depth-First-Search(DFS)-like sequential pattern mining (S-PM) approach for mining spatiotemporal patterns, but in reality only use order information, not temporal information. This is actually one of the first works to propose mining spatiotemporal patterns at multiple spatial granularity levels. It does so simply by joining (sub)regions formed at the original data granularity level, into coarser regions (i.e. of a higher granularity level), and repeating this process as necessary. Featuring multiple spatial granularities is an important consideration for our hierarchical indoor space model that will be described in section 3.3.1.

Lee et al. in [110] propose one of the first online T-PM methods on the basis of incremental sequential patterns, and Hwang et al. in [84] one of the first collective pattern mining methods using the Apriori algorithm and a threshold-based geographic proximity measure of moving object groups. In [35], Huiping et al. propose one of the first works to study trajectory segmentation, in the form of direct lines subsequently matched to spatial regions, and then used to discover sequential patterns based on a substring tree structure and the Apriori technique. Unlike most other works however, their method finds frequent spatiotemporal patterns in a long spatiotemporal sequence, rather than in a set of sequences. This is actually what Mannila et al. [128] as well as subsequent works have referred to as *episode mining*.

In [89], Kalnis et al. propose the problem of extracting a special type of pattern called *moving group*, which is simply an evolving cluster of moving objects, represented as sequences of spatial clusters in consecutive movement snapshots, potentially sharing a number of common objects. The authors then propose three timeslice-based clustering methods to retrieve them.

First attempt at mining museum trajectories. Kanda et al. propose one of the first works [90] to address pattern mining of indoor trajectories, and actually in a museum setting which is particularly interesting to our own use case. Almost all RFID readers used to collect the tracking data were installed in the exhibition space, except for a few mobile ones installed in humanoid robots as part of a separate field trial [160].

First, similar to many other works, the authors transform the coordinate position data into symbolic region data, by applying a k-means algorithm over all input trajectories. Hence, their method divides the indoor space into k areas, rather than use a predefined spatial or semantic partitioning of the indoor space.

Then, they adopt a state-chain representation of trajectories $S^i =< A_1, A_2, ..., A_n >$ where each state is the spatial partition A_i in which the corresponding (x,y)-coordinate point of the trajectory belongs. Thus, a state-chain may include consecutive state repetitions such as $< P1, P1, P1, P2, P1, P3, P2, P2, P4 >$. Subsequently, the authors again apply a k-means clustering method over those state-chains, along with a Dynamic Programming matching approach based on a Levenshtein distance metric. In specific, they use the distance between the centers of the indoor space partitions to calculate this metric.

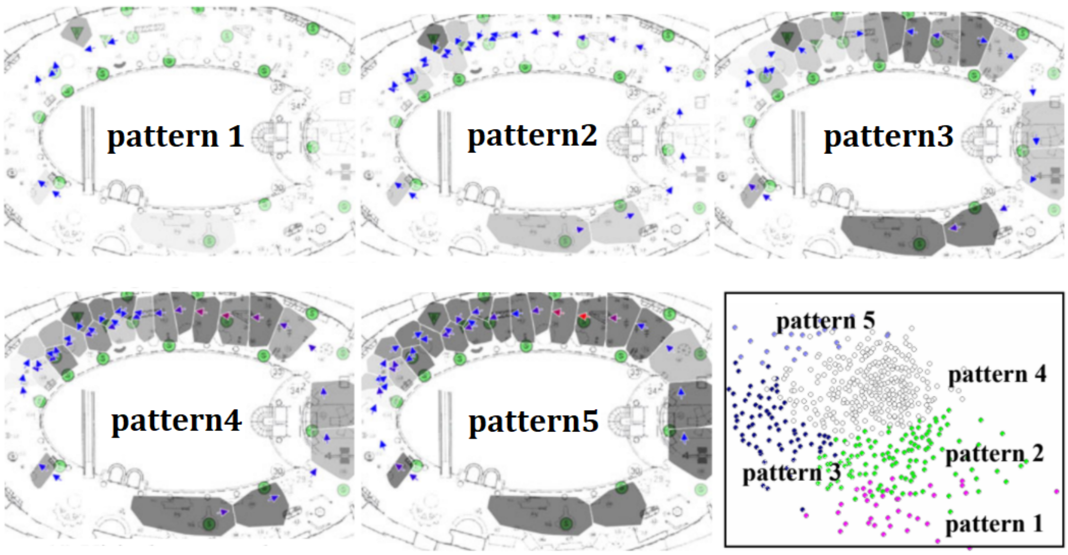


FIGURE 4.1: The five most typical visitor behavior patterns found in [90]: 1) "directly go to robot area", 2) "go around and stay at robot area", 3) go around backwards"' 4)"visit every place", 5) "stay for long time".

Unlike other works however, the authors essentially treat T-PM as a clustering problem where the interpretation of the resulting clusters reveals the overall visiting patterns, as opposed to finding patterns that correspond to trajectory parts, which is what T-PM methods typically do. This is also evident by the output patterns illustrated in Figure 4.1 which describe the entire visit. Unfortunately, time is once again not taken into account, as only the order of regions and the entire visit duration is taken into consideration, and only at the stage of the results' post-processing and interpretation.

t-patterns algorithm. The work of Giannotti et al. in [71] is - to our knowledge - the first to attempt to mine trajectory patterns using properly extended S-PM methods, which capture the temporal aspect of movement instead of relying solely on the order of location information. More specifically, the authors define a *trajectory* as a spatiotemporal sequence of triples:

Definition 4.3.1 (trajectory)
A trajectory is a sequence $T =< (x_0, y_0, t_0), ..., (x_k, y_k, t_k) >$ *where* $t_i < t_{i+1}$ *is a timestamp and* $(x_i, y_i) \in \mathcal{R}^2$ *are Cartesian coordinate points* $(i = 0...k)$.

Then, a *trajectory pattern* or *T-pattern* is defined as follows:

Definition 4.3.2 (T-pattern)
A T-pattern is a pair (S, A) *where* $A =< \alpha_1, \alpha_2, ..., \alpha_k > \in \mathcal{R}^k_+$ *is a sequence of temporal annotations corresponding to a sequence of points* $S =< (x_0, y_0), (x_1, y_1), ..., (x_k, y_k) > \in \mathcal{R}^2$.

Furthermore, by defining the notion of *spatial containment* the authors are able to transform the abovementioned coordinates into symbolic regions of interest, in the same spirit as many other works:

Definition 4.3.3 (spatial containment)
A sequence of spatial points $S =< (x_0, y_0), ..., (x_k, y_k) >$ *is spatially contained in a spatiotemporal sequence* $T =< (x'_0, y'_0, t'_0), ..., (x'_n, y'_n, t'_n) >$ $(S \preceq_N T$ *or* $S \preceq T)$ *IFF* $\exists 0 \leq i_0 < ... < i_k \leq n$ *such that* $\forall 0 \leq j \leq k : (x_j, y_j) \in N(x'_{ij}, y'_{ij})$ *where the neighborhood function* $N : \mathcal{R}^2 \to \mathcal{P}(\mathcal{R}^2)$ *assigns to each point a set of neighboring points.*

As a result, the pattern mining problem is reduced to that of Temporally Annotated Sequential Pattern Mining (TAS-PM) as defined in [70] and studied in detail in the next section, where the sequences' discrete elements (usually from a predefined alphabet) now symbolize the spatial neighborhoods of the input trajectories. Figure 4.2 exemplifies the logic behind matching a candidate subsequence to the input sequences, even when their corresponding transition times are not exactly equal, through the use of a temporal relaxation parameter τ.

Slicing-STS-Miner algorithm. In [83], Huang et al. propose a S-PM algorithm called *Slicing-STS-Miner* for mining spatiotemporal event patterns. It uses a sequence index as a measure of statistical significance of spatiotemporal sequential

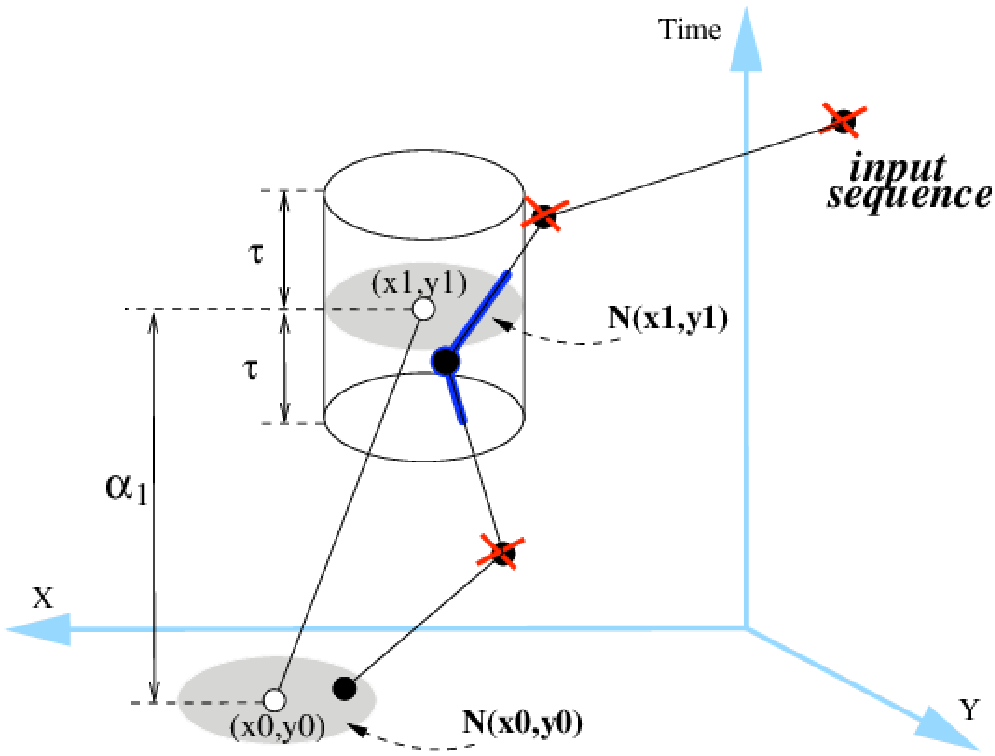


FIGURE 4.2: An occurrence of a trivial T-pattern $(x_0, y_0) \rightarrow (x_1, y_1)$ in an input trajectory instance, given a neighborhood function N and a temporal threshold τ, taken from [71].

patterns. In particular, it builds a pattern tree in which each pattern's *sequence index* is calculated (via node and branch values), and prunes it according to a minimum sequence index threshold value θ. It also partitions the dataset into overlapping temporal slices when the number of events is too large to be processed in memory, processes each slice separately, and then utilizes the unidirectional property of time to recover the patterns across slice boundaries. This helps deal with cases of excessive number of database scans.

Noticeably, Slicing-STS-Miner uses spatiotemporal event data, not trajectory data. An *event* is a happening at a given place and time, which belongs to a specific *event type* (e.g. "car accident", "traffic jam", "chromium-6-polluted water source", "the West Nile disease", "deforestation").

Definition 4.3.4 (spatiotemporal sequential pattern)
A spatiotemporal sequential pattern is a sequence $f_1 \rightarrow f_2 \rightarrow ... \rightarrow f_k$ *of spatial and temporal event types* f *which happens in a specific spatial and temporal manner.*

For example, a West Nile disease sequential transmission path might be $Bird \rightarrow Mosquito \rightarrow HumanBeing$.

Time is taken into account through the use of a spatiotemporal predicate *follow* which requires that, an event happens in the nearby region of another event, and only shortly afterwards that event.

Definition 4.3.5 (follow predicate)

An event e_0 $follows_N$ *another event* e'*, denoted by* e $e0$ *iff* e' *is in the spatiotemporal neighborhood* $N(e)$ *of* e*.*

Definition 4.3.6 (spatiotemporal neighborhood)

A simple spatiotemporal neighborhood of an event e *is* $N(e) = \{p | p \in \mathcal{D}, distance(e.location, p.location) \leq \mathcal{R}, (p.time - e.time) \in [0, \mathcal{T}]\}$

For example, a spatiotemporal neighborhood could be defined by a distance limit of 1.5 miles for any time interval of less than 1 hour.

Whereas the follow predicate is useful when the input dataset consists of a collection of spatiotemporal events, it is not meaningful[2] when it consists of individual trajectories, because then each event corresponds to an already identified moving object. In other words, having already built moving object trajectories, there is no need to "connect" spatiotemporal events to each other.

ST-DMQL query language. In contrast to the previous works, in [24] Bogorny et al. do not propose any specific trajectory data mining algorithm, but aim to offer the mechanisms to support such algorithms. More specifically, they provide the methodology and the primitives for a trajectory data mining query language, as well as a prototype implementation called *ST-DMQL*. Their work targets *semantic trajectories* in particular, but understood merely as trajectories enriched with geographic information based on the concepts of stops and moves. The related semantic trajectory model has already been reviewed in the review of [23] from section 2.2.2.2.

As part of ST-DMQL, they propose two operators for defining the granularity of time and place, namely *timeG* and *stopG*. The first operator converts a timestamp into user-defined granularity labels (e.g. [07:00-09:00] may be labeled as "rushHour", [14:00-18:00] as afternoon, etc.) or predetermined labels (e.g. "WEEKEND-WEEKDAY", "YEAR", "MONTH", "SEASON", and "DAY-OF-THE-WEEK"). The second operator manages the granularity of stops. It can generate two granularity levels, called *feature instance* (e.g. "Centrum Hotel") and *feature type* (e.g. "Hotel"), without any background knowledge such as concept hierarchies or ontologies. This function allows the user to specify stops at different granularities, for example to specify a hierarchy of intermediate granularity levels specifically for some feature types.

This is one of the first times that support for hierarchical spatial information is considered with respect to trajectory data mining. With respect to S-PM in particular, Bogorny et al. envisage queries such as:
SELECT sequentialMoves (item=NameEnd, timeG=[8:00-12:00 AS morning, 18:00-23:00 AS evening], stopG=instance, minsup=0.03)
FROM move
which would return patterns such as

$\{IbisHotel - NotreDame[morning[, EiffelTower - IbisHotel[evening]\}$

[2]On the other hand, requiring spatiotemporal proximity in each item transition might be useful for correcting the input trajectory data.

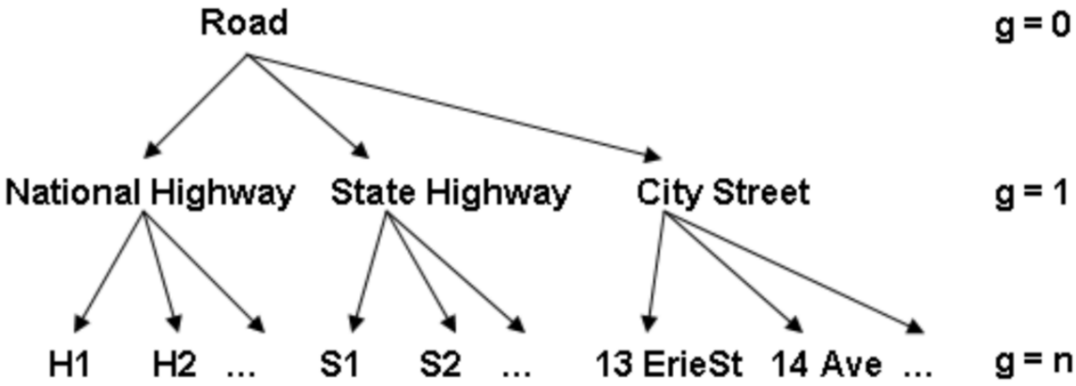


FIGURE 4.3: Example move concept hierarchy from [24] which assumes that the database instances correspond to the lowest level of granularity g=n.

provided of course that a morning move from the Ibis Hotel to Notre Dame happens before an evening move from Eiffel Tower to the Ibis Hotel in at least 4% of the input trajectories.

MTP-ITP and MTP-TEQ algorithms. Similar to [71], Kang et al. in [92] present one of the few works to have identified the inefficient consideration of the temporal information in T-PM literature, in terms of both performance and result quality and interpretability. They propose a method whose first step is - like in the majority of the related works - a transformation of the trajectories from raw form to line segments, in turn clustered into symbolic regions. This region extraction process is a common technique for discretizing continuous location values, and has some obvious advantages over fixed-size grid approaches which risk missing patterns or ending up with redundant locations (depending on cell size).

However, Kang et al. propose two approaches for partitioning the trajectories, leading to two fundamentally different mining methods, namely *MTP-ITP* and *MTP-TEQ*.

The first method (left side of Figure 4.4) involves a spatiotemporal discretization process after which the temporal aspect of a trajectory is abstracted into the resulting spatiotemporal regions. More specifically, a *spatiotemporal pattern* is defined as follows:

Definition 4.3.7 (spatiotemporal pattern)
A spatiotemporal pattern is a sequence $ST =< R_1 R_2 ... R_k >$, $k < n$, *where* $R_i = (s_i, d_i)$ *is a spatiotemporal region,* s_i *is a spatial approximation of points from* t_l *to* t_m *in trajectory* T, $l < m < n$, *and* d_i *is the duration of movements between* t_l *to* t_m.

In this way, the main mining step can be performed using any existing S-PM algorithm and the authors proceed to choose a pattern-growth based one.

The second method (right side of Figure 4.4) instead involves a purely spatial trajectory discretization in the first step, but then the temporal information is explicitly

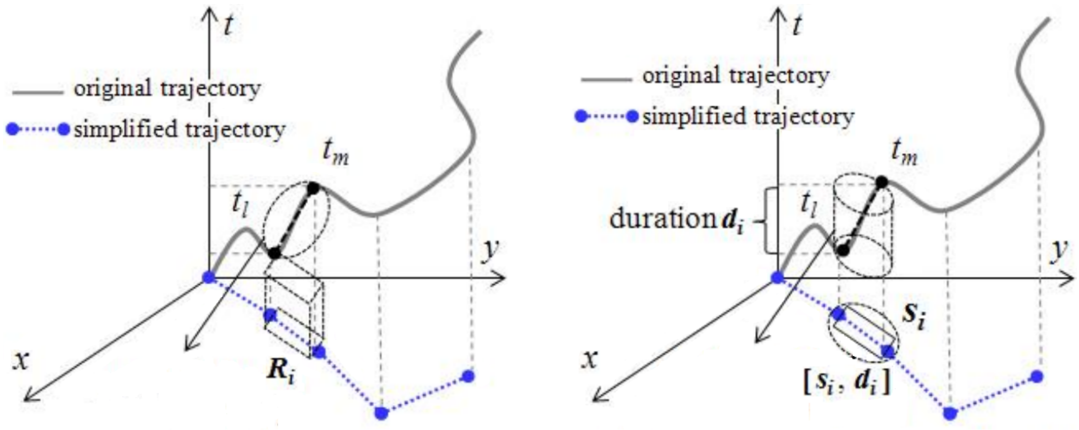


FIGURE 4.4: Alternative approaches of considering time in T-PM as illustrated in [92]: implicitly in the regions R_i (left) or explicitly in the items $[s_I, d_i]$ (right).

normalized and added to the resulting spatial regions. This is akin to extending the items to be mined with an additional dimension representing time, based on the duration of the corresponding segment in the database and using a temporal threshold parameter. More specifically, a *trajectory* is defined as follows:

Definition 4.3.8 (trajectory)
A trajectory is a sequence $ST =< [s_1, d_1][s_2, d_2]...[s_k, d_k] >$, *where* s_i *is the symbolic identifier of a spatial approximation, and* d_i *is the corresponding discrete temporal duration value, for* $0 \leq d_i < 1$.

Indoor spatial region S-PM. The work of Radaelli et al. in [151] is one of the first to highlight the lack of S-PM techniques not taking into account the specificities of indoor space, even when adapted to handle spatial data. It addresses this shortcoming by assuming presence sensor deployment in preselected indoor positions, and as a result is one of the very few works that consider leveraging topological information in the construction of candidate patterns.

More specifically, the authors define an *indoor trajectory* as follows:

Definition 4.3.9 (indoor trajectory)
An indoor trajectory is a time-ordered sequence of a moving object's tracking triples $(sensor_i, t_{si}, t_{ei})$, *extracted with the help of a splitting threshold* T_{split}, *where* $sensor_i$ *is the identifier of the positioning sensor that observed the moving object continuously from time* t_{s_i} *to time* t_{e_i}.

The splitting threshold T_{split} is used to determine based on the transition duration values $t_{s_{i+1}} - t_{e_i}$ whether two consecutive tracking triples belong to the same or to consecutive trajectories. Moreover, an *indoor movement n-pattern* and its *support* are then defined as follows:

Definition 4.3.10 (indoor movement n-pattern)
An indoor movement n-pattern is the sequence of the identifiers of the sensors

that detected the moving object: $< sensor_{i_1} sensor_{i_2}, ..., sensor_{i_n} >$, $n > 1$.

Definition 4.3.11 (indoor movement n-pattern support)

The support of an indoor movement pattern p is the sum of all the times it appears in each trajectory: $support(p, OTT, T_{split}) = \Sigma_{t \in TT(OTT, Tsplit)} OCC(p, t)$ *where* $TT(OTT, Tsplit)$ *is the trajectory table as results from applying the splitting threshold* T_{split} *over the Object Tracking Table OTT (i.e. raw positioning data), and* $OCC(p, t)$ *is the number of occurrences of pattern p in trajectory t.*

The proposed main mining method first obtains the frequent 1-patterns (individual sensor observations) and 2-patterns (transitions), and out of those creates longer candidates. The authors also try to account partially for multi-granular patterns, by extracting region-based patterns as simple spatial aggregations of the finer individual sensor-based patterns in the output. Moreover, they propose a weighted variant of support called *aggregate support*:

Definition 4.3.12 (indoor movement n-pattern aggregate support)

The aggregate support of an indoor pattern is defined as $support_A = support_D + (support_T - support_D)w$, *where* $support_D$ *represents the number of distinct moving objects with a trajectory that upholds the pattern,* $support_T$ *is the total number of times the pattern is upheld by all trajectories in the trajectory table* ($support_T \geq support_D$), *and* $w \in [0, 1]$ *is the controlling weight parameter.*

Thus, in the limit case of w=0 the aggregate support is incremented once per object, whereas in the opposite limit case of w=1 it is incremented once per trajectory. This mechanism proposed by Radaelli et al. is useful for addressing different application needs as long as there exist multiple trajectories per moving object, because sometimes the interest is on how many moving objects present a certain mobility behavior (e.g. a museum visitor getting lost), whereas other times on how often they do it (e.g. a museum visitor stopping in front of an exhibit).

Finally, like in many other works, their method ignores time completely and instead relies solely on sensor order information, but Radaelli et al. propose as interesting future work "to extend the representation of patterns to include temporal information (e.g., duration of the stay in a location) and adapt the mining process to deal with this extra information.", which makes part of our objective in chapter 5.

SPLITTER algorithm. In [197], Zhang et al. propose a way to derive places from a continuous space and then group them together according to their type, in an effort to mine only interesting patterns. To achieve this, their algorithm called *SPLITTER* equates trajectory semantics to the semantics of places, but instead of a fixed space partitioning, it follows a data-driven divide-and-conquer strategy.

Notably, SPLITTER groups all places by category and retrieves a set of so-called *coarse patterns*, each attached with a set of *trajectory snippets* that record the pattern's occurrences in the database. Then, SPLITTER splits a coarse pattern in a top-down divide-and-conquer manner, by clustering its attached trajectory snippets, and then extracting *fine-grained patterns* from the resulting dense (i.e. satisfying the support requirements) and spatially compact snippet clusters. Unqualified snippet clusters form disjoint communities, and small communities that cannot meet the minimum support are pruned to avoid unnecessary splitting.

This process reduces the search space of fine-grained patterns, because any one of those must have one and only one “parent” coarse pattern, and allows the discovery of trajectory patterns at different spatial granularity levels. As far as the temporal aspect of trajectories goes, the authors only consider a threshold-based temporal constraint with regards to the continuity of a trajectory pattern. In other words, they simply impose a Δt upper limit in the transition time values of every candidate pattern.

Object	Semantic Trajectory
o_1	<(p_3, 0), (**p_1**, 10), (**p_7**, 30), (**p_9**, 40)>
o_2	<(p_5, 0), (p_7, 30), (**p_2**, 360), (**p_7**, 400), (**p_{10}**, 420)>
o_3	<(p_3, 0), (p_6, 30)>
o_4	<(p_2, 0), (**p_1**, 120), (p_6, 140), (**p_8**, 150), (**p_{11}**, 180)>
o_5	<(p_{12}, 50), (p_8, 80), (p_{11}, 120), (p_4, 210)>

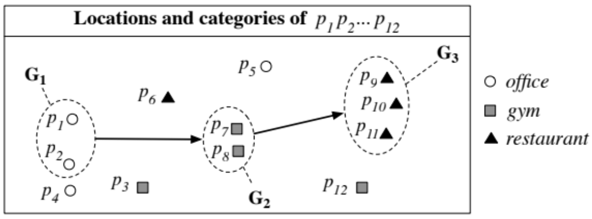


FIGURE 4.5: A fine-grained semantic trajectory pattern example according to the SPLITTER algorithm [83].

Geo-tagged photo trajectory data mining. As previously mentioned, [21, 31–34] focus on mining sequential trajectory patterns from geo-tagged photos. Having already examined the trajectory modeling approach used in this line of works in section 2.2.3.1, let us now look more closely at how they mine trajectory patterns.

In [31], Cai et al. adopt the T-PM framework proposed by Giannotti et al. in [71] and, through the necessary preprocessing, apply it on sequences of locations visited by Flickr photo takers to find patterns relating major cities to tourist hotspots. In [32], they improve its region extraction process while leaving the pattern mining step intact. Towards this end, they propose two algorithms to extract finer and arbitrary-shaped regions of interest. In [21], they extract Regions of Interest (RoIs) taking into account both space and time, and use this approach to discover seasonal visiting patterns in the same type of geo-tagged photo dataset. Finally, in [33], they transition to RoIs represented as tuples of the form $(ownerid, numberofpoints, \{time, latitude, longitude\}$ annotated with place semantics (e.g. hospital, pier, populated place). These are extracted by a hybrid grid-based process which uses a minimum support and a cell size parameter.

Hence, Cai et al. with their work in [34] proceed in mining trajectory behaviors with multiple place semantic dimensions (e.g. contextual information such as place type, weather condition, temporal information) from geo-referenced social media content. They achieve this by combining the PrefixSpan algorithm [141] with the BottomUpCube (BUC) algorithm [22], similar to [146] in turn inspired by [145] which first tried to use these two algorithms in a unified way.

More specifically, the authors distinguish between a *basic semantic trajectory pattern* whose elements are composed of basic geographic semantic annotations each describing the place type of a spatial grid cell, and a *multidimensional sequential pattern* whose elements are composed of the aforementioned basic semantics but also an arbitrary combination of some or all of them as additional semantics. Semantic RoIs are then formed by merging neighboring grid cells exhibiting the same basic place type, as illustrated in Figure 4.6. The authors also take into account transit time annotations, indicating the time intervals between consecutive places in a sequence.

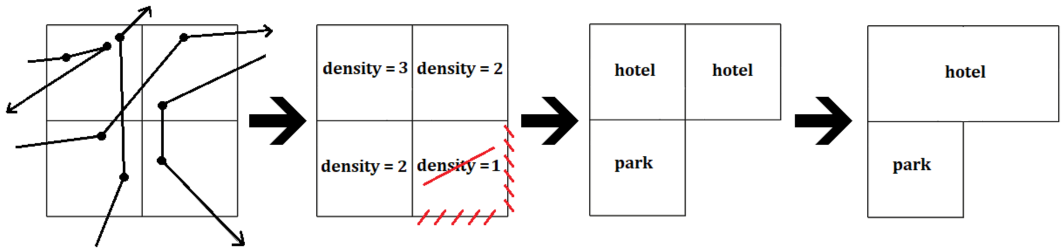


FIGURE 4.6: The merging of grid cells is used to transition from geometric trajectories to semantic RoIs in the approach of [34].

To mine both the semantic and temporal aspects, in [33], Cai et al. propose a new type of sequence containment:

Definition 4.3.13 (dimensional and τ-containment)
A SemT-pattern $(SemS, A) = SemA_0 \xrightarrow{a_1} ... \xrightarrow{a_k} SemA_k$ *is contained in a semantic trajectory* $SemT =< (SemA_0, t_0), ..., (SemA_n, t_n) >$ $((SemS, A) \preceq_{d,\tau} SemT)$ *iff* $\exists$ *a subsequence* $SemT' =< (SemA'_0, t'_0), ..., (SemA'_n, t'_n) >$ *of* $SemT$ *such that:*

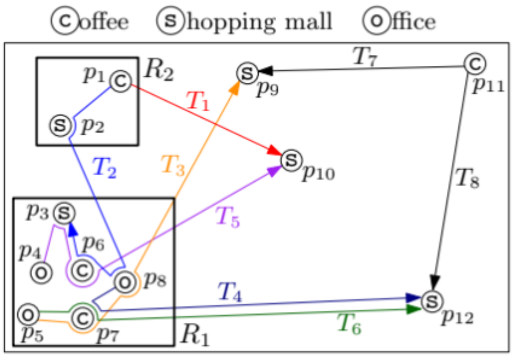
1. $\forall_{0 \leq j \leq k}$: $e_j = e'_j$, $V_j \subseteq V_j$ $SemS \preceq_d SemT'.sequence\_of\_SemA$
2. $\forall_{1 \leq j \leq k}$: $|\alpha_j - \alpha'_j| \leq \tau$ *where* $\forall_{1 \leq k \leq n}$: $\alpha'_j = t'_j - t'_{j-1}$

Whereas for the matching of two time intervals, the gap between them should be smaller than a given tolerance threshold, which is exactly the approach used in the MiSTA algorithm of [69].

Regional pattern mining. [41] proposes a pattern growth strategy-based algorithm for mining sequential so-called *regional* patterns, which are simply sequences of PoI categories. It makes use of a DBSCAN-like clustering technique to find sets of PoIs that play the role of items in the sequential patterns. However, it does not take time into account and semantics are only limited to place categories.

4.3.2 Sequential Pattern Mining

An important part of extracting interesting information from information repositories or databases is discovering patterns in sequences. This process known as Sequential Pattern Mining (S-PM), holds both predictive and descriptive explanatory power as long as the input data are indeed of a sequential nature, as opposed to having been forced into such form. This is why S-PM has been successfully applied in genome analysis, web usage analysis, product sales, alarm data analysis, and other fields.



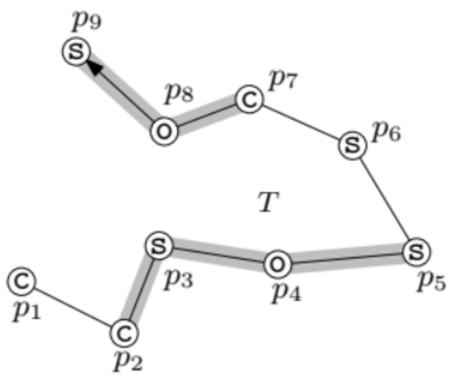


FIGURE 4.7: An example of semantic trajectories (left) and regional patterns (right) according to [41].

Quite fittingly, trajectory data are intrinsically sequential, because at their core they represent successive location information about a moving object. This section looks at the related work from the non-trajectory S-PM research field, in an attempt to identify ideas, mechanisms, and methods transferable to the trajectory data domain, and useful for mining semantic indoor trajectories and especially museum visitor trajectories.

4.3.3 The Standard Problem and Algorithms

The classical problem of S-PM was first defined in [7] as follows:

Definition 4.3.14 (The Sequential Pattern Mining (S-PM) Problem)

Given a database of sequences, where each sequence is a list of transactions ordered by transaction-time, and each transaction is a set of items, discover all sequential patterns with a user-specified minimum support.

Definition 4.3.15 (pattern support)

The support of a pattern is the number of data-sequences that contain the pattern.

Let us assume a set of symbolic items $I = \{i_1, i_2, ..., i_m\}$. These items are the main data components of interest to the analysis and will thus compose the output patterns, irrespective of what they represent in the real world. An itemset X is simply a set of items i.e. a subset of I: $X \subset I$. A sequence $s_X =< X_1, X_2, ..., X_n >$ is an ordered list of itemsets $X_k \subseteq I$, $1 \leq k \leq n$. A sequence database $S_{db} = \{s_1, s_2, ..., s_n\}$ is simply a set of such sequences which represents the analysis input data. Finally, given two sequences $s_X =< X_1, X_2, ..., X_n >$ and $s_Y =< Y_1, Y_2, ..., Y_m >$, we say that s_X is contained in s_Y ($s_X \preceq s_Y$) iff $\exists 1 \leq i_1 \leq i_2 < ... < i_n \leq m$ such that $X_1 \subseteq Y_{i_1}$, $X_2 \subseteq Y_{i_2}$, ..., $X_{i_n} \subseteq Y_{i_n}$, in which case s_X is also said to be a *subsequence* of s_Y.

Based on the above definitions, the S-PM problem lies in finding subsequences that occur frequently "enough" in a given database of sequences, i.e. that appear as subsequences of a large percentage of database sequences. More formally, it is the problem of finding all subsequences s whose support $sup(s) = |s'|s \preceq s', s \in S_{db}|$ surpasses a given threshold value sup_{min} called the *minimum support*: $sup(s) \geq sup_{min}$

4.3.4 Sequential Pattern Mining and Multiple Dimensions

In [145], Pinto et al. propose what can be arguably considered as the first Multi-dimensional Sequential Pattern Mining (MD-S-PM) algorithm, by integrating S-PM methods with multidimensional analysis methods. The result is the proposal of three pattern mining methods named $UniSeq$, $Dim - Seq$, $Seq - Dim$, with the latter being the most efficient one.

Unfortunately, all three methods concern datasets where the sequential part and the multidimensional part are practically independent, as the multidimensional information is practically "appended" to the sequence and not integral part of it. Thus, a multi-dimensional sequence takes the form: $(\alpha_1, ..., \alpha_m, s)$ where $\alpha_i \in (A_i \cup \{*\})$ are the dimension values which "extend" the actual sequence s. To put it more simply using an original example from [145], the sequence $< (business, Boston, middle)(bd)cba) >$ is interpreted simply as the sequence $< (bd)cba >$ with the additional multidimensional information $(business, Boston, middle)$ remaining the same throughout the sequence, and thereby characterising all of it. As a consequence, all three methods use PrefixSpan for the sequential pattern mining part and a BUC-like algorithm for the multidimensional mining part.

The $HYPE$ algorithm [148] extends the approach of the M^2SP algorithm [146] by taking into account hierarchical relations within each sequence. These relations are materialized as taxonomies. In specific, $HYPE$ extracts patterns associated to the lowest possible hierarchical levels, under the assumption that the most specific patterns are typically the most relevant and informative ones. To achieve this, $HYPE$ uses a partitioning of the data dimensions $\mathcal{D}$ into three subsets: 1) a *time* dimension $\mathcal{D}_\sqcup$ associated with a totally ordered domain according to which sequences are constructed, 2) a set of (potentially hierarchical) *analysis* dimensions $\mathcal{D}_\mathcal{A}$, tuples over which form the actual sequential pattern items, and 3) a set of *reference* dimensions $\mathcal{D}_\mathcal{R}$, tuples over which partition the table into a set of blocks $\mathcal{B}(\mathcal{D}_\mathcal{R})$ used for counting the support of candidate sequences in a more efficient way. An example of how the domain of the analysis dimensions may be instantiated in the context of international product orders can be appreciated in Figure 4.8.

The M^3SP algorithm [147] refines $HYPE$'s approach. First, it additionally defines a set of "ignored" dimensions $\mathcal{D}_\mathcal{I}$ to separate all dimensions not used for structuring the sequences nor in the mining process. It also defines a hierarchical pattern inclusion which considers $< \{(France, wine)\}, \{(Germany, beer)\} >$ as a subsequence of $< \{(France, Alcoholic\_drinks), (USA, drinks)\}, \{(EU, Alcoholic\_drinks)\} >$.

At the core of the M^3SP algorithm lies the notion of *item specificity* $\preceq_I$:
α=(d_1, ..., d_m) $\preceq_I$ α'=(d'_1, ..., d'_m) iff $\forall_{1 \leq i \leq l}$: $d'_i \in d_i^{\downarrow}$
where:
$d_i^{\downarrow}$ is the set of specializations of d_i

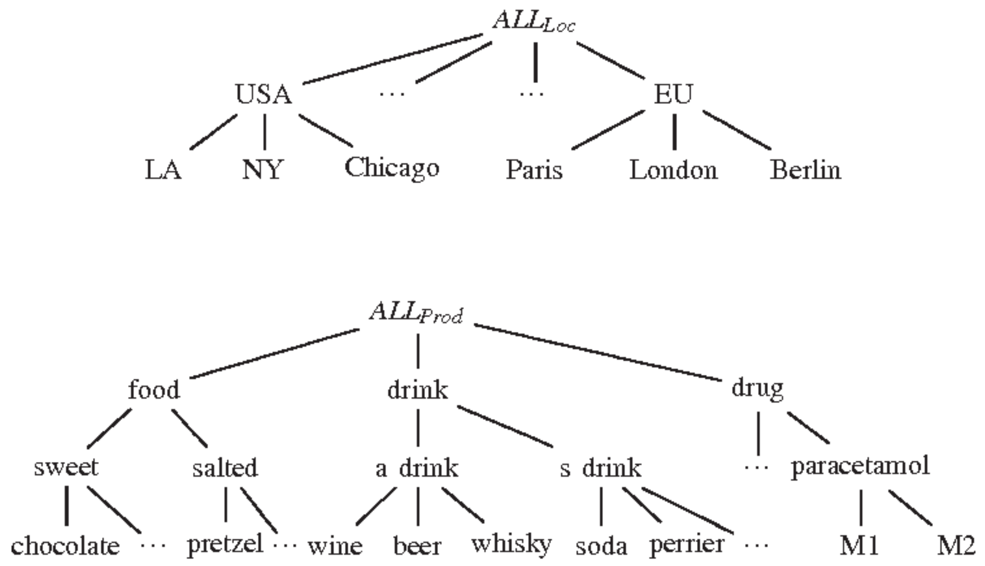


FIGURE 4.8: Example of a hierarchical tree over a "Location" dimension and over a "Product" dimension, taken from [147].

α=(d_1, ..., d_m) and α'=(d'_1, ..., d'_m) are multidimensional items defined over the analysis dimensions $D_i \in \mathcal{D}_{\mathcal{A}}$, $i = 1, ..., m$: $d_i, d'_i \in Dom(D_i)$. In this case, item α' is said to be more specific than item α.

Consequently, a block $B \in \mathcal{B}(\mathcal{D}_{\mathcal{R}})$ supports the sequence $\varsigma = <_1, ..., s_l>$ of itemsets $s = \{\alpha_1, ..., \alpha_k\}$ (i.e. non-empty sets of multidimensional items) if $\exists t_1, ..., t_l \in Dom(D_t)$ such that: 1) $t_1 < ... < t_l$, and 2) $\forall \alpha \in s_i, i = 1, ..., l$: $\exists (t_i, a) \in B$ such that $\alpha \preceq_I a$. Put simply, B supports ς if for each of ς's items it contains (in the proper order) a tuple that is at least as specific as that item. Based on this, a sequence ς of itemsets is frequent if the percentage of blocks supporting it (over the total number of blocks $|\mathcal{B}(\mathcal{D}_{\mathcal{R}})|$) is greater than or equal to the minimum support threshold: *sup(*ς*)* $\geq$ *minsup*.

Moreover, the notion of item specificity $\preceq_I$ gives rise to two other partial orderings, that of itemset specificity $\preceq_{IS}$ and that of sequence specificity $\preceq_S$.

Concerning the latter, a sequence $\varsigma' = <s'_1, ..., s'_{l'}>$ is more specific than another sequence $\varsigma = <s_1, ..., s_l>$ ($\varsigma \preceq_S \varsigma'$) if $\exists 1 \leq i_1 \leq i_2 \leq ... \leq i_l \leq l'$: $s_1 \preceq_{IS} s'_{i_1}$, $s_2 \preceq_{IS} s'_{i_2}$, ..., $s_l \preceq_{IS} s'_{i_l}$.

All three types of specificity are illustrated using examples within the context of international product sales taken from [147]:
1) With respect to item specificity:
$(USA, drink) \preceq_I (USA, soda)$ because $USA \in USA^{\downarrow}$ and $soda \in drink^{\downarrow}$
It can be noticed that not every two items are comparable with respect to $\preceq_I$. For example, $(Paris, wine)$ and $(USA, soda)$ are not comparable because, according to the hierarchy of the "Location" dimension, neither $Paris \in USA^{\uparrow}$ nor $Paris \in USA^{\downarrow}$ holds.

2) With respect to itemset specificity:
$\{(EU, wine)\} \preceq_{IS} \{(Paris, wine)\}$ because $(EU, wine) \preceq_I (Paris, wine)$.
$\{(EU, wine)\} \preceq_{IS} \{(Paris, wine), (USA, soda)\}$ again because $(EU, wine) \preceq_I (Paris, wine)$.
$\{(Paris, wine), (USA, alcoholic\_drink)\} \preceq_{IS} \{(Paris, wine), (Chicago, beer)\}$ because $(Paris, wine) \preceq_I (Paris, wine)$ and $(USA, alcoholic\_drink) \preceq_I (Chicago, beer)$. It is important to notice that an itemset can not include items that are comparable with respect to $\preceq_I$; for example, the set $\{(Paris, wine), (EU, alcoholic\_drink)\}$ does not constitute an itemset because $(EU, alcoholic\_drink) \preceq_I (Paris, wine)$.
3) With respect to sequence specificity:
$<\{(EU, wine)\}> \preceq_S <\{(Paris, wine)\}>$ because $\{(EU, wine)\} \preceq_{IS} \{(Paris, wine)\}$.
$<\{(EU, wine)\}, \{(EU, beer)\}> \preceq_S <\{(Paris, wine), (USA, soda)\}, \{(Berlin, beer)\}>$ because $\{(EU, wine)\} \preceq_{IS} \{(Paris, wine), (USA, soda)\}$ and $\{(EU, beer)\} \preceq_{IS} \{(Berlin, beer)\}$.
$<\{(Paris, alcoholic\_drink)\}> \preceq_S <\{(Paris, wine), (USA, drink)\}, \{(Berlin, beer)\}>$ because $\{(Paris, alcoholic\_drink)\} \preceq_{IS} \{(Paris, wine), (USA, drink)\}$.

Furthermore, M^3SP constructs the sequences to be mined based on *maximal atomic frequent (maf)* sequences, which are defined as atomic sequences $<\{\alpha\}>$ that are frequent and for which there exists no other frequent atomic sequence α' such that $\alpha \preceq_I \alpha'$ (i.e. that is more specfic). M^3SP actually consists of a two-step algorithmic process:

1. mining maf-sequences
2. transforming the dataset (using the maf sequences) and mining frequent sequences (using standard algorithms)

It makes use of the following propositions:

- $\alpha \preceq_I \alpha' \Rightarrow sup(<\{\alpha'\}>) \leq sup(<\{\alpha\}>)$
- $\varsigma \preceq_S \varsigma' \Rightarrow sup(\varsigma') \leq sup(\varsigma)$ (anti-monotonic support of sequences)
- $\forall \alpha = (d_1, ..., d_m) \in Dom(\mathcal{D}_{\mathcal{A}})$:
  $succ(\alpha) = \{(d'_1, ..., d'_m) \mid \exists i \in \{1, ..., m\} : d'_i \in down(d_i) \text{ } and \text{ } (\forall j \neq i : d'_j = d_j)\}$ (set of direct successors of α)
  $gen(\alpha) = \{(d'_1, ..., d'_m) \mid \exists i \in \{\rho(\alpha), ..., m\} : d'_i \in down(d_i) \text{ } and \text{ } (\forall j \neq i : d'_j = d_j)\}$ (set of multidimensional items generated from α)
  where:
  $*$ $down(x)$ denotes the set of all direct specializations of x (the set of all y in $Dom(D_i)$ such that H_i contains an edge from x to y, or the empty set in case x is a leaf)
  $*$ $\rho(\alpha) \in \{0, 1, ..., m\} : d_{\rho(alpha)} \neq ALL_{\rho(\alpha)} \text{ } and \text{ } (\forall j > \rho(\alpha) : d_j = ALL_j)$
  $*$ if $\rho(\alpha) = 0$ then $succ(\alpha) = gen(\alpha)$ (since $\alpha = ALL_{\mathcal{A}}$)
- $\bigcup_{\alpha \in Dom(\mathcal{D}_{\mathcal{A}})} gen(\alpha) = Dom(\mathcal{D}_{\mathcal{A}}) \setminus \{ALL_{\mathcal{A}}\}$
  $\forall \alpha, \alpha' \in Dom(\mathcal{D}_{\mathcal{A}}) : \alpha \neq \alpha' \Rightarrow gen(\alpha) \cap gen(\alpha') = \emptyset$ (for tree hierarchies)
  (i.e. all multidimensional items except $ALL_{\mathcal{A}}$ can be generated in a non-redundant manner)

As illustrated in Figure 4.9, Step 1 consists of a Depth-First traversal of the set $Dom(\mathcal{D}_{\mathcal{A}})$ starting from the most general item $ALL_{\mathcal{A}} = (ALL_1, ..., ALL_m)$ and following a strategy similar to the BUC algorithm proposed in [22]. Namely, for each atomic sequence $<\{\alpha\}>$ all the direct successors $succ(\alpha)$ in the lattice defined by the partially ordered set $(Dom(\mathcal{D}_{\mathcal{A}}), \preceq_I)$ of its multidimensional element α are generated once, and their support is computed against all tuples that are more specialized than α.

Step 2 can then use any typical S-PM algorithm (in [147] the authors applied SPADE [196]) in order to compute non-atomic frequent sequences, because for any multidimensional item α of any candidate sequence ς, $<\{\alpha\}>$ is known to be a maf sequence. Therefore, all the components needed to construct candidate sequences have already been found in Step 1. More specifically, in Step 2 a transformation of the database takes place wherein, each maf sequence $<\{\alpha\}>$ is assigned a unique $id(\alpha)$ and becomes the equivalent of an item in standard S-PM (Figure 4.9), whereas each block $B \in \mathcal{B}(\mathcal{D}_{\mathcal{R}})$ is assigned a unique $ID(B)$ and becomes the equivalent of a sequence $\varsigma(B) = <(\mu_{1,1}, ..., \mu_{1,n_1}), ..., (\mu_{p,1}, ..., \mu_{p,n_p})>$ of itemsets in standard S-PM (right part of figure 4.10) where $\forall j \in \{1, ..., p\}, \forall k \in \{1, ..., n_j\} : \mu_{j,k} = id(\alpha_k)$. Therefore, a standard S-PM algorithm will treat all $\varsigma(B)$ as typical input sequences (unaware that they encode multidimensional information) and will result in transformed frequent sequences, as exemplified in Figure 4.11.

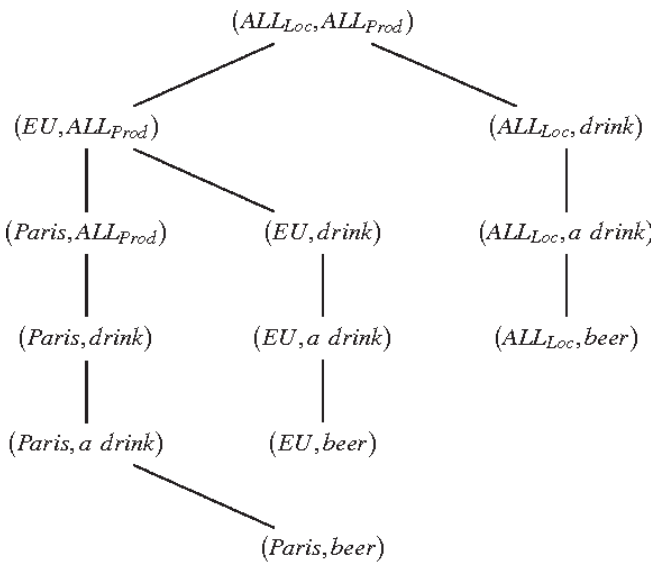


FIGURE 4.9: Example of a tree of generated multidimensional items taken from [147].

4.3.5 Sequential Pattern Mining and the Temporal Dimension

In section 4.3.1 the *t-patterns* algorithm proposed by Giannotti et al. for mining temporally annotated trajectory data is reviewed. $t-patterns$ is actually based upon the (non-trajectory) S-PM algorithm proposed by the same authors in [69]

Maf-sequence $\langle\{a\}\rangle$	$id(a)$
$\langle\{(USA, ALL_{Prod})\}\rangle$	1
$\langle\{(Berlin, ALL_{Prod})\}\rangle$	2
$\langle\{(EU, Pretzel)\}\rangle$	3
$\langle\{(EU, Al.\ Drinks)\}\rangle$	4
$\langle\{(EU, M2)\}\rangle$	5

Example of assigning identifiers to maf sequences taken from [147].

$ID(B)$	$\varsigma(B)$
1	$\langle(2,3,4),(5),(4),(1)\rangle$
2	$\langle(1),(2,3,4),(5)\rangle$
3	$\langle(3,4),(2,5)\rangle$
4	$\langle(1),(2),(1)\rangle$

Example of assigning identifiers to blocks taken from [147].

FIGURE 4.10: Example of the database transformation in the proposal of Plantevit et al. [147].

Transformed frequent sequences	Frequent multidimensional sequences
$\langle(3,4),(5)\rangle$	$\langle\{(EU, pretzel), (EU, a\_drink)\}, \{(EU, M2)\}\rangle$
$\langle(1)\rangle$	$\langle\{(USA, ALL_{Prod})\}\rangle$
$\langle(2)\rangle$	$\langle\{(Berlin, ALL_{Prod})\}\rangle$

FIGURE 4.11: Example of the output of a standard S-PM process run over the transformed database taken from [147].

and called $MiSTA$. $MiSTA$ is actually - to our knowledge - the first algorithm to introduce temporal annotations into the mining step. These annotations characterize the duration of the transitions between consecutive items in the sequences. In chapter 6, $MiSTA$ is applied over the Louvre dataset to discover unknown visiting patterns, and in chapter 5 it is used as a component of a novel T-PM proposal. In the remainder of this section, $MiSTA$'s function is explained in detail.

First, a *Temporally Annotated Sequence (TAS)* is defined as follows:

Definition 4.3.16 (temporally annotated sequence)

A TAS is a couple $T(\bar{s}, \bar{a}) = s_0 \xrightarrow{a1} s_1 \xrightarrow{a2} \ldots \xrightarrow{a_n} s_n$ *where* $\bar{s} =< s_0, s_1, \ldots, s_n >$ *is a sequence of itemsets, and* $\bar{a} =< a_1, a_2, \ldots, a_n >$ *is a sequence of temporal annotations corresponding to the transitions of* $\bar{s}$*.*

In simple terms, a TAS can be considered as a special form of event sequence that includes the transition times between consecutive events.

In terms of its function, MiSTA accepts two parameters: a typical minimum support value s_{min}, and a temporal threshold value τ which specifies the maximally allowed temporal annotation difference for determining whether a pattern occurs in an input TAS or not. Thus, the notion of $\tau - containment$ is introduced:

Definition 4.3.17 (τ-containment)

An n-long TAS $T_1 = (\bar{s}_1, \bar{a}_1)$ *is* $\tau - contained$ *in an m-long TAS* $T_2 = (\bar{s}_2, \bar{a}_2)$ $(n \leq m)$ $(T_1 \preceq_\tau T_2)$ *iff* $\exists 0 \leq i_0 < \ldots < i_n \leq m$ *such that:*

1. $\forall_{0 \leq k \leq n}$: $s_{1,k} \subseteq s_{2,i_k}$

2. $\forall_{1 \le k \le n}$*:* $|a_{a,k} - a_{*,k}| \le \tau$ *where* $a_{*,k} = \Sigma_{i_{k-1} < j \le i_k} a_{2,j}$

As indicated by the use of two different indices k and i_k in the second condition of the above definition, the itemsets of T_1 are considered to be consecutive whereas the corresponding itemsets of T_2 are not necessarily consecutive. In case they are not, all transition times of T_2 that correspond to a single transition of T_1 are summed up before comparing against time threshold τ. In other words, a single annotation of T_1 may match multiple annotations of T_2 added together.

For example, if $T_1 = \{a\} \xrightarrow{2} \{b\} \xrightarrow{6} \{c\}$ and $T_2 = \{a,e\} \xrightarrow{1} \{d\} \xrightarrow{1} \{b,f\} \xrightarrow{5} \{c,f\}$ with $\tau = 1.5$, it holds that T_1 is $\tau - contained$ in T_2, because T_1's pattern $\{a\} \xrightarrow{2} \{b\}$ actually matches perfectly T_2's pattern $\{a,e\} \xrightarrow{1} \{d\} \xrightarrow{1} \{b,f\}$ if the latter's intermediate itemset $\{d\}$ is skipped and the two annotations from either side of it are added together.

Moreover, T_1 is *strictly* τ*-contained* in T_2 ($T_1 \prec_\tau T_2$) if the second condition of Definition 4.3.17 holds with a strict inequality, whereas T_1 is *exactly contained* in T_2 ($T_1 \prec_0 T_2$) if it holds with $\tau = 0$ (in which case containment becomes transitive).

Extending Definition 4.3.17, a TAS T_1 is $\tau - contained$ in a set of TASs D[3] ($T_1 \preceq_\tau D$) iff $\exists T_2 \in D$ such that $T_1 \preceq_\tau T_2$. Based on this, Giannotti et al. introduce the notion of *tau-support* and correspondingly that of a *frequent TAS* in relative terms:

Definition 4.3.18 (τ-support)
The τ*-support of a TAS* T *w.r.t. a set of TASs* D *is equal to* $\tau - support(T) = \frac{|\{T^* \in D | T \preceq_t T^*\}|}{|D|}$

Definition 4.3.19 (frequent TAS)
Given a time threshold τ *and a minimum support threshold* $s_{min} \in [0,1]$*, a TAS* T *is frequent if* τ*-support*$(T) \ge s_{min}$*.*

As expected, a frequent sequence $\bar{s}$ may not correspond to any frequent TAS T if its occurrences in the database are accompanied by highly dispersed annotation values. Simply put, if $\bar{s}$ alone is frequent but has no typical transition times, then $(\bar{s}, \bar{a})$ will probably not be τ-frequent. Interestingly, multiple instances of the same sequence $\bar{s}$ may be τ-contained in a TAS T, with completely different temporal annotations.

As far as finding the frequent TASs goes, Giannotti et al. originally used a two (independent) step mining process in [70], the second step being dedicated to handling the annotations, but $MiSTA$ [69] improves this algorithmic approach, by extending the prefix-projection-based method of PrefixSpan [141]. More specifically, it initializes an evolving set of projections in the form of so-called *T-sequences*. These carry complete information about all useful occurrences of a prefix in the projected sequence. Then, MiSTA recursively performs either *enlargement* (i.e. adding a new element to the last item of the prefix) or *extension* (i.e. adding a new element to the prefix) projections. The recursive usage of these two operations serves to generate all sub-projections of an actual projection.

[3]The authors also refer to D as set of *transactions* according to classic pattern mining terminology.

As already mentioned, the mining mechanism of database projection is not a $MiSTA$ novelty. In general, a projection $D|a$ of the initial dataset D w.r.t. the atomic sequence a as prefix, is simply a set which contains the sequences of D where a appears and where all items up until and including the first appearance of a have been removed. The difference in the case of $MiSTA$ is of course that this projection mechanism needs to be extended with temporal information as well:

Definition 4.3.20 (T-sequence)
A T-sequence of a projected sequence S is a couple (S, A) where:

- $S = S_0|_{s*} =< (s_1, t_1), ..., (s_n, t_n)) >$ *is a temporal (i.e. time-stamped) sequence obtained as a projection of the sequence S_0 w.r.t. prefix $s*$*
- $A =< (a_1, e_1), ..., (a_m, e_m) >$ *is an annotation sequence*
- (a_i, e_i) *represents an occurrence of prefix $s*$ in the original sequence S_0*
- a_i *is the sequence of timestamps of the occurrence*
- e_i *is a pointer to the element of S where the occurrence terminates (or $\emptyset$ if there is no such element in S)*

Given a sequence S, each input TAS can be mapped to a set of annotations corresponding to all possible occurrences of S in that transaction. The resulting $T-$sequences then incorporate the information necessary to find all possible occurrences of S and the exact point within the sequence where each one is concluded.

To use an example from [69], let us assume the following temporal sequence: $S =< (\{a\}, 1), (\{a, b\}, 2), (\{b, c\}, 3), (\{a\}, 4)) >$
Then, the T-sequence of S w.r.t. prefix a is the couple $(S|_a, A|_a)$ where:
$S|_a =< (\{a, b\}, 2), (\{b, c\}, 3), (\{a\}, 4)) >$
$A|_a =< (< 1 >, \emptyset), (< 2 >, \rightarrow 2), (< 4 >, \rightarrow 4) >$
Obviously, $S|a$ is the same as S except for the first itemset having been removed since it contains the first occurrence of the prefix a. Concerning $A|_a$, its first item $(< 1 >, \emptyset)$ denotes that the prefix occurs in the first item of S and has already been erased from the projection, its second item $(< 2 >, \rightarrow 2)$ denotes that the prefix also occurs in the second item of S and ends there, and its third item $(< 4 >, \rightarrow 4)$ denotes that it also occurs in the fourth item of S and ends there. It should be noted that, despite the fact that the annotation sequences A contain timestamps, $MiSTA$ only takes the temporal durations between those timestamps into account.

Then, $S|_a$ can be further projected, for instance w.r.t. to prefix b resulting in the couple $(S|_{ab}, A|_{ab})$ where:
$S|_{ab} =< (\{b, c\}, 3), (\{a\}, 4)) >$
$A|_{ab} =< (< 1, 2 >, \emptyset), (< 1, 3 >, \rightarrow 3), (< 2, 3 >, \rightarrow 3) >$
The two sequences $S|_{ab}$, $A|_{ab}$ can be similarly interpreted as before, with $A|_{ab}$ this time containing the information about all occurrences of the sequence $a \rightarrow b$ in S.

During $MiSTA$'s execution more generally, as projections are recursively performed and the prefix becomes longer, $S|_{prefix}$ becomes shorter, whereas $A|_{prefix}$ becomes longer and eventually contains all the information necessary to retrieve all occurrences of the final prefix in the original TAS dataset.

Algorithm: MiSTA
Input: A dataset D_{in} of time-stamped sequences, a minimum support s_{min}, a temporal threshold τ
Output: A set of couples $(S, \mathcal{D}^*)$ of sequences with annotations

1. $L = 0$, $\mathcal{P}_0 = \{D_{in} \times \{\langle\rangle\}\}$; *//Empty annotations*
2. **while** $\mathcal{P}_L \neq \emptyset$ **do**
3. $\mathcal{P}_{L+1} = \emptyset$;
4. **for each** $P \in \mathcal{P}_L$ **do**
5. **if** $P.length \geq 2$ **then**
6. $\mathcal{A}$= Extract_annotation_blocks(P);
7. $\mathcal{D}$= Compute_density_blocks($\mathcal{A}$);
8. $\mathcal{D}^*$ = Coalesce_density_blocks($\mathcal{D}$);
9. P^* = Annotation-based_prune(P, $\mathcal{D}^*$);
10. Output ($P.prefix$, $\mathcal{D}^*$);
11. **else** $P^* = P$; *//No annotations, yet*
12. **for each** item $i \in P^*$ **do**
13. **if** $P^*.enlarge\_support(i) \geq s_{min}$ **then**
14. $\mathcal{P}_{L+1} = \mathcal{P}_{L+1} \cup \{$ enlarge_proj(P^*, i) $\}$;
15. **if** $P^*.extend\_support(i) \geq s_{min}$ **then**
16. $\mathcal{P}_{L+1} = \mathcal{P}_{L+1} \cup \{$ extend_proj(P^*, i) $\}$;
17. L++;

Algorithm: Extract_annotation_blocks(P)
Input: A projection P
Output: A set of hyper-rectangles, representing the influence areas of each T-sequence in P.

1. $\mathcal{A} = \emptyset$;
2. **for each** T-sequence $T = (S, A) \in P$ **do**
3. $\mathcal{A}_T = \emptyset$;
4. **for each** annotation $(a, e) \in A$ **do**
5. Derive annotation $\bar{\alpha}$ from time-stamps a;
6. h = hyper-cube with center $\bar{\alpha}$ and edge 2τ;
7. Merge h with $\mathcal{A}_T$;
8. Partition $\mathcal{A}_T$ into a set of hyper-rectangles $\mathcal{A}'$;
9. $\mathcal{A} = \mathcal{A} \cup \mathcal{A}'$;
10. **return** $\mathcal{A}$;

Algorithm: enlarge_proj(P,i)
Input: A projection P and an item i
Output: A projection of P w.r.t. i

1. $P' = \emptyset$;
2. **for each** T-sequence $T = (S, A) \in P : i \in T$ **do**
3. $S' = S|_i$ and $A' = \langle\rangle$;
4. **for each** annotation $(a, e) \in A$ **do**
5. **if** e points to element $(s, t) \in S$ and $i \in s$
6. **then** $A' = append(A', (a, e))$;
7. $P' = P' \cup \{(S', A')\}$;
8. **return** P';

Algorithm: Compute_density_blocks($\mathcal{A}$)
Input: A set of hyper-rectangles in $\mathbf{R}^d$
Output: A set of hyper-rectangles and their density.

1. $\mathcal{D} = \emptyset$;
2. Recursive_density($\mathcal{A}, d, \langle\rangle, \mathcal{D}$);
3. **return** $\mathcal{D}$;

Algorithm: Recursive_density($\mathcal{A}, d, \hat{H}, \mathcal{D}$)

1. $B = \{x | [l_1, h_1] \times \cdots \times [l_n, h_n] \in \mathcal{A}, x \in \{l_d, h_d\}\}$;
2. $\hat{B}$ =sorted_sequence(B);
3. **for**$(i = 1; i < |\hat{B}|; i++)$ **do**
4. $\mathcal{A}_i = \{[l_1, h_1] \times \cdots \times [l_n, h_n] \in \mathcal{A} \mid [l_d, h_d] \cap [\hat{B}_i, \hat{B}_{i+1}] \neq \emptyset\}$;
5. **if** $|\mathcal{A}_i| \geq s_{min}$ **then**
6. $\hat{H}' = append((l_d, h_d), \hat{H})$;
7. **if** d=1 **then**
8. $\hat{h} = [l_1, h_1] \times \cdots \times [l_n, h_n]$ given that $\hat{H}' = \langle (l_1, h_1), \ldots, (l_n, h_n) \rangle$;
9. $\mathcal{D} = \mathcal{D} \cup \{\hat{h}\}$;
10. $\hat{h}.density = |\mathcal{A}_i|$;
11. **else**
12. Recursive_density($\mathcal{A}_i, d-1, \hat{H}', \mathcal{D}$);

Algorithm: Coalesce_density_blocks($\mathcal{D}$)
Input: A set of dense hyper-rectangles
Output: A sequence of hyper-rectangles, covering the same volume as $\mathcal{D}$ but yielding a better approximation series.

1. $S = \emptyset$;
2. **while** $\mathcal{D} \neq \emptyset$ **do**
3. Select a random $h \in \mathcal{D}$ and let $\mathcal{D} = \mathcal{D} - \{h\}$;
4. **for each** extension direction dir for h **do**
5. V_{dir} = volume of hyper-rectangle obtained by extending h along direction dir;
6. $\mathcal{D}_{dir}$ = set of hyper-rectangles of $\mathcal{D}$ covered by the extension along dir;
7. **if** one extension was found **then**
8. $ext = \arg\max_{dir} V_{dir}$;
9. $h = h \cup \left(\bigcup_{h' \in \mathcal{D}_{ext}} h'\right)$;
10. $\mathcal{D} = \mathcal{D} - \mathcal{D}_{ext}$;
11. **goto** step 4.
12. $S = append(S, h)$;
13. **return** $\mathcal{A}$;

Algorithm: extend_proj(P,i)
Input: A projection P and an item i
Output: A projection of P w.r.t. i

1. $P' = \emptyset$;
2. **for each** T-sequence $T = (S, A) \in P : i \in T$ **do**
3. $S' = S|_i$ and $A' = \langle\rangle$;
4. **for each** annotation $(a, e) \in A$ **do**
5. **for each** $(s, t) \in S$ s.t. $i \in s \ \wedge \ t > e$ **do**
6. $A' = append(A', (append(a, t), \rightarrow t))$;
7. $P' = P' \cup \{(S', A')\}$;
8. **return** P';

FIGURE 4.12: The $MiSTA$ algorithm as presented in [69].

In Figure 4.12 the $MiSTA$ algorithm from [69] is presented in its entirety. As seen from the pseudocode of Figure 4.12, $MiSTA$ takes as input a dataset of time-stamped sequences $\mathcal{D}_{in}$, a minimum support s_{min}, a temporal threshold τ, and produces as output a set of couples $(S, \mathcal{D}^*)$ of sequences with annotations. First, it initializes a set of projections $\mathcal{P}_L$ that contains all of the input temporal sequences $\mathcal{D}_{in}$ but no annotation sequences yet (lines 1-3). Then, for each projection in the evolving set of projections $P \in \mathcal{P}_L$ (line 4), it performs the following tasks:

1. It handles its annotations (lines 5-11)
2. It generates all of its sub-projections (lines 12-16)

To handle the annotations, $MiSTA$ does the following:

- It extracts annotations from the projection by scanning all annotation sequences (line 6)
- It computes their hyper-cubical areas of influence (line 6)
- It combines those areas thereby partitioning the space of annotations into hyper-rectangles of homogeneous density (line 7)
- It merges the hyper-rectangles together to maximize a quality criterion (line 8)
- It outputs the condensed annotations (line 8)
- It filters the annotation sequences by removing all occurrences whose area of influence is not of any use for computing dense annotations (line 9)

To generate all sub-projections, $MiSTA$ does the following for each item i in the filtered projection P^* (line 12):

- If the support of item i within the projection is greater or equal to s_{min} when counting only occurrences of i that can be used in an enlargement projection, then an enlargement projection takes place (lines 13-14)
- If the support of that item within the projection is greater or equal to s_{min} when counting only occurrences of i that can be used in an extension projection, then an extension projection takes place (lines 15-16)

Given that chapter 6 will be using $MiSTA$ for the algorithmic proposal, let us explain in more detail how the aforementioned pseudocode works.

The actual temporal annotation values of the input TASs (called *dataset points*) are used together with the relaxation parameter τ to build corresponding hyper-cubical influence areas in the annotation space, as depicted in Figure 4.13. Since τ represents the allowed level of temporal similarity relaxation, these influence areas have an edge equal to 2τ. Next, they are merged and partitioned into disjoint hyper-rectangles, which are added to the collection of influence areas outputted by function *Extract_annotation_blocks(P)*. This allows all prefix occurrences whose corresponding dataset points do not contribute to any dense region to be deleted before any new projection.

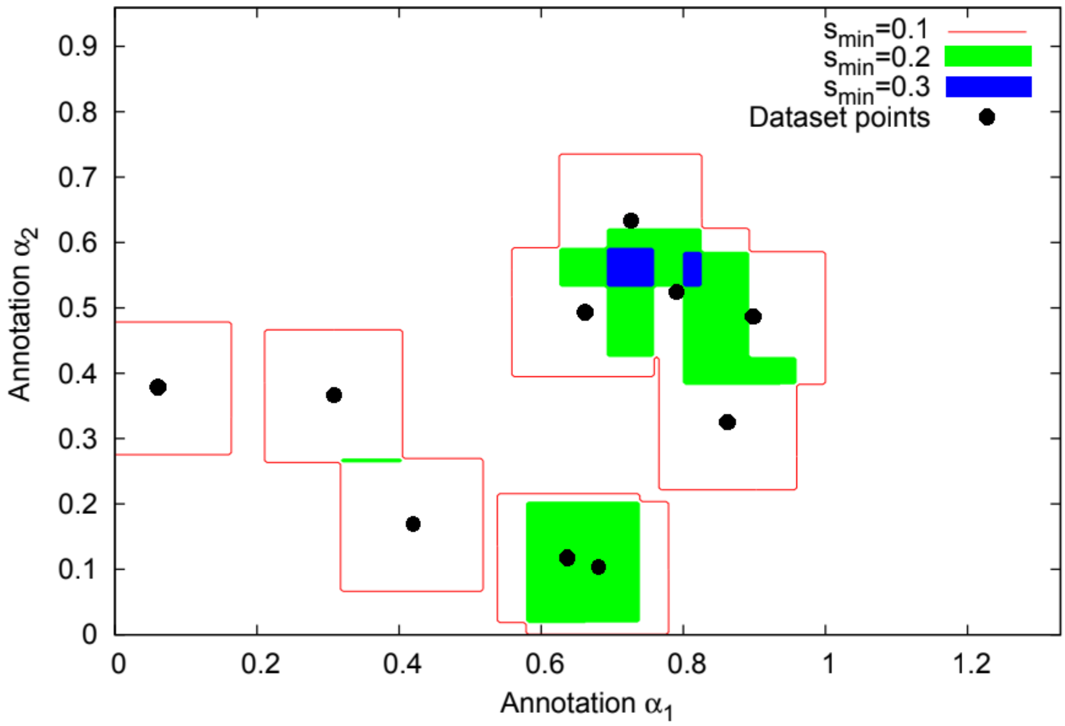


FIGURE 4.13: An example of searching for frequent TASs of length 3 ($a \xrightarrow{a_1} b \xrightarrow{a_2} c$) in the corresponding 2D annotation space, taken from [70].

In the aforementioned illustration provided by Giannotti et al., we can notice how dataset points that are close to each other in the annotation space, i.e. have similar a values, form coloured overlapping areas, inside of which any point would temporally match all of them. For instance, the green surfaces contain all the annotation pairs (a_1, a_2) that would match at least 2 out of the 10 dataset points, whereas the blue surfaces contain all the annotation pairs (a_1, a_2) that would match at least 3 dataset points.

Then, function *Compute_density_blocks(A)* receives as input a set of annotation blocks (i.e. hyper-rectangles) and returns those that have a frequency equal to or higher than s_{min}. It does so, by collecting the extreme coordinates of each block along some dimension d. These denote the boundaries for a split along that d dimension corresponding to homogeneous density (and hence frequency/support). Then a recursive split is performed for each interval defined by two successive boundaries. When all dimensions have been split in this way, all the intervals collected along the recursive calls are combined to extract the hyper-rectangle associated with its density measure. As a result, the annotation space is divided into regions of homogeneous density.

Subsequently, function *Coalesce_density_blocks(D)* is a greedy algorithm that turns a typically large number of small hyper-rectangles of slightly variable density into a sequence of sets of TASs, interpreted as a series of successive approximations of the real set of dense annotations, such that the first outputted provides the best possible approximation in terms of the coverage (i.e. the percentage of annotations in a volumetrical sense represented by the approximation). It does so by randomly choosing an initial dense hyper-rectangle, then repeatedly extending it along the dimension and direction that yields the maximum increase in volume, while merging any other hyper-rectangles that get covered, and then when no more extensions are possible, adding the hyper-rectangle to the output and repeating this process. Information about the density of the coalesced hyper-rectangles is actually dropped or only kept approximately (e.g. avg, min, max) during this process.

Finally, function *Annotation_based_prune(P,$\mathcal{D}^*$)* receives as input a sequence of dense annotation blocks (i.e. hyper-rectangles) and a projection, and produces as output the occurrences of the prefix in the projection that contribute to form dense annotations. More specifically, each occurrence of the prefix corresponds to a dataset point that contributes to the density in the annotation space within its hyper-cubical neighborhood. If no annotation within such a neighborhood is dense, then that dataset point could have been disregarded in Line 7 of the algorithm. In an extension projection, all the annotations of the projection are extended by a temporal component, therefore all dataset points move to a higher-dimensional annotation space where dense regions can become "rarefied" and rarefied regions remain so. As a result, prefix occurrences that cease being interesting at some stage of the computation (i.e. their corresponding dataset points do not contribute to any dense region) will remain useless for the mining process and can be deleted before any new projection.

At the same time, any T-sequence of the projection that does not contain any useful occurrence of the prefix, can only generate larger useless occurrences when projected. As a result, a T-sequence whose dataset points have all been eliminated can itself be deleted. Ultimately, if the projection remains with less than s_{min} T-sequences, then there is no item that can be frequent and the projection process can be stopped altogether.

4.4 Chapter Conclusions

In this chapter, the state-of-the-art research works in the trajectory analysis research domain were described. The state-of-the-art in T-PM more specifically was covered in detail, as well as the necessary background research on S-PM. Chapter 5 will be referring to these works as needed for the design of the proposed T-PM algorithm, which takes into account time, semantics, and topology, as well as hierarchical data dimensions.

5 A Novel Sequential Pattern Mining Approach for Trajectory Pattern Mining

Contents

5.1 Introduction 118
5.1.1 The MultiDimensional Temporally Annotated Sequential Pattern Mining (MD-TAS-PM) Problem 118
5.1.1.1 Temporal Information as Interval Duration Annotations . . . 120
5.1.1.2 Semantic Information as Multiple Dimensions 122
5.1.1.3 Topological Information as Accessibility Constraints 123
5.1.2 Combining Time, Semantics, and Topology: The Case for a Novel Semantic Indoor Trajectory Pattern Mining Algorithm 125
5.1.2.1 Background Definitions 126
5.1.2.2 Compatibility with SITM and Trajectory Data Preprocessing 128
5.1.2.3 SITPE: A MD-TAS Pattern Mining Algorithm 131
5.2 Conclusion 138

5.1 Introduction

Based on the review of trajectory pattern mining research works presented in section 4.3.1, it can be concluded that even though significant advances have been achieved over the last 20 years, especially following the conception of semantic trajectories, there still exist some obstacles on the way to creating museum visitor trajectory analytics systems as envisioned by the author in [100]. The same can also be said for the analysis and mining of trajectories in all other indoor location-based application domains described in section 1.1. More specifically, what is mainly lacking is the necessary mechanisms to intertwine the semantics and the mechanics of movement.

One of the most promising research directions in this respect, showing both predictive and descriptive analytic potential, lies in new specialized trajectory pattern mining algorithms. These can be primarily thought of as extensions and variations of existing general-scope Sequential Pattern Mining (S-PM) algorithms, because the main underlying mathematical approaches (e.g. Apriori principle) remain intact irrespective of the type of data being mined. At the same time however, multidimensional values, hierarchical values, topological restrictions, uncertainty in the data, are additional factors requiring special consideration.

Hence, this chapter first discusses how existing S-PM algorithms can be adapted to trajectory data mining, following the Semantic Indoor Trajectory Model ($SITM$)-based formalism proposed in section 3.3.2. More specifically, the *Semantic Indoor Trajectory Pattern Mining (SIT-PM)* problem is formulated as a *MultiDimensional Temporally Annotated Sequential Pattern Mining (MD-TAS-PM)* problem/ Then, the advantages of $SITM$'s expressive power are highlighted with respect to this problem.

Finally, a new algorithm tackling the SIT-PM problem is proposed, called Semantic Indoor Trajectory Pattern Extractor ($SITPE$). $SITPE$ is described in detail and its implementation process is illustrated. Also, it is explained how it improves upon related trajectory and non-trajectory PM algorithms alike. To the author's knowledge, it is the first algorithm to take into account multiple hierarchical spatiosemantic dimensions, temporal information, and topological constraints. As a result, it outputs more comprehensive patterns and thereby provides a qualitatively richer description of mobility behaviors.

5.1.1 The MultiDimensional Temporally Annotated Sequential Pattern Mining (MD-TAS-PM) Problem

In section 4.3, it was detailed how trajectory pattern mining can be tackled using a combination of different types of techniques and methods, mainly including clustering methods, geographic proximity measures, spatial discretization techniques, data-induced topological constraints, etc. The suitability of each approach depends - among other things - on the type of trajectory data: outdoor or indoor, geometric or symbolic, certain or uncertain, fine-grained or coarse-grained, freely moving or network-constrained, etc.

Nevertheless, S-PM is by far the prevailing approach for trajectory pattern extraction, due to the sequential nature of all types of movement data, with semantic indoor trajectory data of course being no exception. This is why section 4.3 also surveyed how S-PM comprises a broad range of algorithmic methods aimed at finding all or some of the patterns frequently occurring in a given dataset of sequences.

Therefore, a primarily sequential trajectory data model, like the proposed *SITM* (detailed in section 3.3), can serve as the basis for solving the *Semantic Indoor Trajectory Pattern Mining (SIT-PM)* problem.

Let us begin formalizing this problem by defining a *multidimensional temporally annotated sequence (MD-TAS)*:

Definition 5.1.1 (MD-TAS)

A multidimensional temporally annotated sequence is a couple $S = (s, \alpha)$ consisting of two sequences:

1. *An n-long sequence $s = <s_1, s_2, ..., s_n>$ of temporally ordered (according to relation $<_t$) elementary vectors $s_i = (c_{i,1}, c_{i,2}, ..., c_{i,m})$, $i \in [1, n]$ whose components are itemsets $c_{i,j}$, $j \in [1, m]$ composed of one or more items that respectively belong to dimensions $\mathcal{D} = \{D_1, D_2, ..., D_m\}$ in a specific position within their respective domain's hierarchy $\mathcal{H} = \{H_1, H_2, ..., H_m\}$.*

2. *An n-long sequence $\alpha = <\alpha_1, \alpha_2, ..., \alpha_n>$ of real-valued temporal annotations, representing the duration of the respective vectors of s.*

Hence, a MD-TAS can be represented as: $(s, \alpha) = \overset{\alpha_1}{s_1} \rightarrow \overset{\alpha_2}{s_2} \rightarrow ... \rightarrow \overset{\alpha_n}{s_n}$

Moreover, an n-long MD-TAS $S_1 = (s, \alpha)$ is *multidimensionally τ-contained (mdτ-contained)* within another n'-long MD-TAS $S_2 = (s', \alpha')$, $n \leq n'$ ($S_1 \preceq_{md\tau} S_2$) iff $\exists\, 0 \leq i_0 < ... < i_n \leq n'$ such that:

1. $\forall_{0 \leq k \leq n}$: $s_k \leq_{\mathcal{H}} s'_{i_k} \Leftrightarrow \quad c_{k,1} \leq_{H_1} c'_{i_k,1}, c_{k,2} \leq_{H_2} c'_{i_k,2}, ..., c_{k,m} \leq_{H_m} c'_{i_k,m}$

2. $\forall_{0 \leq k \leq n}$: $|\alpha_k - \alpha'_{*k}| \leq \tau$ where $\alpha'_{*k} = \sum\limits_{j=i_{k-1}}^{i_k} \alpha'_j$

Simply put, $S_1 \preceq_{md\tau} S_2$ holds when there is a (potentially non-contiguous) subsequence of S_2, whose itemsets correspond to all the itemsets of S_1 but are *equally or more general* than them (condition 1) according to the respective dimension's hierarchy, and whose annotations differ by at most τ seconds from the corresponding annotations of S_1 (condition 2).

Consequently, the MD-TAS-PM problem is defined as follows:

Definition 5.1.2 (MD-TAS-PM problem)

The MultiDimensional Temporally Annotated Sequential Pattern Mining (MD-TAS-PM) problem is the problem of, given as input a set of MD-TASs S_{in} along with their respective data dimension hierarchies $\mathcal{H} = \{H_1, H_2, ..., H_m\}$, a minimum support value minsup, and a temporal relaxation value τ, to return as output all MD-TAS patterns that are mdτ-contained in S_{in} with a frequency higher than minsup.

Based on the above problem definition and the proposed conceptual trajectory model defined in section 3.3.2, the *Semantic Indoor Trajectory Pattern Mining (SIT-PM) problem* can be defined as follows:

Definition 5.1.3 (SIT-PM problem)
The Semantic Indoor Trajectory Pattern Mining (SIT-PM) problem is the problem of, given as input a set of semantic indoor trajectories T_{in} *each represented according to Definition 3.3.2 along with the respective semantic annotation data hierarchies* $\mathcal{H}_{sem} = \{H_{sem1}, H_{sem2}, ..., H_{semm}\}$*, a multi-layered graph representation* $G = (V, E)$ *of the indoor space represented according to Definition 3.3.1, a minimum support value minsup, and a temporal relaxation value* τ*, to return as output all trajectory patterns which are* $md\tau$*-contained in* T_{in} *with a frequency higher than minsup, and respect the mobility constraints imposed by* G*.*

As evident from the above problem definition, for SIT-PM to be performed properly, order information alone is not enough. Three additional aspects need to be addressed in order to derive more interesting patterns that capture movement phenomena in a much richer way. These are *time*, *semantics*, and *topology.*

5.1.1.1 Temporal Information as Interval Duration Annotations

Among the related trajectory pattern mining works, a few actually do consider the above three aspects of mobility data. First with respect to time, whereas many works ignore it altogether and focus on spatial patterns, others choose to abstract temporal information into the relative order of symbolic locations. This is presumably done to make S-PM algorithms directly applicable to trajectory data.

However, a considerable amount of knowledge remains hidden when, for example, one considers the pattern $A \rightarrow B \rightarrow C$ instead of the pattern $\overset{30sec}{A} \rightarrow \overset{200sec}{B} \rightarrow \overset{1000sec}{C}$. The fact that the moving object remained for 1000 seconds in the spatial region C whereas only for 30 seconds in the region A, is clearly significant for understanding its behavior.

For instance, if A and C were museum exhibition rooms, the duration of visit in each one could be indicative of the level of interest in their corresponding artworks. In such case, it might be concluded that the visitor was much more interested in the artworks of room C.

Some other works including [53, 83, 157], implicitly take time into account in the way they extract spatiotemporal (rather than spatial) regions. This means that each item in a pattern contains temporal information simply by virtue of its existence as a spatiotemporal entity. This also makes standard S-PM algorithms readily applicable.

Unfortunately, this approach is only suited for outdoor environments where the regions or places of interest are not known in advance, and instead are extracted from the movement data with the help of geographic information. On the contrary, for indoor environments, it is not wise to disregard spatial regions simply because they were not visited enough. One reason is that we may wish to find out for example why some parts of a museum are less visited than others, or whether less visited parts actually affect the attendance in the most popular ones.

In addition, spatiotemporal region extraction uses the time of appearance / happening of a spatial event, whereas durations are more helpful for interpreting individual indoor movement parts. To illustrate this point, let us consider which output pattern is more helpful to the museum management for gaining a better understanding of their visitors: $\overset{afternoon}{A} \rightarrow \overset{afternoon}{B} \rightarrow \overset{late\_afternoon}{C}$ or $\overset{30sec}{A} \rightarrow \overset{200sec}{B} \rightarrow \overset{1000sec}{C}$? Naturally, the time of day, or the day of the year, etc. are useful for several types of aggregate analyses or even for periodic pattern detection. At the individual moving object level however, what is more important is the duration of each movement part.

Unfortunately, trajectory pattern mining research has solely focused on temporal information of the former type, mainly due to its focus on outdoor trajectories spanning entire days (e.g. tourist trajectories) or even months (animal trajectories). Ideally, both types of temporal information can be unified in a trajectory model. The former type can even be dealt as any other semantic dimension holding nominal values (i.e. a qualitative description of the temporal context).

Finally, there exist few other works which account for time in more peculiar ways. For example, [83, 151, 197] consider a threshold value acting to restrict the transition time between consecutive symbolic locations. This is used as a splitting threshold in the creation of candidate patterns. The problem with this approach is that it is only relevant for structuring or segmenting the input trajectories, and does not really concern the main mining step, as it relates only to a movement's ending moment (not to the movement itself). This is why applying a splitting threshold makes more sense in the preprocessing phase, rather than in the main mining process step.

To illustrate this in the museum domain for example, item duration threshold constraints would make it impossible to find visiting patterns such as *spending a lot of time in each room at the beginning of the visit, and then moving faster through each room towards the end of the visit.* This is because a trajectory like $\overset{1000sec}{A} \rightarrow \overset{800sec}{B} \rightarrow \overset{900sec}{C} \rightarrow \overset{500sec}{D} \rightarrow \overset{50sec}{E} \rightarrow \overset{40sec}{F}$ would be cut right in the middle for a threshold value of 600 sec for instance. In other words, such a method assumes a homogenized behavior time-wise that is not necessarily realistic.

More generally, it would be useful to be able to extract patterns revealing temporal regularities for which a cut-off value is not enough. In this respect, only [71] and [34] adequately consider time in the main mining step by capturing the duration of item transitions.

Hence, the proposed approach will be using the $MiSTA$ (Mining Sequences with Temporal Annotations) algorithm proposed in [69]. Its trajectory-based counterpart proposed in [71] (for mining so-called *T-patterns*) is not used, because of the symbolic nature of the trajectory data assumed here. The *T-pattern* mining method is identical to $MiSTA$, except for an additional initial grouping of the geometric positional data into geographic regions, that is not useful in the indoor context targeted here.

Finally, whereas $MiSTA$ is indeed one of the few S-PM algorithms accounting for time, it associates the temporal annotations to the transitions from one sequence item to the next. Instead, here they are needed to characterize the items themselves. The reason is simple: it is more interesting how long the moving object spends in a spatial region (e.g. room) than how long it takes for it to transition between consecutive spatial regions (e.g. pass through a door).

This can be easily compensated for at the stage of data pre-processing and does not pose a problem. More importantly, $MiSTA$ does not account for multidimensional item sequences, which leads to the following discussion about capturing the semantics of the trajectories.

5.1.1.2 Semantic Information as Multiple Dimensions

Different Big Data sources can be used to enrich trajectories with complex and heterogeneous semantic information [99]. In such cases, trajectory pattern mining, viewed as a S-PM problem, becomes multidimensional as illustrated in section 1.1.2 for the case of semantic indoor trajectories.

Unfortunately, in comparison to the temporal information discussed in the previous section, here the state-of-the-art practices are much more limited: to the author's knowledge only a couple of works have ever tried to use S-PM to discover interesting patterns in truly semantic trajectories.

Most works are limited to a superficial incorporation of semantics in the sequences, and consequently in the discovered trajectory patterns as well. For example, [197] observes that *"to find frequent sequential patterns in semantic trajectories, one should group similar places together"*, but a place type categorization alone makes only for small part of the semantic wealth that the mining process could be supported with.

Similarly, some works such as [41] claim to retrieve sequential patterns in semantic trajectories, but the semantic information used is only limited to a set of PoIs $P = p_1, p_2, ..., p_{|P|}$ accompanied by a set of semantic categories $C = c_1, c_2, ..., c_{|C|}$, where $p \in P$ is a $2D$ point associated with a category $c \in C$.

Apart from place name semantics[1], semantics of space in general need to be treated differently in indoor environments. For example, congestion of small indoor spaces is something that might affect movement considerably, and thus should be captured in the mined patterns. In addition, there may be semantics related to the movement itself, to the moving object, to other moving objects, to time, etc. depending on the application at hand.

Occasionally, bibliographical trajectory models have even proposed representing very specialized types of semantics, such as the device capturing the tracking data [25]. Thus, all these types of semantics have to be treated as *first-class citizens* and integrated within the core mining process, almost - if not completely - equally to the spatial and temporal data dimensions.

To this end, the focus is momentarily shifted away from trajectory-specific pattern mining in an attempt to derive inspiration from the - admittedly few - existing MultiDimensional Sequential Pattern Mining (MD-S-PM) algorithms.

The goal is to be able to find specific patterns that appear frequently in the input trajectories (e.g. museum visits), taking advantage of multidimensional items representing not only the location information but also the trajectories' semantics. In

[1] Which are really only symbolic unless accompanied by some place hierarchy or ontology.

addition, any existing hierarchies across all data dimensions should be represented in order to capture the different spatiosemantic granularity levels of indoor movement.

As detailed in section 4.3.4, whereas earlier methods proposed by Pinto et al. in their seminal work of [145] are some of the first to account for additional data dimensions extending the main sequence of items, they are not adequate, because the values of the extra dimensions remain static throughout the sequence. In the museum domain for example, these extra static dimensions may represent the visitor's demographic information, profile preferences (e.g. art interests), impairments, etc. but not their level of fatigue, guide content consumption, or any other dynamic characteristic subject to change during the visit. Using the terminology introduced by Mello et al. in [133], these methods can represent any *permanent aspect*, but not any *volatile aspect*, of a moving object's trajectory.

The M^2SP algorithm proposed by Plantevit et al. in [146] is the first to address the aforementioned issue, by completely integrating all extra dimensions into the sequences. This is necessary for discovering truly semantic patterns.

Then, the authors extend it with a mining mechanism that supports hierarchical data values, resulting in the $HYPE$ algorithm of [148]. Their approach is further refined in [147] resulting in the M^3SP algorithm, which detects multidimensional sequential patterns at the highest possible level of specificity. A MD-S-PM method similar to the M^3SP algorithm holds a lot of promise for mining semantic trajectories represented as sequences of multidimensional items.

Unfortunately however, all these methods share a major shortcoming with respect to mobility data applications: they ignore the temporal dimension of the data [30], and thus need to be adapted with respect to the issues considered in the previous section, before being applied on trajectory data. Also, they do not make use of any topological information with regards to transitions between consecutive items in the sequence, nor do they consider any type of spatial data uncertainty, both primary characteristics of moving object trajectory data in indoor environments.

5.1.1.3 Topological Information as Accessibility Constraints

With respect to topological information, whereas it is taken into account in (typically road) network-constrained trajectory data mining approaches, it has been almost totally ignored in the S-PM research landscape. This is partly due to the relevant algorithms not specifically targeting mobility data, let alone indoor trajectories affected by topological restrictions.

To the author's knowledge, the only two works considering topological restrictions in the candidate generation part of an S-PM algorithm are [151] and [50], the first targeting indoor trajectories and the second targeting alarm propagation through a network[2].

In [35], Huiping et al. propose a *connectivity constraint* as a means to reduce the number of candidates, based upon the remark that *"due to continuity of object movement, a spatial region can only connect to some but not all the others..."*. However, this idea is pursued through means of a connectivity graph whose nodes

[2]These can be viewed as metaphorical trajectories of information.

represent all spatial regions and each edge weight reflects the number of occurrences of the corresponding region transition. Then, the idea put forth by the authors is to compare the edge weight to the minimum support, before generating a new candidate sequence. However this is simply equivalent to candidate generation based on frequent 2-sequences.

In [151], Radaelli et al. claim to *"utilize the topology of indoor space to improve performance by pruning irrelevant candidate pattern segments"* but only indirectly capture indoor movement constraints, through the frequent 2-patterns extracted from the trajectory data. However, this approach does not really add anything to the A-Priori principle, according to which if a sequence is not frequent, then neither any of its super-sequences can be frequent. Also from a theoretical point of view, it is equivalent to disregarding impossible transitions based on the assumption that they will not occur frequently enough. This is not necessarily true since a faulty sensor or a misplaced detector might consistently generate erroneous data. Thus, for indoor trajectories it is much preferable that the S-PM algorithm considers explicitly the building's topological structure.

In [50], Devitt et al. consider a network topology while counting sequences as part of their proposed S-PM approach. Although they refer to this as *topographic proximity*, the term *topological* would be more appropriate since the patterns take place over a network topology. Either way, their proposed TP algorithm derives in run-time a proximity score for each candidate sequence. This score acts as an indication of its plausibility in the network context, and based on a minimum threshold score value, determines whether the sequence is discarded or kept. TP is not reliant on any predefined network configuration, but instead considers the network node types and the connections inferred from the data. However, its main idea can be adapted to indoor trajectories as *transition plausibilities* in matrix form, derived either exclusively from the indoor space topology, or in combination with the data.

Taking into consideration the topological information of an indoor space should be done in an explicit way. As long as a building topology is available, there is no point in relying solely on the transitions present in the trajectory dataset, because this is what any standard pattern mining process will do. This is of course unless one is specifically tackling the problem of automated floor plan generation [13, 63, 68], which is not the Thesis' goal.

The proposed model $SITM$ can support such explicit mechanisms, thanks to the edges $e_i \in E_i^{acc}$ representing accessibility relationships between the spatial regions of any layer's NRG $G_i = (V_i, E_i^{acc})$. From those accessibility edges, an *accessibility matrix* of binary values can be derived, or even an *accessibility plausibility matrix* of normalized values, akin to the TP score of [50]. This matrix can then help deal with potential uncertainty with regards to transitions. Finally, inserting topological constraints within the mining process also offers a way to speed up the mining process.

Algorithm: TP mining algorithm

INPUT: seq_1, seq_2: 2 sequences of length $n-1$, e.g. ABC and BCD
OUTPUT: newSeq: sequence of length n, e.g. ABCD
for all o_1 such that o_1 *isa* occurrence of seq_1 **do**
  for all o_2 such that o_2 *isa* occurrence of seq_2 **do**
    Posit occurrence o_{newSeq} from start o_1 to end o_2
    if $Duration\ o_{newSeq} \leq maximumDuration$ **then**
      $TP_{newSeq} = calculateSequenceTP(o_{newSeq})$
      if $TP_{newSeq} \geq TPthreshold$ **then**
        store o_{newSeq}
      end if
    end if
  end for
end for
if $\#occurrences_{newSeq} \leq frequencyThreshold$ **then**
  Prune newSeq
end if

The main TP algorithm uses "topographic" proximity as an additional pruning criterion.

Algorithm: calculateSequenceTP

INPUT: $seq, \{alarm_1, alarm_2 \dots alarm_n\}$
OUTPUT: TPvalue
if $length(seq) == 2$ **then**
  return $calculateTP(alarm_1, alarm_2)$
else
  TP_{seq1} = Retrieve from memory $TP_{alarm_{1\cdots(n-1)}}$
  TP_{seq2} = Retrieve from memory $TP_{alarm_{2\cdots n}}$
  $TP_{new} = calculateTP(alarm_1, alarm_n)$
  return $\frac{TP_{seq1}+TP_{seq2}+TP_{new}}{3}$
end if

Algorithm: calculateTP

INPUT: $alarm_1$, $alarm_2$: two alarm events
OUTPUT: TPvalue
Identify $node_1$, source node type of $alarm_1$
Identify $node_2$, source node type of $alarm_2$
Identify relationship between $node_1$ and $node_2$, given topographical information available
Look up TPvalue in the predefined relationship table.
Return TPvalue for this relationship

Single transitions and entire sequences alike are assigned a plausibility score.

FIGURE 5.1: The TP algorithm proposed by Devitt et al. in [50].

5.1.2 Combining Time, Semantics, and Topology: The Case for a Novel Semantic Indoor Trajectory Pattern Mining Algorithm

In the rest of this chapter, a novel algorithmic method for mining semantic indoor trajectory data is proposed. But before examining it in detail, it needs to be stressed that the techniques designed are actually transferable to the vastly unexplored domain-agnostic field of MD-TAS-PM.

Regarding this, the distinguishing factor between a trajectory-specific approach and any other time-evolving semantic phenomenon is primarily the use of the indoor space model, and secondly also the interval-based nature of time modeling. Besides these two factors, various non-trajectory application domains could benefit from multidimensional sequential patterns encompassing time, such as web activity mining, market prediction, sports performance analysis.

What is completely missing from the related bibliography is an algorithmic method to take all three types of information into account: multiple data dimensions i.e. one topologically constrained spatial dimension and one or more semantic dimensions, along with the duration of stay of the moving object in each spatial region.

At the same time, hierarchical value levels for each of those dimensions should ideally be supported as well. In this way, phenomena characterized by patterns at different levels of granularity may be studied.

As detailed in chapter 3, the proposed model $SITM$ defines a trajectory as a couple comprised of a trace and a set of semantic annotations (Definition 3.3.2), which thanks to the hierarchical indoor space representation, can represent multi-faceted information about any particular presence/stay of the visitor inside of a discrete spatial region of the indoor environment (e.g. museum). These intervals are timestamped and temporally ordered in a non-overlapping manner. Thus, given an input trajectory dataset in $SITM$ form, any standard S-PM algorithm is able to derive trajectory patterns in the form of simple interval subsequences occurring frequently enough in that dataset.

But more importantly, $SITM$ can enable the design of a new algorithm, taking all three aforementioned elements into consideration, effectively solving the corresponding MD-TAS-PM problem. The main goal is to detect qualitatively richer frequent patterns in the output. To quote Cai et al. [34], a semantic trajectory pattern of *"going to a hotel and then going to a park after two hours on a rainy weekday and visiting a beach two days later on a clear weekend"* is much more detailed and meaningful than *"place A to place B"*.

In a similar fashion for indoor settings, let us consider a typical semantic trajectory pattern of Louvre visitors interested in Italian Renaissance paintings. The pattern starts by visiting the *Mona Lisa* for 30 minutes, before proceeding to the *museum shop* located in Salle Denon for 10 minutes and buying the guidebook of the "Leonardo da Vinci" temporary exhibition. Then it continues with the visitor strolling through the Salle Mollien for another 30 minutes while stopping at select paintings such as the *Liberty Leading the People*, before feeling increased art fatigue and deciding to stop for a coffee break at the cafeteria next to the Mollien staircase.

Compared to a simplistic sequential pattern like $SalledesEtats \rightarrow SalleDenon \rightarrow SalleMollien \rightarrow MollienStaircase$, the previous pattern is obviously much more telling about the movement qualities, and in this particular case about the visitor experience. This is despite the fact that the simpler pattern would also qualify as a "semantic" trajectory pattern, simply by virtue of including the names of the visited spatial regions, according to many of the definitions proposed in the related works.

5.1.2.1 Background Definitions

Adopting the notations used in [147], the following notions are defined:

Definition 5.1.4 (value generalizations)

The generalizations $x^{\uparrow}$ of a 1D data value x is the set containing x and all of its predecessors in its corresponding dimension hierarchy H_i.

Definition 5.1.5 (value specializations)

The specializations $x^{\downarrow}$ of a 1D data value x is the set containing x and all of its successors in its corresponding dimension hierarchy H_i.

Definition 5.1.6 (generated MD item)

The set $gen(a)$ of multi-dimensional items generated from a multidimensional item $a = (d_1, ..., d_m)$ is the set of all (non-duplicate) multidimensional items $a' = (d'_1, ..., d'_m)$ that can be created by specializing a across all of its corresponding dimensions' hierarchies $H_1, ..., H_m$.

The above definitions will help simplify the algorithmic notation.

Moreover, by focusing on adapting the notion of *containment* according to their specific needs, most S-PM research works miss out on discussing possible support variants.

For multidimensional sequences for instance, Egho et al. [58] define the support $supp(s, \mathcal{MS}_{\mathcal{DB}})$ of a multidimensional sequence s, as the number of sequences s_i in the input database $\mathcal{MS}_{\mathcal{DB}}$ that are more general than s. In other words, their proposed algorithm $MMISP$ ignores multiple occurrences of a multidimensional pattern in the same input trajectory and only counts them once.

Whereas in [147], Plantevit et al. exclude some of the data dimensions (so-called *reference dimensions*) from the mining step, and use them instead to partition the input sequence database into multiple blocks. In other words, their proposed algorithm M^3SP ultimately counts multiple occurrences of a multidimensional pattern only once per block of input data.

Motivated by this lack of consideration about the core definition of support, this Thesis proposes three variations of the absolute *MD-TAS pattern support*. In the proposed algorithm only the first two (*classic* and *repetitive*) are used. The third one (*block*) is particularly interesting for taking advantage of the *static semantics* of trajectories, or what Mello et al. call in [133] the *long term aspects* of trajectories.

The first variation, *Classic support*, reflects the well-known notion of support as it relates to all types of Pattern Mining: it only considers whether a pattern occurs in an input data instance or not, not how many times it occurs in it. This is for example what Egho et al. adopted to the multi-dimensional case in [58].

In the context of indoor trajectory data, it is relevant for detecting virtually any type of semantic mobility behavior, such as *visiting the Mona Lisa for 30 minutes and then going to the Museum Shop for 20 minutes to buy a Mona Lisa themed handbag* or *drinking a coffee at Louvre's Café Mollien for 15 minutes and then strolling through the Grande Gallerie for 30 minutes while listening to the audio description of various paintings*. Such mobility behaviors typically take place only once or twice in any given visit.

Definition 5.1.7 (Classic MD-TAS Pattern Support)
The classic support $sup(p)$ of an MD-TAS pattern p is equal to the number of input sequences containing p.

The second variation, *Repetitive support*, is oriented more towards detecting mobility micro-behaviors that tend to be repeated numerous times during the lifetime of a single trajectory. As far as the the museum domain is concerned, these could be patterns like *seeing a painting for 10 seconds and then seeing a sculpture for 20 seconds*, or *backtracking to the previous room and then proceeding again to the current room*. These are typically either episodic patterns or simply patterns of lower spatiotemporal granularity. In correspondence to $SITM$, this version of support is more useful for mining patterns at the intra-room RoI leaf layer, as depicted in Figure 3.1 of section 3.3.1.1.

Definition 5.1.8 (Repetitive MD-TAS Pattern Support)
The repetitive support $rsup(p)$ of an MD-TAS pattern p is equal to the total number of times it occurs in any input sequence.

The third variation of support uses a division of the input trajectories into blocks, based on the distinction between their *long term* aspects and their *volatile* aspects

[133]. Since the former stay unchanged throughout the trajectory, they are fit to be used to segment the trajectory dataset into blocks.

A similar approach is used by the M^3SP algorithm in [147], where every block B is eventually assigned a unique identifier $ID(B)$ *"playing the role of the customer identifiers in standard algorithms"*, but it is not justified why this is the case.

Blocking is actually not required for the mining process, but in our view it can be useful for parallelizing it, because input dataset scans need only check a single block at a time, with the overall support value being calculated in a separate step. Hence, a block-based support can speed up the mining process for distributed sequential pattern mining. However, it concerns only application scenarios involving (at least some) static trajectory semantics to act as the *reference dimensions* [147] based upon which the input dataset is partitioned.

Definition 5.1.9 (Block MD-TAS Pattern Support)
The block support $bsup(p)$ of an MD-TAS pattern p is equal to the total number of blocks of input sequences containing p, where the blocks are formed according to the non-changing semantic dimensions of the data.

In the museum domain for example, the semantics describing an entire visit trajectory (A_{traj} from Definition 3.3.2) can be exploited to divide a massive input dataset of visits into smaller blocks: a block of trajectories whose goal is to visit the *Mona Lisa*, another block of trajectories whose goal is to visit the temporary exhibition, etc. Alternatively, moving object semantics such as demographics can be used to partition the input data: a block of trajectories performed by French visitors, another block of trajectories performed by foreign European visitors, and so on.

In the rest of this chapter, the Classic MD-TAS Pattern Support is used, as it generally corresponds to the most wide range of PM application scenarios.

5.1.2.2 Compatibility with SITM and Trajectory Data Preprocessing

As a necessary reminder, the proposed $SITM$ model presented in section 3.3.2 considers a trajectory as a couple comprised of a trace and a set of semantic annotations. The trace consists of a sequence of tuples, each representing information about the (non-overlapping) timestamped presence/stay intervals of the moving object (e.g. visitor) inside of a discrete spatial region of the indoor environment (e.g. museum). The semantic annotations describe both the trajectory in its entirety (the set A_{traj} in Definition 3.3.2) and/or any specific tuple (the set A_{traj} in Definition 3.3.3).

Whereas, specific semantic ontologies or hierarchies are not included in $SITM$ in order to keep it domain-independent, an elaborate indoor space representation suitable for any indoor environment (the layered multigraph G=(V,E) is defined in section 3.3.1).

Hence, $SITM$ provides the analyst with a way to address the semantic aspects of trajectories as additional item dimensions. This in turn, enables the combination of the time-aware prefix-projection generation mechanism of the $MiSTA$ algorithm [69] with the multidimensional item generation mechanism of the M^3SP algorithm [147] (based on the notion of item specificity as explained in section 4.3.4) for mining semantic trajectory patterns. Hence, $SITM$'s trajectory representation can support the information representation needs of the MD-TAS-PM algorithm that will be formalized next. More concretely, the algorithm will make use of the following modeling elements of $SITM$:

1. The *semantic content*: the values d_i of the multidimensional items in the sequences represent the evolving semantic/contextual aspects of the trajectories, and thus correspond to the range of values $a \in A_i$ of $SITM$'s semantic annotation sets A_i, $\forall i \in [1, n]$.

2. The *spatial hierarchy*: the hierarchy H_{sp} of the spatial dimension in the sequences represents the hierarchy of indoor spatial regions where the moving object may be found, and thus corresponds to the binary topological relationships E^{top} of $SITM$'s hierarchical indoor space G=(V,E).

3. The *topology*: the topological restrictions over item transitions in the sequences represent the movement constraints imposed by the indoor environment's architectural or functional properties, and thus correspond to the accessibility relationships E^{acc} of $SITM$'s accessibility NRG G_i=(V_i,$E_{acc,i}$) for any hierarchical level $0 \leq i \leq m$ of the particular spatial data value.

4. The *temporal dimension*: the temporal annotation values of a sequence's multidimensional items represent the duration of the moving object's stay in each corresponding spatial region, and thus correspond to the difference (in seconds) between the two related absolute timestamps: $t_k^{dur} = t_k^{end}$ - t_k^{start}.

Inversely, there are also a few modeling elements in $SITM$ not used by the algorithm, which can nevertheless be taken advantage of by future extensions of the proposed PM approach or other methods inspired by it. These are mentioned next along with how they could potentially be handled:

1. The *edges* e_i: The connections traversed by the moving object in the accessibility NRG G_i=(V_i,$E_{acc,i}$) of any layer $0 \leq i \leq m$ of the hierarchy are not used. Instead, an accessibility matrix is used, which disregards different edges connecting the same two spatial regions. Given that $SITM$'s indoor space representation (section 3.3.1) is multigraph-based, and therefore can support multiple ways to go from one spatial region to another. In that case however, a more complex topological pruning step should be implemented (than the one actually used) that takes edges into account.

2. The *absolute timestamps* t_i^{start} and t_i^{end}: The beginning / ending moments of the presence intervals representing precisely when the moving object entered / exited a spatial region are not used. The $MiSTA$ algorithm actually ignores those. This is a fundamental decision on how to model time in the patterns. A different algorithm, taking absolute timestamps into account, could be used in the final step of the proposed algorithm[3]. The most suitable algorithmic design

[3]Even though as argued in section 5.1.1.1, duration intervals are actually more crucial for interpreting individual movements.

approach to do that would be to combine the intervals with the absolute timestamps, enriching the sequences with a time-related semantic data dimension derived from those timestamps and taking values such as "Morning", "Early Afternoon", "Summer". Then, the algorithm would treat it in the same way as any other categorical semantic data dimension.

3. The set of *semantic annotations* A_{traj}: The semantic aspects of an entire trajectory are not used, because the proposed PM approach is focused on dynamic semantics, rather than on those staying the same throughout the lifetime of a trajectory. However, a block support, as formalized in Definition 5.1.9, can be used to consider these static semantic annotations, and extract patterns that are frequent, not only in the input dataset in general, but also with respect to specific types of trajectories (e.g. visits to see the Mona Lisa, visits of art professionals, guided visits).

Given the above choice of modeling elements, a semantic indoor trajectory

$$T_{ID_{mo},t_{start},t_{end}} = (trace_{ID_{mo},t_{start},t_{end}}, A_{traj})$$

where

$$trace_{ID_{mo},t_{start},t_{end}} = (e_i, v_i, t_i^{start}, t_i^{end}, A_i)_{i\in[1,n]}$$

can be formulated as a MD-TAS (s, α) for which:

- the set of dimensions $\mathcal{D} = \{D_{space}, D_{semantics}\}$ contains a spatial dimension D_{space} which takes its values from the active domain of the edge-node tuples (e_i, v_i) present in the indoor space graph G, and one or more semantic dimensions $D_{semantics} = \{D_{sem_1}, D_{sem_2}, ..., D_{sem_m}\}$ which take their values from the case-specific sets A_i of *trajectory-part* semantic annotations.

- the corresponding set of hierarchies $\mathcal{H} = \{\text{G(V,E)}, H_{semantics}\}$ contains a spatial hierarchy which is the indoor space graph, and one or more semantic taxonomies.

- the elementary vectors $s_i = (v_i, A_{i,1}, ..., A_{i,m})$, $i \in [1, n]$ contain the following items: the node $v_{i,j}$ denoting the value of the spatial dimension (i.e. location information), and one or more items that respectively belong to dimensions $D_{semantics} = \{D_{sem_1}, D_{sem_2}, ..., D_{sem_m}\}$ in a specific position within the respective semantic hierarchy $H_{semantics} = \{H_{sem_1}, H_{sem_2}, ..., H_{sem_m}\}$.

- the annotation sequence $\alpha = <(t_{end_1} - t_{begin_1}), (t_{end_2} - t_{begint_2}), ...,$ $(t_{end_n} - t_{begin_n})>$ is extracted from subtracting the real-valued timestamps representing the start and finish of the corresponding elementary vectors of s.

- the absolute timestamps t_i^{start}, t_i^{end} can (optionally) be used in one of the m semantic dimensions $D_{sem_{temp}}$ to represent the temporal context.

5.1.2.3 SITPE: A MD-TAS Pattern Mining Algorithm

Hereby, the proposed MD-TAS-PM algorithm is formalized. It extracts from an input trajectory dataset in the aforedescribed $SITM$ form, all trajectory patterns occurring "frequently enough" in it, by taking into account multiple data (one spatial and one or more semantic) dimensions, hierarchical value levels for each of those dimensions, and the duration of stay of the moving object in each spatial region. To the best of the author's knowledge, this is the first proposed method for mining sequential patterns based on the multitude of information described in section 5.1.2.

The proposed algorithm called Semantic Indoor Trajectory Pattern Extractor ($SITPE$) comprises a main function (*sitpe_main*) which takes as input a semantic trajectory dataset (according to section 3.3.2), a support threshold value, a temporal threshold value, as well as an indoor space representation (according to section 3.3.1), and the hierarchies of trajectory semantics. It then calculates the frequent trajectory patterns in three steps by calling helper functions and procedures: *count_1D_values*, *discard_1D_values*, *calculate_msfas*(T_{in}, H, I).

Algorithm 1 is the main algorithm and it works in three steps:

1. First, it calculates the frequency with which any 1-dimensional value occurs in the input dataset, and discards from the data dimension hierarchies those values that do not occur frequently enough to potentially belong to a frequent pattern. Then, starting from the set of all frequent multidimensional values of the form $(ALL_1, ..., ALL_{i-1}, d_i, ALL_{i+1}, ..., ALL_m)$, where d_i is a frequent child node value of the value ALL_i, it recursively specializes over all dimensions, until it finds the set of *Most Specific Frequent (Multidimensional) Values (MSFV)*.

2. Secondly, it transforms the input database sequences by using the newly found MSFV set: any item that is not a MSFV is replaced by its corresponding MSFV. Special attention is paid during this generalization step, in order to update accordingly the temporal annotations of the replaced items, but also to fuse together any consecutive items that may appear (as a result of the transformation) having the exact same spatiosemantic multidimensional values.

3. Finally, it uses a TAS-PM algorithm in order to mine the transformed database in a way that takes into account the temporal annotations of the items. More specifically, a variation of the $MiSTA$ algorithm is used in which the τ parameter is not static, but instead adapted to the granularity of the spatial dimension value of each item in the sequence. Moreover, an extra topological pruning criterion is added, which makes use of the indoor space representation as well as the data to derive a transition plausibility matrix. Hence, it is called MultiGranular Topologically Aware MiSTA ($MGTA - MiSTA$).

It is important to notice that not every frequent trajectory pattern is returned due to computational complexity issues. Instead, the algorithm[4] finds only the most specific frequent patterns. For example, if (*D+1:SJ*, $[150{-}300]$*sec*, *low fatigue*) $\rightarrow$

[4] Similar to [147].

Algorithm 1: sitpe_main(T_{in}, H, G, $minsup$, τ)

Data:

- a set of trajectories T_{in} in $SITM$-based form (as defined in section 3.3)
- a minimum support threshold $minsup$
- a temporal relaxation threshold τ (in seconds)
- a set of m hierarchies $H = \{H_i\}, i \in [1, m]$ where H_i is the hierarchy of the i-th semantic data dimension and $H_m = H_{sp} = (V, E^{top})$ is the spatial hierarchy as derived from the indoor space graph representation G
- the indoor space model $G = (V, E)$ (as defined in section 3.3.1)

Result: a set of frequent SIT patterns TP_{out}

`/* Step1 - Calculate the most specific (hierarchically lower) frequent multidimensional values: */`

`/* Step1A - Calculate frequency counts: */`

Counts $\leftarrow$ *count_1D_values*(T_{in}, H) ;

`/* Step1B - Discard infrequent 1D data values: */`

$I \leftarrow$ *discard_1D_values*(H, *Counts*);

`/* Step1C - Find most specific frequent MD values: */`

$MSFV \leftarrow \emptyset$;

forall *item* $\in I$ **do**

  $MSFV \leftarrow$ *calculate_msfv*(T_{in}, H, *item*, *minsup*, $MSFV$) ;

end

`/* Step2 - Transform input data to sequences consisting only of most specific frequent MD items: */`

$T'_{in} \leftarrow$ *transform_dataset*(T_{in}, H, $MSFV$) ;

`/* Step3 - Mine the transformed input sequences taking into account the temporal annotations: */`

$TP_{out} \leftarrow MGTA - MiSTA$($T'_{in}$, G, *minsup*, τ) ;

(*D+1:S*, [50−100]*sec*, *high fatigue*) is a frequent semantic trajectory pattern, then so is (*Denon Wing*, [150-300]sec, $ALL_{fatigue}$) → (*Denon +1 floor*, [50-100]sec, $ALL_{fatigue}$). However, the latter will not be returned, as it is only a generalization of the former.

The eventual output of Algorithm 1 is a set of patterns, that occur frequently in the input semantic trajectory dataset. Unlike the items in the input dataset themselves which contain a specific duration value, the items constituting these sequential patterns contain a range of duration values, due to the temporal relaxation parameter τ that the $MiSTA$ algorithm applies.

This is natural, since not all matching trajectory parts can be expected to contain exactly the same durations in every spatial region. In the previous example for instance, the range for [150-300] followed by the range [50-100], means that remaining in zone *D+1:SJ* for anywhere between 150 and 300 seconds (inclusive), transitioning to zone *D+1:S*, and remaining there for anywhere between 50 and 100 seconds (inclusive), would suffice to produce a frequent pattern no matter the exact number of seconds in each zone. The allowed range is not fixed, instead it depends on the input data.

Next, the helper functions used to implement the main algorithmic steps of $SITPE$ are described.

Algorithm 2: count_1D_values(T_{in}, H)

Data:

- the set of trajectories T_{in} having m semantic data dimensions
- the set of m hierarchies $H = \{H_i\}, i \in [1, m]$ where H_i is the hierarchy of the i-th semantic data dimension and $H_m = H_{sp} = (V, E^{top})$ is the spatial hierarchy as derived from the indoor space graph representation G

Result: The set of frequency counts $Counts(d) = \{counts(d) | d \in H_i, H_i \in H\}$ of all the possible (at any hierarchical level) values d of each data dimension $i \in [1, m]$ is calculated.

/* Initialize occurrence counts to 0: */
for $i \leftarrow 1$ **to** m **do** /* for all item data dimensions */
    forall $d \in H_i$ **do** /* for all 1D values */
        $Counts(d) \leftarrow 0$; /* initialize value count */
    end
end
forall $T \in T_{in}$ **do** /* for all input trajectories */
    for $i \leftarrow 1$ **to** m **do** /* for all item data dimensions */
        forall $d \in H_i$ **do** /* for all 1D values */
            $flag(d) \leftarrow False$; /* initialize flag */
        end
    end
    forall $(d_1, d_2, ..., d_m) \in T$ **do** // for all MD items
        forall $d' \in (d_1, d_2, ..., d_m), d' \neq ALL$ **do** // for each 1D value
            forall $\delta \in d'^{\uparrow}$ **do** // for all predecessors
                if $flag(\delta) = False$ **then**
                    $Counts(\delta) \leftarrow Counts(\delta) + 1$; // count occurrence
                    $flag(\delta) \leftarrow True$; // avoid further increase
                end
            end
        end
    end
end
return $Counts$;

The first auxiliary algorithm is a short helper function, called *count_1D_values* (Algorithm 2). It calculates how often any atomic sequence appears in the input trajectory dataset. Although easy to understand, it contains some less obvious implementation details that make use of the hierarchical representation of the semantic dimensions.

First in line 13, the data value under consideration has to be different than the most general "root" value of the corresponding dimension's hierarchy. This is simply because there is no point in counting these all-encompassing (so-called *ALL* [147]) values, since by definition they are considered to always occur. It would only make sense to consider pruning the *ALL* values in case of missing data under a *closed-world* assumption.

To the contrary, it is assumed that if a semantic dimension value is missing from any of the multidimensional items comprising the input trajectories, then it can be replaced by its corresponding generic *ALL* value. This is a more natural interpretation with regards to most semantic dimensions concerning the museum domain.

For example, if the level of fatigue of a visitor is unknown, it can be assumed that it is either *low*, or *medium*, or *high*. Similarly, if the artistic preferences of another visitor are unknown, then it can be assumed that he or she potentially likes all artwork types and artistic movements (instead of none). Hence, using hierarchical semantic dimensions also serves as a means of managing data uncertainty in the trajectories.

Secondly in line 16, the frequency count is increased not only for the specific one-dimensional value d' that was encountered in the input data, but also for all of its predecessors in the corresponding semantic hierarchy. For example, when encountering the value *Mona Lisa* with respect to the artwork types and art themes, the count of the *Italian Renaissance* value will also be increased, given that it is its parent node in the corresponding semantic hierarchy. This is important to guarantee completeness of the algorithm.

Specifically for the previous example, if the input trajectory dataset explicitly contains the *Mona Lisa* value frequently enough but not the *Italian Renaissance* value, Algorithm 2 will not discard the latter because it will count every appearance of the former as its own appearance as well.

As another example, consider that the input trajectory dataset contains the *Mona Lisa* value but not quite frequently enough, and also contains the *Death of the Virgin* value but again not enough times, then it might be the case that together the occurrences of these two values are enough to qualify their parent value of *Italian Renaissance* as being frequent. Therefore, the appearances of both need to be counted towards that.

Naturally then, the count increase in line 16 needs to adapt to the selected type of support. Because in the atomic value pruning phase in Algorithm 2 each count will be compared to the minimum support. Any discrepancy between the two would lead either to the algorithm keeping atomic values that can never participate in any frequent pattern, or - worse yet - to pruning atomic values that could eventually make it to a frequent pattern.

The choice of support type can of course be parameterized. In line 15 of Algorithm 2, a flag variable is used to calculate the classical support metric (in absolute terms) as defined in Definition 5.1.7. If instead a repetitive support (Definition 5.1.8) is preferred, then multiple occurrences of atomic values can be counted multiple times by removing the *flag* variable as follows:

$\forall T \in T_{in}$: `// for all input trajectories`
  $\forall (d_1, d_2, ..., d_m) \in T$: `// for all` m`-dimensional items`
    $\forall d' \in (d_1, d_2, ..., d_m), d' \neq ALL$: `// for all 1D values`
      $\forall \delta \in d'^{\uparrow}$:
  Counts(δ) $\leftarrow Counts(\delta) + 1$; `// count their occurrence`

Next, a second helper function called *discard_1D_values* is used to discard all atomic values that appear in the input dataset too rarely to have any chance of making it into a frequent subsequence.

Algorithm 3: discard_1D_values(T_{in}, H)

Data:

- the set *Counts* of the number of occurrences of every 1D data value
- a set of m hierarchies $H = \{H_i\}, i \in [1, m]$ where H_i is the hierarchy of the i-th semantic data dimension and $H_m = H_{sp} = (V, E^{top})$ is the spatial hierarchy as derived from the indoor space graph representation G

Result: A set of multidimensional values $I = \{(ALL_1, ..., ALL_{i1}, d_i, ALL_{i+1}, ..., ALL_m)\}$ in which the 1D values d_i are "frequently enough" occurring children of their corresponding dimension's "ALL-encompassing" value ALL_i.

`/* Atomic values occurring less than` *minsup* `times and their successor values are discarded: */`

for $i \leftarrow 1$ **to** m **do** `/* for all item data dimensions */`
  forall $d \in H_i$ **do**
    if *Counts(d)* $<$ *minsup* **then** `/* not "frequent enough" */`
      remove($d^{\downarrow}, H_i$) `// remove value and successor values`
    end
    else if *parent(d)*$=ALL_i$ **then** `/* only keep` *ALL* `values' children */`
      $I \leftarrow I \cup (ALL_1, ..., ALL_{i1}, d, ALL_{i+1}, ..., ALL_m)$;
    end
  end
  return I ;
end

The aforementioned pruning of the hierarchies performed by the function *discard_1D_values(T_{in}, H)* is actually performed in preparation of the next step, whose main idea is rather simple and originates in [147]: starting off with all multidimensional values in the form of $(ALL_1, ..., ALL_{i1}, d, ALL_{i+1}, ..., ALL_m)$, where the 1-dimensional value d occurs frequently enough and is a child node of a corresponding all-encompassing value ALL_i (to avoid unnecessary repetitions), all multidimensional items that occur frequently enough in the dataset can be found by specializing recursively over all data dimensions in a DFS manner.

Now it is apparent why only frequent 1D values of a level below the ALL values (line 6 in Algorithm 3) need to be kept: any more specific values would be exhaustively generated anyway as part of the multidimensional item generation. But out of those multidimensional items, only the most specific ones are needed, assuming of course that it is acceptable to ignore trajectory patterns that are semantically “too general”.

Each of the frequent multidimensional items $(v_1, ..., v_{i1}, v_i, v_{i+1}, ..., v_m)$ that makes it to the set $MSFV$ comprises an atomic sequence $< \{(v_1, ..., v_{i1}, v_i, v_{i+1}, ..., v_m)\} >$ which can then function as the starting point for building the candidate patterns / subsequences of length greater than 1.

The above method actually mirrors the notion of *Maximal Atomic Frequent (MAF) sequences* used in [147] and the practically equivalent notion of *Most Specific Frequent Elementary Vectors (MSFEVs)* used in [58]. In turn, both are inspired from the BUC algorithm proposed by Beyer et al. in [22] tackling the Iceberg-CUBE problem in a bottom-up manner.

Here the term *Most Specific Frequent (MultiDimensional) Values (MSFV)* is preferred, because until the MSFV set is actually used for candidate generation, the algorithm does not yet really consider any *sequences* or *vectors of items*[5]. Hence, those terms might obscure the fact that sequences are only generated in the last step of the main algorithm (Algorithm 1) which is detailed in Algorithm 4.

Algorithm 4 comprises the part of the main algorithm which handles the multiple data dimensions present in the items. It is called multiple times, and each time it take as argument an item in the previously described form $(v_1, ..., v_{i1}, v_i, v_{i+1}, ..., v_m)$ and generates all of each successors, but only according to the pruned hierarchies. If one of these successors occurs more often than $minsup$ then it is added to a set of frequent candidates.

While the $MiSTA$ algorithm proposed by Giannotti et al. in [69] takes into account temporal annotations, it does not consider different hierarchical levels for its items. This raises the issue of how to treat time within sequences whose tuples contain data values that belong to different levels of granularity, especially concerning the spatial dimension. For instance, two indoor spatial regions that differ considerably in size would be expected to relate to different presence interval values: shorter durations for smaller regions and longer durations for bigger regions.

However, $MiSTA$'s τ parameter is not adjustable to any symbol semantics. Therefore, a τ value that is too low (strict restriction) risks losing patterns that contain large regions, whereas a τ value that is too high (loose restriction) risks over-representing patterns that contain small regions.

[5]Nor of itemsets as in the case of [58].

Algorithm 4: calculate_msfv(T_{in}, H, $item$, $minsup$, $MSFV$)

Data:

- the set $Counts$ of the number of occurrences of every 1D data value
- a set of m hierarchies $H = \{H_i\}, i \in [1, m]$ where H_i is the hierarchy of the i-th semantic data dimension and $H_m = H_{sp} = (V, E^{top})$ is the spatial hierarchy as derived from the indoor space graph representation G
- a MD item $item = (ALL_1, ..., ALL_{i1}, d_i, ALL_{i+1}, ..., ALL_m)$ in which the 1D value d_i is "frequent enough" and a child of its corresponding ALL_i value
- the minimum support threshold $minsup$
- the current set $MSFV$ of MD values each occurring frequently enough in the input data and having no other more specific MD value that does so

Result: an updated set $MSFV$ (at the end of the recursion it coincides with the Most Specific Frequent MD Values)

/* For the input *item*, we generate all specializations traversing the pruned hierarchies in a DFS manner: */

forall $v = (v_1, v_2, ..., v_m) \in gen(item,H)$ **do** /* for all MD values generated from item */
    if $sup(v) \geq minsup$ **then** /* if candidate value is frequent */
        $FreqCand \leftarrow FreqCand \cup v$; /* add it to set of frequent candidates */
    end
end
if $FreqCand = \emptyset$ **then** /* if item generated no frequent candidate */
    if $\nexists v' \in MSFV : v \preceq_I v'$ **then** /* if no more specific value already in */
        $MSFV \leftarrow MSFV \cup v$; /* add value v in MSFV */
    end
end
else
    forall $v'' \in FreqCand$ **do** /* recursively calls itself */
        $MSFV \leftarrow calculate\_msfv(T_{in}, H, v'', minsup, MSFV)$;
    end
end
return $MSFV$;

Moreover, as already seen, virtually all S-PM algorithms do not account for topological restrictions which are nonetheless of paramount importance for indoor trajectories. Therefore, this additional pruning criterion needs to be added into $MiSTA$'s prefix-based projection method explained in section 4.3.1.

Hence, a variation of the $MiSTA$ algorithm called $MGTA-MiSTA$ (MultiGranular Trajectory $MiSTA$) is proposed, which has three main distinctions with respect to the original proposal of Giannotti et al.:

1. It assigns the duration annotation not to the transition between two items, but

to the item of departure. This makes the algorithm applicable according to the proposed trajectory model $SITM$. From a technical implementation point of view, it mainly suffices to add a last "dummy" ending ($EXIT$) state in each input sequence, in order not to miss its last annotation.

2. Instead of relaxing the temporal annotation value by τ in a static manner, it relaxes it by $\alpha * \tau$ where α is a normalization coefficient used to reflect that stays in large spatial regions deserve a looser temporal pattern matching requirement.

3. It implements an additional pruning criterion (completely orthogonal to the temporal criterion) in the form of a transition plausibility matrix, which essentially is a merging of a transition matrix and an accessibility matrix.

A final technical issue that needs to be addressed without affecting the algorithmic process of $MiSTA$ is that when used with trajectory datasets, instead of sequences of itemsets supported by $Mista$, $MGTA-MiSTA$ is restricted to work with sequences of multidimensional items. In practice however, only the input file needs to be restricted, because a S-PM for sequences of itemsets will run just as fine with sequences of items. Also, only the symbol representing the spatial dimension is considered by $MGTA - MiSTA$, because the extra semantic dimensions have already been dealt with by the previous steps of the proposed algorithm.

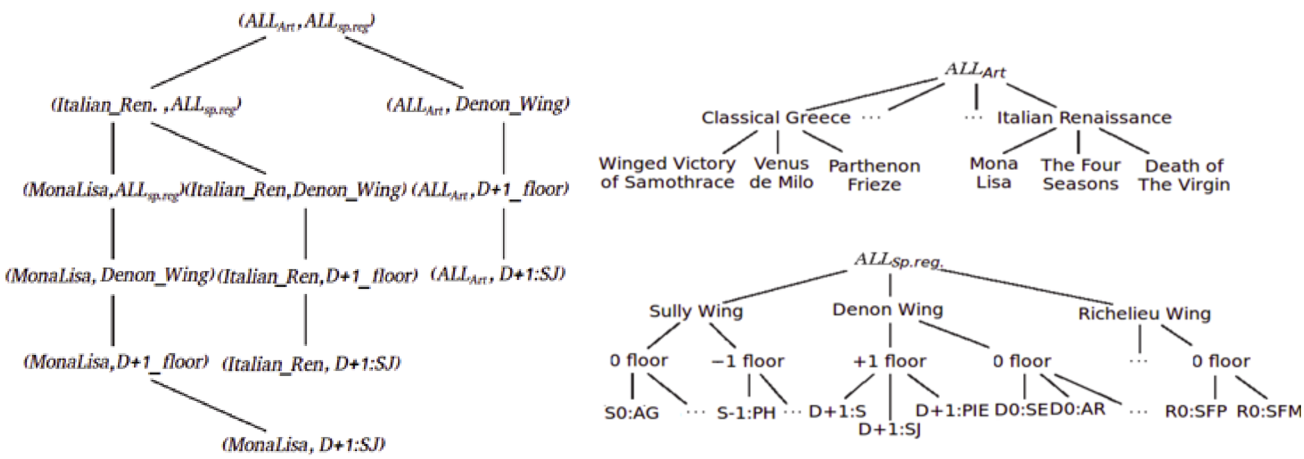


FIGURE 5.2: Data dimension hierarchies corresponding to a Louvre visit dataset.

5.2 Conclusion

This chapter was consecrated to the design of a novel algorithmic approach, with respect to S-PM specifically for the domain of semantic indoor trajectories. First,the case was made for why it is important to consider semantics, time, and topology in a unified manner. Then, the corresponding pattern mining problem was formulated, and the necessary definitions provided. Finally, an indicative algorithmic method for solving this problem was proposed, based on combining the time-aware prefix-projection generation mechanism of the $MiSTA$ algorithm [69] with the multidimensional item generation mechanism of the M^3SP algorithm [147] and the accessibility plausibility idea inspired from [50]. The specific algorithmic implementation of this method is called $SITPE$. Unfortunately, in the next chapter, experimental results from the application of $SITPE$ are not included.

6

Trajectory Model Validation and the Louvre Case Study

Contents

6.1 Introduction . . . 142
6.2 The Museum Application Domain . . . 142
6.3 The Louvre Case Study . . . 145
6.3.1 The Louvre Setting . . . 146
6.3.2 Trajectory Dataset and Model . . . 148
6.3.2.1 Trajectory Dataset Description & Data Quality Issues . . . 148
6.3.2.2 Smartphone Application Usage Dataset Description . . . 151
6.3.2.3 Trajectory Data Model Instantiation . . . 156
6.4 Louvre Visitor Experimental Results . . . 165
6.4.1 Trajectory Data Preprocessing . . . 165
6.4.2 Preliminary Statistical Analysis . . . 167
6.4.3 Visualization & Standard Sequential Pattern Mining . . . 168
6.4.4 SITM-based State-Of-The-Art Sequential Pattern Mining . . . 171
6.5 Chapter Conclusion . . . 178

6.1 Introduction

In this chapter, the usefulness of the proposed trajectory modeling proposal from chapter 3 is illustrated, with a real-world case study concerning the Louvre Museum, in an effort to provide a pragmatic view of what it represents and how it can be used for the analysis of trajectories. The rest of this chapter is dedicated to experiments conducted in order to analyze the mobility behavior of the Louvre visitors, either based on statistical analysis or on Trajectory Pattern Mining (T-PM) methods applied over trajectories represented according to the proposed $SITM$. Our main aim is to validate how state-of-the-art mining algorithms can be applied on indoor trajectory data, outline their advantages and limitations, and gain a clearer perspective on how our novel Trajectory Pattern Mining algorithm proposed in chapter 5 should be implemented. Equally importantly, this chapter offers a rare account of current and future trajectory data analytics practices in a museum case study, unique in terms of exhibition size, architectural environment, visit rates, and cultural significance. Hence, it lends itself perfectly for identifying the real-world difficulties related to data acquisition and data quality, which may not be present in artificially generated or smaller scale case studies. Therefore, this chapter also bears a high educational value.

6.2 The Museum Application Domain

As stated in section 1.1.3, museums invest a lot of effort in studying the *visiting experience* of their public, as it relates to the motivations of their visit, the way in which they engage with the exhibition and the individual exhibits, the stimulation of their curiosity and the satisfaction of their expectations, even the fulfillment of the museums' educational mission towards them. Thus, they have long been conducting observational studies of their visitor's mobility behavior[171], a particularly effective method of acquiring insight about their overall experience of the visit.

Whereas similar observational studies do not scale well when performed in person, nowadays observational data of visitor movement can be automatically and massively amassed, thanks to the prevalence of low-cost indoor tracking infrastructure [127]. Especially, for internationally renowned museums, the collected trajectory datasets can quickly become very large, given their sheer surface size (e.g. over 652,000 square feet for the Louvre) and the daily tracking of tens of thousands of visitors. At the same time, the required infrastructure mainly includes the deployment of low-cost sensors based on diverse wireless indoor positioning, even communication, technologies such as IR, RFID, WiFi, Bluetooth, UWB, Zigbee, etc., as well as the use of smartphones and portable electronic guides.

Actually, the idea of using handheld devices in conjunction with sensors deployed in the museum environment, dates back to the start of this century. For instance, in [96], Kindberg et al. proposed the usage of portable digital assistants (PDAs)[1]

[1] A mobile device type now displaced by smartphones.

Goal Type	Goal	Beneficiary	Indicative Use Case
Visitor Experience	G1: Visitor experience personalization	Individual Visitor	Adapting the delivery of multimedia content according to visitor location.
	G2: Accessibility promotion (i.e. meeting the needs of atypical visitors)		Designing specialized itineraries avoiding large crowds for visitors with autism or social anxiety disorders.
	G3: Dynamic tour proposal		Using streams of location data of other visitors in combination with current visitor preferences to update in real-time the itineraries proposed to that visitor.
	G4: Social interaction promotion		Exploiting demographics and visitor preferences to propose individual visitors with similar interests to form groups.
Managerial Decision Making	G5: Intragroup visitor dynamics study	Museum Organization	Identifying groups of visitors and then profiling them according to how often or far they are likely to split during the visit.
	G6: Location-based services evaluation		Comparing different indoor positioning technologies to be embedded in the museum's electronic guide.
	G7: Visitor Profiling		Identifying types of visitors based on how they move and comparing them to existing ones obtained through conventional observation studies.
	G8: Visitor behavior quantification		Deriving new metrics such as visitor resting time in proportion to total visit time.
Crowd Management	G8: Emergency response planning	Visitor Crowd	Improving the evacuation routes based on mobility patterns from past emergency occasions.
	G9: Visitor flow control		Changing the accessibility of spaces to strengthen particular visitor flows and prevent bottlenecks.
	G10: Optimization of the exhibition spaces' spatial organization/arrangement		Changing the accessibility of spaces to strengthen particular visitor flows and prevent bottlenecks.

TABLE 6.1: Our classification of visitor trajectory data mining and analysis goals in the museum domain [100]

to receive URLs correlated with a physical object (e.g. artwork) from wireless infrared transceivers. Two decades later, while it is true that sensors like proximity beacons are affordable even for small museums, due to financial pressure and lack of technical skills, only museums of a certain size have so far afforded to offer complete digital services to their visitors (e.g. website, applications, multimedia guides and tools) [61]. For those museums in particular, traditional museum audio guides have gradually transformed into full-fledged multimedia devices, bearing multiple functionalities, including location-based services. Thus, the device positioning data can be extracted, collected, and used to structure individual visit trajectories. Furthermore, these trajectories may then be enhanced with semantic and contextual information, again originating from the electronic guide.

This is why in [100], the author proposes a first-ever classification of the types of analyses that can be achieved through computational analysis of museum visit trajectory data. Both the museum's visitors and management can benefit from such practices, as long as museum professionals are aware of their potential benefits. As shown in Table 6.1, the identified analytical goals serve both the visitors and the museum management. Unfortunately however, there is a gap between the current state-of-the-art in indoor trajectory analysis (reviewed in chapter 2) and a principled holistic approach of trajectory data mining in general, let alone in the museum context.

More specifically, contextual/semantic data should be used to add meaning to the spatiotemporal visit data. Similar to other domains, the additional information may target all aspects of the trajectory model:

- the moving object (e.g. a visitor's demographics, fatigue level, favorite artists, reduced mobility)
- the spatial entities(e.g. an artwork's type, a museum zone's theme, an exhibition room hosting a small shop, a room's level of congestion)
- the temporal dimension (e.g. lunch break, free-entry day, time-slot allocated tickets)
- the movement itself or any part of it (e.g. group following, hasty visit, resting, being lost, leaving the premises)

Even less important modeling elements may be semantically enriched, such as the observer of the movement (e.g. tracking technology, sensor accuracy, sensor range), or the connection of the spatial entities (e.g. mean transition time, vertical connection type). The plurality of possible semantic aspects is actually closely connected to the heterogeneity of their sources: text annotations, vocabularies, categorical/attribute data, ontologies, linked open data, etc.

What is more, in many cases the localisation process suffers from imprecision or other quality issues. This may happen due to the museum's architecture and the various mobile obstacles found in it, particularly other visitors or even the exhibits themselves. It is especially true for museums whose original function was not related to the housing of art collections, and whose architectural design is therefore not really compatible with their role.

Those museums also face an additional challenge related to the function of space often being ambivalent. For example the Louvre, originally built as a fortress and subsequently turned into a palace, before becoming a museum in 1793, contains many spaces that serve multiple functions (Figure 6.1). For both cases illustrated, it is nearly impossible to analyze properly or interpret correctly the visitor trajectories without taking into account the semantics of space.

FIGURE 6.1: Visitors often use The Daru staircase surrounding the Winged Victory of Samothrace to rest (left). The Salon Denon has a triple function: it hosts a museum shop, it serves as a major junction point, and it houses many large-scale neoclassical paintings (right).

Finally, due to the volume and heterogeneity of the visit data, the museum visitor movement analysis problem falls right within the emerging field of Big Trajectory Data analytics [179]. Within this context riddled with challenges but also opportunities for innovation, *SITM* will be implemented to the particular case of the world's most frequented museum [122], the Louvre Museum in Paris.

By intertwining a semantic model of visitor trajectories with a hierarchical model of the museum space, more elaborate types of movement analysis can be supported. Especially, the graph-based representation of the museum space can serve to model permanent space semantics in the form of node classes or attributes, whereas the semantic annotations can serve to model the more dynamic semantics, related to the moving object or to the evolving movement.

6.3 The Louvre Case Study

This section introduces the broader context of the Louvre application case, which motivated our work and served as a testbed for the *SITM* trajectory model, but also as an opportunity to aid the museum management in its study of the visitors' mobility data. The analysis results (both positive and negative) presented in this section also formed part of an internal tech report [178], conducted by the museum for the scope of identifying visitor data sources and practices, their potential, and their limitations.

6.3.1 The Louvre Setting

Our research partnership with the Louvre coincides with the museum's 2016-2020 research plan [104] which focuses on the exploration of cultural Big Data and the formation of multidisciplinary strategies for developing, supporting, and making transparent the collection and usage of such data. This of course includes automated tracking data. From the Louvre's perspective, the in-depth study of Big Data is expected to improve its knowledge about the dynamics of attendance, the related tools, and their future evolution.

With respect to the study of automatically collected visitor movement data, the Louvre has been considering it since 2006, when according to a report [49] on the modernization of the museum's audioguide system it was deemed that *"the development of mobile visit aid systems in exterior environments is currently favoured thanks to the stability of the GSM (satellite) network"* but that *"this technology is inoperable inside the museum, whose buildings need to be equipped with an internal wireless communication network (WiFi, Bluetooth) or RFID tags."*. Hence, even for a museum of the Louvre's magnitude, the necessary infrastructure was a prohibitive factor fifteen years ago.

Four years later, in April 2010, a small set of Bluetooth proximity sensors was temporarily installed as part of a series of visitor mobility research works which have derived interesting conclusions about visiting behaviors [191–195]. For instance, it was verified that, the length of stay in the museum tends to decrease towards the closing hours of the museum, that the earlier a visitor enters the museum the longer that visitor can be predicted to stay, and that short-stay visitors exhibit stronger patterns than long-stay visitors. These works have been based upon a statistical analysis of symbolic trajectories, where each sensor corresponds to a so-called *node*. These nodes are simply the spatial regions of fixed size illustrated in Figure 6.2.

In specific, the aforementioned works by Yoshimura et al. calculate metrics such as the following: node stay duration, museum visit duration, node-to-node travel time and transition rates, node sequence length (i.e. number of visited nodes) and probability, number of unique visited nodes, number of devices per node and per path, node-to-node distribution/transition rates. The authors also compare these measurements with respect to each other and to the hour of the day. In [194] in specific, the authors consider how frequent sequential patterns in the trajectory data compare to a random walk simulation model, induced from a simple graph representation of the museum's structure, but nevertheless do not apply sequential pattern mining methods.

Since then, permanent data collection mechanisms have been installed in the Louvre's premises. In April 2012, the Nintendo 3DS console system became the Louvre's official electronic guide, containing photographs, audio commentary, high resolution images and 3D models of the exhibits, all aimed at increasing appreciation of the artworks. It also functions as a navigation device showing the quickest way to an artwork of interest. The position of the visitors is tracked thanks to the installation of approximately 500 WiFi beacons (top right in left part of Figure 6.3) scattered across the permanent exhibition premises. The consoles' usage data are being stored in a

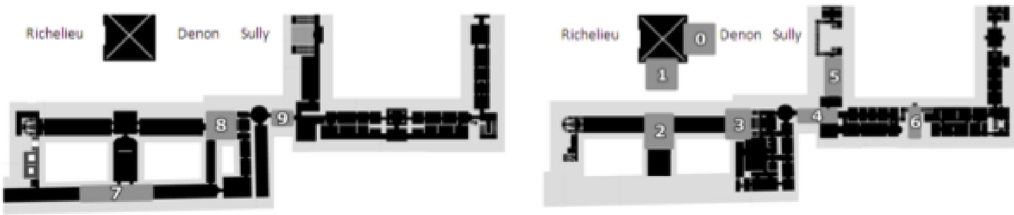


FIGURE 6.2: The symbolic spatial regions used in past analysis studies of the Louvre's visitor movement data [191–193, 195]. Area coverage has since improved considerably.

big relational database silo, but unfortunately this dataset is not addressed within the scope of this Thesis. The usage of the Nintendo 3DS console guide has also been studied with respect to visitor profiling and the study results have been made public in [154], but this study did not consider the visitors' mobility data.

More potential sources of visitor tracking data do exist such as a network of 150 Wi-Fi public hotspots, which goes to show the heterogeneity of tracking data sources in big museums. More generally, as newer technological solutions become available, it can be expect from forward thinking museums to adopt technological innovations that will enable the study of visitor mobility, while at the same time respecting privacy restrictions [137].

Bluetooth (smartphone app) and WiFi (Nintendo 3DS) beacons installed throughout the Louvre.

Usage of the Louvre's Nintendo 3DS and smartphone application guides is what enables its visitors' location detection.

FIGURE 6.3: Visitor tracking data acquisition infrastructure used in the Louvre.

In July 2016, the Louvre launched its official *My Visit to the Louvre*[2] smartphone application, which takes advantage of a very large Bluetooth Low Energy (BLE)

[2]The french version of the application was named *Louvre: Ma Visite*.

beacon infrastructure[3] depicted in Figure 6.3, and the smartphone's accelerometer and compass, in order to estimate the visitors' precise $(lat, long)$ coordinate position within the museum. This is accomplished via BLE Received Signal Strength Indicator (RSSI)-based trilateration, extended Kalman and particle filtering techniques. Then, the positions get reported every second, as long as the device has internet connection, or else they get stored in the app and pushed to the cloud after Internet access has been restored. The application visualizes the position over a locally stored version of the museum map for navigation purposes. It was discontinued and stopped being available on 01-10-2019.

6.3.2 Trajectory Dataset and Model

Having described the data sources, this section presents the trajectory dataset used for the experimental analysis and the validation of our model's usefulness. Coming from the world's most frequented museum [117–122], it is also of practical value for identifying problems related to the large-scale collection and retention of trajectory, such as issues of data quality and integration problems.

6.3.2.1 Trajectory Dataset Description & Data Quality Issues

In the obtained JSON dataset, raw geometric positions have already been spatially aggregated into 52 non-overlapping zones. Each zone corresponds to a large polygonal area of the museum, as can be seen in Figure 5.4, specified by the museum administration in such a way so as to reflect a single exhibition theme (e.g. *Italian paintings*) but also only extend within a single floor.

Big museums can be a vast source of Big Trajectory Data, especially in terms of their volume and variety. The conceptual trajectory model presented in section 3.3.2 aims at supporting the development of analysis techniques making use of such large datasets, produced on a daily basis by thousands of visitors. In the analysis of historic trajectory datasets for example, the trajectory traces (Definition 3.3.3) themselves can be used at the pre-processing phase, in order to annotate semantically the trajectories with the level of congestion in the corresponding spatial region.

However, the results presented in this Thesis are based on a more modest dataset consisting of 4,945 visits, continuously collected from 19-01-2017 to 29-05-2017, each composed of a sequence of timestamped *zone detections* i.e. detections of the visitor's smartphone inside a certain zone. The duration of a visit ranges from 0 sec (considered as an error) to 7 hours 41 min and 37 sec, whereas the duration of a zone detection ranges from 0 sec (considered as an error) to 5 hours 39 min and 20 sec. The visits were performed by 3,228 different visitors using both the iPhone and Android versions of the application. Out of those, 1,227 were returning visitors who made 1,717 second or third visits (not necessarily on different days). The dataset includes 20,245 zone detections and 15,300 (intra-visit) zone transitions in total.

[3] In specific, 1800 beacons were installed across all five floors of the museum, considerably improving tracking coverage and continuity, with respect to all earlier tracking infrastructure.

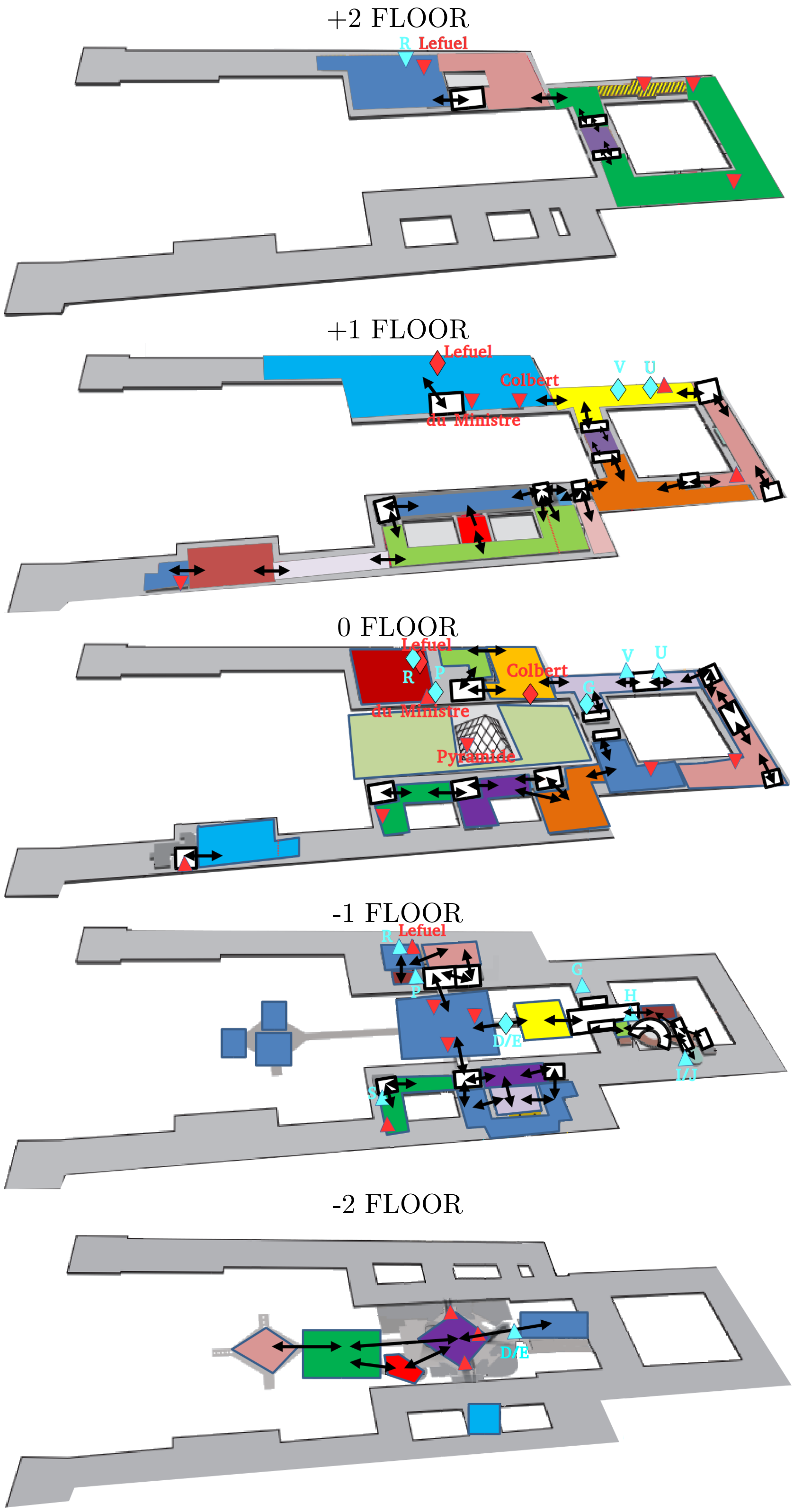


FIGURE 6.4: Thematic zones of the Louvre Museum's five floors, hand-annotated with accessibility relations within or across floors. White zones are not covered by the Bluetooth beacons but are included in our graph-based representation of indoor space.

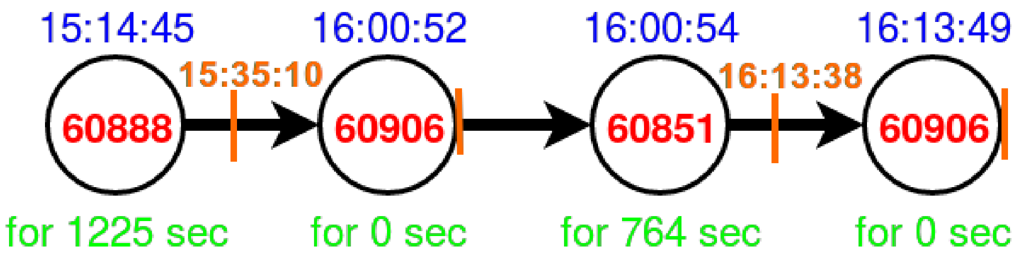


FIGURE 6.5: Visual representation of a real Louvre visit trajectory, structured as a sequence of detection intervals inside zones (red). Since "Begin" detection timestamps (blue) do not coincide with their respective previous "End" timestamps (orange), there are detection gaps.

Unfortunately, only 30 out of the 52 zones appear in the movement dataset, with the -1 floor completely missing. Apart from some beacons not recording data during that particular period, additional factors that may explain the movement dataset's sparsity include the following:

- A visitor may launch and/or close the application mid-visit (due to battery depletion, sporadic navigation-only usage, etc.) resulting in its partial recording[4].
- A visitor may deactivate the phone's Bluetooth while the application is running.
- The period of data collection is characterized by lower adoption rates and potentially transitory phenomena, because it coincides with the operational launch of both the app and the beacon infrastructure.
- 10.55% of the zone detections have a duration value equal to 0, forcing us to filter them out as detection errors.

The sparsity of the data is partly expressed by the power law distribution of the length of visit presented in Figure 6.6. It can be seen there that 53.55% of the visits actually degenerate into a single zone detection. Out of that percentage, only 2.61% is due to erroneous detections of zero duration, therefore practically one out of every two visits has a length of only 1. Obviously then, the dataset quality is not the desired.

What is more, the raw tracking dataset also includes periods of non-detection and periods of double detection, of the moving object. For example, for visit #1485530085 performed by visitor number #4575878 and depicted in Figure 6.5, there exist temporal gaps between each "End" timestamp and the next "Begin" timestamp. These gaps denote that in the meantime the visitor was not being tracked. As another example in Figure 6.7, by projecting the nominal value of the museum zone on the y-axis, we can see in a more pronounced way how a visitor's zone sequence 60907 → 60908 → 60907 is characterized by a double detection: while detected in zone 60908, the visitor was also continuing to be detected in zone 60907, which is of course impossible given that the zones are non-overlapping.

[4] This is supported by the fact that most of the frequent visits in the dataset do not satisfy the constraint of beginning in an entry zone and ending in an exit zone.

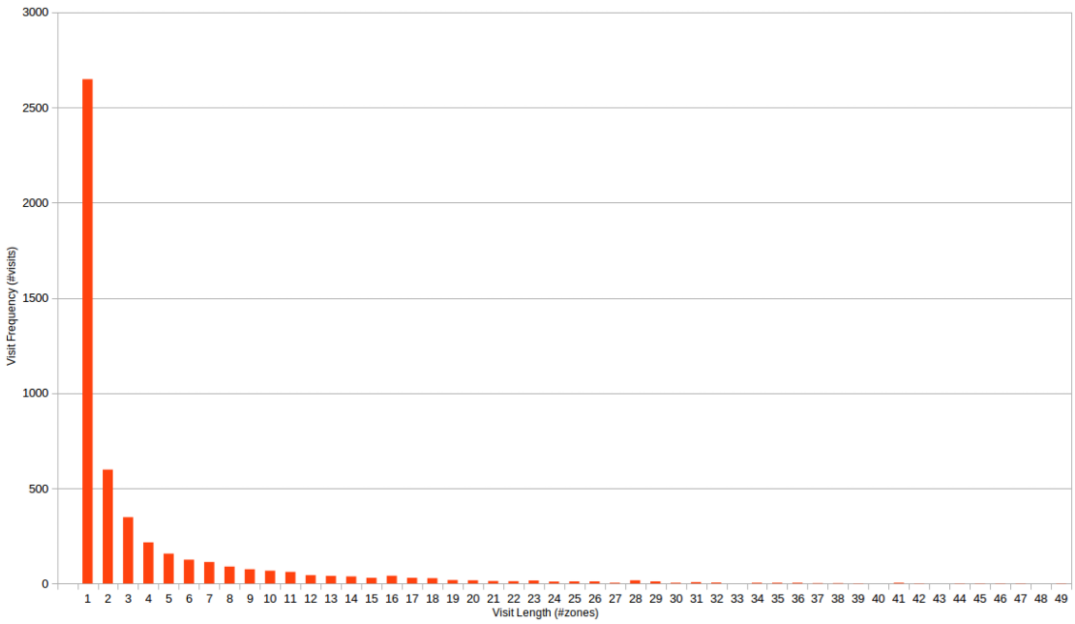


FIGURE 6.6: Bar chart illustrating the trajectory dataset's distribution of visit length.

Different approaches for correcting multiple detection data can be implemented using our $SITM$, each corresponding to a different assumption. For example, they can be based on the duration t_{end} - t_{start} (Definition 3.3.2) of each detection interval, and lead to the conclusion that in this particular case the visitor did not really pass from zone 60908, because 10 seconds is not long enough in comparison to the surrounding zone detections. Or alternatively, the accessibility relationships E_i^{acc} from the zone layer NRG (Definition 3.3.1) can be checked, in which case the zone sequence 60907 → 60908 → 60907 is valid, because it is indeed possible to transition between the two zones. Perhaps the visitor really entered 60908 for a little while and then returned.

Similarly, for cases of non-detection, one can simply assign the visitor to either of the two spatial regions where he or she was right before or right after the "disappearance", or instead consult the building's topology, for a more informed decision, especially in the cases of longer detection gaps. The hierarchical multigraph of $SITM$ allows even to complete such gaps in the trajectories with location information at a coarser level (e.g. floor, wing)) of spatial granularity, in case there is enough confidence about it, but not about the specific position.

6.3.2.2 Smartphone Application Usage Dataset Description

In addition to the trajectory dataset described in the previous section, the author obtained access to CSV log files recorded from the Yahoo Flurry analytics platform used in the *My Visit to the Louvre* smartphone application. This platform actually captures various metrics and statistics, but except for the aforementioned event logs, all other are aggregates (hourly at best) that can not be mapped to individual visits.

```
{
    "zone": {
        "id": 60907,
        "name": "N1-Sully-Antiquités égyptiennes",
        "tags": [
            "stats"
        ],
        "level": 1
    },
    "begin_at": "2017-05-15T07:53:48.000Z",
    "duration": 101,
    "nextZone": 60906
},
{
    "zone": {
        "id": 60906,
        "name": "N1-Sully-Antiquités grecques et romaines",
        "tags": [
            "stats"
        ],
        "level": 1
    },
    "begin_at": "2017-05-15T07:55:17.000Z",
    "duration": 10,
    "nextZone": 60907
},
{
    "zone": {
        "id": 60907,
        "name": "N1-Sully-Antiquités égyptiennes",
        "tags": [
            "stats"
        ],
        "level": 1
    },
    "begin_at": "2017-05-15T07:55:30.000Z",
    "duration": 38,
    "nextZone": 60891
},
```

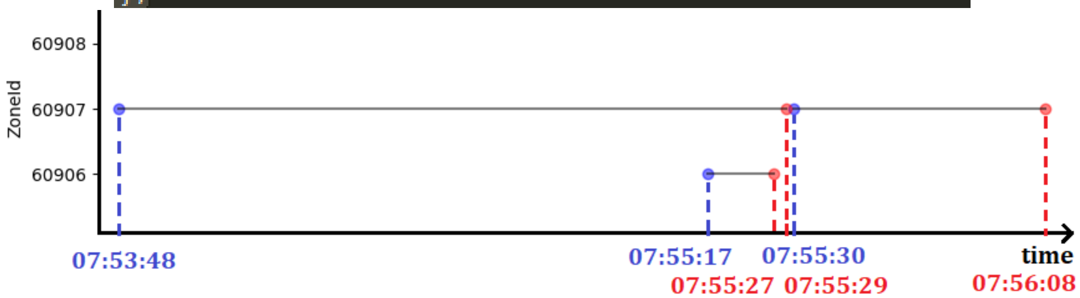



FIGURE 6.7: Part of a real Louvre visit showcasing a detection overlap example.

Timestamp	**Session Index**	**Event**	**Description**	**Version**
Nov 07, 2017 01:17 AM	2	map_popup	-	2.0.26
Platform	**Device**	**User Id**	**Params**	
Android	Motorola Moto G4 Plus	396f8c54db4ceb67	{ poi_idx : 92}	

TABLE 6.2: A single log record created by the *map_popup* event of Figure 6.8.

Unfortunately, this includes demographic information which is of particular interest for museum visitor studies[5], since it allows them to identify target groups based on cultural, ethnic, or social affiliation, educational level, leisure preferences, etc[164].

The logs contain records of *app usage events* grouped into *sessions* which describe how users interact with the app, irrespective of whether they are visiting the museum at the time of usage or not. More specifically, they contain 28 different types of events for the iOS version and 23 different types of events for the Android version. Out of those, 18 are common (including their parameters) to both app versions.

The records begin in 30-01-2017 and therefore a subset corresponding to the period from that day to 29-05-2017 was used, in order to match the trajectory dataset's collection period as closely as possible, given our primary objective to semantically enrich the trajectories. Hence, there were 159,742 *sessions* left containing 717,580 *events* (belonging to 33 different *event types*) performed by 21,935 unique *devices*. Out of those, 3,406 unique Android User Id's[6] were identified in the museum at the time of usage. This is a reasonable number given the 3,228 unique Android and iPhone visitors of the trajectory dataset corresponding to that period. Some event types (e.g. *profile_validation*, *map_popup*, *poi_audio_time*) included in the logs hold useful information in the form of parameter values, which mainly indicate what the visitor was interested in learning about, or which exhibit the visitor was trying to locate in the museum. In Figure 6.8, such a *map_popup* event is illustrated from the application usage's point of view. Such usage will result in a log record getting created:

Most event types however are not at all interesting with respect to trajectory enrichment, as they typically concern User Interface navigation and control actions (e.g. *home_burger*, *skip_intro*, *map_keypad*).

Of particular interest is the *app_launch_localisation* event which exists only in the Android version of the application, and captures the GPS coordinates of the device at the time of app launch. These coordinates are only captured when inside a circle of $\approx$ 1km radius centered near the Glass Pyramid of the Louvre. For the time period in question, there exist 8,336 such localization events (less than the total number of sessions since these are Android only and on-site only events).

In Figure 6.10, the temporal distribution of these events signaling the launch of the smartphone application is drawn. As expected, there are very few people launching the application on Tuesdays, since that is the only day of the week that

[5]Except for a "countryISO" code ("FR", "US", etc.) field assigned to each session individually.

[6]The number of iOS User IDs was even greater but contained an unknown number of duplicates, because any time the app is uninstalled and re-installed again on an iPhone, a new User ID was generated.

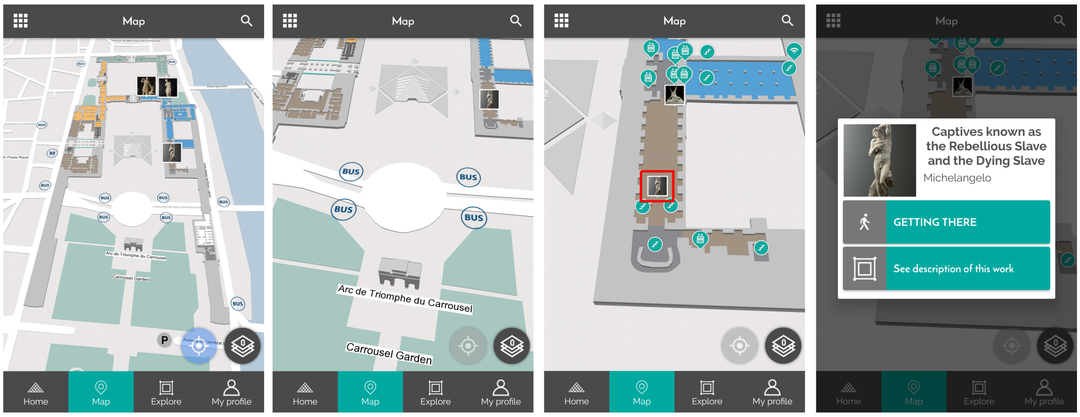
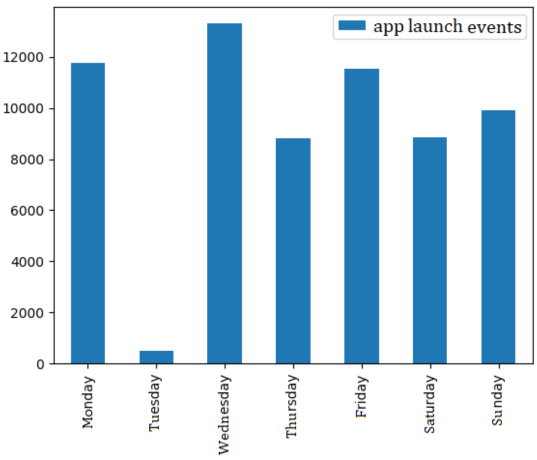
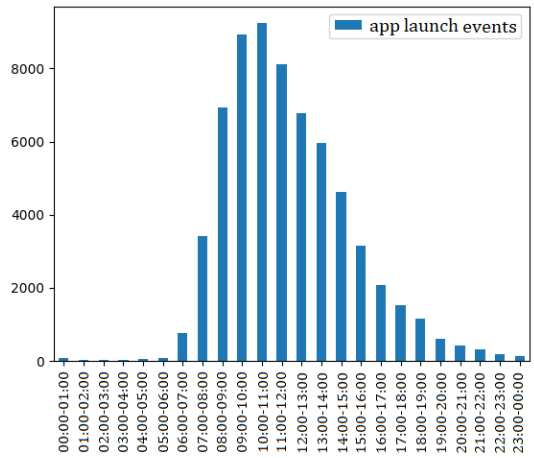


FIGURE 6.8: While in the map tab of the *My Visit to the Louvre* app, the visitor can find out more about individual exhibits, in this case Michelangelo's famous Slave sculptures.

the Louvre is closed to the public. Also, the peak hour in terms of launching the app is from 10am to 11am, which is in agreement with the museum's perception of peak arrival times. Moreover, whereas after reopening from the COVID-19 pandemic the museum's opening hours are from 9am to 6pm, back in 2017 the closing time on Wednesdays and Fridays was moved to 21:45 to offer nocturnal visits. This explains the slightly increased attendance levels of those two days. It also justifies the existence of a small but not insignificant amount of application launches at night.



App launch localization events per day.



App launch localization events per hour.

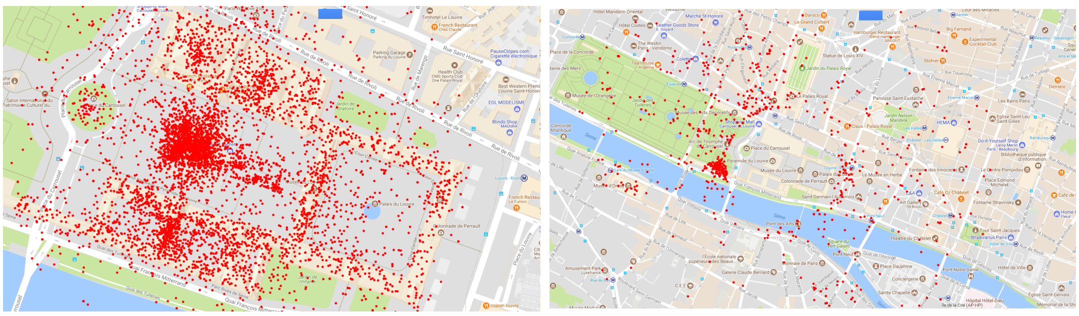
FIGURE 6.9: Temporal distribution of the smartphone application's launch events.

In Figure 6.10, the GPS coordinates of the same event type's instances as before are visualized over a map of the area. It can be seen that the visitors start using the smartphone application in different parts of the Museum, but there are two clearly visible clusters: around the Glass Pyramid in the Napoleon Court and in the middle of the Denon wing. The latter happens to be exactly where the Mona Lisa is located, although the positional data lacks any floor information or z-coordinate value to verify any such correlation. A reasonable explanation however, is that perhaps visitors get the chance to launch the application while standing in line, waiting either to enter the Museum or to see the Mona Lisa. On the other hand, for the case of the Glass Pyramid in particular, we can not be certain whether they are indeed out in the open or have already entered through the Pyramid when launching the app, which would just record the last known location of the device in case it can't properly receive a satellite signal from within the museum.

Still, despite a few more smaller clusters, such as in the Carrousel area, or near the Porte des Lions (where perhaps visitors regroup before entering the museum), the overall distribution of the points indicates that the visitors do not consistently launch the application exactly at the start of the visit. Instead, they do so both in advance and during their visit, since the entire museum premises are covered, not just the entry points. Here, it is important to mention that according to the Flurry platform, if the app pauses or moves to the background for more than 10 seconds, then the next time it runs, the Flurry agent will automatically create a new session and end the previous one. Therefore there are signs of discontinued usage of the application.

Moreover, a small number of people seems to be launching the application from the Musée d'Orsay which is not surprising, given that essentially it is the same type of public who is interested in visiting both museums. A typical behavior is for them to visit the Louvre on Mondays when the Musée d'Orsay is closed, and visit the the Musée d'Orsay on Tuesdays when the Louvre is closed, or a few may even combine both visits on the same day. What is more, out of all the app launches that took place in or around the museum, 908 (or 10.893%) took place around it, which suggests that at least one in ten visitors that use the app, starts doing so clearly in preparation to the visit (or less likely after it).

Finally, it should be noted since GPS signals are blocked and reflected by walls, it is often impossible to calculate the location when inside the museum due to the insufficient signal strength. In that case, More specifically, the 42,473 events initiating a session correspond to 34,460 unique coordinate pairs, duplicates are perhaps due to the GPS system reporting the last known coordinates when new ones are unavailable).



88.3% app launches took place in the museum.

11.7% app launches took place around the museum.

FIGURE 6.10: The GPS coordinate positions where the Louvre visitors launched the app.

6.3.2.3 Trajectory Data Model Instantiation

This section describes how each visit can be modeled as a trajectory, so that visiting behaviors can be discovered and studied in the form of elements or parts of such trajectories.

Indoor space representation.
In order to instantiate the $SITM$ that was proposed in section 3.3.2 for the Louvre case study, a representation of the museum's indoor spaces according to the graph-based structure formalized in Definition 3.3.1 is needed. Although the Louvre's multi-layered graph is prohibitively large to be shown, its correspondence to Figure 3.1 is explained here in a top-to-bottom fashion:

- Layer 4 is instantiated as the whole *Louvre Museum*: it represents a level above any specific building, denoting presence in the museum in general.

- Layer 3 is instantiated as the museum's three wings (*Richelieu*, *Denon*, and *Sully*) as well as the *Napoleon* area that they surround and which contains the Glass Pyramid: it represents the museum's main structural parts as separate buildings, given that their spaces and usage are those of a typical building.

- Layer 2 is instantiated as a wing's five different floors (-2, -1, 0, +1, +2), except for the *Napoleon* area which is only built underground and thus does not have +1, +2 floors.

- Layer 1 is instantiated as a floor's rooms and halls (hundreds in total).

- Layer 0 is instantiated as a room's most important exhibits in the form of Regions of Interest (several hundreds in total): it represents predefined fully-navigable (without any holes) spatial areas of engagement with each exhibit, outside of which a visitor is certain not to be paying attention to it.

Moreover, a semantic layer representing the *thematic zones* present in the available trajectory dataset (described in section 6.3.2.1) is added. This layer happens to fall right between Layer 2 and Layer 1. Both intra-floor (e.g. door, ramp) and inter-floor (e.g. staircases, elevators) zone accessibility topology was extracted on site (Figure 6.4) and used to derive the zone layer NRG (Figure 6.11). It does not however include zones missing from the data, nor any additional indoor areas needed to completely cover the navigable space.

This brings forth an interesting space modeling decision concerning whether or not to assume that the spatial region represented by a node in layer i+1 is *fully covered* by the union of the spatial regions represented by its child nodes in layer i. For example, is a floor fully covered by the rooms it contains? Similarly, is a room fully covered by the RoIs it contains? Or are there coverage gaps as in Figure 6.4?

Although not explicitly stated, the IndoorGML standard and related works (e.g. [91]) adhere to a *full space coverage* assumption. The same is considered to be true

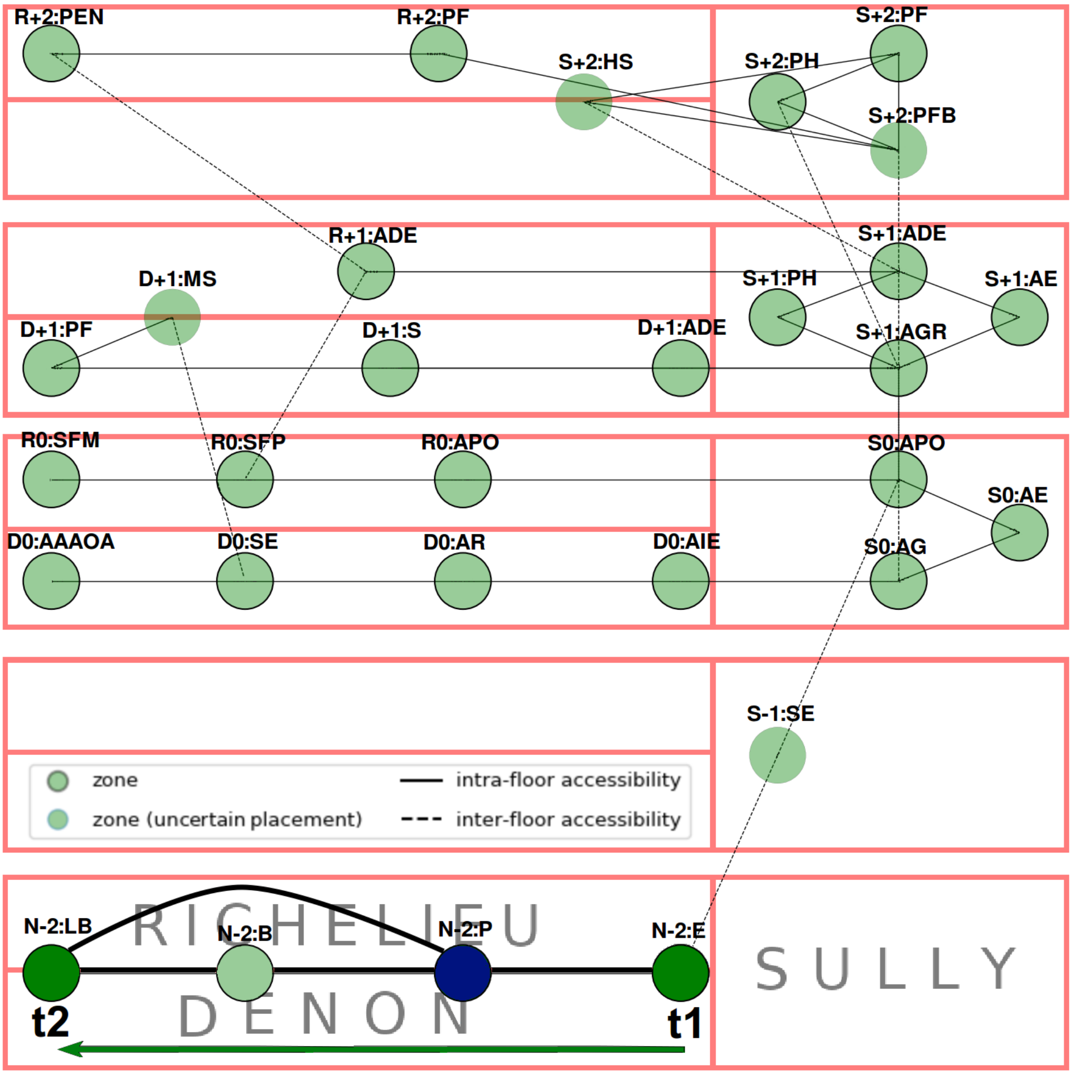


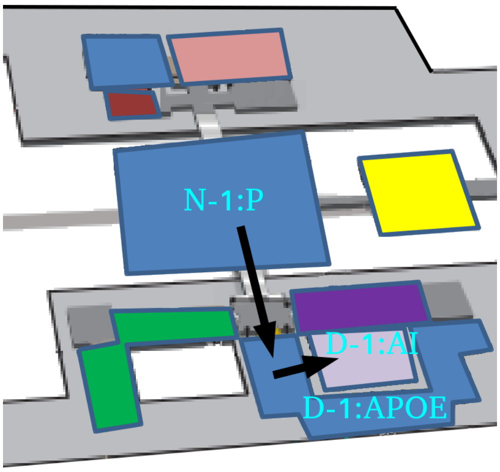
FIGURE 6.11: Based on the chain topology of zones (denoted by alphanumeric IDs), a visitor's presence in the blue zone can be inferred, even when undetected.

with respect to the three-layer core hierarchies. This has the advantage that accessibility relations need only be captured at the lowest possible level of the hierarchy, from where they can be exhaustively inferred for the higher levels.

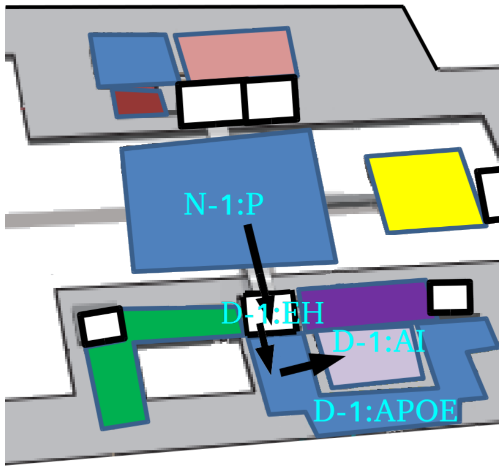
In the Louvre example, if each wing's floor (parent node) is assumed to be fully covered by its zones (child nodes), then the available zone-level accessibility topology is enough to automatically deduce the floor-level one, because if two floors contain no zones that are directly accessible one from the other, then neither can these floors themselves be reciprocally directly accessible.

The full space coverage assumption is closely related to a stronger *full movement detection* assumption, which requires that, not only does the indoor space representation (i.e. our graph model) completely cover the areas where the moving object may find itself in, but also that each of those areas is observable with respect to the detection data acquisition mechanism.

The proposed model does not make this often unrealistic (e.g. for proximity detection devices [124]) assumption. Simply put, if at some point the moving object is not detected anywhere, then its position is considered to be *unknown* and open to further estimation. A complete indoor space topology is then enough to repair the trajectory data, by inferring the presence of the moving object in non-observable - yet modeled - areas (Figure 6.12), just as it is enough to filter out impossible transitions as already explained in the example of Figure 6.7.



Holes may arise when only part of the navigable indoor space is represented.



Under full indoor space coverage, trajectories with holes can be restored.

FIGURE 6.12: Taking into account non-observable areas in the modeling of space can help obtain trajectories that are more faithful to the actual real-world movement.

However, at the lowest level of spatial granularity full coverage is not always easy to implement. Let us assume that RoIs represent the displayed exhibits or even some facility installations like "you-are-here" maps (Figure 6.15). These RoIs will almost certainly incompletely cover the surface of the room that they belong to. In a room full of paintings for example, there most of the times there are at least some spots near the center of the room, where a visitor is not standing close enough to focus on any exhibit. Similarly, when a visitor transitions from the *Beatrice d'Este* RoI to the *Battle Scene* RoI within Room 403 (Figure 6.13), his/her trace is briefly lost, because the two regions are disjoint and thus not directly accessible from each other.

If deemed necessary, *SITM* can address this by adding to the RoI-layer NRG, a single *complement* node representing the spatial area of the room excluding all areas covered by all of its RoI children nodes. It is then up to the application-level logic to infer whether the visitor is in the remaining area of the room (i.e. in the complement) or if perhaps left the room altogether. Obviously, the RoI topology plays a critical role in determining this.

An alternative approach would be to simply switch to a geometric representation at the intra-room level and encode all RoIs' surface geometries within it, effectively adopting a hybrid indoor space representation [5]. Despite the advantage of increased precision, this approach would need to rely on a correspondingly precise data acquisition infrastructure. Which of the two approaches is preferable depends highly on the particular case. It may well be enough even to simply assume the visitor's presence in the room containing the last known RoI detection, until further re-appearance in another RoI.

Semantic indoor trajectories representation.

Having instantiated the Louvre's indoor space representation, $SITM$ is used to extract from the zone detection dataset, the visit trajectories as sequences of presence intervals in the museum's thematic zones. For example, a complete one hour visit trajectory spanning three floors in the museum (Figure 6.16) can be encoded as the couple $T_{ID_{vis},12:00:00,13:00:00} = (trace_{ID_{vis},12:00:00,13:00:00}, \emptyset)$ whose trace is the following sequence of 19 presence intervals:

$trace_{ID_{vis},12:00:00,13:00:00} = \{$
(01) (*pyramid_control*, *"N-2:P"*, 12:00:00, 12:03:00, ∅),
(02) (*door001*, *"N-2:B"*, 12:03:00, 12:08:00, ∅),
(03) (*door001*, *"N-2:P"*, 12:08:00, 12:09:00, ∅),
(04) (*D_electric_stairs*, *"N-1:P"*, 12:09:00, 12:10:00, ∅),
(05) (*D_ticket_control*, *"D-1:EH"*, 12:10:00, 12:12:00, ∅),
(06) (*opening001*, *"D-1:APOE"*, 12:12:00, 12:16:00, ∅),
(07) (*door002*, *"D-1:AI"*, 12:16:00, 12:20:00, ∅),
(08) (*door003*, *"D-1:AG"*, 12:20:00, 12:30:00, ∅),
(09) (*opening002*, *"D-1:DS"*, 12:30:00, 12:31:30, ∅),
(10) (*Daru_stairs_-1_0*, *"D0:DS"*, 12:31:30, 12:36:00, ∅),
(11) (*opening003*, *"D0:AIE"*, 12:36:00, 12:38:00, ∅),
(12) (*opening004*, *"S0:AG"*, 12:38:00, 12:40:00, ∅),
(13) (*opening004*, *"S0:HIIS"*, 12:40:00, 12:41:00, ∅),
(14) (*HenryII_stairs_-1_0*, *"S-1:HIIS"*, 12:41:00, 12:42:00, ∅),
(15) (*opening004*, *"S-1:EH"*, 12:42:00, 12:44:00, ∅),
(16) (*opening005*, *"N-1:E"*, 12:44:00, 12:45:00, ∅),
(17) (*S_ticket_control*, *"N-1:P"*, 12:45:00, 13:46:00, ∅),
(18) (*S_electric_stairs*, *"N-2:P"*, 12:46:00, 13:47:00, ∅),
(19) (*opening006*, *"N-2:LB"*, 12:47:00, 13:00:00, ∅) }

The zones comprising this visit trajectory are described in detail in Table 6.3. It can be noticed that the beacon infrastructure does not cover all of them: tuples 5, 9, 10, 13, 14, and 15 represent *inferred* (rather than directly *observed*) visitor presence in the corresponding areas. Inferred tuples are derived thanks to the topology of indoor space. Alternatively, the representation can be limited to the actual observation data, in which case the trajectory will contain temporal gaps in their place.

In the above trajectory example, only its spatial and temporal dimensions were

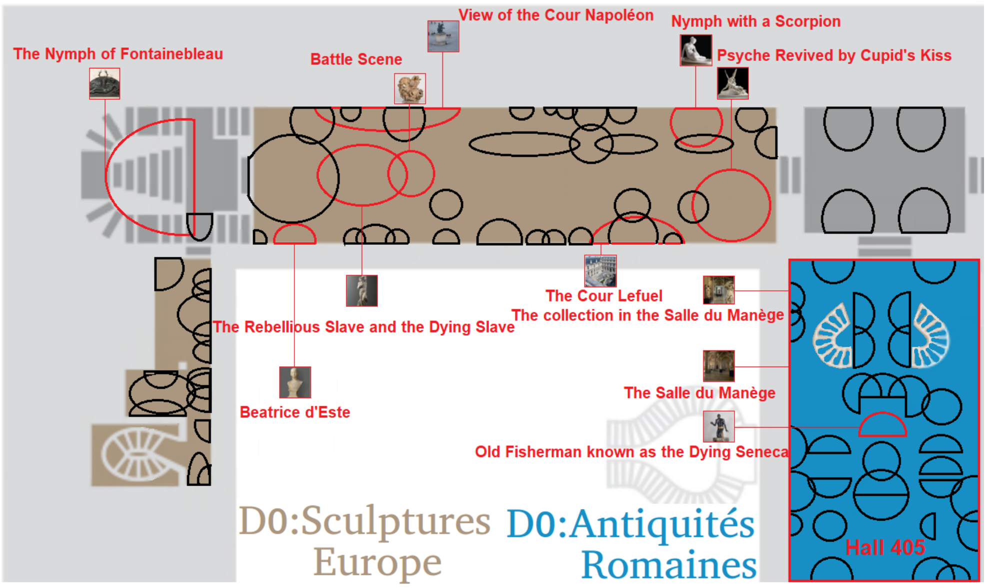


FIGURE 6.13: Indicative Regions of Interest belonging to two different thematic zones.

Abbrev.	Thematic Zone	Wing	Floor	Obser.	Adm.	Tuples
"N-2:P"	*Pyramide*	*Napoleon*	-2	Yes	Free	1, 3, 18
"N-2:B"	*Billetterie*	*Napoleon*	-2	Yes	Free	2
"N-1:P"	*Pyramide*	*Napoleon*	-1	Yes	Free	4, 17
"D-1:EH"	*Entrance Hall*	*Denon*	-1	No	Ticket	5
"D-1:APOE"	*Art du Proche Orient et de l'Egypte*	*Denon*	-1	Yes	Ticket	6
"D-1:AI"	*Art de l'Islam*	*Denon*	-1	Yes	Ticket	7
"D-1:AG"	*Antiquités Grecques*	*Denon*	-1	Yes	Ticket	8
"D-1:DS"	*Daru Staircase*	*Denon*	-1	No	Ticket	9
"D0:DS"	*Daru Staircase*	*Denon*	0	No	Ticket	10
"D0:AIE"	*Antiquités Italiques et Étrusques*	*Denon*	0	Yes	Ticket	11
"S0:AG"	*Antiquités Grecques*	*Sully*	0	Yes	Ticket	12
"S0:HIIS"	*Henry II Staircase*	*Sully*	0	No	Ticket	13
"S-1:HIIS"	*Henry II Staircase*	*Sully*	-1	No	Ticket	14
"S-1:EH"	*Entrance Hall*	*Sully*	-1	No	Ticket	15
"N-1:E"	*Exhibition*	*Napoleon*	-1	Yes	Ticket	16
"N-2:LB"	*Librairie, Boutiques*	*Napoleon*	-2	Yes	Ticket	19

TABLE 6.3: Information about the zones encountered in the trajectory of Figure 6.13.

taken into account[7]. However, the semantics of places also offer us valuable insight about the visitor's trajectory. For instance, beginning the visit in zone *"N-2:P"* is known to be normal because this is one of the Louvre's entrance zones (either through the Glass Pyramid or through the Richelieu passage). In general, validation against *entrance/exit zones* allows us to distinguish truncated visits from complete visits. As another example, zones *"D-1:DS"* and *"D0:DS"* represent the big *Daru staircase* which also serves as a resting place for visitors [193] (Figure 6.1), so it shouldn't come as a surprise that the visitor of the previous trajectory example spent nearly 5 minutes there.

More generally, all of the above are examples of *static* place semantics remaining the same throughout any given trajectory. These can be represented by node classes, since each spatial region corresponds to a node from the set V of Definition 3.3.1. For example, zones *"D0:DS"* and *"D-1:DS"* would both belong to a *staircase* and a *resting space* class of nodes. This way, what might at first seem as an inexplicably long time spent transitioning from one floor to the other, can now be appropriately interpreted and treated in the analysis. The Louvre actually contains numerous eponymous staircases whose various details can even be appreciated as artworks on their own [74] (e.g. *HenryII*, *HenryIV*, *Lefuel*, *Mollien*, *du Midi*, *de la Colonnade*). These merit to be represented as nodes in the room-layer NRG, instead of edges which is more fitting for smaller or insignificant staircases.

Another type of space semantics is *zone admissibility*, which can function as a criterion for dividing our previous example's main trajectory into three episodes:

- *arrival* (tuples $1-4$): presence in freely accessible zones
- *main visit* (tuples $5-16$): presence in zones requiring a ticket
- *departure* (tuples $17-19$): presence in freely accessible zones

Just like every semantic subtrajectory (Definition 3.3.4), an episode is assigned a semantic annotation set that reflects its overall meaning. For example:

- the *arrival* episode can be enriched with:
  $A'_{traj} = \{activities : [\text{“}buy\_ticket\text{”}, \text{“}enter\_permanent\_exhibition\text{”}]\}$
- the *main visit* episode can be enriched with:
  $A''_{traj} = \{activities : [\text{“}visit\_greek\_antiquities\text{”}], goals : [\text{“}visit\_Victoire\text{”}]\}$
- the *departure* episode can be enriched with:
  $A'''_{traj} = \{activities : [\text{“}buy\_souvenir\text{”}], goals : [\text{“}leave\_Louvre\text{”}]\}$

It is important to clarify here that the role of the trajectory model is to support such semantics, not necessarily to provide them. The semantic information itself may be either *explicitly given* in the form of additional data, or *derived implicitly* from the spatiotemporal movement data. In either case, the trajectory model has to enable and

[7] Timestamps are rounded for illustrative purposes

facilitate this process, but not define it, since the semantic enrichment of trajectories poses a task of its own[8].

In our previous example, we may explicitly know that the visitor bought a *normal ticket*, or we may derive it from the fact that he entered the permanent but not the temporary exhibition, which is hosted in zone *"N-2:E"* and requires a separate ticket. Similarly, we may explicitly know the visitor's interest in *Ancient Greek art* as stated in the mobile application's profile section, or we may infer it from the proportionally larger amount of time spent in the respective zones.

It is now also more apparent why our $SITM$ allows for *overlapping episodes* instead of requiring mutually exclusive episode predicates like for instance the model of Yan et al. [190].

Firstly, given the multiple spatiotemporal granularity levels at which movement can be characterized, the essence of any movement segment may be quite different if examined at a macroscopic or at a microscopic level. In the trajectory example of Figure 6.16, the segment consisting of tuples 1-5 corresponds to entering the Louvre's permanent exhibition space. However, the first part of it (tuples 1-3) corresponds more specifically to buying a ticket, and can therefore also meaningfully stand on its own.

More generally, one may wish to model situations where one or more episodes (e.g. *buy ticket*) are contained within a broader episode (e.g. *enter exhibition*), in turn taking place within an overarching episode (e.g. *see Mona Lisa*). There may even be multiple such containment instances within the same trajectory. This is true in the previous example, where apart from beginning of the trajectory, there is also an overlap at the end of it: tuples 17-19 correspond to exiting the museum premises (i.e. departure episode), whereas tuple 19 alone corresponds to shopping at the museum shop (i.e. *buy souvenir* episode).

Secondly, by allowing semantic hierarchies potentially independent from each other, cases can be modeled where an episode defined on the basis of one semantic dimension overlaps with an episode defined on the basis of another semantic dimension. To illustrate this, let us consider the trajectory example in Figure 6.14. The visit starts with the visitor spending very little time in highly congested rooms housing Italian Renaissance paintings (tuples 1-2), whereas next the visitor stays a lot longer in rooms housing Ancient Greek artworks (tuples 4,6). Thus, if both semantic dimensions (i.e. *congestion level* and *artwork theme*) are taken into account as well as the temporal dimension (i.e. period of stay in each room), it can be reasonably inferred that a *crowd-avoidance* behavior was driving the visit at first, followed by a particular interest in *Ancient Greek art*.

Let us now look more closely at the trace of the example visit:
$trace_{vis0058,16:00:00,16:45:00} = \{$
(01) (*door010*, *"Room710"*, 16:00:00, 16:01:00, {"high-congestion"}),
(02) (*door011*, *"Room709"*, 16:01:00, 16:02:00, {"high-congestion"}),
(03) (*door012*, *"D1:DS"*, 16:02:00, 16:05:00, {"low-congestion"}),

[8] We classify it under Trajectory Data Preprocessing tasks in Table tab:2-Classification of trajectory data mining tasks.

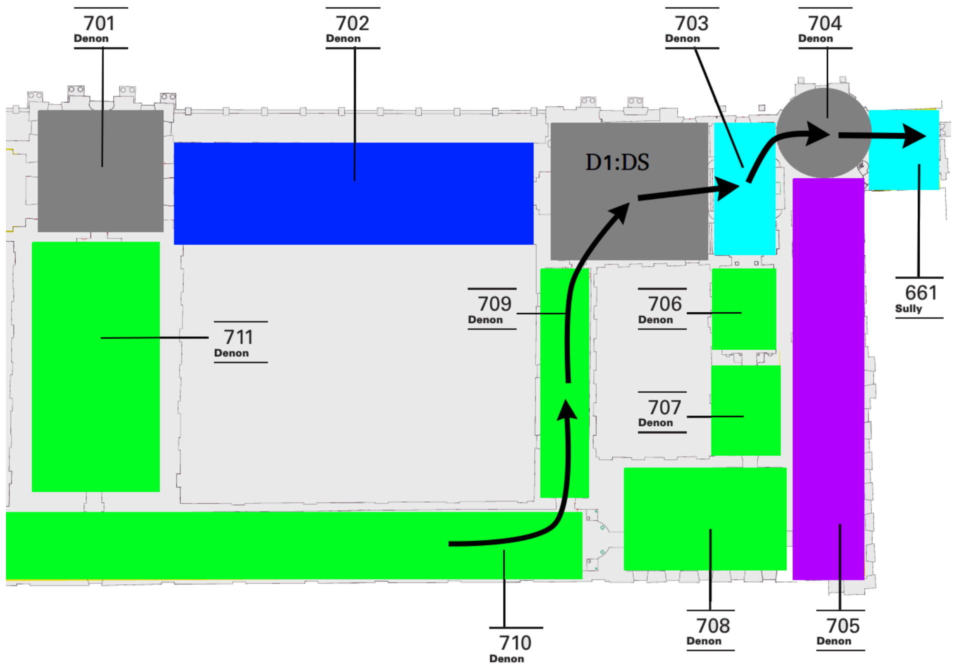


FIGURE 6.14: $T_{vis0058,16:00:00,16:45:00} = (trace_{vis0058,16:00:00,16:45:00}, \emptyset)$ is a subtrajectory example composed of 6 presence intervals in several rooms of the Louvre's Denon wing. Green rooms house Italian Renaissance paintings. Cyan rooms house Ancient Greek sculptures.

(04) (*door013*, *"Room703"*, 16:05:00, 16:20:00, {"low-congestion"}),
(05) (*door014*, *"Room704"*, 16:20:00, 16:23:00, {"high-congestion"}),
(06) (*door015*, *"Room661"*, 16:23:00, 16:45:00, {"high-congestion"}) }

In specific, the first two transitions *"Room710"* → *"Room709"* → *"D1:DS"* can be assigned to a "crowd avoidance" episode, since the visitor quickly passes through large crowds visiting Italian Renaissance paintings, and the last two transitions *"Room703"* → *"Room704"* → *"Room661"* can be assigned to a "visit Ancient Greek sculptures" episode, since the visitor now slowly strolls through rooms filled with Ancient Greek marble artworks, despite the congestion. However, it is not apparent exactly at which point the former behavior gave its place to the latter: the transition *"D1:DS"* → *"Room703"* may well have been due to the visitor finding *"Room703"* to be both less occupied and at the same time more interesting (thematically) than the equally accessible *"Room706"* and *"Room702"*. Therefore, it applies to both types of episodes, as the two behaviors coexist for a certain amount of time.

More generally, any part of a MO's trajectory might correspond to multiple episodes, goal-related or other. The main analytical advantage of allowing overlapping episodes is the quality of the produced results. For example, we can now distinguish between three different trajectory segments:
(*crowd-avoidance*) → (*crowd-avoidance, visit Greek art*) → (*visit Greek art*),
instead of just two, therefore enabling a more subtle interpretation of the visitor's

mobility data. Such distinctions can make a big difference for museum curators who are more interested in a qualitative interpretation of experimental results. The disadvantage in doing so is that the order of the episodes is no longer assured by the model, in contrast to the order of the MO's physical presence in space, and it is up to the analysis method (e.g. the particular pattern mining algorithm) to deal with the additional complexity.

As detailed in section 3.3.2, individual presence intervals can also be enriched with semantic annotations. For example, the *"buy_souvenir"* tag specifically characterizes tuple 19 in Figure 6.16. Similarly, based on the specific zone (i.e. ticket office), on the time spent in it (i.e. 5 minutes), and on the zones that follow it (i.e. permanent exhibition), the visitor's activity could be inferred and tuple 2 could be enriched with the annotation set $A_2 = \{activities : [\text{“}buy\_ticket\text{”}]\}$.

Naturally, semantics of individual tuples can potentially be the ones that give rise to semantics of (sub)sequences. For example, if the visited rooms are highly congested (e.g. based on specific threshold values) over a long period of time, then a significant part of, or even the whole visit may be characterized by the average congestion levels. Similarly, if a zone subsequence contains numerous Italian Renaissance-themed zones, then it may be characterized as a *visit Italian art* episode. Such semantics typically characterize the movement itself, but possibly even the moving object (e.g. visitor tiredness level), the spatial entities i.e. nodes (e.g. room congestion level), the connections between spatial entities i.e. edges (e.g. zone closure), etc.

Finally, application domain semantics can be matched to the indoor space hierarchy and to the trajectory elements. In general, there are various advantages in using ontologies for context modeling [175] such as their hierarchical structure and their enabling of inferring new information. In [175], the authors propose an ontology-based method which combines cross-domain behavior primitives (activities, locations, emotions) referred to as *low-level contexts*, in order to infer more complex and abstract human *high-level contexts* that need low-level ones in order to be identified. Activity context needs to be specialized according to the particular application domain: museums, shopping malls, subway stations, etc. Related to the museum domain in particular, besides activity semantics, the CIDOC Conceptual Reference Model (CRM) [106] is an ISO standard that provides a semantic framework for describing concepts and relationships used in cultural heritage documentation. It can be used to implement an ontological hierarchy that structures the semantic content of the museum space, as illustrated in Figure 6.15. In specific, the E18 *Physical Thing* concept is adopted, which comprises "all persistent physical items with a relatively stable form, man-made or natural" in order to represent the area of engagement with individual exhibits as a RoI. The E4 *Period* concept is also adopted. This is often used to describe prehistoric or historic periods such as the *Neolithic Period*, the *Ming Dynasty* or the *McCarthy Era*, in order to model the historical context and style of the artworks, and from that, of the groupings of artworks as well, such as at the level of zones.

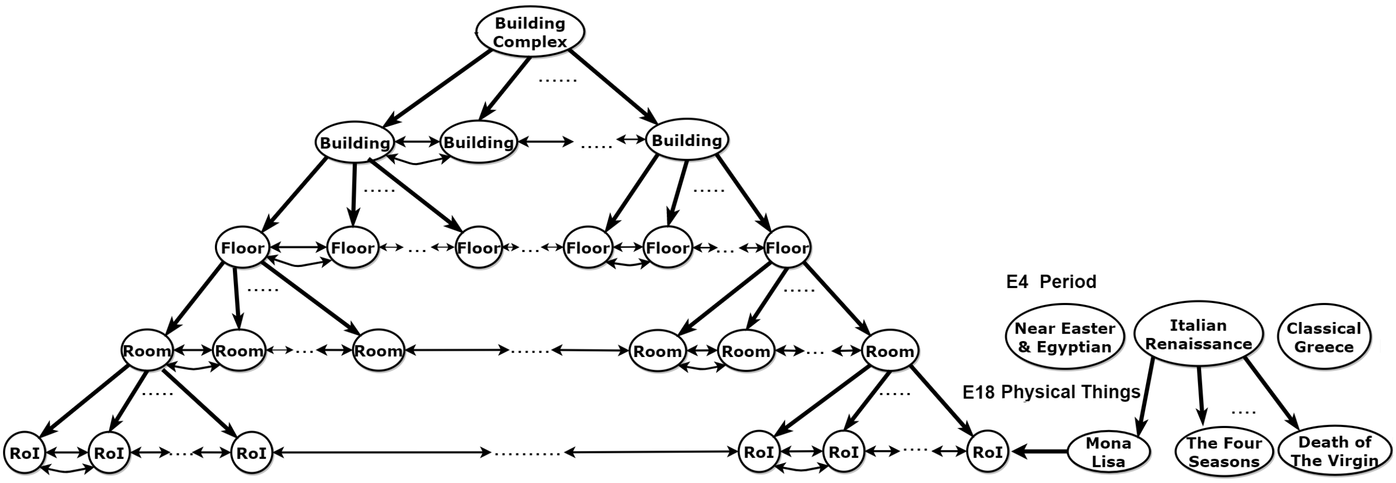


FIGURE 6.15: A simple 2-level domain-specific instantiation of CIDOC concepts maps the RoIs to exhibits providing structure to the interpretation of indoor space.

6.4 Louvre Visitor Experimental Results

This section presents the analysis results of the Louvre visitor trajectory dataset described in detail in section 6.3.2.1. First, some standard statistical and pattern mining results are derived. Those offer a much deeper understanding of the dataset. Then the temporal aspect of the visits is introduced into the mining process by using the $MiSTA$ algorithm [69] over the trajectory dataset represented accorind to $SITM$.

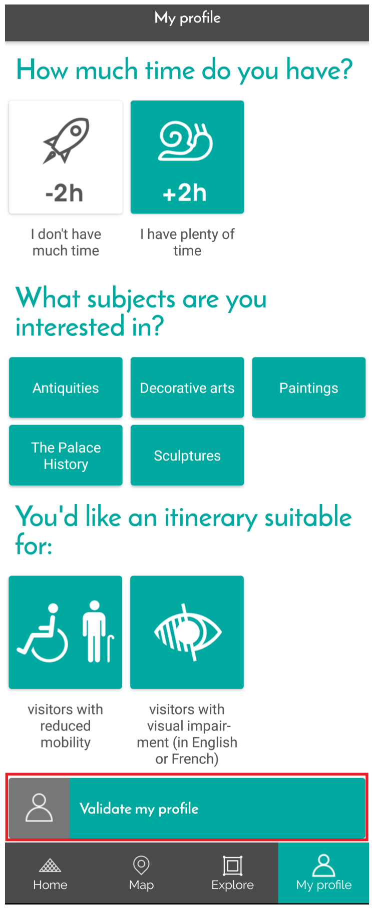
6.4.1 Trajectory Data Preprocessing

As explained in the previous section, the initial aim of our cooperation with the Louvre Museum was to apply semantic trajectory analytics methods using the semantics collected by the museum's official smartphone application. These data are collected in near real-time as the visitors use the app.

However, after examining the related data in detail, a few discrepancies were found, such as the fact that the event logs of the app usage dataset have a minute-level temporal granularity, whereas the movement dataset contains timestamps at the level of seconds. More importantly however, the user identifier (i.e. "UserID" field in Flurry) in the app usage dataset was completely independent of the visitor identifier in the movement dataset. The latter were generated by the company providing the detection data, after applying an anonymization / randomisation process. As a result, the movement of any given individual visitor could not be matched to his or her usage of the app.

The possibility of tackling the problem of matching the visitors in the two datasets based on temporal coincidence and spatiosemantic similarity was considered. In other words, to treat the identifier mismatch as a particular instance of the *moving entity resolution / linking* problem, for which specific bibliography is only recently starting to appear [88, 94, 98]. Unfortunately, there is no way to verify the accuracy of any

A complete zone-level visit trajectory in three different floors of the Louvre Museum.



User-specified profile options may constitute a source of moving object semantics.

FIGURE 6.16: UI action logs (right) can in principle enrich trajectory data (left).

matching estimation, given the lack of ground truth data, and the high numbers of visitors co-existing in the museum spaces at the same time. Only through on-site experimentation would it have been possible to derive some limited ground-truth data, in order to develop a matching method and then apply it to the two historic datasets.

As a result the focus will now be solely on the trajectory dataset, for which a TAS-PM algorithm will be applied in the next section. But first, the trajectory data need to be brought into the proper *SITM*-based form. This includes the following main transformation actions implemented in the Pandas Python open source framework:

1. Transform all timestamps from UTC (“yyyy-MM-dd’T’hh:mm:ss.sss’Z’”) to local time in France (CET in the winter and CEST in the summer) according to the “date_time” format and ISO 8601.

2. Calculate duration values as the difference between a “Begin” detection timestamp and its previous “End” detection timestamp.

3. Remove all detections of zero duration.

4. Split visits according to a temporal threshold of half hour.

v_id	u_id	v_dur	#pos	z_id	z_name	z_lev	z_beg	z_dur	z_end	z_next
14..78	45..78	3544	1535	**60888**	“S2-Nap”	-2	**27-01 15:14:45**	**1225**	**27-01 15:35:10**	60906
14..78	45..78	3544	1535	~~**60906**~~	~~“N1-Sul”~~	~~1~~	~~**27-01 16:00:52**~~	~~**0**~~	~~**27-01 16:00:52**~~	~~60851~~
14..78	45..78	3544	1535	**60851**	“RC-Sul”	0	**27-01 16:00:54**	**764**	**27-01 16:13:38**	60906
14..78	45..78	3544	1535	~~**60906**~~	~~“N1-Sul”~~	~~1~~	~~**27-01 16:13:49**~~	~~**0**~~	~~**27-01 16:13:49**~~	~~-~~
14..27	45..27	938	28	60887	“S2-Nap”	-2	13-02 09:30:07	938	13-02 09:45:45	-
14..18	45..18	351	4	60909	“N1-Sul”	1	17-03 12:06:02	351	17-03 12:11:53	-
14..18	45..18	2783	608	60888	“S2-Nap”	-2	01-03 09:59:35	881	01-03 09:14:16	60889
14..18	45..18	2783	608	60889	“S2-Nap”	-2	01-03 10:27:33	65	01-03 10:28:38	60908
14..18	45..18	2783	608	60908	“N1-Sul”	1	01-03 10:28:40	73	01-03 10:29:53	60891
14..18	45..18	2783	608	60891	“N2-Sul”	2	01-03 10:30:49	909	01-03 10:27:02	-
...	...	...	...	...	...	...	...	...	...	-

Table 6.4: The preprocessed trajectory dataset in tabular form: each row corresponds to a single zone detection. Detections of zero duration are discarded.

In Table 6.4, it is evident in the structured trajectory dataset that a visit consists of a group of rows (i.e. zone detections). The color coding corresponds to the representation of temporal information in Figure 6.5. Detections of zero duration are discarded as errors. The “#pos” column concerns the indoor positioning process and is not of concern to the analysis of the trajectories.

6.4.2 Preliminary Statistical Analysis

Although this Thesis focuses on individual trajectory data modeling and mining, a simple statistical analysis of the aggregate spatial data offers a good macroscopic understanding of the visits. Also, since similarly extensive visit trajectory datasets have so far very rarely been available to museums (and never before for the Louvre Museum in particular), there is great motivation in helping the museum gain some additional insight into its visitors’ behaviors thereby contributing in museum visitor studies more generally.

As part of exploring the Louvre visitor trajectory dataset, Figure 6.17 visualizes the Louvre’s thematic zones, each shaded in proportion to the absolute number of times a visitor was detected in it. As explained in section 6.3.2.1, the whole -1 floor is omitted and any other zone missing from the dataset is displayed as striped.

It can be noted that the most frequented zone is unsurprisingly the main *Pyramid* hall (*“N-2:P”*) of the -2 floor, located right under the Glass Pyramid (Figure 6.18). This is due to the fact that virtually all visitors need to pass by there after entering the museum if they want to proceed to the exhibition spaces. Moreover, the zones in the southern part of the Louvre (*Denon* wing and southern half of *Sully* wing) are more frequented than the ones in the northern part (*Richelieu* wing). An exception is the *Arts Décoratifs Européens* zone (*“R+1:ADE”*) on the +1 floor of the Richelieu

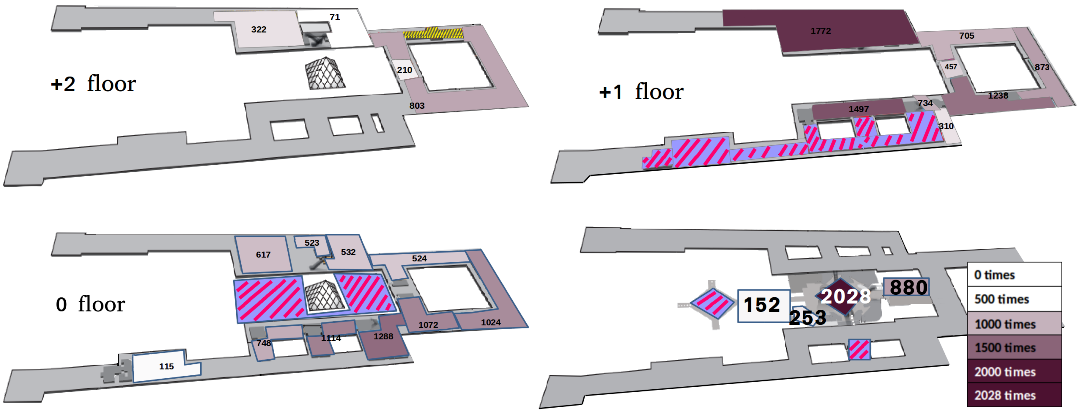


FIGURE 6.17: Choropleth map of the Louvre's zones (-1 floor missing from the data).

wing being the second most visited zone, but at the same time much larger than most others. Finally, as evident from its light color in Figure 6.17, the +2 floor is considerably less visited than the rest.

This spatial imbalance of visitor attendance during the first half of 2017 may still be relevant today. After a record-breaking attendance of 10.2 million visitors in 2018 [120], the Louvre Museum implemented an online time-slot booking system which helped spread its 9.6 million visitors in 2019 [121] throughout the day, but not throughout the exhibition spaces. For instance, the *Leonardo da Vinci* temporary exhibition has managed to attract over a period of four months more than 1 million visitors [122] in zone *"N-2:E"* alone. Therefore, in light of the recent COVID-19 pandemic, an attempt at re-balancing the attractiveness of the different areas could be envisaged, although more direct measures such as, limiting the daily number of visitors, modifying visitor reception processes, and regulating more heavily the visitor flow, are certainly easier to implement in the short term and expected to have a more controllable effect.

6.4.3 Visualization & Standard Sequential Pattern Mining

Now let us move beyond aggregate statistical analysis. Once the zone detection data are properly structured in the form of individual visitor trajectories according to $SITM$, traditional itemset and sequential pattern mining algorithms may be applied. But first, we would like to have the means of visualizing individual visits for two reasons. First, it will allow us to be able to quickly grasp the main structure of each visitor's mobility behavior and compare it to the others if desired. More importantly perhaps, it will allow us to intuitively check the quality of our dataset, by making detection gaps and detection overlaps directly visible.

For these reasons, it was decided to visualize the *space-time* graph [14] of each visitor, a two-dimensional display with the horizontal axis representing time and the vertical axis representing space as a finite linearly ordered set of locations. The proposed model $SITM$ can support such types of visual analytics, since it is both sequential

Few detections take place in the Pyramid Hall at night.

Throughout the day most visitors pass at least once from the Pyramid Hall.

FIGURE 6.18: The main hall of the Louvre Museum is its most frequented area.

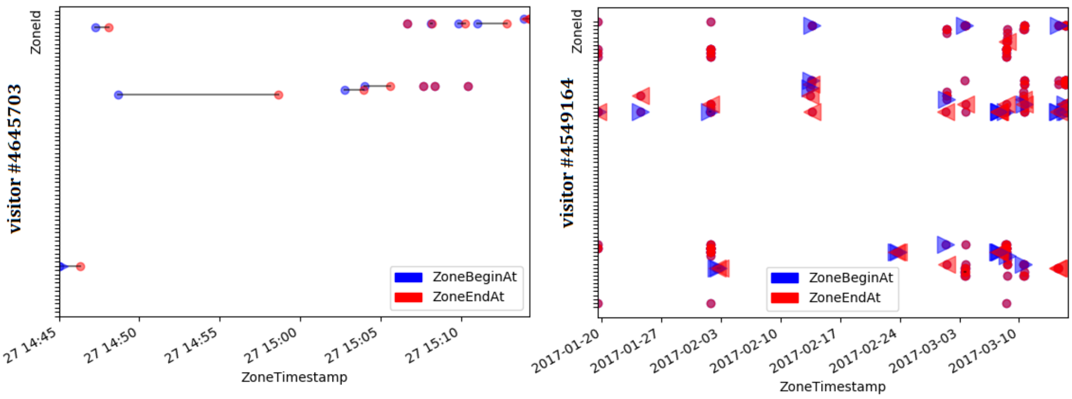


FIGURE 6.19: Space-time graphs [14] of the trajectories of two real Louvre visitors: single-visit visitor #4645703 (left) and returning visitor #4549164 (right).

and primarily symbolic. Figure 6.19 depicts the space-graphs for two different Louvre visitors. Their presence intervals inside of zones are shown as horizontal lines, with points denoting the start and end of a single detection, and triangles denoting the start and end of a whole visit.

Notice how in the right part of Figure 6.19, visitor #4549164 is actually a returning visitor who went to the Louvre on multiple days. By visualizing the space-time graph, the analyst can zoom in on any particular visit and get a clearer image of the corresponding trajectory, in what would resemble the graph of visitor #4645703 shown in the left part of the same Figure 6.19.

Next, after filtering out visits of length equal to 1, we are left with 2,297 visits on which two conventional pattern mining algorithms are applied, namely FPGrowth [78] for zone co-occurence mining and GSP [165] for zone sequence mining. Table 6.5 contains the support values of the most frequent patterns. Noticeably, there is no 3-zone (or longer) sequence that is more frequent than either of the ten most frequent 2-zone sequences (i.e. transitions). Not surprisingly, most frequent patterns take place in the southern part of the museum, just as expected given their spatial distribution in Figure 6.17.

Zone Co-occurence	**support**	**Zone Transition**	**support**
"S0:AG", "D0:AIE"	19.98%	"S0:AG" → "D0:AIE"	6.40%
"D0:AIE", "D0:AR"	18.46%	"D0:AR" → "D+1:PF"	5.27%
"D0:AIE", "D+1:S"	17.07%	"D0:AIE" → "D+1:S"	4.61%
"D0:AR", "D+1:PF"	16.33%	"D0:SE" → "D+1:PF"	4.53%
"D+1:S", "S+1:AGR"	14.80%	"D0:AIE" → "D0:AR"	3.79%
"D+1:PF", "D0:AIE"	13.93%	"D0:AR" → "D0:SE"	3.79%
"S+1:AGR", "S0:AG"	13.80%	"N-2:E" → "N-2:P"	3.70%
"D+1:S", "D0:AR"	13.76%	"S0:AG" → "S+1:AGR"	3.57%
"S+1:AGR", "D0:AIE"	13.71%	"D0:AR" → "D+1:S"	3.48%
"D0:SE", "D0:AR"	13.67%	"D+1:S" → "S+1:AGR"	3.48%

Table 6.5: Ten most frequent Louvre zone co-occurences and zone transitions.

While it is natural for shorter sequences to populate the output of the mining process, the extent to which this happens here suggests that, either the visits quickly diverge, or the dataset is indeed fragmented (or perhaps both). The latter is also supported by the fact that combining most of the frequent transitions, a *dominant visitor flow* emerges as illustrated in Figure 6.20.

In terms of floor transitioning, it is evident that the general movement trend is *going upwards* which is not surprising: visitors are more prone to be using their smartphones while entering deeper into the museum's exhibition spaces, whereas once they decide to leave, they might close the application before starting to descend. This is sensible due to the app-based navigation service being of much greater help for finding a particular artwork, rather than for finding the exit, which is relatively easier thanks to the related signage. Of course, this hypothesis can neither be proved or disproved without a corresponding observational experiment.

In terms of the general movement trend, visitors tend to proceed right-to-left at the 0 floor and *left-to-right at the +1 floor* of the Denon wing. Unfortunately, the neighboring *Peintures Salle Joconde* and *Peintures Italie Est* zones are missing from the data, which makes deriving any conclusions risky. It seems however to be the case that visitors who arrive at the *Daru* staircase at the 0 floor, coming from the *Sully* wing, tend to continue all the way until the *Mollien* staircase before going up, instead of directly climbing the *Daru* staircase to visit the *Winged Victory of Samothrace.* Yet, there is a short pattern opposite to this main flux of visitors, namely *"D0:AR" → "D+1:S"* which corresponds to the 9th most frequent transition in the museum (Table 6.5). This is normal because the *Winged Victory* is one of the museum's most iconic masterpieces and hence can be expected to attract people to the *"D+1:S"* zone from the lower floor level.

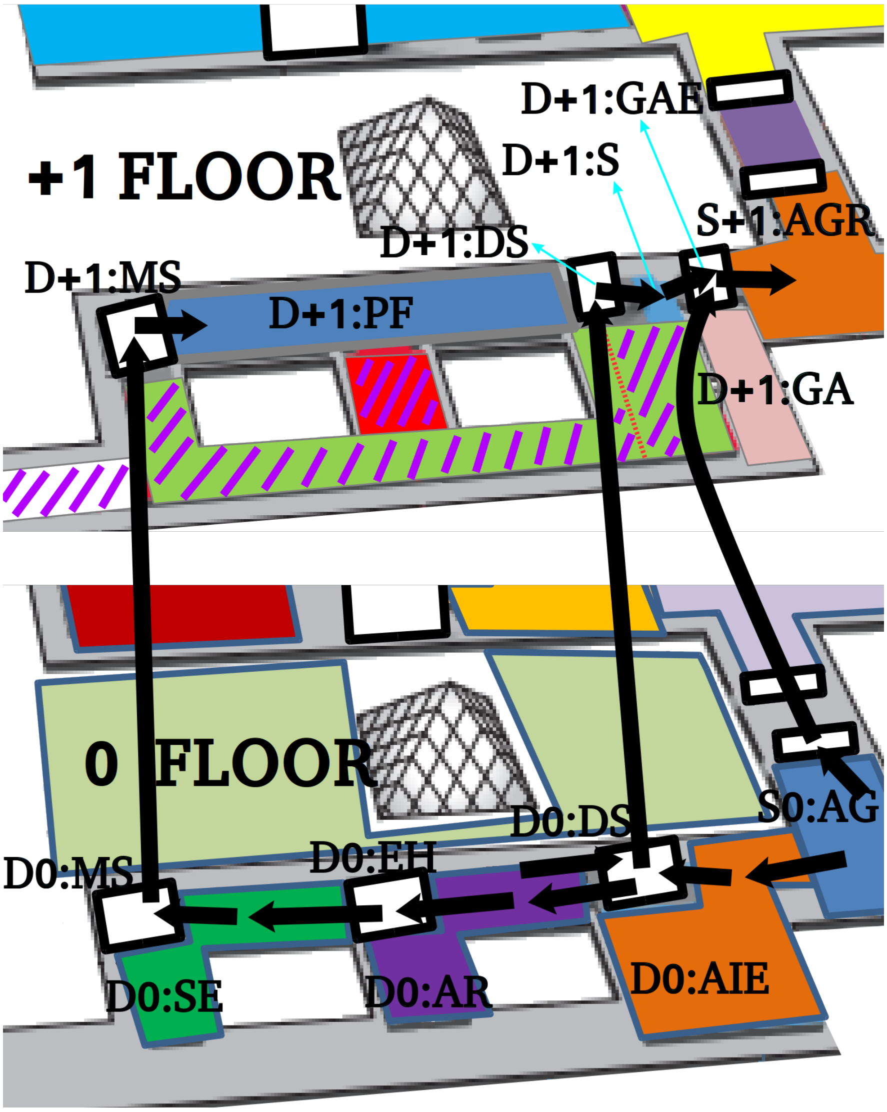


FIGURE 6.20: Most frequent zone transitions happen in the southern part of the Louvre.

6.4.4 SITM-based State-Of-The-Art Sequential Pattern Mining

This section studies how the proposed conceptual model for the representation of semantic indoor trajectories supports and enhances the process of trajectory mining. To this end, two pattern mining approaches are evaluated with respect to the Louvre case study, namely Multidimensional Sequential Pattern Mining (MD-S-PM), reviewed in section 4.3.4, and Temporally Annotated Sequential Pattern Mining (TAS-PM), reviewed in section 4.3.5. Their limitations are identified, and it is reported which are the elements that make their combination a promising approach for extracting even more interesting patterns.

SITM-based MultiDimensional Sequential Pattern Mining

In the previous section 6.4.3, interesting visiting patterns in the Louvre Museum were derived by using standard S-PM algorithms. While these were applied on trajectories represented according to $SITM$, they do not take advantage of its full expressiveness (e.g. trajectory semantics, contextual information, indoor topology, temporal duration), since they rely exclusively on the presence or sequence of the detection records. However, $SITM$ contains extra information that can be used by more advanced mining methods, or even inspire the design of new ones. With regards to the former, MD-S-PM methods are especially interesting because they enable the analysis to benefit from the integration of contextual information.

Studied in section 4.3.4, the mining methods proposed by Pinto et al. [145] consider the multidimensional part to be independent from the main sequence. The same type of method can be applied over $SITM$-based museum visit trajectories. For instance, if static visitor information (e.g. profile settings) is retrieved from the mobile guide application (Figure 6.16), then the trajectory example from section 3.3.2 can be enriched by adding more semantic annotations in A_{traj} describing each individual visitor's declared interests and time availability:

$T_{vis0042,11:30:00,13:30:00} = (trace_{vis0042,11:30:00,13:30:00}, A_{traj})$

$A_{traj} = \{goals : [\text{“}visit\_temporary\_exhibition\text{”}], regularity : \text{“}First - Timer\text{”},$

$subjects : [\text{“}Antiquities\text{”}, \text{“}Sculptures\text{”}], time : [\text{“} > 2hours\text{”}]\}$

This allows us for instance to find frequent visiting patterns per type of visitor, instead of only for all visitors indiscriminately.

Again as examined extensively in section 4.3.4, Plantevit et al. [146] generalized the above approach to account for multiple dimensions within the sequence itself. Data are stored in a relational table T as a finite set of tuples $t = (d_1, ..., d_n)$ whose values belong to the domain of several data dimensions $d_i \in dom(D_i)$, $i = 1, ..., n$.

The same or similar methods (studied in chapter 2) can be applied over $SITM$. For example, assuming that *dynamic visitor information* (e.g. access records of the application's educational content) is available, then it can be used to annotate each presence record independently. More concretely, a particular interval of the visitor's presence in spatial area v_j together with its corresponding annotations A_j can combine for a multidimensional item i_j, and the application of MD-S-PM methods becomes straightforward.

For instance, in the previous trajectory example from section 3.3.2, the 5th and last tuple of the trace may be enriched based on the *audio description playback* that the visitor listened to and the *textual description* that he read, while being detected in the *Inverse Pyramid Hall*:

($opening002$, “IPH”, 13:28:30, 13:30:00, $\{audio : [\text{“}Fountainhead\text{”} : 03'15'',$

$\text{“}Lady\_of\_Auxerre\text{”} : 00'32''], text : [\text{“}Al\_Mughira's\_Pyxis\text{”}]\}$)

On a more practical side, the semantic enrichment of visitor trajectories would be easier if the movement data were annotated automatically at the moment of collection, but for now this is not being done in the museum domain, and most probably neither in any other application domain. In the specific case of the Louvre's smartphone application, a *poi_audio_time* event record gets created whenever a visitor listens to an artwork's audio description, and such records would produce some of the semantic

Timestamp	Session Index	Event	Version
Nov 07, 2017 03:06 AM	4	poi_audio_time	2.0.26
Platform	**Device**	**User Id**	**Params**
Android	Motorola Moto G4 Plus	396f8c54db4ceb67	{ poi_idx:92; total_time:214,274; stop_time:213,862}

TABLE 6.6: A single log record created by the *poi_audio* event of Figure 6.21.

annotations to be included in the trajectories.

For instance, Figure 6.21 and Table 6.6 describe the User Interface action for listening to the audio description of Michelangelo's *Slave sculptures* and the corresponding event record that is created in the log files of the Flurry platform. Noticeably, the event is timestamped and also includes the *total listening time* and the *playback stoppage time.* Also, the parameter "poi_idx" uniquely identifies the related artwork. Such information is ideal to add to the trajectory represented in *SITM* form a semantic annotation to the particular presence interval within which this particular event occurred. This could for example help us find out whether visitors are generally located within an artwork's RoI (according to *SITM*) when they choose to listen to that artwork's audio description, or whether they instead tend to listen to a description before / after visiting the corresponding artwork.

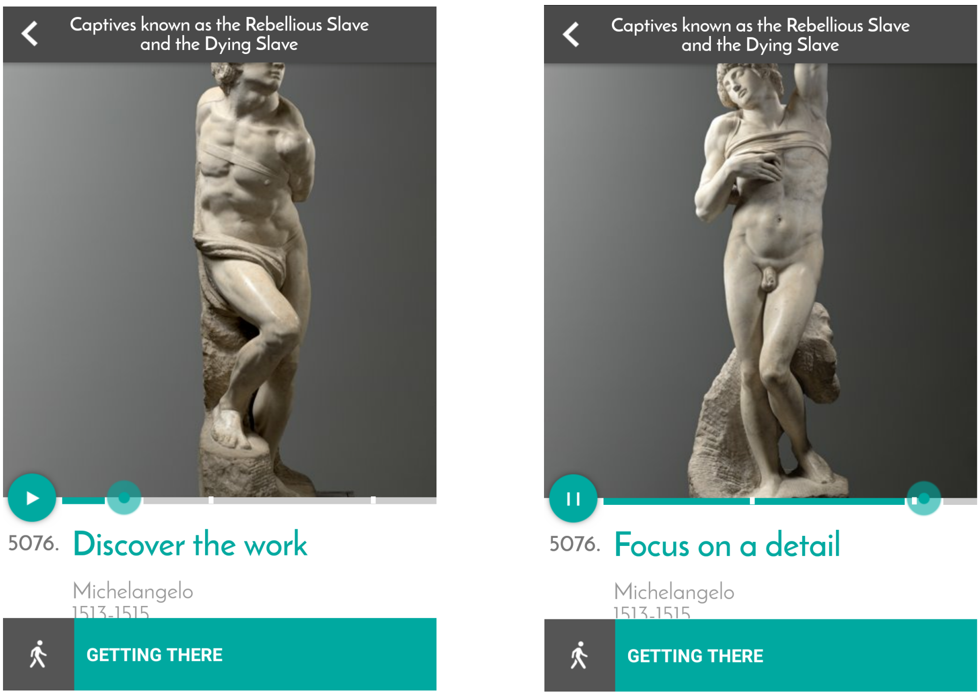


FIGURE 6.21: Multimedia content consumption from the smartphone application can theoretically provide the semantic enrichment of the trajectories of the Louvre museum visitors.

Furthermore, in [147, 148], Plantevit et al. allow for several hierarchical levels to be mixed within the same sequence, and the extracted patterns to be automatically associated to the lowest possible level of granularity. To achieve this, a new type

of hierarchical pattern inclusion was defined based on the notion of *item specificity* described in detail in section 4.3.4. $SITM$'s trajectory representation enables the application of such hierarchical pattern mining methods, by using annotations that belong to multiple levels of a semantic taxonomy. For instance, the previous presence interval may be changed to have the consumed textual description information at the level of artwork-type, but keep the audio description playback information at the lower level of specific artworks:
(*opening002*, *"IPH"*, 13:28:30, 13:30:00, $\{audio : [\text{"}Fountainhead\text{"} : 03'15'',$
$\text{"}Lady\_of\_Auxerre\text{"} : 00'32''], text : [\text{"}Sculptures\text{"}]\})$

In this way, the methods of [147, 148] can be applied in order to extract semantic trajectory patterns formulated as multidimensional and at the same time multigranular sequences:
... → (*"IPH"*, *"low-congestion"*, *"Spanish_Islamic"*) →
(*"AG"*, *"high-congestion"*, *"Classical_Greece"*, *"Sculptures"*) →
(*"SE"*, *"normal-congestion"*, *"Mona_Lisa"*) →*...*
Extracting such patterns will have great value for museum professionals because it brings some of the qualitative elements of *traditional observation studies* into the realm of *Big Trajectory Data* analytics, with all of the advantages that this entails.

SITM-based Temporally Annotated Sequential Pattern Mining

Most mining approaches described in section 4.3 completely ignore the temporal aspect of movement data and at best only account for their sequential nature. Unfortunately, this is problematic for mining trajectory patterns, given that the resulting patterns are much poorer this way. In the particular context of this chapter, we would like to find out whether the Louvre's visitors show any representative behaviors with respect to how much time they spend in each zone as they move within the museum.

In this regard, it would be useful to adopt a Temporally Annotated Sequence (TAS) form for the visit sequences, introduced in [70], where Giannotti et al. propose the *MiSTA* algorithm for calculating the most frequent TAS patterns given an input TAS dataset. $MiSTA$ was studied in detail in section 4.3.5, and it was found that, whereas it lacks support for multidimensional or hierarchical pattern mining, it does have the ability to account for duration intervals, which is of paramount importance for mining trajectory data.

In the follow-up work of [71], the elements of a TAS sequence S are assumed to constitute coordinate pairs, thus defining a trajectory pattern mining problem, and Giannotti et al. introduce the *t-patterns* algorithm to solve it. As explained in detail in section 4.3.1, t-patterns consists simply of the MiSTA algorithm preceded by a transformation step which groups the raw geometric data into regions, based on a neighborhood function, so that they take a symbolic form that can be handled by MiSTA.

Given that our conceptual trajectory model requires symbolic location data, like the ones available from the Louvre case study, and not necessarily geometric data, the focus is solely on the MiSTA algorithm and not on the t-patterns algorithm.

In order to be able to apply MiSTA' mining process on $SITM$'s trajectories, the following two assumptions are necessary:

- MiSTA only considers sequences of itemsets and their corresponding duration annotations, and cannot integrate in the mining process other elements. Thus, the $SITM$-based trajectories need to be restricted to the following elements:

$$T_{ID_{mo},t_{start},t_{end}} = (trace_{ID_{mo},t_{start},t_{end}}, \cancel{A_{traj}})$$

$$trace_{ID_{mo},t_{start},t_{end}} = (\cancel{e_k}, v_k, \overset{t_k^{duration}}{\cancel{t_k^{start}, t_k^{end}}}, \cancel{A_k})_{k\in[1,n]}$$

- MiSTA's input data are restricted into consisting solely of sequences of items rather than sequences of itemsets, because $SITM$ assumes that a moving object can not be present in multiple symbolic spatial entities at the same time. As a consequence, only extension projections will take place during MiSTA's execution, and not enlargement ones. This assumption does not induce any changes in the MiSTA algorithm itself.

- Instead of attributing each temporal annotation to the transition between two consecutive items, one can attribute it to the duration of stay in the first item, which represents the spatial entity of departure. This assumption simply serves to correct the interpretation of TASs according to the $SITM$.

For instance if τ=20, then the TAS $(S, A) = \overset{60}{\textit{“N-1:P”}} \rightarrow \overset{120}{\textit{“D-1:EH”}} \rightarrow \overset{240}{\textit{“D-1:APOE”}}$ τ-contains the TAS pattern $(S_1, A_1) = \overset{130}{\textit{“D-1:EH”}} \rightarrow \overset{255}{\textit{“D-1:APOE”}}$, but not the TAS pattern $(S_2, A_2) = \overset{150}{\textit{“D-1:EH”}} \rightarrow \overset{240}{\textit{“D-1:APOE”}}$, because their corresponding annotations differ by more than 20 seconds in at least one case.

Extracting Louvre visitor trajectory patterns

Before the actual mining process, some pre-processing of the original trajectory dataset is needed to transform it into TAS form. First, any zone detection record with duration equal to 0 is filtered out. This leads to the deletion of 2135 out a total of 20245 records. Secondly, in 1080 of those cases, the previous and the subsequent (to the deleted one) zones actually coincide, which constitutes further indication that the deleted zones are indeed errors. Those are merged into a single zone in order to avoid any identical items appearing consecutively in the input TASs. Thirdly, all trajectories containing less than 3 zones are filtered out.

We opt for a low threshold value, taking into account the coarse spatiotemporal granularity of the available trajectory data, as well as their length distribution being left-skewed (Figure 6.6). Finally, even though $SITM$ can represent temporal gaps in the trajectories, MiSTA does not allow for gaps in the sequences. Therefore, all periods of visitor non-detection need to be erased, or it is assumed that the visitor is actually continuously located in the last known zone until he or she is re-detected in a different one. Given that the duration values are on the low side of what would normally be expected for a museum, the second approach is adopted as more realistic.

Next, in order to choose a proper τ parameter value, the normal distribution of the zone detection duration value is calculated. As expected and confirmed by the curve's push to the right (Figure 6.22), taking into account the detection gaps in the original trajectory dataset increases the values of the temporal annotations in MiSTA's input TAS dataset. After trying out lower (i.e. stricter) and higher (i.e. more relaxed) values, we opted for τ=117 sec, equal to the median zone stay duration value. In practice, this means that MiSTA will count a projected pattern's occurrence in the input TAS data only as long as all of the corresponding annotations differ by less than 2 minutes.

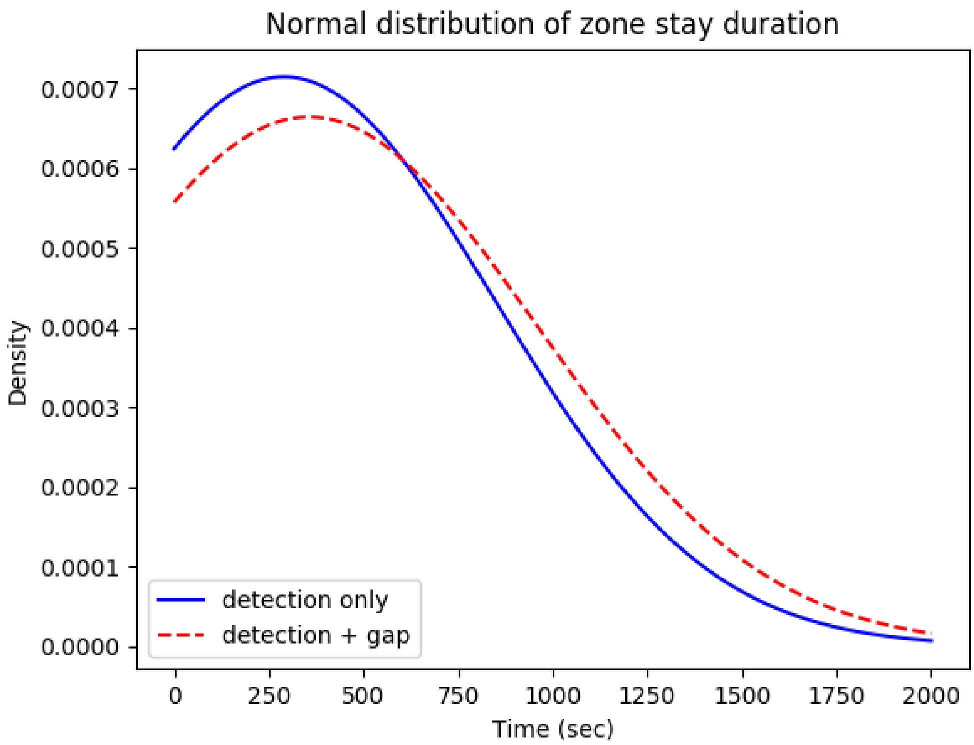


Figure 6.22: Normal distribution of the Louvre visitors' duration of stay in each zone, under two different interpretations of the detection gaps.

Figure 6.23 contains the frequent TAS patterns of length equal to 3, enabling us to derive additional insight compared to the purely sequential approach used previously. Interestingly, all four patterns involve two types of movement taking place in two different parts of the museum, both characterized by a floor-switching *back and forth* type of behavior (Figure 6.24). In addition, they do not take place in the busiest parts of the museum like the ones shown in Figure 6.20. This means that their support values are not much higher than their τ-support values.

However, due to the fact that MiSTA involves a lossy step of merging frequent annotation intervals, specific τ-support values are not reported because there is no approximation guarantee other than their being higher than $supp_{min}$=5%. Instead, the contiguous sequence support is presented, which corresponds to requiring direct transitions only (Figure 6.23).

Temporally Annotated Zone	**cont. support**	**support**
(*"S+1:AE"*) $\xrightarrow{\alpha_1}$ (*"S0:AE"*) $\xrightarrow{\alpha_2}$ (*"S+1:AE"*) (α_1,α_2): ([50,50],[89,104]),([51,55],[39,120]),([56,72],[29,121]), ([73,73],[17,123]),([74,82],[14,124]),([83,99],[0,125]), ([100,100],[0,126]),([101,121],[0,127]),([122,122],[0,126]), ([123,125],[0,125]),([126,127],[0,124]),([128,128],[17,123]), ([129,131],[29,122]),([133,134],[89,121])	7.15%	8.01%
(*"S0:AE"*) $\xrightarrow{\alpha_1}$ (*"S+1:AE"*) $\xrightarrow{\alpha_2}$ (*"S0:AE"*) (α_1, α_2): ([29,31],[89,101]),([32,47],[50,120]),([48,65],[48,121]), ([66,69],[31,122]),([70,81],[19,122]),([82,91],[15,122]), ([92,118],[9,124]),([119,119],[19,122]),([120,121],[60,120]), ([122,124],[89,101])	7.02%	8.23%
(*"R+1:ADE"*) $\xrightarrow{\alpha_1}$ (*"R0:SFM"*) $\xrightarrow{\alpha_2}$ (*"R+1:ADE"*) (α_1, α_2): ([67,85],[116,118]),([86,91],[110,121]), ([92,118],[94,122]),([119,120],[110,121]),([121,122],[116,118])	7.15%	7.66%
(*"R0:SFM"*) $\xrightarrow{\alpha_1}$ (*"R+1:ADE"*) $\xrightarrow{\alpha_2}$ (*"R0:SFM"*) (α_1, α_2): ([42,59],[126,129]),([60,61],[124,130]), ([62,64],[101,130]),([65,84],[77,131]),([85,85],[77,134]), ([86,119],[59,134]),([120,122],[101,131]),([123,127],[126,130]), ([128,128],[126,129]),([129,130],[128,129])	6.06%	6.25%

FIGURE 6.23: The four frequent Louvre TAS patterns of length 3, for $supp_{min}$=5% and τ=117sec.

It can be noticed that the two patterns in the *Arts décoratifs européens* and *Sculptures France Marly* zones of the Richelieu wing (*"R+1:ADE"* and *"R0:SFM"*) are almost always contiguous, whereas the two patterns in the *Antiquités Égyptiennes* zones of the Sully wing (*"S0:AE"* and *"S+1:AE"*) more often include intermediate transitions. Also interestingly, the former two patterns contain more restricted duration intervals than the latter two, which suggests that visitors spend a more specific amount of time in that part of the Richelieu wing. Further interpretation of the reported time intervals is outside the scope of this illustrative experiment.

Finally, let us report on a few important implementation details. First, MiSTA counts each TAS pattern only once per input TAS, even when it appears multiple times in the same input TAS. In some application cases, this may not be the most desired way to calculate τ-support, particularly true for datasets containing long

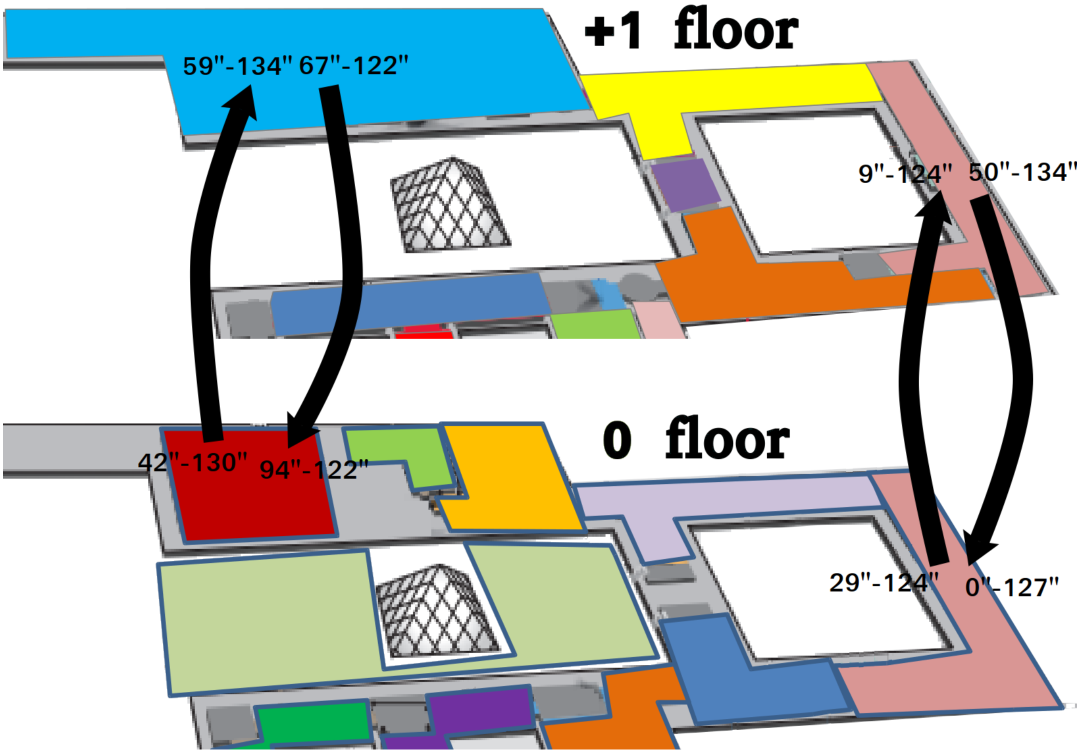


FIGURE 6.24: The four frequent Louvre TAS patterns of length 3, for $supp_{min}$=5% and τ=117sec.

episodic trajectories. Secondly, as already explained, due to how τ-containment is defined [70], the TAS patterns reported by MiSTA are not necessarily contiguous. In such case, the corresponding annotation describes how long the visitor stayed in all of the zones combined, leading up to the next zone in the pattern. Again, a contiguous variation may be interesting depending on the case. Thirdly, since MiSTA's annotations were originally designed to describe the transitions and the last item of each TAS is not followed by any transition, the duration of stay in the last zone of any trajectory is lost during its transformation to an input TAS. Technically, this issue can be easily solved by adding an *"EXIT"* item at the end of each TAS. This does not alter the mining output, apart from adding any frequent trajectory-ending patterns that contain the newly included interval.

6.5 Chapter Conclusion

This chapter presented in detail the Louvre case study, especially as it relates to semantic indoor trajectory modeling and analysis. First, the context and motivation for this case study were presented, as well as the related past research studies and future research opportunities, stemming from the museum's digital innovation strategy. Many more museums will certainly follow this line of practice in the near future. Then, the specific datasets deemed useable for this Thesis were presented in detail, along with the implementation of the proposed trajectory model, and the analysis results obtained by applying state-of-the-art sequential pattern mining methods. Along with [194], this is to the author's knowledge the first time that sequential patterns have been extracted and studied in the context of a museum visitor study at this scale. Unfortunately, Bluetooth detection data alone cannot disclose the semantics of movement, neither was it possible in this case to match the trajectories to the application's usage data, in order to apply the envisaged multidimensional sequential pattern mining techniques over semantic trajectories.

7
Conclusion and Perspectives

Conclusion

Thesis and Results Synopsis

The goal of this Thesis has been to explore new ways to extract useful insight about an indoor movement phenomenon under examination, given a raw tracking dataset of individual moving objects, and given any type of semantic information related either to those moving objects or to the movement itself. The Thesis was developed in cooperation with the Louvre Museum and under the support of the Foundation for Cultural Heritage Science. Thus, the initial motivation was both academic and of an applied nature, aiming to explore the potential of novel trajectory-based methods for studying visitor trajectories in the Louvre Museum.

Naturally then, this Thesis is related to the ever-expanding research field of *Human Mobility Computational Analytics* and more specifically to *Trajectory Data Modeling and Representation*, *Trajectory Data Mining and Analysis*, *Indoor Trajectories*, *Sequential Pattern Mining*, and *Museum Visitor Mobility Studies*.

The intention from the beginning and throughout its course has been, on the one hand to help the Louvre better understand its visitors by confirming or disproving currently held beliefs, discovering previously unknown behavioral patterns, etc., but on the other hand to contribute to trajectory-based research in a domain-agnostic fashion. Hence, the proposed model and mining algorithm were kept relevant to any type of indoor mobility data or scenario, instead of only museum visits or any specific type of built environment.

Furthermore, the start of the Thesis coincided with the museum's launch of its official *My Visit to the Louvre* smartphone application, which was meant to provide the necessary data concerning this prestigious case study. While these data were still being collected, the author embarked on a broad overview study of the whole Trajectory Data Mining and Analytics state-of-the-art landscape, presented in chapter 2. This led to a classification of the trajectory data mining tasks being proposed, according to the state-of-the-practice in Table 4.1.

With the focus on semantic trajectories and indoor trajectories (both separately and in combination), an exhaustive review of the respective literature was provided. The last part of chapter 2 was dedicated to the research works related to the study of museum visitor movement represented in trajectory form, and reveals the importance of both movement semantics and the indoor context for modeling these particular types of trajectories. Related to this, chapter 6 also reviewed a series of computational mobility analysis research works regarding the Louvre visitors specifically. Apart from relating directly to our own case study, this line of works is actually one of the few attempts at analyzing real-world museum visitor mobility data automatically collected from a sensor network.

As argued in chapter 2, while the field of Semantic Trajectory Modeling has produced interesting concepts and worthwhile approaches to the representation of meaningful movements, it has also largely ignored indoor environments. Perhaps even more importantly, the formal models proposed in the literature have not yet been adopted significantly by the more application-oriented trajectory data mining research community. Specifically with respect to trajectory pattern mining, this is presumably also done to fit the standard S-PM methods that are more readily available. Therefore, chapter 3 proposed a new conceptual trajectory model, called $SITM$, specifically meant to address the needs of semantic trajectory data mining in indoor environments.

The proposed model takes into account any type of available semantic information related to the various aspects of the movement phenomenon. $SITM$ is based upon a representation of indoor space inspired by and abiding to the IndoorGML standard. It does however feature its own specifications, and deviates from the standard's guidelines whenever deemed necessary, without breaking compatibility. Over its multilayered multigraph-based representation of indoor space, the trajectories are principally modeled as sequences of presence intervals in the spatial regions represented by the graph's nodes. At the same time, each interval contains a set of semantic annotations, not bound to any particular vocabulary or ontology of movement. The model is instantiated and validated in chapter 6 with respect to the Louvre case study.

Moreover, since S-PM methods in particular have often been successfully implemented for trajectory data mining, those were studied in further detail in chapter 2, outside of the trajectory domain as well. Just like trajectory pattern mining most existing works neglect the semantic aspects of movement, apart from a thin-layer of geographic semantics. Similarly in non-trajectory S-PM, not many methods consider enriching the item sequences with dynamic contextual/semantic information. As a result, such methods, including ours, cannot take advantage of the semantic load present in most semantic trajectory models proposed in the bibliography.

What is more, most times these S-PM methods also ignore the temporal dimension of the trajectory data, relying instead solely on the sequential order of items that represent the location of the moving object. This is of paramount importance since time is central to movement. And of course, the vast majority of the related literature has been dedicated to mining outdoor trajectories, given the prevalence of GPS data, therefore ignoring the effect that the building has over movement.

Thus, having identified the aforementioned undesirable traits of state-of-the-art methods and algorithms for finding patterns in trajectory datasets, chapter 5 proposed a novel trajectory mining method which embeds semantic information in the mining process, along with temporal and topological information. It is influenced by the following algorithms:

- M^3SP[147] for the spatiosemantic dimensions;
- $MiSTA$[69] for the temporal dimension;
- TP[50] for the topological information.

The relevant algorithm, called $SITPE$, produces in its output frequent trajectory patterns that are qualitatively richer compared to the state-of-the-art methods. For example, as explained in chapter 1, with the case of a shopping mall client, instead of obtaining simple patterns such as:

$$Yves\_Rocher \rightarrow H\&M \rightarrow Promod \rightarrow MacDonald's$$

we obtain more indicative temporally and semantically aware patterns like:

$$\{Yves\_Rocher, disc\_buy, reg\_cust\} \xrightarrow{15min} \{Promod, prod\_replacement, cas\_cust\} \xrightarrow{5min} \{MacDonald's, reg\_buy, new\_cust\}$$

informing us that a typical shopping behavior involves customers that tend to regularly visit $Yves\_Rocher$ to take advantage of its discounts, followed by $Promod$ to replace products that they have already bought, before having a seated meal for the first time in this particular $MacDonald's$ outlet.

Finally, in chapter 6 an intriguing case study was presented, involving real-world trajectory data from the visitors of the Louvre Museum. The Louvre has actually long been one of the leading museums in terms of innovative technology adoption. For instance, in 1995 it was one of the first cultural institutions to offer a website to its visitors [61]. Thus, it comes as no surprise that since the aforementioned works of Yoshimura et al., it has installed its own permanent tracking infrastructure, which has so far been used only for location-based services offered to the visitors. This Thesis exploits for the first time this infrastructure through a computational trajectory-based analysis of the visits.

More specifically, the author was given access to a historic dataset spanning a four month period in 2017, as well as an additional dataset of app usage events over the same period, both emanating from the official (at the time) smartphone application offered by the museum. Thus, the analysis and data exploration results were presented. The initial goal was to enrich the trajectories with semantics derived from the app usage data. However, this was found to be impossible due to the unresolvable mismatch of user identifiers.

In terms of PM, two existing algorithms were run, namely GSP and $FPGrowth$ which revealed some interesting but not surprising patterns. Also, $MiSTA$ was applied, which takes transition durations into account. First, the dataset had to be adapted to the proposed trajectory model and to the proper interpretation of the patterns. The resulting patterns were unexpected and merit further investigation. Data quality issues were also identified and discussed in chapter 6, as well as considered by the Louvre for improving its technological solutions in the future [178].

Finally, apart from learning more about the visitors' mobility behaviors and evaluating the potential of trajectory-based analysis methods, the Louvre case study served as a validation of the practicality and usefulness of the proposed trajectory representation.

Contributions

The main contribution of this Thesis is organized around the proposal of a new conceptual model of semantic trajectories specifically for indoor environments called $SITM$. This model is used to represent spatiotemporal trajectories enriched with semantic annotations, taking place in indoor environments. Moreover, it is compatible with the IndoorGML standard and it was validated by being instantiated for the case of the Louvre Museum, and used to run pattern mining methods over real-world Louvre visitor trajectories.

Another contributing element of this Thesis is the definition of a new type of pattern mining problem, called the MultiDimensional Temporally Annotated Sequential Pattern Mining (MD-TAS-PM) problem, and the definition of a new type of trajectory pattern mining problem, called the Semantic Indoor Trajectory Pattern Mining (SIT-PM) problem. Also, an algorithmic method and a first corresponding algorithm called $SITPE$ are proposed to help solve the SIT-PM problem. $SITPE$ combines and extends three different state-of-the-art sequential pattern approaches, in an attempt to mine qualitatively rich trajectory patterns.

In total, the contributions of this Thesis (both primary and secondary) can be summarized as follows:

- An overview and a practical classification of trajectory data mining tasks.
- An extensive survey of semantic and/or indoor trajectory modeling.
- A new conceptual model of spatiotemporal indoor trajectories enriched with semantic annotations, called Semantic Indoor Trajectory Model ($SITM$) [102].
- A formal definition of the problem of mining multidimensional and temporal patterns from input sequences.
- A formal definition of the problem of mining semantic indoor trajectory patterns from input trajectories.
- A new algorithmic approach called Semantic Indoor Trajectory Pattern Extractor (SITPE), which extends and combines state-of-the-art general-scope sequential pattern mining algorithms, in order to solve the SIT-PM problem.
- An identification and classification of the museum goals w.r.t. the computational analysis of their visitors' trajectory data.

Finally, throughout this Thesis the author cooperated with researchers and professionals from the Louvre Museum, especially with regards to the extraction and exploration of the datasets, but also the interpretation of the analysis' results.

This experimentation phase produced the following results:

- A validation of $SITM$ through its instantiation in the context of the Louvre Museum case study, and its usage to support standard and state-of-the art pattern mining algorithms applied over real-world visit trajectory data.
- A unique study of the Louvre visitors' mobility patterns and dynamics of attendance.
- A practical study and report of how the Louvre's trajectory data acquisition and wrangling processes, and corresponding data quality issues, constrain or otherwise impact semantic trajectory modeling and analysis, raising the awareness of museum management for the future development of digital tools.

Perspectives

Emerging Trends in Trajectory-Based Research

After extensively reviewing the Trajectory Data Modeling and Mining literature, as well as experiencing recent developments in Museum Visitor Mobility Studies and experimenting on a unique museum visit trajectory dataset, in this final section the most promising lines of research for advancing the respective fields are identified. The proposals are in no way limited to the specific domain of museums. Specific issues that need to be addressed for the advances to take place are also discussed here.

First, with respect to the future of Trajectory Data Modeling and Mining, there is an emerging trend of adopting vector representations and Deep Learning methods [9, 40, 46, 48, 64, 125, 176]. While this is promising in many respects, it is so far best suited to trajectory prediction problems, whereas cultural institutions such as museums are much more interested in obtaining qualitative descriptions of their visitors' behavioral patterns. This is due to the fact that organizations in the Arts and Humanities are typically less interested in maximizing attendance or profitability, and more interested in maintaining a high level of quality of their visitors' experience.

Hence, even though the use of Neural Network-based methods can be expected to transition into other trajectory data mining tasks (e.g. generating realistic artificial trajectories), there is also a lot of untapped potential in developing S-PM methods that will leverage properly represented semantic trajectories for descriptive types of analyses.

This brings us to the field of Semantic Trajectory Modeling, for which there is significant bibliographical content and multiple well developed approaches, but at the same time adoption is lacking by part of the trajectory data analytics literature. When faced with a particular trajectory-related problem instance, the analyst is tempted to adopt ad-hoc representations, rather than study a properly structured trajectory model in detail, and spend time to adhere to it. Perhaps, an industry standard for representing semantic trajectories would be useful in addressing this.

Additionally, concentrated efforts by the trajectory research community to clarify the terminology and homogenize the conceptualization of semantic trajectory representations would also considerably help. This is particularly true for indoor trajectories, because at least for outdoor trajectories there is a certain level of established consensus over concepts and practices, as a result of the longer period that they have been studied for and of the lesser variance of the raw trajectory data form.

Moreover, even though the $SITPE$ algorithm proposed in section 5.1.2.3 takes a big step in this direction, our vision is to combine even more closely together indoor space, semantics, and time, within the scope of trajectory analytics. This is a complex task that merits further investigation. Even the semantic nature of indoor trajectories alone is still largely unexplored. Let us provide examples of future research directions in relation to this:

- Examine the interaction between trajectory data research and the Space Syntax approach [3].
- Explore how time-series analysis can complement current trajectory data mining

practices (e.g. to detect patterns in the evolution of the number of moving objects present in each indoor spatial region).

- Study in more depth the implications of allowing overlapping trajectory episodes as proposed in *SITM* (e.g. how should S-PM methods handle them?, how should pattern containment change to reflect those?).

- Study in more depth ways to use the spatiosemantic hierarchies in order to handle trajectory data uncertainty issues (e.g. methods to infer presence in coarser spatial regions).

- Develop a new notion of trajectory pattern containment which takes into account recent advances in semantic similarity research should be developed.

- Further explore the difference between *actual movement* and *potential or intended movement* (e.g. detect Louvre visitors who move in order to see the Mona Lisa but end up skipping doing so because they don't want to wait in line).

- Use semantic information to improve the quality of indoor positioning / location estimation (e.g. use multimedia content access logs from museum guides to infer the location of the visitor at a level of granularity finer than that of the original data).

- Use the degree of the nodes of the indoor space graph representation to develop a new type of trajectory similarity metric (thereby distinguishing whether two trajectories are similar by "free-will" or as a result of indoor architectural constraints.).

- Transfer network-constrained outdoor trajectory analysis practices to the indoor space case

- Explore how landmarks affect indoor navigation (popular artworks serve the museum visitor as a reference for mentally re-estimating his current position).

- Develop methods to distinguish between *systematic non-detections* (i.e. coverage gaps or malfunctioning sensors), and *random non-detections* (i.e. occasional tracking errors) based on the relative frequency of the spatial entities involved and methods to deal with the former (e.g. graph rearrangement).

More specifically with respect to the proposed modeling work, whereas it was shown how the proposed *SITM* can handle certain modeling aspects such as topological constraints (accessibility NRGs), indoor space semantics (node classes and properties, parallel hierarchies), location uncertainty (spatial hierarchy, accessibility NRGs), there are other issues that were left untreated, such as changes in the indoor topology (e.g. dynamic graph), group patterns (e.g. overlap of presence intervals), etc.

There are also new interesting concepts to consider. For example, whereas in outdoor (people) trajectories, movement often consists of long travel distances between PoIs which clearly separate the activity-rich *stops* (taking place in those PoIs) from the activity-poor *moves* (taking place over transportation networks), for indoor mobility this is not the case. The corresponding distinction would be that between periods of *walking* and periods of *standing*, which means that activities can be more equally assigned to either interval type. The proposed model is interval-based and can therefore support this by annotating each item in the sequence as a "standing" or "walking" presence interval. Spatiotemporal threshold-based approaches can be used for doing so.

However, also the semantics of space have a major role to play in this. For instance, quoting [155]: *"...it is not considered as a movement if she/he just is moving around in a store. However, if she/he is walking in a corridor, we consider it as a movement. It means that the decision of movement depends on the granularity and classification of location."*. Hence, spatiotemporal threshold-based approaches could be used in synergy with space semantics. For example, a *stationarity* temporal threshold could be higher in a "TransitionSpace" with respect to an "ExhibitionSpace" to reflect the intuition that it is harder for a valid stop to take place there.

Future developments to be pursued after completion of the Thesis include the following:

- Generate artificial semantic trajectories using characteristics of the real Louvre visit trajectories.
- Validate $SITPE$ over those trajectories.
- Modify it to derive proper τ values automatically from the data in combination with domain expertise (not necessarily the size of the spatial region).
- Apply our analysis over a more complete Louvre dataset (i.e. without so many missing zones or such coarse spatial granularity) and validate if the discovered patterns still apply.
- Ensure that $SITM$ conforms to the new version of the IndoorGML standard (IndoorGML 2.0).
- Add a temporal dimension to the edges of the indoor space graph (i.e. $G = G(t)$) to model dynamic accessibility restrictions with dynamic/temporal graphs [134] (e.g. unavailability of exhibits, spaces under restoration).

Towards a Better Understanding of Museum Visitor Behavior

This Thesis illustrated how $SITM$ can be applied to an intriguing real-world case study, and validated it using state-of-the-art pattern mining algorithms in order to analyze real-world museum visitor trajectory data. Due to data accessibility and quality issues, it was not possible to validate experimentally the proposed novel pattern mining algorithm $SITPE$ (defined in section 5.1.2.3).

This is precisely why the next step of our reasearch group is to apply the $SITPE$ with the help of artificially generated museum visit trajectory data. The existing trajectory dataset is planned to be used to derive transition probabilities and duration distributions in this respect. Also, given that the user mismatch problem is insurmountable, the app usage dataset will enable the mimicking of user behavior in order to generate realistic semantic trajectories.

At the same time, it would be useful to repeat the experiments over an extended version of the Louvre dataset, because at least some of the quality issues that encountered in this Thesis have been improved over the course of time. Also, it is worth pursuing the mining of patterns from the app usage dataset on its own. After all, S-PM is often used to mine web activity which closely resembles app usage. Despite the fact that this is not related to trajectories, it would be innovative to map the usage events to their corresponding physical locations in the Louvre's indoor space graph, study the derived *virtual* trajectories in comparison to the actual space, and quantify how much they deviate from the typical physical trajectories. Correlating app usage data to indoor space has been recently proposed in [93], albeit in a different manner.

In any case, the Louvre visit patterns discovered in this Thesis, despite being simpler (i.e. only space and time are taken into account in the mining process), are still helpful and show that the proposed model is polyvalent in terms of trajectory mining and analysis support. In addition, after closely studying similar past research works in the museum domain, and with the help of our Louvre partners, this Thesis has identified the most important trajectory analysis tasks for any museum, presented in Table 6.1. This can serve as a basis for contemplating how to best take advantage of the digital tools at their disposal for a variety of purposes.

Finally, trajectory-based mining and analysis, especially related to semantic and indoor trajectories, will eventually be a key technology for improving the visitor experience, as it offers an insightful look at how they behave, and at the same time can scale to the numbers of visits encountered in the world's biggest museums. Trajectory pattern mining in particular will have a prominent role in this effort, especially if more progress is made around the integration of semantics and time in sequential patterns.

Glossary

Abbreviation	Explanation
2D	Two Dimensional
3D	Three Dimensional
BTD	Big Trajectory Data
BUC	BottomUpCube
GPS	Global Positioning System
IndoorGML	Indoor Geographic Markup Language
MD-S-PM	MultiDimensional Sequential Pattern Mining
MD-TAS-PM	MultiDimensional Temporally Annotated Sequential Pattern Mining
MLSM	Multi Layered Space Model
NRG	Node Relation Graph
PoI	Point of Interest
RoI	Region of Interest
RSSI	Received Signal Strength Indicator
S-PM	Sequential Pattern Mining
SITM	Semantic Indoor Trajectory Model
SITPE	Semantic Indoor Trajectory Pattern Extractor
SIT-PM	Semantic Indoor Trajectory Pattern Mining

Table 7.1: Glossary of special terms and abbreviations used in the present manuscript.

References

References

[1] CityGML homepage.

[2] Geography Markup Language | OGC.

[3] The Space Syntax approach.

[4] Henry Adams. What a physics student can teach us about how visitors walk through a museum. https://www.smithsonianmag.com/arts-culture/what-a-physics-student-can-teach-us-about-how-visitors-\walk-through-a-museum-96782621/, 5 2012. Accessed: 01-10-2021.

[5] Imad Afyouni, Cyril Ray, and Christophe Claramunt. Spatial models for context-aware indoor navigation systems: A survey. *Journal of Spatial Information Science*, 1(4):85–123, May 2012.

[6] Imad Afyouni, Cyril Ray, and Christophe Claramunt. *Representation: indoor spaces*, pages 1–12. American Cancer Society, 2017.

[7] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, pages 3–14, 1995.

[8] Shamal Al-Dohuki, Yingyu Wu, Farah Kamw, Jing Yang, Xin Li, Ye Zhao, Xinyue Ye, Wei Chen, Chao Ma, and Fei Wang. Semantictraj: A new approach to interacting with massive taxi trajectories. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):11–20, 2017.

[9] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, 2016.

[10] Lívia Almada Cruz, Karine Zeitouni, and José Antonio F. de Macedo. Trajectory prediction from a mass of sparse and missing external sensor data. In *2019 20th IEEE International Conference on Mobile Data Management (MDM)*, pages 310–319, 2019.

[11] Luis Otavio Alvares, Vania Bogorny, Bart Kuijpers, Jose Antonio Fernandes de Macedo, Bart Moelans, and Alejandro Vaisman. A Model for Enriching Trajectories with Semantic Geographical Information. In *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems*, GIS '07, pages 22:1–22:8, New York, NY, USA, 2007. ACM.

[12] Luis Otavio Alvares, Vania Bogorny, Bart Kuijpers, Bart Moelans, Jose Antonio Fern, ED Macedo, and Andrey Tietbohl Palma. Towards semantic trajectory knowledge discovery. *Data Mining and Knowledge Discovery*, 12, 2007.

[13] Moustafa Alzantot and Moustafa Youssef. Demonstrating crowdinside: A system for the automatic construction of indoor floor-plans. In *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 321–323, 2013.

[14] Gennady Andrienko, Natalia Andrienko, Peter Bak, Daniel Keim, Slava Kisilevich, and Stefan Wrobel. A Conceptual Framework and Taxonomy of Techniques for Analyzing Movement. *Journal of Visual Languages & Computing*, 22(3):213 – 232, 2011.

[15] Gennady Andrienko, Natalia Andrienko, and Marco Heurich. An Event-based Conceptual Model for Context-Aware Movement Analysis. *International Journal of Geographical Information Science*, 25:1347–1370, 2011.

[16] Gowtham Atluri, Anuj Karpatne, and Vipin Kumar. Spatio-temporal data mining: A survey of problems and methods. *ACM Comput. Surv.*, 51(4), August 2018.

[17] Marco Aurelio Beber, Carlos Ferrero, Renato Fileto, and Vania Bogorny. Towards Activity Recognition In Moving Object Trajectories From Twitter Data. November 2016.

[18] Marco Aurelio Beber, Carlos Andres Ferrero, Renato Fileto, and Vania Bogorny. Individual and group activity recognition in moving object trajectories. *Journal of Information and Data Management*, 8(1):50–66, 2017.

[19] Christian Becker and Frank Dürr. On Location Models for Ubiquitous Computing. *Personal Ubiquitous Comput.*, 9(1):20–31, January 2005.

[20] Thomas Becker, Claus Nagel, and Thomas H. Kolbe. Supporting Contexts for Indoor Navigation using a Multilayered Space Model. In *2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, pages 680–685, 2009.

[21] Luke Bermingham and Ickjai Lee. Spatio-temporal sequential pattern mining for tourism sciences. *Procedia Computer Science*, 29:379–389, 2014. 2014 International Conference on Computational Science.

[22] Kevin Beyer and Raghu Ramakrishnan. Bottom-up computation of sparse and iceberg cube. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, SIGMOD '99, page 359–370, New York, NY, USA, 1999. Association for Computing Machinery.

[23] Vania Bogorny, Carlos Alberto Heuser, and Luis Otavio Alvares. A Conceptual Data Model for Trajectory Data Mining. In *Proceedings of the 6th International Conference on Geographic Information Science (GIScience'10)*, Lecture Notes in Computer Science, pages 1–15, Berlin, Heidelberg, 2010. Springer-Verlag.

[24] Vania Bogorny, Bart Kuijpers, and Luis Otavio Alvares. ST-DMQL: a Semantic Trajectory Data Mining Query Language. *International Journal of Geographical Information Science*, 23(10):1245–1276, October 2009.

[25] Vania Bogorny, Chiara Renso, Artur Ribeiro de Aquino, Fernando de Lucca Siqueira, and Luis Otavio Alvares. CONSTAnT - A Conceptual Data Model for Semantic Trajectories of Moving Objects. *Transactions in GIS*, 18(1):66–88, 2014.

[26] Matteo Bohm, Mirco Nanni, and Luca Pappalardo. Quantifying the presence of air pollutants over a road network in high spatio-temporal resolution. 2020.

[27] Mariem Brahem, Hafsa E.L. Hafyani, Souheir Mehanna, Karine Zeitouni, Laurent Yeh, Yehia Taher, Zoubida Kedad, Ahmad Ktaish, Mohamed Chachoua, and Cyril Ray. 12 - data perspective on environmental mobile crowd sensing. In Siddhartha Bhattacharyya, Naba Kumar Mondal, Jan Platos, Vaclav Snasel, and Krömer, editors, *Intelligent Environmental Data Monitoring for Pollution Management*, Intelligent Data-Centric Systems, pages 269–288. Academic Press, 2021.

[28] Sotiris Brakatsoulas, Dieter Pfoser, and Nectaria Tryfona. Modeling, storing and mining moving object databases. In *Proceedings. International Database Engineering and Applications Symposium, 2004. IDEAS '04.*, pages 68–77. IEEE, 2004.

[29] Barry Brumitt and Steven Shafer. Topological World Modeling Using Semantic Spaces. In *In UbiComp 2001 Workshop on Location Modeling for Ubiquitous Computing*, pages 55–61, 2001.

[30] Julie Bu Daher, Armelle Brun, and Anne Boyer. A Review on Heterogeneous, Multi-source and Multi-dimensional data mining. Technical report, LORIA - Université de Lorraine, January 2018.

[31] Guochen Cai, Chihiro Hio, Luke Bermingham, Kyungmi Lee, and Ickjai Lee. Mining frequent trajectory patterns and regions-of-interest from flickr photos. In *2014 47th Hawaii International Conference on System Sciences*, pages 1454–1463, 2014.

[32] Guochen Cai, Chihiro Hio, Luke Bermingham, Kyungmi Lee, and Ickjai Lee. Sequential pattern mining of geo-tagged photos with an arbitrary regions-of-interest detection method. *Expert Systems with Applications*, 41(7):3514–3526, 2014.

[33] Guochen Cai, Kyungmi Lee, and Ickjai Lee. Mining semantic sequential patterns from geo-tagged photos. In *Proceedings of the 2016 49th Hawaii International Conference on System Sciences (HICSS)*, HICSS '16, page 2187–2196, USA, 2016. IEEE Computer Society.

[34] Guochen Cai, Kyungmi Lee, and Ickjai Lee. Mining semantic trajectory patterns from geo-tagged data. *Journal of Computer Science and Technology*, 33:849–862, 07 2018.

[35] Huiping Cao, N. Mamoulis, and D.W. Cheung. Mining frequent spatio-temporal sequential patterns. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8 pp.–, 2005.

[36] Pablo Samuel Castro, Daqing Zhang, Chao Chen, Shijian Li, and Gang Pan. From taxi gps traces to social and community dynamics: A survey. *ACM Comput. Surv.*, 46(2), December 2013.

[37] Pietro Centorrino, Alessandro Corbetta, Emiliano Cristiani, and Elia Onofri. Managing crowded museums: Visitors flow measurement, analysis, modeling, and optimization. *J. Comput. Sci.*, page 101357, 2021.

[38] Georgios Chatzigeorgakidis, Kostas Patroumpas, Dimitrios Skoutas, and Spiros Athanasiou. A visual explorer for geolocated time series. In *Proceedings of the 28th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '20, page 413–416, New York, NY, USA, 2020. Association for Computing Machinery.

[39] Juan Chen, Anthony G. Cohn, Dayou Liu, Shengsheng Wang, Jihong Ouyang, and Qiangyuan Yu. A Survey of Qualitative Spatial Representations. *The Knowledge Engineering Review*, 30(1):106–136, 2015.

[40] Yile Chen, Cheng Long, Gao Cong, and Chenliang Li. Context-aware deep model for joint mobility and time prediction. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20, page 106–114, New York, NY, USA, 2020. Association for Computing Machinery.

[41] Dong-Wan Choi, Jian Pei, and Thomas Heinis. Efficient mining of regional movement patterns in semantic trajectories. *Proc. VLDB Endow.*, 10(13):2073–2084, September 2017.

[42] C. Claramunt, C. Ray, L. Salmon, E. Camossi, M. Hadzagic, A-L. Jousselme, G. Andrienko, N. Andrienko, Y. Theodoridis, and G. Vouros. Maritime data integration and analysis: Recent progress and research challenges. *Advances in Database Technology - EDBT*, 2017:192–197, January 2017.

[43] Anthony Cohn, Brandon Bennett, John Gooday, and Mark Gotts. Qualitative Spatial Representation and Reasoning with the Region Connection Calculus. *GeoInformatica*, 1:275–316, 1997.

[44] Lara Corona. Museums and communication: The case of the louvre museum at the covid-19 age. *Humanities and Social Science Research*, 4(1):p15–p15, 2021.

[45] Christophe Cruz. Semantic Trajectory Modeling for Dynamic Built Environments. In *Special Session on Environmental and Geo-spatial Data Analytics (EnGeoData 2017), The 4th IEEE International Conference on Data Science and Advanced Analytics*, Tokyo, Japan, October 2017.

[46] Lívia Almada Cruz, Karine Zeitouni, Ticiana Linhares Coelho da Silva, José Antônio Fernandes de Macêdo, and José Soares da Silva. Location prediction: a deep spatiotemporal learning from external sensors data. *Distributed Parallel Databases*, 39(1):259–280, 2021.

[47] Zhan Cui, Anthony G. Cohn, and David A. Randell. Qualitative and Topological Relationships in Spatial Databases. In *Proceedings of the Third International Symposium on Advances in Spatial Databases (SSD '93)*, Lecture Notes in Computer Science, pages 296–315, Berlin, Heidelberg, 1993. Springer-Verlag.

[48] Anooshmita Das, Emil Stubbe Kolvig-Raun, and Mikkel Baun Kjærgaard. Accurate trajectory prediction in a smart building using recurrent neural networks. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, UbiComp-ISWC '20, page 619–628, New York, NY, USA, 2020. Association for Computing Machinery.

[49] Sophie Deshayes. Projet de renouvellement de l'audioguidage au louvre. Technical report, Laboratoire C2So, ENS-LSH, Lyon, Lyon, France, Mars 2006.

[50] Ann Devitt, Joseph Duffin, and Robert Moloney. Topographical Proximity for Mining Network Alarm Data. In *Proceedings of the 2005 ACM SIGCOMM Workshop on Mining Network Data*, MineNet '05, page 179–184, New York, NY, USA, 2005. Association for Computing Machinery.

[51] A. A. Diakité, S. Zlatanova, A. F. M. Alattas, and K. J. Li. Towards indoorgml 2.0: Updates and case study illustrations. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B4-2020:337–344, 2020.

[52] Abdoulaye Diakité, Sisi Zlatanova, and Ki Joune Li. About The Subdivision Of Indoor Spaces In IndoorGML. In *12th 3D Geoinfo Conference*, pages 41–48, 2017.

[53] Maja Dimitrijević. *Mining for co-occurring motion trajectories: sport analysis.* PhD thesis, University of British Columbia, 2001.

[54] Xin Ding, Lu Chen, Yunjun Gao, Christian S. Jensen, and Hujun Bao. Ultraman: A unified platform for big trajectory data management and analytics. *Proc. VLDB Endow.*, 11(7):787–799, March 2018.

[55] Max Egenhofer and John Herring. Categorizing Binary Topological Relations Between Regions, Lines and Points in Geographic Databases. *The 9-Intersection, Formalism and Its Use For Natural-Language Spatial Predicates*, 94:1–28, 1992.

[56] Max J. Egenhofer. Reasoning about binary topological relations. In *Proceedings of the Second International Symposium on Advances in Spatial Databases*, SSD '91, pages 143–160, London, UK, UK, 1991. Springer-Verlag.

[57] Max J. Egenhofer and Franzosa Robert D. Point-set topological spatial relations. *International Journal of Geographical Information Systems*, 5(2):161–174, 1991.

[58] Elias Egho, Chedy Raïssi, Nicolas Jay, and Amedeo Napoli. Mining Heterogeneous Multidimensional Sequential Patterns. Research Report RR-8521, INRIA, April 2014.

[59] Hafsa El Hafyani, Mohammad Abboud, Jingwei Zuo, Karine Zeitouni, and Yehia Taher. *Tell Me What Air You Breath, I Tell You Where You Are*, page 161–165. Association for Computing Machinery, New York, NY, USA, 2021.

[60] Golnaz Elmamooz, Bettina Finzel, and Daniela Nicklas. Towards understanding mobility in museums. In Bernhard Mitschang, Daniela Nicklas, Frank Leymann, Harald Schöning, Melanie Herschel, Jens Teubner, Theo Härder, Oliver Kopp, and Matthias Wieland, editors, *Datenbanksysteme für Business, Technologie und Web (BTW 2017) - Workshopband*, pages 127–136, Bonn, 2017. Gesellschaft für Informatik e.V.

[61] Yves Evrard and Anne Krebs. The authenticity of the museum experience in the digital age: the case of the louvre. *Journal of Cultural Economics*, 42:353–363, 01 2018.

[62] John H. Falk and Lynn D. Dierking. *The Museum Experience Revisited.* Routledge, 1 edition, 2013.

[63] Ziquan Fang, Lu Chen, Yunjun Gao, Lu Pan, and Christian S Jensen. Dragoon: a hybrid and efficient big trajectory management system for offline and online analytics. *The VLDB Journal*, 30(2):287–310, 2021.

[64] Shanshan Feng, Gao Cong, Bo An, and Yeow Meng Chee. Poi2vec: Geographical latent representation for predicting future visitors. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 102–108. AAAI Press, 2017.

[65] Zhenni Feng and Yanmin Zhu. A survey on trajectory data mining: Techniques and applications. *IEEE Access*, 4:2056–2067, 2016.

[66] Esther Calvo Fernández, José Manuel Cordero, George Vouros, Nikos Pelekis, Theocharis Kravaris, Harris Georgiou, Georg Fuchs, Natalya Andrienko, Gennady Andrienko, Enrique Casado, et al. Dart: a machine-learning approach to trajectory prediction and demand-capacity balancing. *SESAR Innovation Days, Belgrade*, pages 28–30, 2017.

[67] Carlos Ferrero, Luis Alvares, and Vania Bogorny. Multiple aspect trajectory data analysis: Research challenges and opportunities. 11 2016.

[68] Ruipeng Gao, Bing Zhou, Fan Ye, and Yizhou Wang. Knitter: Fast, resilient single-user indoor floor plan construction. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pages 1–9, 2017.

[69] Fosca Giannotti, Mirco Nanni, and Dino Pedreschi. Efficient mining of temporally annotated sequences. volume 2006, 04 2006.

[70] Fosca Giannotti, Mirco Nanni, Dino Pedreschi, and Fabio Pinelli. Mining Sequences with Temporal Annotations. In *Proceedings of the ACM Symposium on Applied Computing*, volume 1, pages 593–597, 01 2006.

[71] Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. Trajectory Pattern Mining. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 330–339, New York, NY, USA, 2007. ACM.

[72] Nicholas Giudice, Lisa Walton, and Michael Worboys. The informatics of indoor and outdoor space: A research agenda. pages 47–53, 01 2010.

[73] Fragkiskos Gryllakis, Nikos Pelekis, Christos Doulkeridis, Stylianos Sideridis, and Yannis Theodoridis. Spatio-temporal-keyword pattern queries over semantic trajectories with hermes@ neo4j. In *EDBT*, pages 678–681, 2018.

[74] Guillaume Fonkenell. Les escaliers du Louvre, December 2020.

[75] Ralf Hartmut Güting, Thomas Behr, and Christian Düntgen. *SECONDO: a platform for moving objects database research and for publishing and integrating research implementations*, volume 356 of *Informatik-Berichte*. FernUniversität in Hagen, Hagen, 2010.

[76] Ralf Hartmut Güting, Michael H. Böhlen, Martin Erwig, Christian S. Jensen, Nikos A. Lorentzos, Markus Schneider, and Michalis Vazirgiannis. A foundation for representing and querying moving objects. *ACM Trans. Database Syst.*, 25(1):1–42, March 2000.

[77] Ralf Hartmut Güting, Fabio Valdés, and Maria Luisa Damiani. Symbolic trajectories. 1(2), July 2015.

[78] Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. *Data Mining and Knowledge Discovery*, 8(1):53–87, 1 2004.

[79] Benjamin Harbelot, Helbert Arenas, and Christophe Cruz. LC3: A spatio-temporal and semantic model for knowledge discovery from geospatial datasets. *Journal of Web Semantics*, 35:3 – 24, 2015. Geospatial Semantics.

[80] Zhixiang He, Chi-Yin Chow, and Jia-Dong Zhang. Stcnn: A spatio-temporal convolutional neural network for long-term traffic prediction. In *2019 20th IEEE International Conference on Mobile Data Management (MDM)*, pages 226–233, 2019.

[81] Stephen C. Hirtle and John Jonides. Evidence of hierarchies in cognitive maps. *Memory & Cognition*, 13(3):208–217, 1985.

[82] Haibo Hu and Dik Lun Lee. Semantic Location Modeling for Location Navigation in Mobile Environment. In *5th IEEE International Conference on Mobile Data Management (MDM 2004), 19-22 January 2004, Berkeley, CA, USA*, pages 52–61. IEEE Computer Society, 2004.

[83] Yan Huang, Liqin Zhang, and Pusheng Zhang. A framework for mining sequential patterns from spatio-temporal event data sets. *IEEE Transactions on Knowledge and Data Engineering*, 20(4):433–448, 2008.

[84] San-Yih Hwang, Ying-Han Liu, Jeng-Kuen Chiu, and Ee-Peng Lim. Mining mobile group patterns: A trajectory-based approach. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 713–718. Springer, 2005.

[85] S. Ilarri, D. Stojanovic, and C. Ray. Semantic management of moving objects: A vision towards smart mobility. *Expert Syst. Appl.*, 42:1418–1435, 2015.

[86] C. S. Jensen, H. Lu, and B. Yang. Graph Model Based Indoor Tracking. In *2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, pages 122–131, May 2009.

[87] Christian Søndergaard Jensen, Hua Lu, and Bin Yang. Indoor - A New Data Management Frontier. *IEEE Computer Society Data Engineering Bulletin*, 33(2):12–17, 2010.

[88] Fengmei Jin, Wen Hua, Thomas Zhou, Jiajie Xu, Matteo Francia, Maria Orowska, and Xiaofang Zhou. Trajectory-based spatiotemporal entity linking. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2020.

[89] Panos Kalnis, Nikos Mamoulis, and Spiridon Bakiras. On discovering moving clusters in spatio-temporal data. In Claudia Bauzer Medeiros, Max J. Egenhofer, and Elisa Bertino, editors, *Advances in Spatial and Temporal Databases*, pages 364–381, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[90] Takayuki Kanda, Masahiro Shiomi, Laurent Perrin, Tatsuya Nomura, Hiroshi Ishiguro, and Norihiro Hagita. Analysis of people trajectories with ubiquitous sensors in a science museum. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 4846–4853, 2007.

[91] Hae-Kyong Kang and Ki-Joune Li. A Standard Indoor Spatial Data Model - OGC IndoorGML and Implementation Approaches. *ISPRS International Journal of Geo-Information*, 6(4):1–25, 2017.

[92] Juyoung Kang and Hwan-Seung Yong. Mining spatio-temporal patterns in trajectory data. *JIPS*, 6:521–536, 12 2010.

[93] Shinjin Kang and Soo Kyun Kim. Behavior analysis method for indoor environment based on app usage mining. *The Journal of Supercomputing*, pages 1–21, 2021.

[94] Tung Kieu, Bin Yang, Chenjuan Guo, and Christian S. Jensen. Distinguishing trajectories from different drivers using incompletely labeled trajectories. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 863–872, New York, NY, USA, 2018. Association for Computing Machinery.

[95] Taehoon Kim, Kyoung-Sook Kim, Jiyeong Lee, and Ki-Joune Li. How to extend indoorgml for points of interest. In Sergio Di Martino, Zhixiang Fang, and Ki-Joune Li, editors, *Web and Wireless Geographical Information Systems*, pages 3–13, Cham, 2020. Springer International Publishing.

[96] T. Kindberg, J. Barton, J. Morgan, G. Becker, D. Caswell, P. Debaty, G. Gopal, M. Frid, V. Krishnan, H. Morris, J. Schettino, B. Serra, and M. Spasojevic. People, places, things: Web presence for the real world. In *Proceedings Third IEEE Workshop on Mobile Computing Systems and Applications*, pages 19–28, October 2002.

[97] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. Multilayer Networks. *Journal of Complex Networks*, 2(3):203–271, Jul 2014.

[98] Dániel Kondor, Behrooz Hashemian, Yves-Alexandre de Montjoye, and Carlo Ratti. Towards matching user mobility traces in large-scale datasets. *IEEE Transactions on Big Data*, 6(4):714–726, 2020.

[99] X. Kong, M. Li, K. Ma, K. Tian, M. Wang, Z. Ning, and F. Xia. Big Trajectory Data: A Survey of Applications and Services. *IEEE Access*, 6:58295–58306, 2018.

[100] Alexandros Kontarinis, Claudia Marinica, Dan Vodislav, Karine Zeitouni, Anne Krebs, and Dimitris Kotzinos. Towards a Better Understanding of Museum Visitors' Behavior through Indoor Trajectory Analysis. In *Seventh International*

Conference on Digital Presentation and Preservation of Cultural and Scientific Heritage (DiPP2017), volume 7, pages 19–30, Burgas, Bulgaria, September 2017.

[101] Alexandros Kontarinis, Karine Zeitouni, Claudia Marinica, Dan Vodislav, and Dimitris Kotzinos. Towards a Semantic Indoor Trajectory Model. In *2nd International Workshop on "Big Mobility Data Analytics" (BMDA) with EDBT 2019*, Lisbon, Portugal, March 2019.

[102] Alexandros Kontarinis, Karine Zeitouni, Claudia Marinica, Dan Vodislav, and Dimitris Kotzinos. Towards a semantic indoor trajectory model: application to museum visits. *GeoInformatica*, 25:311–352, 05 2021.

[103] Manolis Koubarakis, Timos K. Sellis, Andrew U. Frank, Stéphane Grumbach, Ralf Hartmut Güting, Christian S. Jensen, Nikos A. Lorentzos, Yannis Manolopoulos, Enrico Nardelli, Barbara Pernici, Hans-Jörg Schek, Michel Scholl, Babis Theodoulidis, and Nectaria Tryfona, editors. *Spatio-Temporal Databases: The CHOROCHRONOS Approach*, volume 2520 of *Lecture Notes in Computer Science*. Springer, 2003.

[104] Anne Krebs. Musée et "Mégadonnées": Partenariats de Recherche au Musée du Louvre. *Grande Galerie: le Journal du Louvre, LA RECHERCHE AU LOUVRE 2019 HORS-SÉRIE*, pages 10–17, 06 2020.

[105] Tsvi Kuflik and Eyal Dim. Integrating signals for reasoning about visitors' behavior in cultural heritage. In Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe, editors, *Multimodal Behavior Analysis in the Wild*, Computer Vision and Pattern Recognition, pages 159 – 169. Academic Press, 2019.

[106] Patrick Le Boeuf, Martin Doerr, Christian Emil Ore, and Stephen Stead. Definition of the CIDOC Conceptual Reference Model, 2019. Accessed on 05.05.2020.

[107] Jiyeong Lee. A Spatial Access-Oriented Implementation of a 3-D GIS Topological Data Model for Urban Entities. *Geoinformatica*, 8(3):237–264, September 2004.

[108] Jiyeong Lee, Ki-Joune Li, Sisi Zlatanova, Thomas H. Kolbe, Claus Nagel, and Thomas Becker. Requirements and Space-Event Modeling for Indoor Navigation. Technical report, Open Geospatial Consortium, 2010.

[109] Jiyeong Lee, Ki-Joune Li, Sisi Zlatanova, Thomas H. Kolbe, Claus Nagel, and Thomas Becker. OGC IndoorGML. Technical report, Open Geospatial Consortium, 2018.

[110] JW LEE, TW Heo, KS Kim, and JH Lee. Continuous moving pattern mining approach in lbs platform. In *Proceedings of the KSRS Conference*, pages 597–599. The Korean Society of Remote Sensing, 2003.

[111] Luiz André P. Paes Leme, Chiara Renso, Bernardo P. Nunes, Giseli Rabello Lopes, Marco A. Casanova, and Vânia P. Vidal. Searching for Data Sources for the Semantic Enrichment of Trajectories. In *Web Information Systems Engineering - WISE 2016*, Lecture Notes in Computer Science, pages 238–246. Springer, Cham, November 2016.

[112] Ruiyuan Li, Huajun He, Rubin Wang, Sijie Ruan, Yuan Sui, Jie Bao, and Yu Zheng. Trajmesa: A distributed nosql storage engine for big trajectory data. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 2002–2005, 2020.

[113] Yuxuan Liang, Kun Ouyang, Junkai Sun, Yiwei Wang, Junbo Zhang, Yu Zheng, David Rosenblum, and Roger Zimmermann. Fine-grained urban flow prediction. In *Proceedings of the Web Conference 2021*, WWW '21, page 1833–1845, New York, NY, USA, 2021. Association for Computing Machinery.

[114] Thomas Liebig, Gennady Andrienko, and Natalia Andrienko. Methods for Analysis of Spatio-Temporal Bluetooth Tracking Data. *Journal of Urban Technology*, 21:27–37, April 2014.

[115] L. Liu, S. Zlatanova, Q. Zhu, and K. Li. Towards The Integration Of INDOORGML And INDOORLOCATIONGML For Indoor Applications. In *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume IV-2-W4, pages 343–348. Copernicus GmbH, September 2017.

[116] Liu Liu and Sisi Zlatanova. Towards a 3d network model for indoor navigation. In *Urban and Regional Data Management, UDMS Annual 2011 - Proceedings of the Urban Data Management Society Symposium 2011*, pages pp.79–92. September 2011.

[117] Louvre Museum Press Room. The Louvre Welcomes 8.6 Million Visitors in 2015, 2015. Accessed on 01.10.2021.

[118] Louvre Museum Press Room. 7.4 million visitors to the Louvre in 2016, 2016. Accessed on 01.10.2021.

[119] Louvre Museum Press Room. 8.1 million visitors to the Louvre in 2017, 2017. Accessed on 01.10.2021.

[120] Louvre Museum Press Room. 10.2 million visitors to the Louvre in 2018, 2018. Accessed on 01.10.2021.

[121] Louvre Museum Press Room. 9.6 million visitors to the Louvre in 2019, 2019. Accessed on 01.10.2021.

[122] Louvre Museum Press Room. 1.1 million visitors for "Leonardo da Vinci" A new record for an exhibition at the Louvre, 2020. Accessed on 01.10.2021.

[123] H. Lu, X. Cao, and C. S. Jensen. A Foundation for Efficient Indoor Distance-Aware Query Processing. In *2012 IEEE 28th International Conference on Data Engineering*, pages 438–449, April 2012.

[124] Hua Lu, Chenjuan Guo, Bin Yang, and Christian S. Jensen. Finding Frequently Visited Indoor POIs Using Symbolic Indoor Tracking Data. In *Proceedings of the 19th International Conference on Extending Database Technology (EDBT2016)*, pages 449–460, 2016.

[125] Jianming Lv, Qing Li, Qinghui Sun, and Xintong Wang. T-conv: A convolutional neural network for multi-scale taxi trajectory prediction. In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 82–89, 2018.

[126] Takuya Maekawa, Kazuya Ohara, Yizhe Zhang, Matasaburo Fukutomi, Sakiko Matsumoto, Kentarou Matsumura, Hisashi Shidara, Shuhei J. Yamazaki, Ryusuke Fujisawa, Kaoru Ide, Naohisa Nagaya, Koji Yamazaki, Shinsuke Koike, Takahisa Miyatake, Koutarou D. Kimura, Hiroto Ogawa, Susumu Takahashi, and Ken Yoda. Deep learning-assisted comparative analysis of animal trajectories with deephl. *Nature Communications*, 11, 2020.

[127] Giovanni Malnati, Irene Rubino, Andrea Giuseppe Bottino, Jetmir Xhembulla, and Andrea Martina. MusA: Using Indoor Positioning and Navigation to Enhance Cultural Experiences in a museum. *MyScienceWork*, January 2013.

[128] Heikki Mannila, Hannu Toivonen, and A. Inkeri Verkamo. Discovery of frequent episodes in event sequences. 1(3):259–289, January 1997.

[129] Nikola Marković, Przemyław Sekuła, Zachary Vander Laan, Gennady Andrienko, and Natalia Andrienko. Applications of trajectory data from the perspective of a road transportation agency: Literature review and maryland case study. *IEEE Transactions on Intelligent Transportation Systems*, 20(5):1858–1869, 2019.

[130] Paul F. Marty and Katherine Burton Jones. *The Museum Experience Revisited.* Routledge, 1 edition, 2008.

[131] Paul F. Marty and Katherine Burton Jones. *Museum Informatics: People, Information, and Technology in Museums.* Taylor & Francis, 2008.

[132] Jean Damascene Mazimpaka and Sabine Timpf. Trajectory data mining: A review of methods and applications. *Journal of Spatial Information Science*, 13:61–99, 12 2016.

[133] Ronaldo dos Santos Mello, Vania Bogorny, Luis Otavio Alvares, Luiz Henrique Zambom Santana, Carlos Andres Ferrero, Angelo Augusto Frozza, Geomar Andre Schreiner, and Chiara Renso. MASTER: A Multiple Aspect View on Trajectories. *Transactions in GIS*, 23(4):805–822, 2019.

[134] Othon Michail. *An Introduction to Temporal Graphs: An Algorithmic Perspective*, pages 308–343. Springer International Publishing, Cham, 2015.

[135] A. Mousavi, A. Sheikh Mohammad Zadeh, M. Akbari, and A. Hunter. A new ontology-based approach for human activity recognition from gps data. *Journal of AI and Data Mining*, 5(2):197–210, 2017.

[136] Hassan Noureddine, Cyril Ray, and Christophe Claramunt. Multiple views of semantic trajectories in indoor and outdoor spaces. In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '21, page 99–102, New York, NY, USA, 2021. Association for Computing Machinery.

[137] Georgios Papaioannou and Ioannis Sarakinos. The general data protection regulation (gdpr, 2016/679/ee) and the (big) personal data in cultural institutions: Thoughts on the gdpr compliance process. In Milena Dobreva, Annika Hinze, and Maja Žumer, editors, *Maturity and Innovation in Digital Libraries*, pages 201–204, Cham, 2018. Springer International Publishing.

[138] Simon Michael Papalexiou, Francesco Serinaldi, and Emilio Porcu. Advancing space-time simulation of random fields: From storms to cyclones and beyond. *Water Resources Research*, 2021.

[139] Christine Parent, Stefano Spaccapietra, Chiara Renso, Gennady Andrienko, Natalia Andrienko, Vania Bogorny, Maria Luisa Damiani, Aris Gkoulalas-Divanis, Jose Macedo, Nikos Pelekis, Yannis Theodoridis, and Zhixian Yan. Semantic Trajectories Modeling and Analysis. *ACM Comput. Surv.*, 45(4):42:1–42:32, 2013.

[140] Christine Parent, Stefano Spaccapietra, and Esteban Zimányi. Spatio-temporal conceptual models: Data structures + space + time. In *Proceedings of the 7th ACM International Symposium on Advances in Geographic Information Systems*, GIS '99, page 26–33, New York, NY, USA, 1999. Association for Computing Machinery.

[141] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In *Proceedings of the 17th International Conference on Data Engineering*, pages 215–224, USA, 2 2001. IEEE Computer Society.

[142] Nikos Pelekis, Stylianos Sideridis, Panagiotis Tampakis, and Yannis Theodoridis. Simulating our lifesteps by example. 2(3), October 2016.

[143] Donna J. Peuquet. *Representations of Space and Time.* Guilford Press, June 2002.

[144] Gregory Piateski and William Frawley. *Knowledge Discovery in Databases.* MIT Press, Cambridge, MA, USA, 1991.

[145] Helen Pinto, Jiawei Han, Jian Pei, Ke Wang, Qiming Chen, and Umeshwar Dayal. Multi-dimensional Sequential Pattern Mining. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 81–88, New York, NY, USA, 2001. ACM.

[146] Marc Plantevit, Yeow Wei Choong, Anne Laurent, Dominique Laurent, and Maguelonne Teisseire. M2SP: Mining Sequential Patterns Among Several Dimensions. In Alípio Mário Jorge, Luís Torgo, Pavel Brazdil, Rui Camacho, and João Gama, editors, *Knowledge Discovery in Databases: PKDD 2005*, volume 3271, pages 205–216, Berlin, Heidelberg, October 2005. Springer-Verlag.

[147] Marc Plantevit, Anne Laurent, Dominique Laurent, Maguelonne Teisseire, and Yeow Wei Choong. Mining Multi-Dimensional and Multi-Level Sequential Patterns. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4:1–37, 2010.

[148] Marc Plantevit, Anne Laurent, and Maguelonne Teisseire. HYPE: Mining Hierarchical Sequential Patterns. In *Proceedings of the 9th ACM International Workshop on Data Warehousing and OLAP*, DOLAP '06, pages 19–26, New York, NY, USA, 2006. ACM.

[149] Iulian Sandu Popa, Karine Zeitouni, Vincent Oria, and Ahmed Kharrat. Spatiotemporal Compression of Trajectories in Road Networks. *GeoInformatica*, 19(1):117–145, 2015.

[150] T. Prentow, A. Thom, H. Blunck, and J. Vahrenhold. Making Sense of Trajectory Data in Indoor Spaces. In *2015 16th IEEE International Conference on Mobile Data Management*, volume 1, pages 116–121, June 2015.

[151] Laura Radaelli, Dovydas Sabonis, Hua Lu, and Christian S. Jensen. Identifying typical movements among indoor objects - concepts and empirical study. In *Proceedings of the 2013 IEEE 14th International Conference on Mobile Data Management - Volume 01*, MDM '13, page 197–206, USA, 2013. IEEE Computer Society.

[152] Chiara Renso, Vania Bogorny, Konstantinos Tserpes, Stan Matwin, and Jose Antonio Fernandes de Macedo. Multiple-aspect analysis of semantic trajectories (master). *International Journal of Geographical Information Science*, 35(4):763–766, 2021.

[153] Jochen Renz. *Qualitative Spatial Reasoning with Topological Information.* Springer-Verlag, Berlin, Heidelberg, 2002.

[154] GFK-Custom Research. Publics, usages et réception de l'audioguide sur console nintendo 3ds du musée du louvre. Technical report, Centre de Recherche - Direction de la Recherche et des Collections Etudes et Recherches sur les Publics, Paris, France, March 2014.

[155] Hyung-Gyu Ryoo, Taehoon Kim, and Ki-Joune Li. Comparison Between Two OGC Standards for Indoor Space: CityGML and IndoorGML. In *Proceedings of the Seventh ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness*, ISA '15, pages 1:1–1:8. ACM, 2015.

[156] Stefan Schabus, Johannes Scholz, and Thomas J Lampoltshammer. A space in a space: Connecting indoor and outdoor geography. In *Proceedings of 18th AGILE Conference on Geographic Information Science*, pages 1–5, 06 2015.

[157] Stan Sclarrof, George Kollios, Margrit Betke, and Romer Rosales. Motion mining. In Maurizio Tucci, editor, *Multimedia Databases and Image Communication*, pages 16–30, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[158] Mohammad Sharif and Ali Asghar Alesheikh. Context-aware movement analytics: Implications, taxonomy, and design framework. 8(1):n/a, January 2018.

[159] Shelley Bernstein. The Realities of Installing iBeacon to Scale, 2015.

[160] Masahiro Shiomi, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. Interactive humanoid robots for a science museum. *IEEE Intelligent Systems*, 22(2):25–32, 2007.

[161] Alina Sirbu, Gennady Andrienko, Natalia Andrienko, Chiara Boldrini, Marco Conti, Fosca Giannotti, Riccardo Guidotti, Simone Bertoli, Jisu Kim, Cristina Ioana Muntean, et al. Human migration: the big data perspective. *International Journal of Data Science and Analytics*, page 1–20, 2020.

[162] Stefano Spaccapietra and Christine Parent. Adding Meaning to Your Steps. In *Proceedings of the 30th International Conference on Conceptual Modeling*, ER'11, pages 13–31, Berlin, Heidelberg, 2011. Springer-Verlag.

[163] Stefano Spaccapietra, Christine Parent, Maria Luisa Damiani, Jose Antonio de Macedo, Fabio Porto, and Christelle Vangenot. A Conceptual View on Trajectories. *Data and Knowledge Engineering*, 65(1):126–146, 2008.

[164] Flavia Sparacino. The museum wearable: real-time sensor-driven understanding of visitors' interests for personalized visually-augmented museum experiences. *INSTITUTION Archives and Museum Informatics, Pittsburgh, PA. ISBN ISBN-1-885626-25-1*, 17:41, 2002.

[165] Ramakrishnan Srikant and Rakesh Agrawal. Mining Sequential Patterns: Generalizations and Performance Improvements. In *Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '96, pages 3–17, London, UK, UK, 1996. Springer-Verlag.

[166] Edgar-Philipp Stoffel, Bernhard Lorenz, and Hans Jürgen Ohlbach. Towards a Semantic Spatial Model for Pedestrian Indoor Navigation. In Jean-Luc Hainaut, Elke A. Rundensteiner, Markus Kirchberg, Michela Bertolotto, Mathias

Brochhausen, Yi-Ping Phoebe Chen, Samira Si-Saïd Cherfi, Martin Doerr, Hyoil Han, Sven Hartmann, Jeffrey Parsons, Geert Poels, Colette Rolland, Juan Trujillo, Eric Yu, and Esteban Zimányie, editors, *Advances in Conceptual Modeling – Foundations and Applications*, pages 328–337, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[167] Edgar-Philipp Stoffel, Korbinian Schoder, and Hans Jürgen Ohlbach. Applying Hierarchical Graphs to Pedestrian Indoor Navigation. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '08, pages 54:1–54:4, New York, NY, USA, 2008. ACM.

[168] Shan-Yun Teng, Wei-Shinn Ku, and Kun-Ta Chuang. Toward mining stop-by behaviors in indoor space. *ACM Trans. Spatial Algorithms Syst.*, 3(2), August 2017.

[169] John M.A. Thompson. *Manual of Curatorship: A Guide to Museum Practice.* Routledge, 2 edition, 2012.

[170] IIias Tsoukatos and Dimitrios Gunopulos. Efficient mining of spatiotemporal patterns. In Christian S. Jensen, Markus Schneider, Bernhard Seeger, and Vassilis J. Tsotras, editors, *Advances in Spatial and Temporal Databases*, pages 425–442, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[171] Kali Tzortzi. Movement in museums: mediating between museum intent and visitor experience. *Museum Management and Curatorship*, 29(4):327–348, 2014.

[172] Fabio Valdés and Ralf Güting. A framework for efficient multi-attribute movement data analysis. *The VLDB Journal*, 28, 08 2019.

[173] Fabio Valdés and Ralf Güting. Efficient multi-attribute analysis for trajectories: A case study for aircraft. 11 2017.

[174] Brandon Victor, Aiden Nibali, Zhen He, and David L. Carey. Enhancing trajectory prediction using sparse outputs: application to team sports. *Neural Computing and Applications*, 33(18):11951–11962, Mar 2021.

[175] Claudia Villalonga, Muhammad Asif Razzaq, Wajahat Ali Khan, Hector Pomares, Ignacio Rojas, Sungyoung Lee, and Oresti Banos. Ontology-Based High-Level Context Inference for Human Behavior Identification. *Sensors*, 16(10):1–26, 2016.

[176] John Violos, Stylianos Tsanakas, Maro Androutsopoulou, Georgios Palaiokrassas, and Theodora Varvarigou. Next position prediction using lstm neural networks. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 232–240, New York, NY, USA, 2020. Association for Computing Machinery.

[177] G. Vouros, C. Doulkeridis, G. Santipantakis, A. Vlachou, N. Pelekis, H. Georgiou, Y. Theodoridis, K. Patroumpas, E. Alevizos, A. Artikis, G. Fuchs,

M. Mock, G. Andrienko, N. Andrienko, C. Ray, C. Claramunt, E. Camossi, A-L. Jousselme, D. Scarlatti, and J. M. Cordero. Big data analytics for time critical maritime and aerial mobility forecasting. *Advances in Database Technology - EDBT*, 2018:612–623, January 2018. © 2018 Copyright held by the owner/author(s).

[178] Thomas Vroylandt. Cartographie des données du louvre. Technical report, Unité des études et des recherches socio-économiques de la direction de la recherche et des collections, Paris, France, February 2021.

[179] Di Wang, Tomio Miwa, and Takayuki Morikawa. Big trajectory data mining: A survey of methods, applications, and services. *Sensors*, 20(16), 2020.

[180] Gangani Geethika Wijewardena, Maria Vasardani, and Stephan Winter. Towards indoor localization and navigation independent of sensor based technologies. In *Proceedings of the Eighth ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness*, ISA '16, page 19–26, New York, NY, USA, 2016. Association for Computing Machinery.

[181] M. F. Worboys. A Unified Model for Spatial and Temporal Information. *The Computer Journal*, 37(1):26–34, January 1994.

[182] Michael Worboys. Modeling indoor space. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness*, ISA '11, page 1–6, New York, NY, USA, 2011. Association for Computing Machinery.

[183] Jing Wu, Christophe Claramunt, and Min Deng. An integrated qualitative and boundary-based formal model for a semantic representation of trajectories. *SIGSPATIAL Special*, 7(1):35–42, May 2015.

[184] Qing Xiong, Qing Zhu, Sisi Zlatanova, Liang Huang, Yan Zhou, and Zhiqiang Du. Multi-dimensional Indoor Location Information Model. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, volume 40, December 2013.

[185] Jianqiu Xu, Zhifeng Bao, and Hua Lu. A framework to support continuous range queries over multi-attribute trajectories. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2021.

[186] Jianqiu Xu, Hua Lu, and Ralf Hartmut Güting. Understanding human mobility: A multi-modal and intelligent moving objects database. In *Proceedings of the 16th International Symposium on Spatial and Temporal Databases*, SSTD '19, page 222–225, New York, NY, USA, 2019. Association for Computing Machinery.

[187] Zhixian Yan, Dipanjan Chakraborty, Christine Parent, Stefano Spaccapietra, and Karl Aberer. SeMiTri: A Framework for Semantic Annotation of Heterogeneous Trajectories. In *Proceedings of the 14th International Conference on*

Extending Database Technology, EDBT/ICDT '11, pages 259–270, New York, USA, 2011. ACM.

[188] Zhixian Yan, Dipanjan Chakraborty, Christine Parent, Stefano Spaccapietra, and Karl Aberer. Semantic Trajectories: Mobility Data Computation and Annotation. *ACM Transactions on Intelligent Systems and Technology*, 4(3):49:1–49:38, 2013.

[189] Zhixian Yan, Jose Macedo, Christine Parent, and Stefano Spaccapietra. Trajectory Ontologies and Queries. *Transactions in GIS*, 12(s1):75–91, 2008.

[190] Zhixian Yan, Christine Parent, Stefano Spaccapietra, and Dipanjan Chakraborty. A Hybrid Model and Computing Platform for Spatio-semantic Trajectories. In *The Semantic Web: Research and Applications*, volume 6088 of *Lecture Notes in Computer Science*, pages 60–75, Berlin, Heidelberg, May 2010. Springer Berlin Heidelberg.

[191] Yuji Yoshimura, Fabien Girardin, Juan Pablo Carrascal, Carlo Ratti, and Josep Blat. New tools for studying visitor behaviours in museums: A case study at the louvre. In Matthias Fuchs, Francesco Ricci, and Lorenzo Cantoni, editors, *Information and Communication Technologies in Tourism 2012*, pages 391–402, Vienna, 2012. Springer Vienna.

[192] Yuji Yoshimura, Anne Krebs, and Carlo Ratti. An analysis of visitors' length of stay through noninvasive bluetooth monitoring in the louvre museum. *CoRR*, abs/1605.00108, 2016.

[193] Yuji Yoshimura, Anne Krebs, and Carlo Ratti. Noninvasive Bluetooth Monitoring of Visitors' Length of Stay at the Louvre. *IEEE Pervasive Computing*, 16(2):26–34, 2017.

[194] Yuji Yoshimura, Roberta Sinatra, Anne Krebs, and Carlo Ratti. Analysis of visitors' mobility patterns through random walk in the Louvre Museum. *Journal of Ambient Intelligence and Humanized Computing*, 08 2019.

[195] Yuji Yoshimura, Stanislav Sobolevsky, Carlo Ratti, Fabien Girardin, Juan Pablo Carrascal, Josep Blat, and Roberta Sinatra. An analysis of visitors' behavior in the louvre museum: A study using bluetooth data. *Environment and Planning B: Planning and Design*, 41(6):1113–1131, 2014.

[196] Mohammed J. Zaki. Spade: An efficient algorithm for mining frequent sequences. *Mach. Learn.*, 42(1–2):31–60, January 2001.

[197] Chao Zhang, Jiawei Han, Lidan Shou, Jiajun Lu, and Thomas La Porta. Splitter: Mining fine-grained sequential patterns in semantic trajectories. *Proc. VLDB Endow.*, 7(9):769–780, May 2014.

[198] Yu Zheng. Trajectory Data Mining: An Overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3):1–41, 2015.

[199] Qing Zhu, Yun Li, Qing Xiong, Sisi Zlatanova, Yulin Ding, Yeting Zhang, and Yan Zhou. Indoor Multi-Dimensional Location GML and Its Application for Ubiquitous Indoor Location Services. *ISPRS International Journal of Geo-Information*, 5(12):220, November 2016.