



Sixth Framework Programme  
Information Society Technologies  
Future Emerging Technologies



Project Acronym:

**GeoPKDD**

Project full title:

**Geographic Privacy-aware Knowledge Discovery and Delivery**

Project Number: FP6-014915

Instrument: Specific Target Research Project

Project Deliverable D3.3

**First report on methods for reasoning and visualisation of spatio-temporal patterns**

Date of preparation: 05/12/2006

Revision: final

Operative commencement date of Contract: 01/12/2005

Project Coordinator: Fosca Giannotti

Project institute coordinator: Knowledge Discovery and Delivery-LAB / ISTI-CNR

**Programme Name:** .....IST  
**Project Number:**.....014915  
**Project Title:** .....GEOPKDD

**Document Number:** .....D3.3  
**Work-Package:**.....WP3  
**Document Type:** .....Deliverable  
**Contractual Date of Delivery:** .....01/12/2006  
**Actual Date of Delivery:** .....05/12/2006  
**Title of Document:** .....First report on methods for reasoning and visualisation of spatio-temporal patterns

**Author(s):** .....Gennady Andrienko, Natalia Andrienko, Chiara Renso, Bart Kujpers, Hasselt University & TUL, Belgium Monica Wachowicz, Arend Ligtenberg,

**Dissemination Level** .....Public

#### GeoPKDD consortium participants

Partic. Role	Partic. N.	Participant name	Short name	Country
CO	1	<b>KDD Lab.</b> , joint research group of <b>ISTI-CNR</b> , Istituto di Scienza e Tecnologie dell'Informazione, Pisa. <a href="http://www-kdd.isti.cnr.it/">http://www-kdd.isti.cnr.it/</a> and <b>Univ. Pisa</b> , Dept. of Computer Science <a href="http://www.di.unipi.it">http://www.di.unipi.it</a>	<b>KDDLAB</b>	I
CR	2	<b>Univ. of Hasselt</b> , Theoretical Computer Science Group. <a href="http://alpha.uhasselt.be/research/groups/theocomp/">http://alpha.uhasselt.be/research/groups/theocomp/</a>	<b>HASSELT</b>	B
CR	3	<b>EPFL - Ecole Polytechnique Fédérale de Lausanne</b> , Lab. DB, Lausanne. <a href="http://lbdwww.epfl.ch/e/">http://lbdwww.epfl.ch/e/</a> and University of Milan - Computer Science and Communication Department <a href="http://www.dico.unimi.it/">http://www.dico.unimi.it/</a>	<b>EPFL</b>	CH
CR	4	<b>Fraunhofer Institute for Autonomous Intelligent Systems</b> , Sankt Augustin. <a href="http://www.ais.fraunhofer.de/">http://www.ais.fraunhofer.de/</a>	<b>FAIS</b>	D
CR	5	<b>Wageningen UR</b> , Centre for GeoInformation. <a href="http://cgi.girs.wageningen-ur.nl/">http://cgi.girs.wageningen-ur.nl/</a>	<b>WUR</b>	NL
CR	6	<b>Research Academic Computer Technology Institute</b> , Research and Development Division. <a href="http://www.cti.gr/">http://www.cti.gr/</a> and Univ. Piraeus, Dept. of Informatics <a href="http://www.unipi.gr">http://www.unipi.gr</a>	<b>CTI</b>	GR
CR	7	<b>Sabanci University</b> , Faculty of Engineering and Natural Sciences. <a href="http://www.sabanciuniv.edu/">http://www.sabanciuniv.edu/</a>	<b>UNISAB</b>	TK
CR	8	<b>WIND Telecomunicazioni SpA</b> , Direzione Reti Wind Progetti Finanziati & Technology Scouting. <a href="http://www.wind.it">http://www.wind.it</a>	<b>WIND</b>	I

## Table of contents

SUMMARY .....	3
1. INTRODUCTION .....	4
2. TYPES OF PATTERNS IN MOVEMENT DATA.....	5
2.1 Abstract model of movement data .....	5
2.2 Behaviour and pattern .....	5
2.3 Pattern types and pattern instances .....	7
2.4 Connection patterns.....	8
2.5 Descriptive pattern types for movement data.....	8
2.6 Connectional patterns in movement data .....	10
2.7 Pattern properties .....	10
References .....	11
3. REQUIREMENTS AND CHALLENGES .....	12
3.1 Achieve a synergy of human and computer .....	12
3.2 Scalability issues.....	13
3.3 Privacy issues.....	14
3.4 Levels of analysis.....	14
3.5 Multiple complementary views.....	15
3.6 Linking exploration with validation.....	15
3.7 Support of knowledge capture and manipulation.....	15
References .....	16
4. VISUAL ANALYTICS METHODS FOR PATTERN DETECTION .....	17
4.1 Data manipulation .....	17
4.1.1 Aggregation .....	17
4.1.2 Other data transformation techniques .....	19
4.2 Exploring the behaviour of the IMBs over the set of entities .....	21
4.2.1 Use of clustering.....	21
4.2.2 Visualisation of clustering results.....	22
4.3 Exploring the behaviour of the MCB over time.....	26
4.4 Looking for connectional patterns .....	32
4.5 Summary of techniques for supporting pattern detection .....	35
4.6 Visualisation of patterns.....	37
4.7 Conclusion .....	38
References .....	38
5. SPATIO-TEMPORAL REASONING.....	40
5.1 The role of metaphors in reasoning to discover patterns .....	40
5.2 The multi-tier ontological framework .....	43
5.2.1 Tier 0 – The Reality Space .....	44
5.2.2 Tier 1 – The Positioning Space.....	45

---

5.2.3	Tier 2 – The Geographic Space .....	46
5.2.4	Tier 3 – The Social Space .....	47
5.2.5	Tier 4 – The Cognitive Space .....	49
5.3	Conclusions .....	49
	References .....	51
6.	CONCLUSION .....	53

## SUMMARY

This document represents the current state of research on supporting human analysts in the process of exploration, analysis, and interpretation of large volumes of movement data, which is conducted in workpackage WP3 of the project GeoPKDD. The goals of analysing movement data referring to multiple entities may be formulated as *describe and compare dynamic collective behaviours and relate them to properties of space, properties of time, properties and activities of the moving entities, and relevant external phenomena*. Consequently, the goal of WP3 is to define a set of interactive instruments that would allow a *human analyst* to achieve these goals. We emphasise the active role of human analyst since WP3 has a user-centred character.

The major research goal in WP3 corresponds to the definition of a new research discipline referred to as Visual Analytics. Visual Analytics is defined as *the science of analytical reasoning facilitated by interactive visual interfaces*. This is a multidisciplinary field that focuses on analytical reasoning techniques, visual representation and interaction techniques, data representations and transformations, and techniques to support production, presentation, and dissemination of the results of analysis. All this is aimed at achieving a truly synergetic work of human and computer.

Consequently, much attention in WP3 is given to visual information displays and user interaction with these displays and the underlying information. However, since exploration and analysis of large datasets cannot be done using purely visual means, visualisation is combined with database technologies, computerised data processing, and computational analysis methods. We undertake a systematic approach to defining what methods and techniques and what ways for linking them can appropriately support visual analysis of massive collections of movement data. The main idea is that software tools prepare and visualise the data so that the human analyst can detect various types of patterns by looking at the visual displays. In order to facilitate the detection of patterns, it is necessary to understand what types of patterns may exist in the data. We define the possible types of patterns in such a kind of movement data on the basis of an abstract model of the data as a mathematical function that maps entities and times onto spatial positions. Then, we look for data transformations, computations, and visualisation techniques that can facilitate the detection of these types of patterns and are suitable for very large datasets, possibly, not fitting in the computer memory. Under such a constraint, visualisation is applied to data that have been previously aggregated and generalised by means of database operations and/or computational techniques. Another reason for aggregation and generalisation is preserving privacy.

Besides detection and description of patterns, an important part of analysis is analytical reasoning, when the analyst establishes essential links between various phenomena and between different aspects of the same phenomenon. Different types of inferences play a different role accordingly to what a domain expert want to infer. We present three modes of reasoning using a multi-tier ontological framework. They are deductive, inductive, and abductive modes of reasoning. We emphasise the role of metaphors, which help the comprehension of what makes one pattern structurally and meaningfully different from another.

In the deductive mode of reasoning, the geographic knowledge discovery process involves the search for common attributes among a set of mobile trajectories, and then the arrangement of these trajectories into classes, clusters, or patterns according to a meaningful metaphor. The focus is on applying statistical approaches. In the inductive mode of reasoning, the geographic knowledge discovery process is based on learning as the reduction of uncertainty in knowledge. Several techniques have been developed, such as rule induction, neural networks, genetic algorithms, case-based learning and analytical learning (theorem proving). In the abductive mode of reasoning, the importance of cognitive tacit knowledge needs to be considered. The inevitable challenge facing the research community at the moment is directed toward a more complete integration of these modes of reasoning and their association to movement metaphors within a geographic knowledge discovery process.

## 1. INTRODUCTION

The strategic goal of WP3 is to support human analysts in the process of exploration, analysis, and interpretation of large volumes of movement data. When data are massive, it is insufficient to use only visual displays but it is necessary to involve the database technologies and computational methods of data processing and analysis. Still, the visualisation plays the central role since it allows the innate perceptual and cognitive capabilities and background knowledge of a human analyst to be utilised in the process of data exploration and analysis. These capabilities and knowledge cannot be replaced by purely machine processing. Hence, the combination of the visualisation with computer operations makes a ground for a truly synergetic work of human and computer.

The goals of analysing movement data referring to multiple entities may be formulated as *describe and compare dynamic collective behaviours and relate them to properties of space, properties of time, properties and activities of the moving entities, and relevant external phenomena*. To describe the behaviour means to represent it by an appropriate *pattern*. In deliverable D3.1, we have defined the types of patterns that could be detected in movement data and between movement data and data about other phenomena. However, to make this document self-contained, we repeat in Section 2 the definition of the patterns. We also repeat in Section 3 the analysis of requirements and challenges to methods and tools intended to support analysis of movement data.

In Section 4, we undertake a systematic approach to defining what methods and techniques and what ways for linking them can appropriately support visual analysis of massive collections of movement data. The main idea is that software tools prepare and visualise the data so that the human analyst can detect various types of patterns by looking at the visual displays. The tools must be suitable for very large datasets, possibly, not fitting in the computer memory. Under such a constraint, visualisation is applied to data that have been previously aggregated and generalised by means of database operations and/or computational techniques. Another reason for aggregation and generalisation is preserving privacy.

Besides proposing tools for supporting pattern detection, we consider the problem of pattern visualisation, which is currently far from being solved and requires further research. We outline some approaches, which we deem to be promising.

Section 5 is dedicated to analytical reasoning in the process of analysing movement data. This section introduces a geographic knowledge discovery process, in which the primary goal of identifying, associating, and understanding patterns is used to infer the location, identity, and relationships among mobile entities, and their respective trajectories in a spatial environment. In this case, the different types of inferences play a different role accordingly to what a domain expert want to infer, that is, the location, changes, properties, identity, or relationship among the appropriate metaphors. We emphasise the role of metaphors, which help the comprehension of what makes one pattern structurally and meaningfully different from another. It is the metaphor, and only after it makes sense that an unknown set of patterns can be interpreted and understood by a domain expert. Basically, three modes of reasoning are presented using a multi-tier ontological framework. They are deductive, inductive, and abductive modes of reasoning.

The inevitable challenge facing the research community at the moment is directed toward a more complete integration of these modes of reasoning and their association to movement metaphors within a geographic knowledge discovery process.

## 2. TYPES OF PATTERNS IN MOVEMENT DATA

The types of patterns that could be detected in movement data and between movement data and data about other phenomena have been defined in deliverable D3.1. We repeat this definition here in order to make this document self-contained. The definition of the pattern types is based on an abstract model of movement data.

### 2.1 Abstract model of movement data

On an abstract level, movement data can be viewed as consisting of the three principal components:

- time: a set of moments;
- population (this term is used in statistical rather than demographic sense): a set of entities that move;
- space: a set of locations that can be occupied by the entities.

A trajectory may be viewed as a function mapping time moments onto positions in space. Analogously, movement of multiple entities may be seen as a function mapping pairs <time moment, entity> onto positions. This is a very abstract data model, which is independent of any representation formalism (of course, there may be other models; for example, a database-oriented view would consider the same data as a table of tuples with at least three attributes Entity, Time, and Space). The time and population of entities play the role of “independent variables”, or *referential components*, and the space plays the role of “dependent variable”, or *characteristic component*.

A combination of values of the referential components is called a *reference*. In our case, a reference is a pair consisting of a time moment and an entity. The set of all possible references is called the *reference set*. Values of the characteristic components corresponding to the references are called *characteristics* of these references.

The state of a moving entity at a selected time moment can be characterised not only by its position in space but also by additional characteristics such as speed, direction, acceleration, etc. These characteristics can be viewed as secondary since they can be derived from the values of the principal components. Nevertheless, we can extend our concept of movement data and see it as a function mapping references <time moment, entity> onto combinations of characteristics (position, speed, direction, ...).

Locations, time moments, and entities may have their own characteristics. For example, locations may be characterised by altitude, slope, character of the surface, etc.; entities may be characterised by their kind (people, vehicles, animals, ...), age, gender, activity, and so on. Such characteristics are independent of the movement, that is, do not refer to pairs <time moment, entity> but to individual values of the three principal components, time, population, and space. Note that the space plays the role of a referential component for altitude, slope, and so on. The characteristics of time moments, entities, and locations will be further called *supplementary characteristics*. The characteristics of the pairs <time moment, entity> (including the secondary ones) will be called *characteristics of movement*.

### 2.2 Behaviour and pattern

We introduce the notion of *behaviour*: this is the configuration of characteristics corresponding to a given reference (sub)set. The notion of behaviour is a generalisation of such notions as distribution, variation, trend, dynamics, trajectory, etc. In particular, a trajectory of a single entity is a configuration of locations (possibly, in combination with the secondary characteristics of movement) corresponding to a time interval. We say “configuration” rather than “set” meaning that the characteristics are arranged in accordance with the structure and properties of the reference

(sub)set and the relations between its elements. Thus, since a time interval is a continuous linearly ordered set, a trajectory is a continuous sequence of locations ordered according to the times they were visited.

The term “behaviour” is used here in a quite general sense and does not necessarily mean a process going on in time. Thus, the spatial distribution of a set of entities at some time moment is also a kind of behaviour although it does not involve any temporal variation.

Since a population of entities is a discrete set without natural ordering and distances between the elements, it does not impose any specific arrangement of the corresponding characteristics. Still, the corresponding behaviour is not just a set of characteristics. Thus, one and the same characteristic or combination of characteristics can occur several times, and these occurrences are treated as different while in a set each element may occur only once. A behaviour over a set of entities may be hence conceptualised as the frequency distribution of the characteristic values over this set of entities.

The absence of natural ordering and distances on a population of entities does not mean that ordering and distances between entities cannot exist at all. Thus, a set of participants of a military parade is spatially ordered and has distances between the elements. However, the ordering and distances are defined in this case on the basis of certain characteristics of the entities, specifically, their spatial positions. The characteristics that define ordering and/or distances between entities can be chosen, in principle, quite arbitrarily. Thus, participants of a parade can also be ordered according to their heights, or weights, or ages. In data analysis, it may be useful to consider different orderings of the entities and the corresponding arrangements of characteristics. In such cases, the behaviours are not just frequency distributions but more complex constructs where characteristic values are positioned according to the ordering and/or distances between the entities they are associated with.

The collective movement behaviour of a population of entities over a time period is a complex configuration built from movement characteristics of all entities at all time moments, which has no arrangement with respect to the population of entities and has a continuous linear arrangement with respect to the time.

The goals of analysing movement data referring to multiple entities may be formulated as *describe and compare dynamic collective behaviours and relate them to properties of space, properties of time, properties and activities of the moving entities, and relevant external phenomena*. To describe the behaviour means to represent it by an appropriate *pattern*. A pattern may be defined as a representation of a behaviour in some language, e.g. natural, mathematical, graphical, etc. This agrees with the definition of a pattern in the data mining literature: “a pattern is an expression  $E$  in some language  $L$  describing facts in a subset  $F_E$  of a set of facts  $F$  so that  $E$  is simpler than the enumeration of all facts in  $F_E$ ” (Fayyad et al. 1996). Note that the latter definition emphasises the synoptic nature of a pattern: a pattern does not simply enumerate some facts but describes them all together as a whole. We extend the notion of pattern to all kinds of representation, including schematic drawings and mental constructs built in analyst’s mind.

As should be clear from the definition, different patterns (e.g. focusing on different aspects) may represent one and the same behaviour. A pattern may be compound, i.e. composed of other patterns. For example, the description “most of the people tend to move towards the city centre in the morning and outwards in the evening” is a compound pattern including two simpler patterns, inward and outward movement. Patterns representing movement behaviours of individual entities (i.e. trajectories) and collective movement behaviours of sets of entities base first of all on the characteristics of movement but may also involve supplementary characteristics. Thus, our example pattern concerning the movement of people describes first of all the direction of the movement but also mentions such supplementary characteristics as the character of the moving entities (people), the character of a location (city centre), and the character of the times (morning or evening).

In a pattern describing the movement behaviour on a set of references, one may include various summary values derived from the individual characteristics of the references, for instance, the average speed, prevailing direction, or frequency of turns.

### 2.3 Pattern types and pattern instances

A pattern may be viewed as a statement in some language; however, the language may be chosen quite arbitrarily (e.g. natural language, mathematical formulas, graphical language); hence, the syntactic and morphological features of a pattern are irrelevant to data analysis. What is relevant is the meaning, or semantics. It is natural to assume that representations of the same behaviour in different languages have a common meaning. Hence, the constructs of the different languages refer to the same system of basic language-independent elements from which various meanings can be composed. By analogy with meanings of words in a natural language, we can posit that the basic semantic elements for building various patterns include *pattern types* and *pattern properties*. A specific pattern is an *instantiation* of one or more pattern types. This is analogous to the specialisation of a general notion by means of appropriate qualifiers. In the case of patterns, the qualifiers are specific values of the pattern properties. For example, the pattern ‘entities  $e_1, e_2, \dots, e_n$  moved together during the time period  $T$ ’ instantiates the pattern type “joint movement” by specifying what entities and when moved in this manner.

It is quite reasonable to assume that the possible pattern types exist in the mind of a data analyst as mental schemata. Moreover, these schemata are likely to drive the process of visual data analysis, which is generally believed to be based on pattern recognition: the analyst looks for constructs that may be viewed as instantiations of the known pattern types. Therefore, for the design of proper analysis methods for movement data, it is important to define the pattern types relevant to such data.

On a very general level, pattern types are introduced in the book by Andrienko & Andrienko (2006). Descriptive patterns, which characterise behaviours, are distinguished from connectional patterns, which characterise relations between phenomena. The basic types of descriptive patterns are similarity, difference, and arrangement, where the latter type embraces such concepts as trend, sequence, periodicity, symmetry, etc. From instances of the basic pattern types, compound patterns are built as is shown graphically in Fig.2.1.

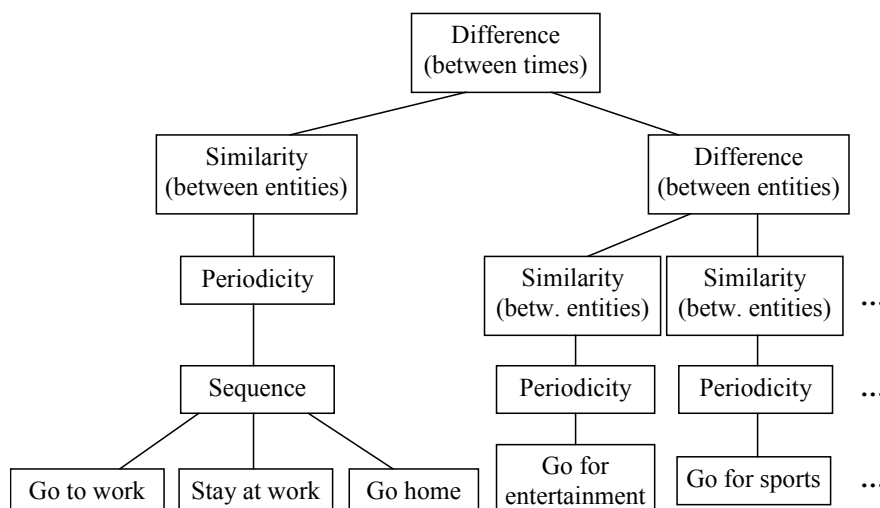


Fig. 2.1. An illustration of a compound pattern with several levels of nested sub-patterns.

In defining the pattern types relevant to movement data, we focus on collective movement behaviours of multiple entities. Moreover, our aim in GeoPKDD is to design such analysis techniques that will preclude the access of the user to any data about individual entities for privacy

reasons. Hence, the techniques must allow the user to see only secondary data resulting from aggregation or generalisation of the individual data. Besides the privacy, another reason for this is the potentially large data size.

## 2.4 Connection patterns

When studying a phenomenon, an analyst is interested not only in describing or summarising its behaviour but also in explaining it. The analyst wishes to find out the reasons or driving forces that make the phenomenon behave in the way observed. These forces may be internal or external. Internal forces originate from the inherent structure of the phenomenon and interactions between its structural components. External forces originate from interactions between the phenomenon and other phenomena. Hence, the goal is to determine what components and/or phenomena interact and how they interact. Thus, concerning the movement of entities, an analyst may be interested to know whether and how the movement is related to various spatial, temporal, and spatio-temporal phenomena such as weather, events (e.g. traffic jams or accidents), opening hours of shops, activities of people, etc. The analyst may also wish to detect interactions between parts of the overall movement behaviour, e.g. between the behaviours of traffic and of pedestrians, or between properties of movement, e.g. direction and speed. A result of such activity is a description in some language of the connection that has been discovered. We call such a description a *connection pattern* while a connection, or interaction, may be viewed as a ‘mutual behaviour’ of two or more phenomena or parts of the same phenomenon.

In data analysis, the following types of connections are typically looked for:

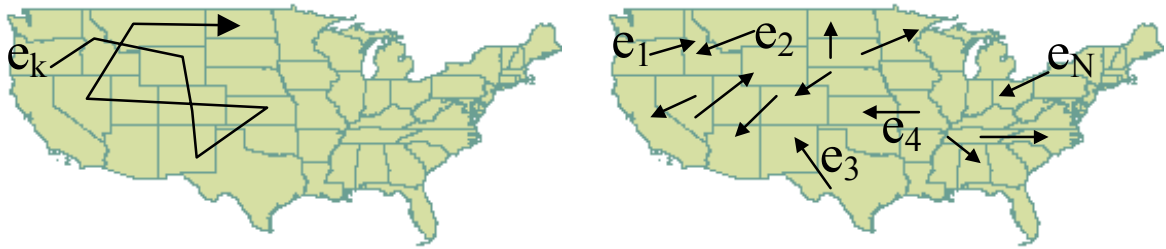
- *Correlation*: an undirected, or symmetrical, connection. This includes not only the statistical correlation between two numeric variables but also all cases of regular co-occurrence of characteristics or behaviours, possibly, with a temporal and/or spatial lag. For example, working in the centre of a city may correlate with using the public transport or a bike for getting to the workplace.
- *Dependency, or influence*: a directed connection; for example, the use of a car or a bike for getting to the workplace depends on the weather (or, in other words, the weather influences whether a car or a bike is used).
- *Structural connection*: an observed movement behaviour results from a composition of two or more different movements performed simultaneously, like the observed movement of the planets is the result of a combination of their own movement and the movement of the Earth.

## 2.5 Descriptive pattern types for movement data

Our ultimate goal is to define pattern types for collective movement behaviours of multiple entities. In order to achieve this, it is necessary to consider the following “slices”, or “projections”, of this overall behaviour:

- individual movement behaviour, i.e. movement of a single entity over time, and
- distribution of movement characteristics (position, speed, direction, etc.) over the set of entities at a single time moment. For the sake of brevity, we shall call it “momentary collective behaviour”.

Fig.2.2 gives a visual illustration of these two notions.



**Fig. 2.2.** An illustration of individual movement behaviour (left) and momentary collective behaviour (right).

We shall use the following abbreviations: **IMB** for individual movement behaviour, **MCB** for momentary collective behaviour, and **DCB** for dynamic collective behaviour, i.e. behaviour of multiple entities during a time interval. A DCB can be viewed from two different perspectives:

- As a construct formed from the IMBs of all entities, i.e. the behaviour (variation) of the IMB over the set of entities;
- As a construct formed from the MCBs at all time moments, i.e. the behaviour (variation) of the MCB over time.

These two views are called aspectual behaviours. Aspectual behaviours exist in data having two or more referential components, or independent variables. Movement data have two referential components, entity and time, which yield two aspectual behaviours. The aspectual behaviours are essentially different and described in terms of different types of patterns.

The behaviour of the IMB over the set of entities can be described by means of similarity and difference patterns, i.e. as groups of entities having similar IMBs, which differ from the IMBs in other groups of entities. It may happen that some entities have quite peculiar IMBs, which differ from the IMBs of all other entities. Such peculiar IMBs are also described by means of difference patterns. Arrangement patterns are usually not relevant to the behaviour of the IMB over the set of entities because the set of entities has no natural ordering and no distances between the elements (Andrienko & Andrienko 2006).

What does it mean that IMBs of several entities are similar? There are diverse possible meanings, and all of them may be relevant to analysis of movement data:

- Similarity of the overall characteristics: geometric shapes of the trajectories, travelled distances, durations, movement vectors, etc.
- Co-location in space, i.e. the trajectories of the entities consist of the same positions or have some positions in common:
  - ordered co-location: the common positions are attained in the same order;
  - order-irrelevant co-location: the common positions may be attained in different orders;
  - symmetry: the common positions are attained in the opposite orders.
- Synchronisation in time:
  - full synchronisation: similar changes of movement characteristics occur at the same times;
  - lagged synchronisation: changes of the movement characteristics of entity  $e_1$  are similar to changes of the movement characteristics of entity  $e_0$  but occur after a time delay  $\Delta t$ .
- Co-incidence in space and time:
  - full co-incidence: same positions are attained at the same times;
  - lagged co-incidence: entity  $e_1$  attains the same positions as entity  $e_0$  but after a time delay  $\Delta t$ .

Let us now consider the other aspectual behaviour, that is, the behaviour of the MCB over time. Mathematically, time is a continuous set where ordering and distances exist between the elements, i.e. time moments. Hence, besides similarity and difference patterns, arrangement patterns are relevant. An arrangement pattern describes changes in the MCB with respect to the ordering and

distances between the corresponding time moments. Here are the pattern types for describing the behaviour of the MCB over time (we note in parentheses the basic pattern types that have been specialised):

- Constancy (similarity): the MCB was the same or changed insignificantly during a time interval.
- Change (difference): the MCB significantly changed from moment  $t_1$  to moment  $t_2$ .
- Trend (arrangement): consistent changes of the MCB during a time interval.
- Fluctuation (arrangement): irregular changes of the MCB during an interval.
- Pattern change or pattern difference (difference): the behaviour of the MCB during time interval  $T_1$  differs from that during time interval  $T_2$ . The term “pattern change” applies when  $T_1$  and  $T_2$  are adjacent. For example, a trend can change for constancy or for a different trend. The term “pattern difference” applies to non-adjacent time intervals.
- Sequence (arrangement): patterns follow one another in a specific order.
- Repetition (similarity): occurrences of the same patterns or pattern sequences on different time intervals.
- Periodicity, or regular repetition (similarity and arrangement): occurrences of the same patterns or pattern sequences on regularly spaced time intervals.
- Symmetry (similarity and arrangement): opposite trends; pattern sequences where the same patterns are arranged in opposite orders.

## 2.6 Connectional patterns in movement data

The types of *correlation* and *influence* patterns are similar since the relation of influence differs from the relation of correlation only in its being directed: from two related things, it is specified which one influences the other. Correlations or influences may exist

- between different movement characteristics, e.g. direction and speed;
- between movement characteristics and supplementary characteristics (see Chapter 1), which include characteristics of entities, characteristics of time moments, and characteristics of spatial locations;
- between individual or collective movement behaviours on different time intervals, e.g. after slow movement in a traffic jam, drivers tend to move faster than usual;
- between collective movement behaviours of different subsets of entities, e.g. different teams in a football game;
- between individual or collective movement behaviours and supplementary characteristics;
- between individual or collective movement behaviours and behaviours of external phenomena like weather, various types of events, etc.

What concerns *structure* patterns, these may be compositions of movement behaviours with regard to different temporal cycles like daily, weekly, and annual. For example, working people go to and from their work every day except weekends and go shopping on Saturdays. On Sundays, they usually stay at home in winter time and go to the countryside in summer time.

## 2.7 Pattern properties

When a user detects, for example, a pattern of similarity of IMBs of multiple entities, he/she is interested to know how many entities have this common behavioural pattern. Likewise, when the user detects a pattern of synchronisation or a trend, he/she measures the duration of the time interval when the entities moved synchronously or the trend lasted. These properties of the patterns taken as examples may be generalised as *support base*, that is, the size of the reference set on which this pattern takes place. Hence, for a pattern describing movements of multiple entities, the support base is the size of the subset of entities (i.e. the number of entities), and for a pattern

describing movement on a time interval, the support base is the size (length) of the interval. Logically, the support base of a pattern describing movements of multiple entities during a time interval includes both the number of entities and the length of the interval. Besides the absolute support base, an important property is the *relative support base*, i.e. the size of the reference subset where the behaviour corresponds to the pattern in relation to the size of the whole reference set.

Not only is the length of the time interval of a pattern interesting for a user but also the *temporal localisation*, i.e. the position of the interval on the time scale. Likewise, it may be interesting to know which particular entities behave according to a pattern, in addition to the number of such entities. However, this pattern property may not always be accessible either because of a very large number of entities or because of privacy constraints.

Besides these general properties, there are also more specialised properties, which are relevant either to particular types of patterns or to characteristics involved in pattern definition. We shall not try to list exhaustively all these specialised properties but instead give a few examples. Thus, a change pattern may be characterised in terms of the magnitude and direction of the change while a periodic pattern is characterised by the length of the period between the repetitions. An important property of a similarity pattern describing movement of multiple entities along the same route (i.e. visiting the same locations in space) is the *spatial localisation*, i.e. where in space this common behaviour takes place. The properties of a pattern describing some behaviour in terms of the speed of movement include summarised speed characteristics such as average and maximum speed.

## References

1. Andrienko, N. and Andrienko, G. (2006): *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*, Springer, Berlin
2. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. (1996): From data mining to knowledge discovery: an overview. In: *Advances in Knowledge Discovery and Data Mining*, ed. by Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (AAAI Press/MIT Press, Menlo Park 1996) pp. 1–36

### 3. REQUIREMENTS AND CHALLENGES

The definition of the task and pattern types allowed us to formulate requirements to interactive tools intended to support exploration and analysis of movement data and analytical reasoning about patterns discovered in movement data and essential links between characteristics of the movement and various phenomena and events. We have also identified the major challenges that need to be met in the further research.

The goals of analysing movement data referring to multiple entities may be formulated as *describe and compare dynamic collective behaviours and relate them to properties of space, properties of time, properties and activities of the moving entities, and relevant external phenomena*. Consequently, the goal of WP3 is to define a set of interactive instruments that would allow a *human analyst* to achieve these goals. We emphasise the active role of human analyst since WP3 is not intended to design fully automated analysis methods but has a user-centred character.

#### 3.1 Achieve a synergy of human and computer

Visualisation is very important for supporting data analysis by humans. According to a dictionary, one of the meanings of the word “visualise” is “to make perceptible to the mind or imagination”. This definition provides a full justification for the primary importance of visualisation as a tool for data analysis: in order to be able to think about data, the human mind needs to perceive the data. No thinking is possible without prior perception. And it is clear that the perception must be, on the one hand, correct with respect to the data, and on the other hand, opportune for reasoning. This imposes very high requirements upon data visualisation tools.

However, visualisations alone may be insufficient when massive data collections need to be explored and analysed. This is not only the matter of technical limitations such as the screen size and resolution or the speed of rendering but also of the natural perceptual and cognitive limitations of the humans who need to view and interpret the visual displays. Hence, there is a need in combining visualisation with computational analysis methods, database queries, data transformations, and other computer-based operations. The goal is to create visual analytics environments for synergetic work of humans and computers on solving complex problems where the computational power amplifies the human capabilities such as pattern recognition, imagination, association, and analytical reasoning and is, in turn, directed by human’s background knowledge and insights gained. This goal entirely corresponds to the definition of visual analytics (Thomas and Cook 2005), which is defined as *the science of analytical reasoning facilitated by interactive visual interfaces*. This is a multidisciplinary field that focuses on analytical reasoning techniques, visual representation and interaction techniques, data representations and transformations, and techniques to support production, presentation, and dissemination of the results of analysis. All this is aimed at achieving a truly synergetic work of human and computer.

Movement data present an appropriate target for such a multidisciplinary approach. On the one hand, purely computational methods of analysis are insufficient for dealing with the spatial component of the data. These methods operate with discrete numbers and symbols, which cannot adequately represent a continuous two- or three-dimensional geographical space with its heterogeneity and the multitude of spatial relations. An adequate representation is a map or a three-dimensional display, which needs to be examined by a human analyst. On the other hand, movement data are often quite voluminous, characterising movements of many entities and/or referring to many time moments. Even two trajectories represented on a map or in a space-time cube may be difficult to analyse if they have common locations or segments (or even a single long trajectory with loops or repeated segments), and a display of ten trajectories may be completely illegible. Computational techniques may compensate for the limitations of the visual analysis methods.

### 3.2 Scalability issues

Irrespective of the dataset size, the visualisation and analysis of movement data are quite difficult because of the quite complex data structure, which involves time, space, multiple entities, and multiple movement characteristics. Is there any way to put all this information on a display so that the representation is comprehensible to a human viewer? The representation of two-dimensional space requires two display dimensions, and the representation of time requires one more dimension. This may be the third spatial dimension, as in a space-time cube, or the temporal dimension in an animated display.

From the representation of individual trajectories by lines in an interactive 3D display it is possible to estimate the positions, speeds, directions, and other movement characteristics at different times. Similarities and differences between individual movement behaviours (IMBs) are well noticeable. The use of a movable plane, as suggested by Kraak (2003), helps in exploring the momentary collective behaviours (MCBs) at different moments and the behaviour of the MCB over time. However, all these benefits fade away with increasing the number of moving entities, the length of the time period, and/or the geometric complexity of the trajectories. The data do not need to be really massive: a space-time cube with ten trajectories will already look like a bowl of spaghetti, from which one can hardly extract any useful information. Analogously, an animated display of individual movements is quite appropriate when the entities are few and the time period is not very long but decreases in utility with increasing the number of moving entities and/or different time moments in the data. The upper limit may be higher for an animated display than for a space-time cube: in an animated display, the information is presented in portions, which makes the display at each moment simpler and easier to perceive than a space-time cube, which portrays all information at once. However, this slight increase of the applicability limit does not solve the problem in general. Besides, the portion-wise representation of information has clear disadvantages: no overview of the whole dataset is possible as well as no comparison between states at different moments.

Hence, the limitations of visual displays concerning the data size come much earlier than the size becomes too big for the computer memory. Therefore, some methods for reducing the data size have to be applied prior to the visualisation. Possible approaches include aggregation, filtering, and clustering. When the data are too big for the human perception but not yet too big for the computer, high interactivity may compensate for the indispensable information losses resulting from the data reduction. Suppose, for example, that the visualisation shown in Fig.3.4 is an interactive display on the computer screen that allows the user to click on the lines in order to select the corresponding entities. In response, the movements of the selected entities are shown in the display by lines of a different colour. Through further interaction with the display, the user may modify the selection and immediately receive visual feedback. Moreover, several displays of aggregated data providing complementary views of the dataset may be linked by means of brushing, similarly to the linked histograms in Attribute Explorer (Spence and Tweedy 1998, Spence 2001). For example, the display with tiered maps as in Fig.3.4 may be linked to a bar chart showing the numbers of tourists coming from different countries. When the user selects a subset of tourists through the tiered map display, the bar chart shows how many of these tourists come from each country by special colouring of the corresponding bar segments. The user may also select the tourists coming from a particular country by clicking on the respective bar chart. In response, the tiered map display will show the flows of these tourists.

The things become much more complicated when the original data cannot be kept and processed in the computer memory. This means that aggregation, filtering, clustering, selection, and brushing can only be done with the involvement of database operations, which may take much time. Hence, the visual displays can no more be interactive in the same way as in the case of smaller datasets. It is necessary to devise new methods of interaction that could perform reasonably well when datasets are huge.

Because of the challenges arising from large data volumes, Daniel Keim argues that Ben Shneiderman’s Information Seeking Mantra “Overview first, zoom and filter, and then details-on-demand” (Shneiderman 1996) should be replaced by the Visual Analytics Mantra: “Analyse First - Show the Important - Zoom, Filter and Analyse Further - Details on Demand” (Keim 2005). The Visual Analytics Mantra stresses the fact that fully visual and interactive methods do not work with big datasets. It is necessary to start with database operations and computations (“Analyse<sup>1</sup> First”) and apply visualisation to the results obtained (“Show the Important”). The user may interact with the visualisation and the secondary data it represents (i.e. the outcomes of the analysis but not the original data), in particular, zoom and filter, and trigger further analysis, which, again, requires visualisation of the results. In this way, visual analytics is an iterative process involving three major steps, computational analysis, visualisation of the results of the computational analysis, and interactive visual analysis of these results. A detailed consideration (“Details on Demand”) is possible for small data portions when they require, for some reason, a special attention of the analyst. This does not necessarily happen at the end of the process.

Hence, visual analytics tools for movement data need to be designed in accord with the Visual Analytics Mantra, when database technologies and computational analysis are applied prior to visualisation and iteratively re-applied during the process of data analysis.

### 3.3 Privacy issues

In designing methods and tools for helping users to recognise various patterns, we must comply with a crucial constraint: detecting patterns must be done without seeing any information about individual entities, for preserving their privacy. Hence, we must design such techniques and tools for analysis that will preclude the access of the user to any data about individual entities. This means that the tools must present to user’s view only secondary data resulting from aggregation or generalisation of the individual data. Besides the privacy, another reason for this is the potentially large data size, as discussed above.

### 3.4 Levels of analysis

The role of data aggregation is not limited to hiding individual information and reducing the amount of data. While information reduction means substantial information loss, there is also a positive side, specifically, the possibility to omit “high-detail noise” and focus on characteristic features of the phenomenon under study. We may say that aggregation and generalisation helps us to see forest for trees.

The degree of data aggregation and generalisation matters a lot in data analysis. This is not only the matter of the size of the resulting data and the amount of information lost. This is also the matter of the scale on which the data are considered. Depending on the scale, the user sees the data differently and detects different patterns. Thus, in movement data, there may be local patterns like a flock (synchronous movement of several entities having close positions and same speed), or there may be larger scale patterns like massive movement towards industrial or commercial areas in mornings, or, on a yet larger scale, the difference of collective movement patterns on weekdays and weekends, and so on.

Hence, the appropriate degree of data aggregation and generalisation is not just a good trade-off between the simplification gained and the amount of information lost. The aggregation and generalisation must be adequate to analysis goals, i.e. the patterns of what scale the user is interested to detect. If the interests of the user include patterns of different scales, it is necessary to consider the data on different levels of aggregation. The tools for interactive analysis must thus enable the user to do this.

---

<sup>1</sup> D.Keim uses the term “analyse” in a wide sense including not only purely analytical procedures but also data aggregation and other ways of data processing.

### 3.5 Multiple complementary views

When data have complex structure, there is no way to represent all the information in a single display. Any visualisation technique suitable for movement data shows only a certain aspect of the data. Therefore, different techniques need to be combined for a comprehensive visual exploration of the data. Moreover, these different techniques need to be used in parallel; otherwise, the user will not be able to relate, for example, the distribution of the entities in space at a particular moment to changes of their positions and other movement characteristics. The user should be provided interactive facilities to support establishing connections between different views, in particular, finding elements in different displays corresponding to same spatial positions, same times, same groups of entities, and/or same values of movement characteristics.

Brushing has been mentioned several times as a technique that supports establishing links between displays. Besides brushing, there are other methods of display linking such as propagation of a division of the set of entities into classes (each class is assigned its specific colour, which is consistently used in all displays) and simultaneous reaction of all displays to interactive filtering of the data; a review may be found in Andrienko & Andrienko (2006). It should be noted that most of the currently existing techniques for coordination of multiple displays involve dealing with individual data items and are therefore not scalable to very large datasets. There is a need in new technical solutions, which could properly work in a situation when all displays show only aggregated data (moreover, differently aggregated data!) while the original individual data are not present in the computer memory.

### 3.6 Linking exploration with validation

Traditionally, geovisualisation and information visualisation focus on developing methods and tools to support discovery of patterns and relationships in data. However, it is known that visual displays are not always productive but may be misleading. Therefore, it is necessary to find ways for linking exploratory visualisation with validation of patterns and relationships detected in data by means of visualisation. How to ensure that the analyst is not misled by graphical presentation and does not jump to conclusions? How to support and, possibly, even prompt immediate testing of what appears as a pattern in data?

### 3.7 Support of knowledge capture and manipulation

What an analyst gets from viewing visual information displays is impressions, images, and ideas that appear in his or her mind. How can such impressions and ideas be put in a form suitable for later reviewing, for communicating to others, and for use in the further analysis and in the following phases of the decision process? In particular, how can spatial and spatio-temporal patterns and relations discovered in data be represented in an explicit, preferably visual, form?

The need to visualize patterns and relationships discovered in data arises not only when visual analysis tools are used but also in case of automated search, or “mining”, for patterns, which can be used complementarily to interactive methods. The use of automatic methods requires the results to be presented to the user in a way allowing the user to interpret and evaluate them. In other words, the patterns need to be made perceptible to human mind, that is, visualised.

However, it is not sufficient to capture and visualise patterns. According to a dictionary, “to analyse” means “to separate (a material or abstract entity) into constituent parts or elements”. Because of data size and complexity, the analyst has to look at different aspects of the dynamic collective behaviour of the moving entities, to decompose it into slices, to divide the data into subsets, and to view the data on multiple levels of aggregation and abstraction. From the examination of each aspect, slice, subset, or view, the analyst gains some bit of knowledge, which is expected to bring him/her closer to obtaining overall knowledge of the dynamic collective behaviour and its links to other phenomena. However, the overall knowledge is not a mere arithmetic sum of all the bits and pieces obtained by means of the analysis, as a three-dimensional

shape is not a sum of its two-dimensional projections. The overall knowledge is obtained by means of integrative, synthetic actions which involve not only building a structure where each bit has its proper place but also generalisation, abstraction, induction, and deduction.

Hence, it is necessary to support human analysts both in analytic and in synthetic activities. To our best knowledge, all currently existing systems and toolkits for data exploration and analysis address only the analysis side, and none of them can support, at least to a small extent, knowledge synthesis. Moreover, there is no clear understanding in the research community what kind of support is needed and how it could be provided. This remains a challenging research problem.

## References

1. Andrienko, N. and Andrienko, G. (2006): *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach* (Berlin: Springer)
2. Keim, D. A.: 2005, *Scaling visual analytics to very large data sets*, presentation at the Workshop on Visual Analytics, June 4th, 2005, Darmstadt, Germany, <http://infovvis.uni-konstanz.de/index.php?region=events&event=VisAnalyticsWs05>
3. Kraak, M.-J. (2003): *The space-time cube revisited from a geovisualization perspective*, in: Proc. of the 21st International Cartographic Conference, Durban, South-Africa, 10-16 August 2003, pp. 1988-1995
4. Shneiderman, B. (1996): The eyes have it: a task by data type taxonomy for information visualizations. In: *Proceedings of the 1996 IEEE Symposium on Visual Languages*, ed. by Burnett, M., Citrin, W. (IEEE Computer Society Press, Piscataway 1996) pp.336–343
5. Spence, R. (2001): *Information Visualisation*, Addison-Wesley, Harlow, 2001
6. Spence, R., and Tweedy, L. (1998): The Attribute Explorer: information synthesis via exploration. *Interacting with Computers* 11, pp. 137-146
7. Thomas, J.J., and Cook, K.A., editors (2005): *Illuminating the Path. The Research and development Agenda for Visual Analytics*, IEEE Computer Society, 2005

## 4. VISUAL ANALYTICS METHODS FOR PATTERN DETECTION

This section presents an attempt of a systematic design of a toolkit that could support visual exploration and analysis of massive collections of movement data. When data are massive, it is insufficient to use only visual displays but it is necessary to involve the database technologies and computational methods of data processing and analysis. Still, the visualisation plays the central role since it allows the innate perceptual and cognitive capabilities and background knowledge of a human analyst to be utilised in the process of data exploration and analysis. These capabilities and knowledge cannot be replaced by purely machine processing. Hence, the combination of the visualisation with computer operations makes a ground for a truly synergetic work of human and computer.

We shall use the following abbreviations, which were introduced in Section 2: **IMB** for individual movement behaviour, **MCB** for momentary collective behaviour, and **DCB** for dynamic collective behaviour, i.e. behaviour of multiple entities during a time interval.

### 4.1 Data manipulation

#### 4.1.1 Aggregation

One of the most important data manipulation methods is aggregation. As any other method for data reduction, it involves substantial information loss but has also a positive side, specifically, the possibility to generalise, i.e. omit “high-detail noise” and focus on characteristic features of the phenomenon under study. The degree of data aggregation and generalisation matters a lot in data analysis. This is not only the matter of the size of the resulting data and the amount of information lost. This is also the matter of the scale on which the data are considered. Depending on the scale, the analyst sees the data differently and detects different patterns. Thus, in movement data, there may be local patterns like a flock (synchronous movement of multiple entities having close positions and same speed), or there may be larger scale patterns like massive movement towards industrial or commercial areas in mornings, or, on a yet larger scale, the difference of collective movement patterns on weekdays and weekends, and so on.

Hence, the appropriate degree of data aggregation and generalisation is not just a good trade-off between the simplification gained and the amount of information lost. The aggregation must be adequate to analysis goals, i.e. the patterns of what scale the analyst seeks to detect. If the interests of the analyst include patterns of different scales, it is necessary to consider the data on different levels of aggregation. The tools for visual analysis must thus enable the user to do this.

Aggregation consists of two operations: 1) grouping individual data items or, in other words, dividing the data into subsets and 2) deriving characteristics of the subsets from the individual characteristics of their members. Typically, various statistical summaries are used as characteristics of the subsets: number of elements, mean, median, minimum, maximum values of characteristics, mode, percentiles, etc. It is also important to know the degree of variation of the characteristics within the aggregates. For this purposes, such statistical measures as variance (or standard deviation) or inter-quartile distance are computed. Aggregates with high variation of characteristics of the members should not be used in data analysis since they may lead to wrong conclusions concerning the data.

Grouping/division may be necessary not only for data aggregation but also for other kinds of data processing, for instance, clustering. Movement data involve two referential components, the set of entities and the time. Grouping/division may be applied to any of them or to both. Thus, the time may be divided into equal-length intervals, e.g. 10 minutes, 1 hour, or 1 week. Depending on the data and analysis goals, it may also be useful to divide the time into slightly unequal intervals corresponding to calendar units such as months, quarters, or years, or to apply other division principles, for example, divide a year into semesters and holidays. Furthermore, it may be

reasonable to divide the time into subsets consisting of non-contiguous intervals, in particular, according to one or more of the temporal cycles. Thus, the user may wish to group all Mondays, all Tuesdays, and so on. Hence, the data analytics toolkit should include a tool for time partitioning where the user can flexibly define the principles of division.

A similar tool is needed for dividing the set of entities. This set has no distances that could provide a basis for division, as in the case of the time. It can be instead divided on the basis of the characteristics of the entities (e.g. age or occupation in case of people) or characteristics of their movement (e.g. position in space, speed, direction, etc.). This means that entities with close values of the selected characteristics are grouped together. For the grouping, either computational methods (clustering) or interactive techniques can be applied. The groups (clusters) of entities resulting from computational methods may be quite difficult to interpret. An appropriate visualisation of the characteristics of the entities forming the clusters may be helpful.

For interactive grouping, the user chooses the characteristics and specifies equivalence classes between their values, i.e. which values must be treated as close. The way of defining equivalence classes depends on the type of a characteristic. Thus, for numeric values, the user divides the whole value range into intervals. If the values of a qualitative characteristic are not too numerous, groups are formed from entities with equal values; otherwise, the user may wish to divide the values into classes according to their semantic closeness. For positions in space, the user may divide the space into compartments. In particular, these may be cells of a regular grid with the cell size and, possibly, shape (e.g. rectangular or hexagonal) chosen by the user. These may also be units of an administrative or other existing territory division or regions specified interactively according to any appropriate criteria such as surface type, way of use, accessibility, or other relevant properties of the space (see the list given in the section “Problem statement”). The visual analytics tools should support such arbitrary divisions of the space. Thus, the user may define space compartments by interacting with a map display or by applying database search operations like retrieving the locations of schools, shops, etc.

As it was mentioned, entities may be grouped according to values of their movement characteristics. Since these values change over time, interactive grouping can be done on the basis of values at selected time moments or on the basis of aggregated values on selected time intervals. Unfortunately, selection of each additional time moment or interval multiplies the number of groups and causes difficulties for the visualisation and visual exploration of the results of the aggregation.

Besides values at selected time moments, entities can also be grouped on the basis of *changes* of the values that occurred between two time moments. A change involves several aspects:

- the original value and the resulting value;
- the amount or degree of change, i.e. the absolute or relative distance between the original and resulting values (in a case when distances between the values exist);
- the direction of change: increase or decrease for numeric or ordinal values and the spatial direction for positions.

Any of these aspects may be taken as a basis for aggregation. Suppose, for example, that the user wishes to aggregate entities according to changes of their speed from moment  $t_x$  to moment  $t_y$ . The user may divide the whole range of speeds into intervals (say, 3 intervals: low, medium, and high speed) and build aggregates on the basis of all possible pairs (i.e. low-low, low-medium, low-high, medium-low, and so on). The user may also find the range of speed change, i.e. from the maximum decrease (taken as a negative number) to the maximum increase, and aggregate the entities by dividing this range into suitable intervals. The user may also divide the entities into 3 groups depending on whether the speed increased, decreased, or remained the same. It is clear that these approaches to aggregation are not equivalent in terms of the information that may be gained in the result. Analogously to the example with the speeds, grouping of the entities according to changes of their spatial positions may be done on the basis of the possible pairs composed of a

source position and a destination position, on the basis of the distances between the original positions and the destination positions, or on the basis of the spatial directions in which the destinations lie with respect to the source positions.

The methods for dividing (grouping) movement data are summarised in Table 4.1.

**Table 4.1.** A summary of methods for the division/grouping of movement data.

What is divided / division principle	Way of division	Examples
Time / inherent ordering and distances	Regular intervals	10 min; 2 hours
	Existing division	Days; months
	Temporal cycles	Time of day; day of week
	“Semantic” division	Day and night; workday and weekend
Entities / numeric characteristics	Regular intervals	Speed: 0-10, 10-20, ..., 190-200 km/hour
	“Semantic” intervals	Age: 0-15, 16-24, 25-64, 65 and over
Entities / qualitative characteristics	Individual values	Vehicle type: bike, motorbike, car, truck
	“Semantic” groups of values	Travel purpose: business (work, study), shopping and services, leisure (sports, walk, entertainment)
Entities / spatial positions	Regular sections	Rectangular grid
	Existing division	Administrative districts; cities
	Space properties	Water, forest, field, built area, ...
	“Semantic” division	City centre, residential area, shopping area, industrial area, ...
Entities / changes	Original and resulting values	From France to Germany, from Germany to France, from France to the UK, from the UK to France, ... (see Fig.2)
	Amount or degree of change	Travelled distance: 0-0.01km (no change), 0.01-100 km, 100-500km, ...
	Direction of change	North, northeast, east, ...

#### 4.1.2 Other data transformation techniques

Aggregation is not the only useful data transformation, and we shall briefly discuss some other data manipulation techniques that may increase the comprehensiveness of analysis and give additional insights into the data. One of them is the computation of the amounts or degrees and directions of changes, which is valuable not only for grouping of the entities by also by itself. Thus, it may be useful to look at change maps portraying (in a generalised manner) the changes of the MCB from one moment to another.

From other possible methods, especially useful may be transformations of space and time from absolute to relative. Thus, similarities between temporally and/or spatially separated behaviours can be more easily detected when these behaviours are somehow aligned in time and/or in space. To align behaviours in time, the “objective”, absolute time of each behaviour (i.e. the calendar dates and times) is ignored and only its “internal” time is considered, i.e. the time relative to the moment when this behaviour began. An example may be seen in the representation of the tourist movement in New Zealand, which was considered in deliverable D3.1; see also Fig.4.2. The tourists come to New Zealand on different days; however, the data are presented in such a way as if all the tourists came simultaneously. For this purpose, the designers of the visualisation

transformed the absolute dates into the day numbers starting from the day of arrival to New Zealand.

In this example, the analysts superposed the starting times of the IMBs of different tourists. It may also be useful to superpose both starting and ending times. In this case, the absolute time moments in each IMB are transformed into their distances from the starting moment divided by the durations of the behaviours (i.e. the lengths of the intervals between the starting and ending moments). This facilitates detecting similarities between movements performed with different speeds. Such an approach could be useful, for example, in comparing movements of migratory animals in different years.

Moreover, there may be cases when non-uniform transformation of the time of each IMB is reasonable. For example, an analyst exploring daily movements of people may be interested in excluding the times when the people stay in the same place for extended time (e.g. at work, in a shop, at home, etc.) and adjusting the times when they move. In this case, time transformation is done separately for each interval of movement.

Analogous ideas can be applied for spatial alignment of IMBs initially disjoint in space. An analyst may try to bring a set of IMBs (trajectories) to a common origin and search for coincidences between them. Furthermore, the analyst may be interested in disregarding the movement directions and considering only changes of the direction (turns). For this purpose, the trajectories are “rotated” until the initial movement directions coincide. Coincidences between further trajectory fragments indicate similarities. It may also be useful to “stretch” or “shrink” the trajectories to adjust their lengths.

In looking for co-locations between trajectories where positions are specified as points in the space, it may be reasonable to apply a kind of “spatial coarsening”, i.e. replace the original points by regions (areas), for example, circles with some chosen radius around the points. The resulting trajectories are treated as similar when there is an overlap between their “expanded” positions while there may be no sharp co-incidence between the original positions.

In studying MCBs and their behaviours over time, it may be appropriate to treat the space as a discrete set of coarsely defined “places” rather as a continuous set consisting of dimensionless points. For this purpose, one uses the methods for space partitioning, which has been discussed before in relation to data aggregation. Such a transformation may be called “space discretisation”. Furthermore, it may be useful to transform the geographical space into a kind of “semantic” space consisting of such locations as home, working place, shopping site, sport facility, etc. Then, each trajectory is transformed into a sequence of movements between pairs of these locations, and the analyst looks for similar sub-sequences occurring in different trajectories.

Table 4.2 indicates what types of patterns various data transformations may help to detect.

**Table 4.2.** Some types of patterns in movement data and data transformations that may support pattern detection

Pattern type	Data transformations
Full synchronisation of IMBs (same changes at same times)	Change computing: original values transformed into changes of positions, speeds, directions, ...
Lagged synchronisation of IMBs	Change computing (see above) Temporal alignment: superposition of the starting moments
Order-irrelevant co-location of IMBs Co-incidence in space and time	Spatial coarsening (disregards minor differences in positions)
Lagged co-incidence of IMBs	Spatial coarsening Temporal alignment: superposition of the starting

	moments
Ordered co-location of IMBs	Spatial coarsening Temporal alignment: superposition of the starting and ending moments (disregards differences in speed)
Geometrically similar trajectories	Spatial alignment: superposition of the origins and destinations Spatial coarsening Temporal alignment: superposition of the starting and ending moments
Constancy, change, trend in the MCB	Change computing Space discretisation

When we say that the analyst looks for similarities between IMBs, we do not really mean that the IMBs are presented to the analyst as individual items, without any aggregation and generalisation. As we have discussed before, the large data size precludes this way of analysis. Hence, similarities between IMBs need to be somehow detected without the analyst seeing the IMBs. This can only be done by involving computational analysis methods.

## 4.2 Exploring the behaviour of the IMBs over the set of entities

### 4.2.1 Use of clustering

In order to analyse IMBs without seeing them, the analyst can apply clustering methods, which divide entities into groups so that the entities within a group are as similar as possible and differ as much as possible from the entities in the other groups. If a clustering method can group the moving entities according to similarities and differences of their IMBs, the analyst can then look at various aggregated characteristics and aggregated behaviours of the groups instead of looking at the individual behaviours.

In its work, a clustering method computes some numeric values expressing the degree of similarity between entities. These values are usually called distances (in an abstract sense): the smaller the distance is, the more similarity between the entities exists. Hence, to group moving entities according to their IMBs, it is necessary to find a way to express numerically the degree of similarity between two IMBs, or, in other words, to define a method for computing distances between IMBs. Such a method will be further referred to as “distance function”.

As we have noted, two or more IMBs may be similar in various diverse ways, and any sort of similarity may be of interest. Each sort of similarity requires a different distance function. Thus, the degree of spatial and temporal co-incidence is computed from the distances between the spatial positions at the corresponding moments. The same function would suit for the lagged co-incidence after applying temporal alignment of the IMBs (see Table 2.2). The degree of order-irrelevant co-location may be computed from the distances between each position on one trajectory and the nearest position on the other trajectory. For the ordered co-location, the corresponding function must find common (overlapping) positions and check whether they were reached in the same order. This method is also suitable for the estimation of the degree of similarity of trajectory shapes after the trajectories have been spatially aligned.

Hence, it is reasonable to devise a clustering tool where the distance function is replaceable. In this case, the analyst could choose the appropriate distance function, depending on his/her current interests, and let the clustering tool run with the use of this function. A library of appropriate distance functions can be created in advance, as well as a library of data transformation methods.

It should be also borne in mind that the existence and the types of similarity patterns between IMBs depend on the temporal resolution chosen for looking at the data. Thus, fine movements of entities, which are made in the scale of minutes or hours, may be quite different while there may be a clear similarity between the behaviours of the same entities considered in the scale of days or weeks. Hence, it makes sense to run a clustering method several times with the same distance function but with different degrees of aggregation and generalisation of the data with respect to the time, i.e. with the time partitioned into intervals of different lengths.

A serious technical problem in applying clustering algorithms is that they can work effectively only when the data are in the computer memory. The reason is the necessity of numerous repeated distance computations. Not only pair-wise distances between entities need to be computed but also, as clusters are built, the distances between the current clusters (which change over time) and the entities that has not yet been attached to any cluster. When the data are too big for the computer memory, clustering may require too much time.

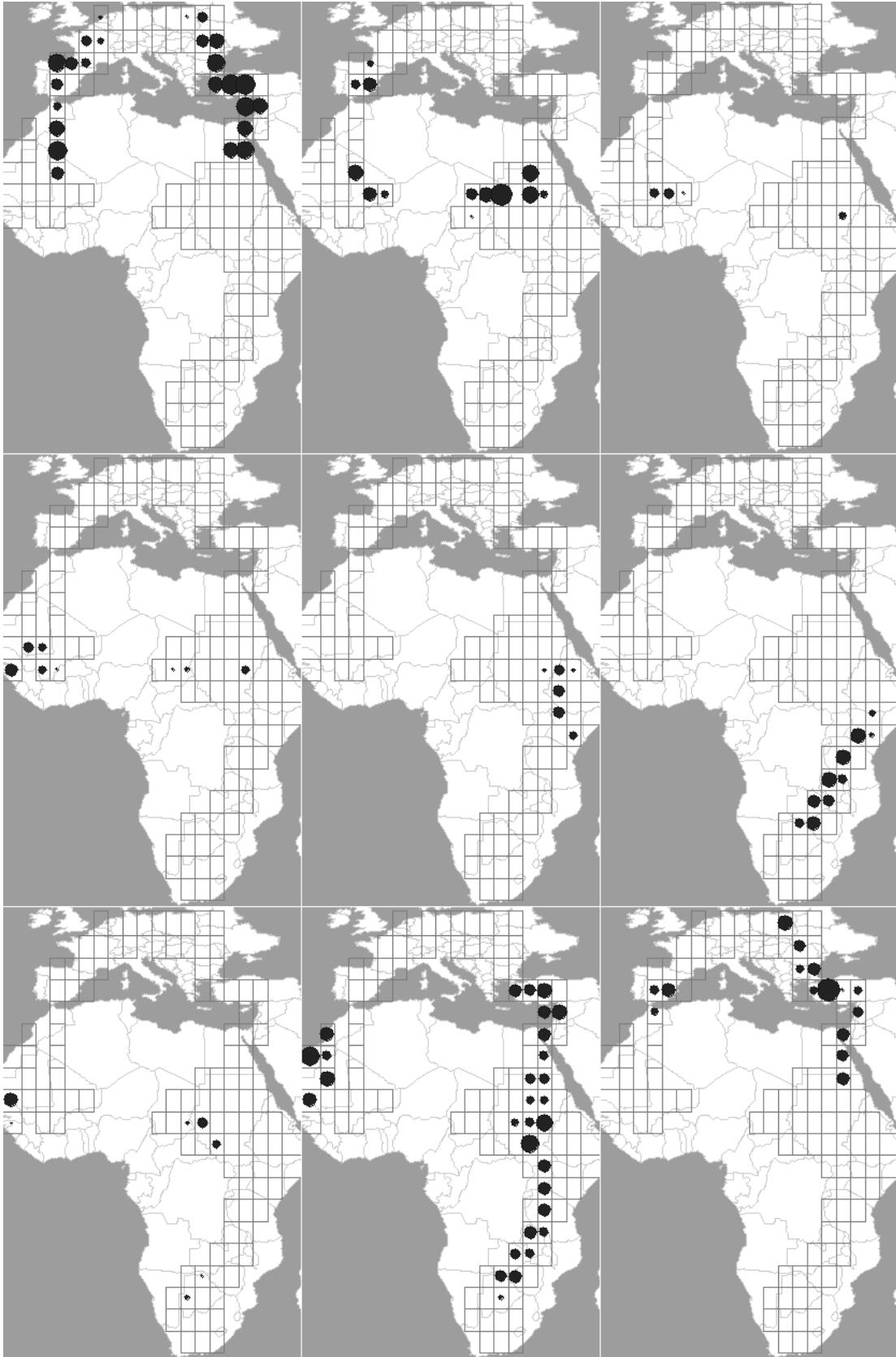
A possible way to cope with this problem is based on sampling. The idea is that a subset of entities is sampled from the whole set of entities so that the corresponding movement data have a size suitable for effective clustering. Depending on the specifics of the data and the goals of the analysis, it may also be reasonable to sample fragments of IMBs. For example, from data about people movements during many days, fragments corresponding to one-day movements of individuals can be sampled.

Once a manageable subset of IMBs or fragments of IMBs has been extracted, clustering is applied to this subset. After the clusters are built, the distances between them and each of the remaining IMBs or fragments can be computed using the same distance function as for the clustering. This requires a single run through the database. On this basis, each IMB or fragment is attached to the closest cluster or, if it is too distant from all clusters, selected for further application of clustering or for a detailed consideration by the user (this may be an outstanding behaviour).

After the clustering is done, the results need to be visualised so that the analyst could interpret and investigate them.

#### **4.2.2 Visualisation of clustering results**

Basically, the visualisation must allow the user to see the common features of the IMBs in each cluster as well as the degree of variation. Unfortunately, clustering algorithms do not provide any general description of the clusters built. The clusters are defined extensionally, i.e. by listing the elements they consist of. Hence, any information about the common features of the IMBs in each cluster has to be extracted from the data that were used as the input of the clustering method. A realistic way to do this is to obtain various statistics about the movement characteristics of the members of a cluster by means of database operations and to visualise these statistics. By comparing the statistics for different clusters, the analyst can understand what is in common between the members of each cluster and how they differ from the members of the other clusters. Andrienko and Andrienko (2006) demonstrated how histograms can be used to interpret clusters built on the basis of numeric characteristics of entities. In case of movement data, appropriate statistics and visualisations are chosen depending on how the similarity between the IMBs has been defined for the clustering (i.e. what kind of distance function has been used).



**Fig. 4.1.** A visualisation technique applicable for the display of results of clustering made on the basis of co-location of trajectories. The sizes of the circles show in how many trajectories each location appears. A separate map is built for each cluster.

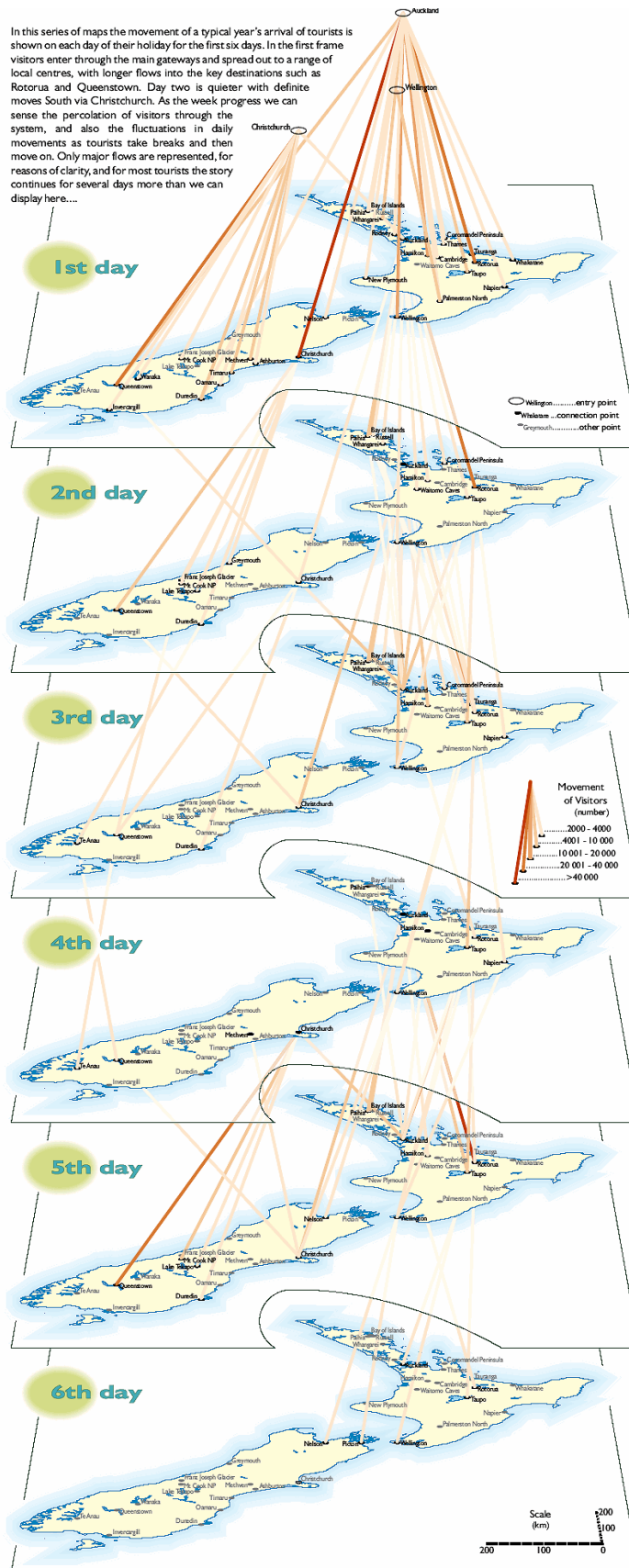
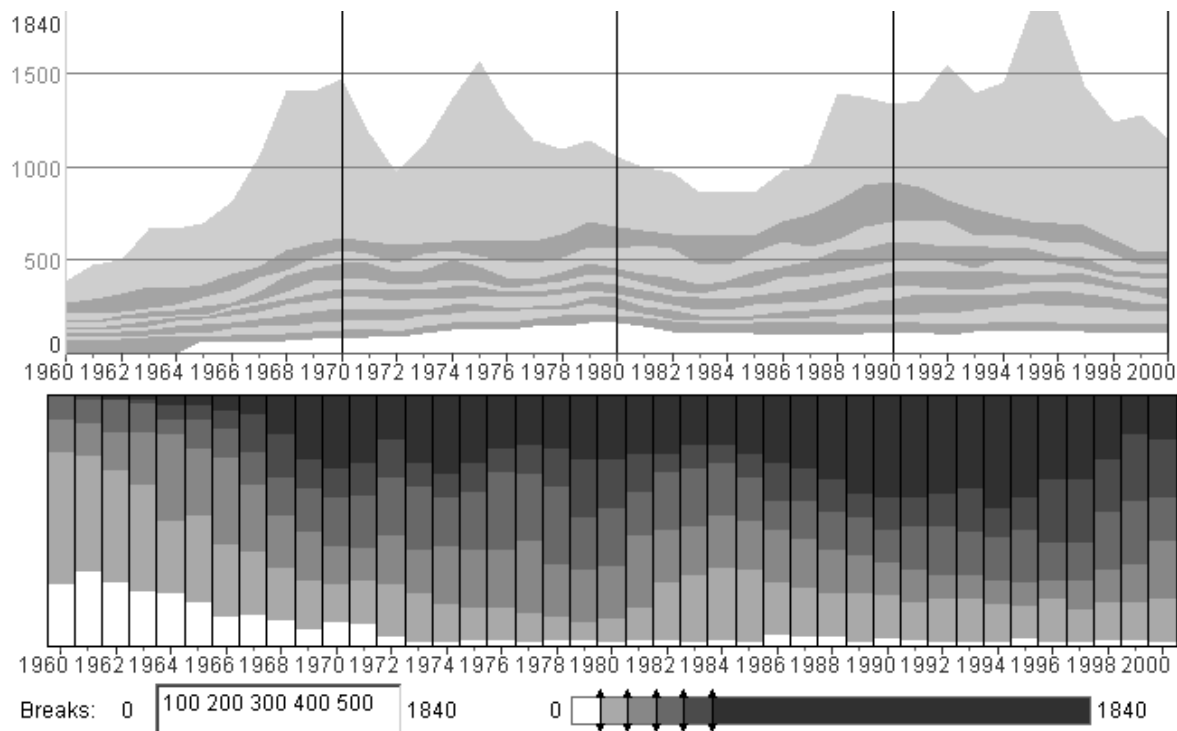


Fig. 4.2. The visualisation of the major flows of tourists in New Zealand by Drecki and Forer (2000).

Thus, when IMBs have been clustered on the basis of co-location of the trajectories, a suitable visualisation would be a map where for each location of the space (resulting from space discretisation or original, if there are not too many different locations in the source data) it is shown in how many trajectories it appears. Graduated symbols or graduated shading are suitable for this purpose. A separate map is built for each cluster, which enables comparison of the clusters. The maps may look like is shown in Fig.4.1.

For ordered co-location and for spatio-temporal co-incidence, it is reasonable to compute for each pair of locations  $x$  and  $y$  and time interval  $T$  how many cluster members moved from  $x$  to  $y$  during the interval  $T$ , where  $T$  results from an appropriate partitioning of the time (which may be previously transformed as discussed in the previous section). A good way to visualise this statistics is using tiered maps, as in Fig.4.2 (this visualisation was discussed in deliverable D3.1). In case of ordered collocation, the third (temporal) dimension reflects the temporal order, and in case of spatio-temporal co-incidence, either full or lagged, the third dimension reflects also the temporal distances.

When clustering is used in order to group IMBs according to the derivative movement characteristics rather than positions, other types of visualisation are appropriate. For example, the variation of the speeds may be shown in an aggregated way on a modified time graph, as is illustrated in Fig.4.3.



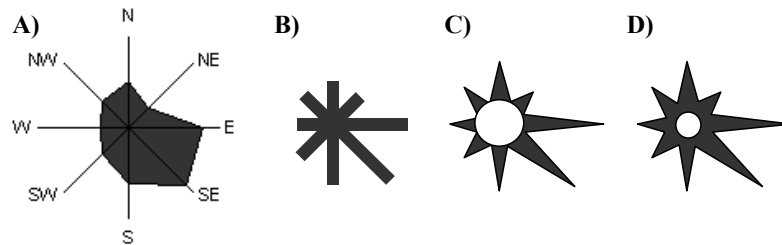
**Fig. 4.3.** Two modifications of the time graph technique show aggregated rather than individual data.

The upper display in Fig.4.3 represents positional statistical measures (specifically, deciles) for each time moment. The graph area is divided into shaded stripes alternating between light grey and dark grey. The boundary between a dark grey area and an adjacent light grey area corresponds to one of the deciles and shows the dynamics of this decile over time. In this way, the display gives an idea of the statistical distribution of the values at each moment and shows how the distribution changes over time.

The idea of the lower display is that the user divides the value range of the characteristic into intervals, and the tool counts how many values fall at each moment into each of the intervals. The counts are represented by the heights of the shaded segments of the bars where each bar

corresponds to one moment. Both methods can be applied to the whole population of entities and to clusters. The user may compare the variations of characteristics in different clusters by viewing several aggregated time series displays in parallel.

The distribution of movement directions in each cluster at different time moments or summarised over time intervals may be represented using diagrams as shown in Fig.4.4.



**Fig. 4.4.** Possible methods for representing numbers of entities moving in different directions.

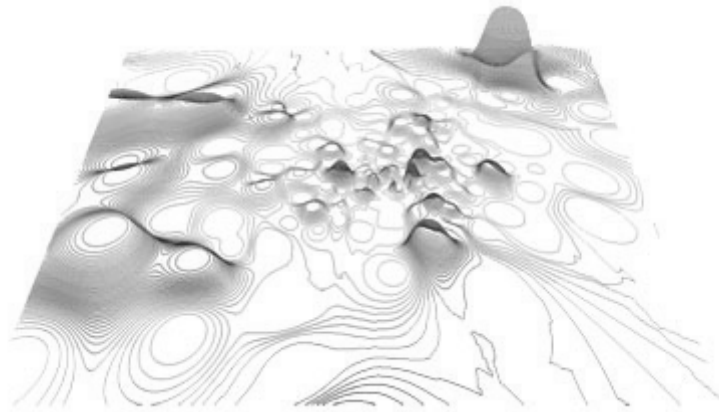
Besides the features of the IMBs of the cluster members, the analyst should be informed about the number of members in each cluster and the statistics of their static characteristics, if available in the data. The analyst should also be able to obtain any statistics concerning the movement of the entities, such as average and maximal speed or total travelled distance.

Apart from the computational clustering and visual examination of the results, the user may be interested in having a close look at subsets of IMBs with specific features, for instance, at the trajectories where the entities move out from the city centre in the morning and towards the city centre in the evening. For this purpose, interactive query tools are necessary. A challenge is to design effective methods for data retrieval and visualisation ensuring an acceptable reaction time. It is also important to design a proper user interface taking into account that quite perceptible delays are indispensable with large datasets, especially when the data are out of the computer memory. Thus, the principle of Dynamic Query (Ahlberg et al. 1992), when the tool immediately reacts to any slight user interaction with the query device such as moving a slider by one pixel, is not applicable to this case. However, the tool should enable the mode of the work when the user does not formulate a complete query at once but iteratively refines the query depending on the results received.

### 4.3 Exploring the behaviour of the MCB over time

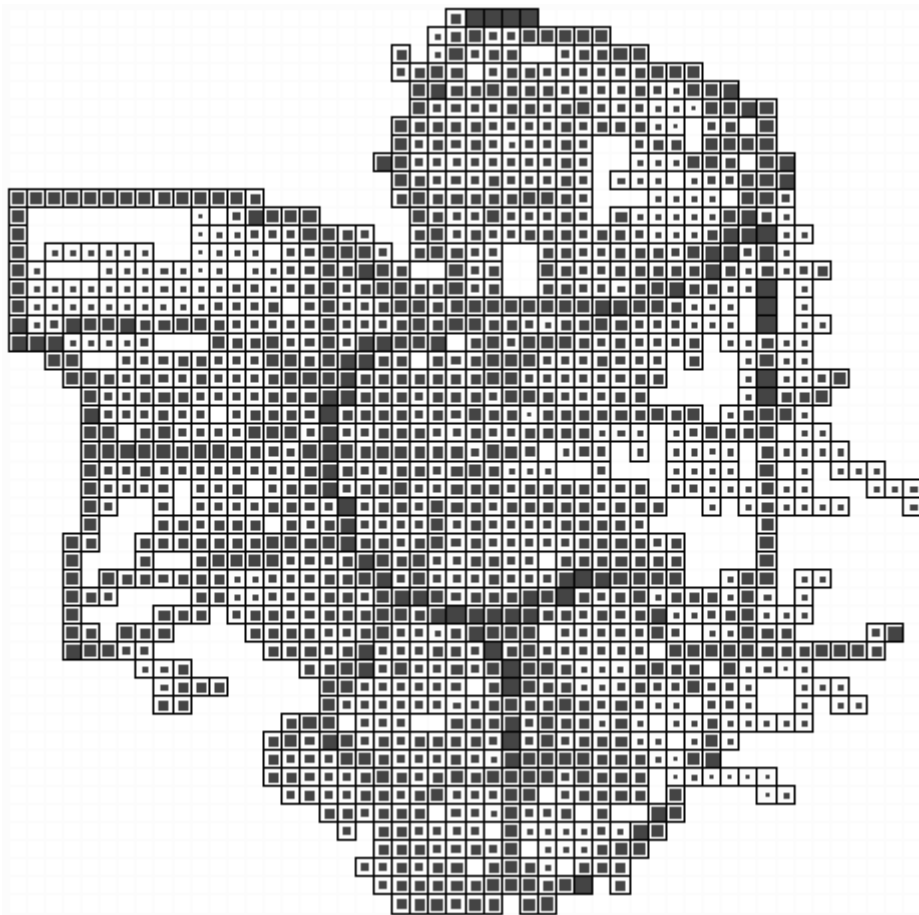
In order to explore the behaviour of the MCB over time, the analyst needs visualisations that show him/her the MCB at different time moments or, in a summarised way, on different intervals into which the whole time of movement is divided. There are two basic ways to do this: an animated display (map, diagram, or graph, depending on the information to be portrayed) and multiple uniform displays, or “small multiples”, in terms of E.Tufte (Tufte 1983). An example of “small multiples” may be seen in Fig.4.1. We shall not discuss here the advantages and disadvantages of each approach (in our opinion, they are complementary and should be used in combination) but focus instead on the content of a single animation frame or a single display in “small multiples”, which corresponds to the MCB at a single time moment or on a single interval.

In order to look at the spatial distribution of the moving entities at a selected time moment (interval), it is natural to use a map. Since the entities are very numerous, the positions have to be shown in an aggregated manner, i.e. as densities. There are approaches when densities are visualised as smooth surfaces, which are built using kernel methods or other computational techniques. Such surfaces are represented by colouring or shading, by contour lines (isolines), or in 3D views, which are rather appealing visually (Fig.4.5).



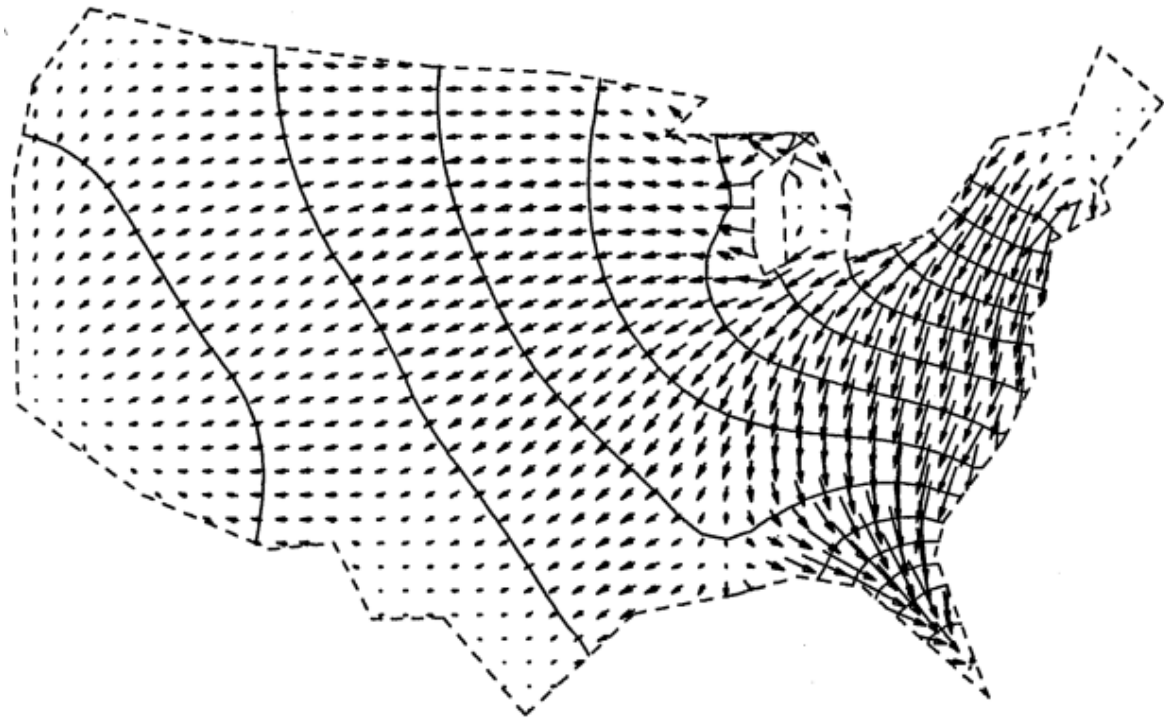
**Fig. 4.5.** A perspective view of the density surface of the mobile phone activity in the city of Graz, Austria (Graz 2005, Ratti et al. 2005).

Another approach is “binned” visualisation of the densities, when the map area is divided into regular “bins”, or cells (e.g. squares), and the number of entities fitting in each cell is shown by colouring, shading, or graduated symbols (Fig.4.1 and 4.6). Such a visualisation can be built using database operations. The user can vary the size of a cell in order to look at the data at different levels of aggregation (of course, re-aggregation of a big database may require some time).



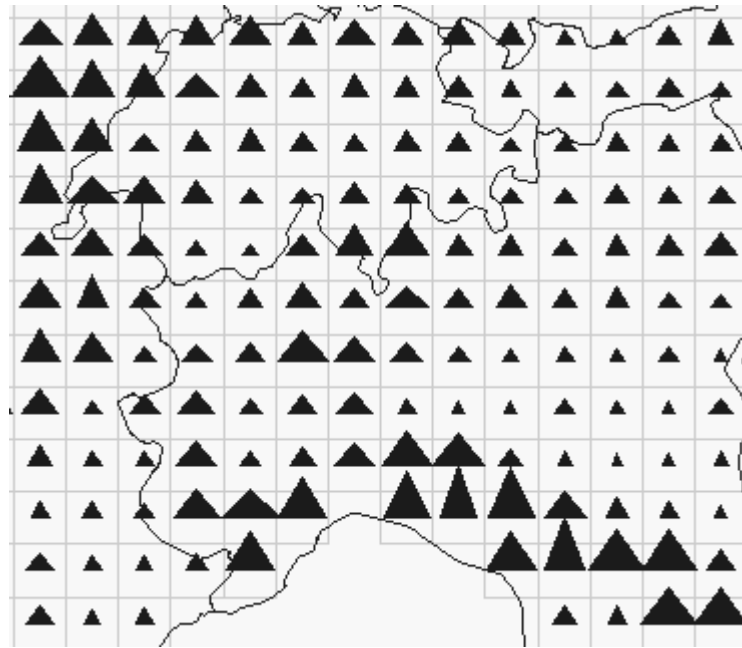
**Fig. 4.6.** A display of vehicle movement data aggregated by spatial cells.

Maps are suitable for showing not only the positions of the entities but also various movement characteristics associated with these positions, such as speed and direction of movement. Again, in the case of large dataset, these characteristics need to be aggregated. The “binning” approach is quite appropriate here: it is possible to compute and visualise various summary statistics for the cells such as the average, minimum, and maximum speed or the number of entities moving in each direction. A single value (such as average speed) may be represented by colouring, shading, or graduated symbols, as in Figs. 4.1 and 4.6. Prevailing movement directions can be indicated by arrows, as in Fig.4.7. A vector map may show not only the prevailing direction in each place (by vector orientation) but also how many entities moved in that direction (e.g. by vector length) and how much this direction prevails in relation to the other directions (e.g. by vector shade or colour).



**Fig. 4.7.** A possible visualisation for prevailing movement directions (source: Tobler 2005).

Several values (e.g. numbers of entities moving in different directions) require the use of diagrams, as, for example, in Fig.4.4. The sizes of the diagrams should not exceed the sizes of the cells where they are placed, and hence the cells need to be large enough for the diagrams to be legible. Representation of average, median, or most frequent values should be accompanied with the display of appropriate statistics expressing the degree of variance. An example is given in Fig.4.8, where triangle symbols are used to represent simultaneously the mean values and the variances of the values in the cells.



**Fig. 4.8.** The heights and the widths of the triangle symbols encode the mean values and the variances, respectively, computed for the cells.

As a complement to maps and perspective views of the (geographical) space, non-cartographic displays are used in order to look at the statistical distribution of various movement characteristics at different time moments or on different intervals. Frequency histograms provide aggregated information about the statistical distribution of numeric values; statistics about qualitative values can be shown by bar charts where each bar corresponds to one value and the size of the bar is proportional to the number of occurrences of this value.

There is an extension of the histogram technique that can be applied to a pair of characteristics or a pair of time moments. This extension is known as two-dimensional histogram or binned scatterplot. The area of the plot is divided into regular compartments (bins) as is shown in Fig.4.9. In the compartments, the frequencies of corresponding value combinations are shown by symbol sizes, shading, or colouring. A variation of two-dimensional histogram is a transition matrix, where rows and columns correspond to different spatial locations while the symbol or colour in each cell shows how many entities moved from one of the respective locations to the other between the selected time moments. Like traditional histograms, two-dimensional histograms may allow interactive selection of groups of entities by clicking on cells.

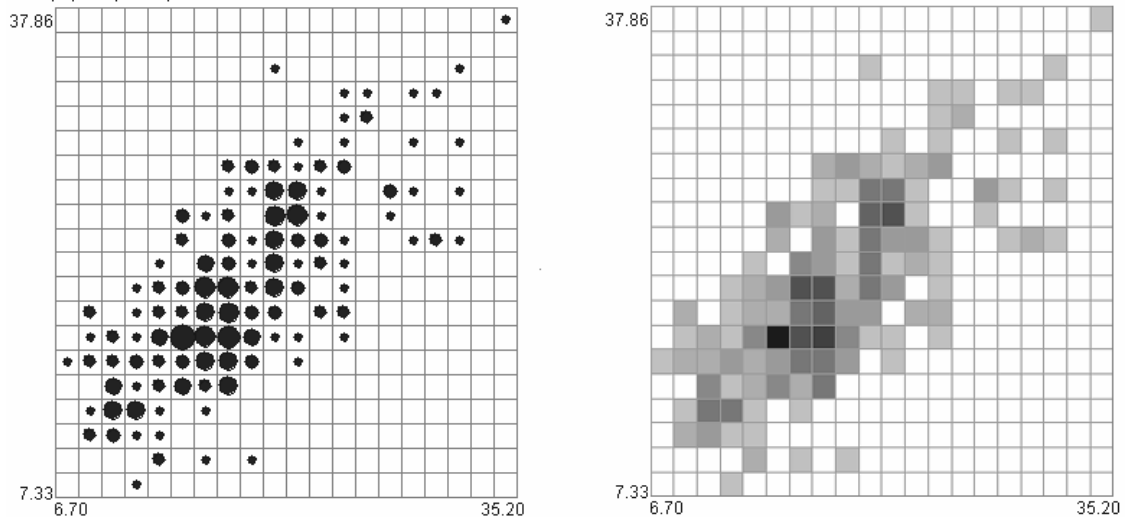


Fig. 4.9. Two-dimensional histogram, or binned scatterplot.

It seems desirable to extend the idea of two-dimensional histogram, which is good for analysing changes that occurred between two time moments, so that more than two time moments could be considered. One option is to build a series of two-dimensional histograms. However, this takes much screen space. Besides, the user may have difficulties in establishing correspondences between elements of different displays. Another option could be to borrow the idea of the parallel coordinate plot (Inselberg 1985). Parallel axes can correspond to different time moments or intervals while positions on the axes may have the same meanings as in an ordinary histogram, i.e. encode values or value intervals of a time-variant characteristic. Two positions representing values  $v_i$  and  $v_j$  on neighbouring axes, which correspond to successive times  $t_k$  and  $t_{k+1}$ , are connected by a line if there are at least  $N$  entities ( $N$  is a parameter;  $N > 1$ ) that had the value  $v_i$  at time  $t_k$  and value  $v_j$  at time  $t_{k+1}$ . The number of such entities is encoded in line thickness, or shade, or colour. This idea is similar to that of the tiered map display as in Fig.4.2. To our knowledge, no implementation of such a display technique (which can be called “*change histogram*”) currently exists. The technique needs to be implemented before its effectiveness can be judged.

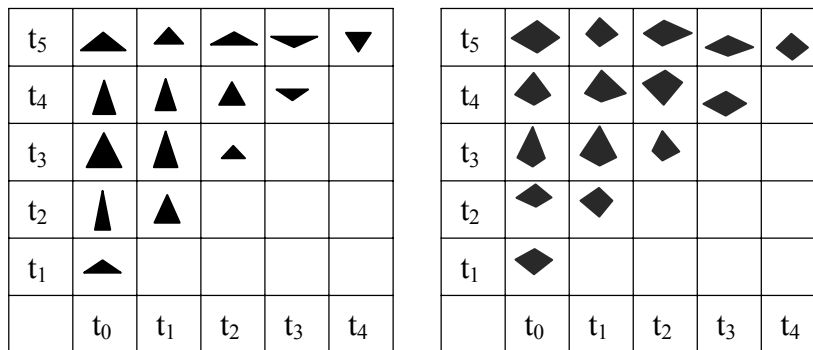
For statistics about movements in different directions, it may be convenient to use radial bar charts, where the orientation of the bars corresponds to spatial directions: north, northeast, east, and so on (Fig.4.4). Analogously to the spatial views, such displays are built for each moment (interval) in time and presented simultaneously as “small multiples” or in temporal sequence (animation).

Another possibility is to represent the time using one of the display dimensions, as in a time graph. For example, this may be a display where the horizontal dimension represents the whole time period divided into intervals. For each interval, there is a segmented bar showing the frequencies of different values of some movement characteristic (Fig.4.3 bottom), i.e. the number of occurrences of each value (for qualitative values) or the number of occurrences of values from each of intervals previously specified by the user (for numeric values).

To facilitate detection of significant changes of the MCB, it is useful to compute and visualise the changes that took place from one moment (interval) to another, in particular, changes with respect to the previous moment or interval. For example, with “binned” maps, the differences between the values in the cells at consecutive moments (intervals) may be computed and represented by cell colouring. It is reasonable to use a diverging colour scale (Brewer 1994), where one hue represents decrease and another hue represents increase. It is also useful to compute changes in the movement characteristics of the individual entities and represent them in an aggregated way on maps and non-cartographic displays. For example, the visualisation technique with temporally ordered segmented bars, which has been discussed above, may be used to

represent changes of the speeds: how many entities decreased their speeds (by more than  $x_1\%$ , by  $x_1$  to  $x_2\%$ , etc.), how many entities kept the same speed, and how many entities increased their speed (by more than  $x_1\%$ , by  $x_1$  to  $x_2\%$ , ...). While the primary focus of the analysis is the collective movement behaviour rather than individual movements, significant changes in the statistical and/or spatial distribution of individual movement characteristics may indicate changes in the collective behaviour.

A modified time graph, as in Fig.4.3, can represent the changes at each moment with respect to the previous moment or the differences with respect to one selected time moment. In order to see simultaneously the differences between the values for all time moments, the visualisation technique of time-time plot, or T-T plot (Imfeld 2000), may be used. Traditionally, T-T plots represent information about individual entities. It has two time axes and represents changes of a certain characteristic of the movement, such as the speed, travelled distance, or direction, between moments  $t_x$  and  $t_y$  by symbols placed in the positions corresponding to  $x$  and  $y$  or by colouring or shading of the cells, in which the plot area is divided. Since representing individual data is inappropriate in our case, the technique needs to be modified to show aggregated information. Instead of the changes of the individual values, the plot will show statistical summaries of the changes over the whole population or a group of entities, such as the means or medians. Additionally to this, it is advisable to give the user an idea about the variability among the degrees or amounts of change. The designs presented in Fig.4.10 may be helpful for this purpose.



**Fig. 4.10.** Two design options for aggregated T-T plot.

The design on the left may be used to represent simultaneously the mean values of the changes and the variances or standard deviations. The mean values are encoded in the heights of the triangles and the variances in the widths. The upward or downward orientation of a triangle reflects the sign of the mean: positive or negative, i.e. increase or decrease of the values. In the design on the right, the positions, sizes, and shapes of the quadrilaterals may encode the positional statistical measures: medians, quartiles, minimums, and maximums. Thus, in each cell, the vertical position of the left and right vertices of the quadrilateral may reflect the value of the median change and the vertical positions of the lower and upper vertices may encode the values of the first and third quartiles, respectively. The width of the quadrilateral may represent the distance between the minimum and maximum values of change. Note that the scale in the horizontal dimension is not necessarily the same as the scale in the vertical dimension. We suggest quadrilaterals rather than box-and-whiskers plots, which are traditionally used for representing positional measures (Tukey 1977), since we believe that multiple small quadrilaterals may be better perceived than multiple small box-and-whiskers plots. However, it should be tested whether users can easily understand such displays as in Fig.4.10 and extract useful information from them. If it turns out that people experience significant difficulties, it may be reasonable to use a simplified variant of aggregated T-T plot, which shows only summarised changes or only measures of variance but not both of them simultaneously. In such a simplified plot, the cells may be shaded or coloured or contain graduated symbols, like in two-dimensional histograms (Fig.4.9).

Like time-series displays, T-T plots may be built from temporally aggregated data. In this case, the rows and columns of a plot correspond to time intervals rather than individual moments.

It should be noted that aggregated displays may be used not only for viewing the data but also as direct manipulation query devices where the user may select subsets of data through the selection of the aggregates representing them, e.g. cells on a map, bars in a histogram, or segments in segmented bars. To support such a kind of interaction, the displays must “remember” how each aggregate has been produced and be able to transform user’s actions into appropriate database queries. However, it must be taken into account that noticeable time may be needed to fulfil the queries in case of massive data. Therefore, immediate reaction of the tool to any user’s click or slight mouse movement may be inappropriate. Instead, the user should be able to make and modify selections without triggering any queries and, when the selection process is finished, to signify this explicitly.

#### **4.4 Looking for connectional patterns**

Direct manipulation query interfaces are especially convenient for brushing, when the user interactively selects a subset of data and, in response, graphical elements in different displays corresponding to this subset are similarly marked (highlighted). Brushing helps the analyst to establish links between two or more displays providing complementary information. This, in turn, may be helpful in a search for connectional patterns, i.e. for correlations, influences, and structural links between characteristics, phenomena, processes, events, etc.

For example, the analyst may use a map to select areas with high density of moving entities at some time moment  $t_1$ . From maps corresponding to other moments, the analyst will learn whether the densities are always higher in these areas than on the remaining territory, which may indicate a link between the number of moving entities and the properties of the space where they move. From speed histograms, the analyst may see whether there is any relation between the areas of high density and the variation of the speed of movement such as high speeds of the entities before entering the areas of high density, low speeds inside the areas, and high speeds after exiting the areas. Furthermore, simultaneously with the displays of the movement data, the statistical distribution of the static properties of the moving entities and/or their activities, by time intervals, may be visualised. Then, the analyst may look whether the entities in the areas of high density have any particular properties and/or perform any particular activities. If a display of various events is available (this may be a map display if the events are spatially located or a calendar display otherwise), the analyst may look whether the times and places of high density are related to any events.

It should be noted that direct manipulation and brushing, while being convenient and easy to use, are not strictly necessary for such a kind of analysis. Other query interfaces are also possible. Thus, the user may benefit from a temporal query tool capable of extracting data that refer to the same relative positions or sub-intervals in different cycles and aggregating the data across the cycles. For example, the user may be interested to extract all people movements made from 6AM to 9AM in all days and have the extracted data aggregated by the days of the week. Then, the aggregated morning hours movements on Mondays, Tuesdays, Wednesdays, and so on should be appropriately presented to the user so that the user could see the differences between the movements on weekdays and weekends. Moreover, the movements on Monday mornings may differ from the movements in the mornings of other weekdays and movements on Saturday mornings may differ from those on Sunday mornings. Similar queries can be applied to other hours of the day in order to understand in the result how the daily and weekly cycles interact in people movement.

A disadvantage of using queries for the search of connectional patterns is that each query provides information about a subset of the data, and hence, the procedure has to be repeated for other subsets. In our example, the analyst would have to select other places on the same map, the same places on maps corresponding to other time moments, as well as other places on other maps.

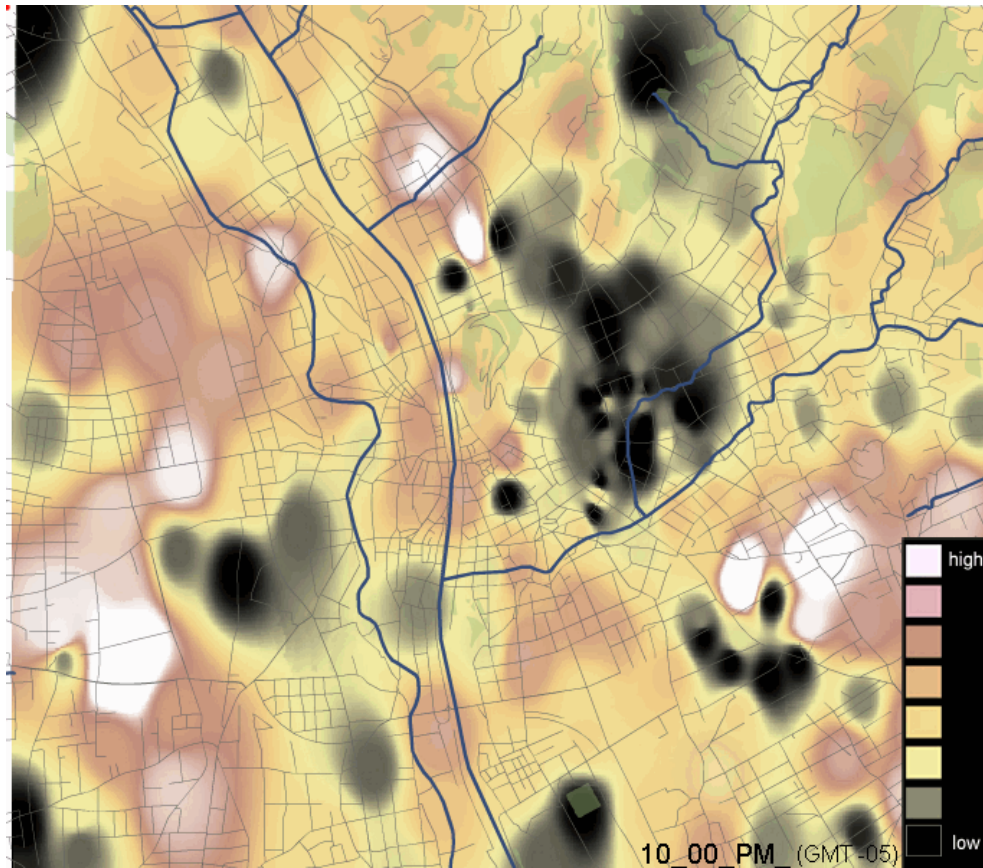
Moreover, data subsets can be selected using various criteria (space, time, speed, direction, means of the movement, activity, ...) and combinations of criteria, which makes the number of possible selections infinite. Hence, the use of querying is reasonable when it is necessary to investigate particular cases, especially outliers such as extreme values or extreme changes.

A better way to search for correlations and dependencies is to divide the whole dataset into subsets (rather than select a single subset) on the basis of various characteristics and to obtain for each of the subsets appropriate statistics of other characteristics. These statistics are then compared, possibly, visually; significant differences between them may indicate the presence of some links between the two (groups of) characteristics. For example, all the movements may be partitioned into subsets according to their positions within a temporal cycle, such as the days of the week. Then, the analyst may look at visualisations of aggregated positions, speeds, movement directions, etc. in each subset in order to see whether the movement characteristics are related to temporal cycles. Another example is division of the entities according to their static characteristics or their activities and looking at statistics of their movement characteristics. Such divisions as well as computing statistics can be done by means of database operations.

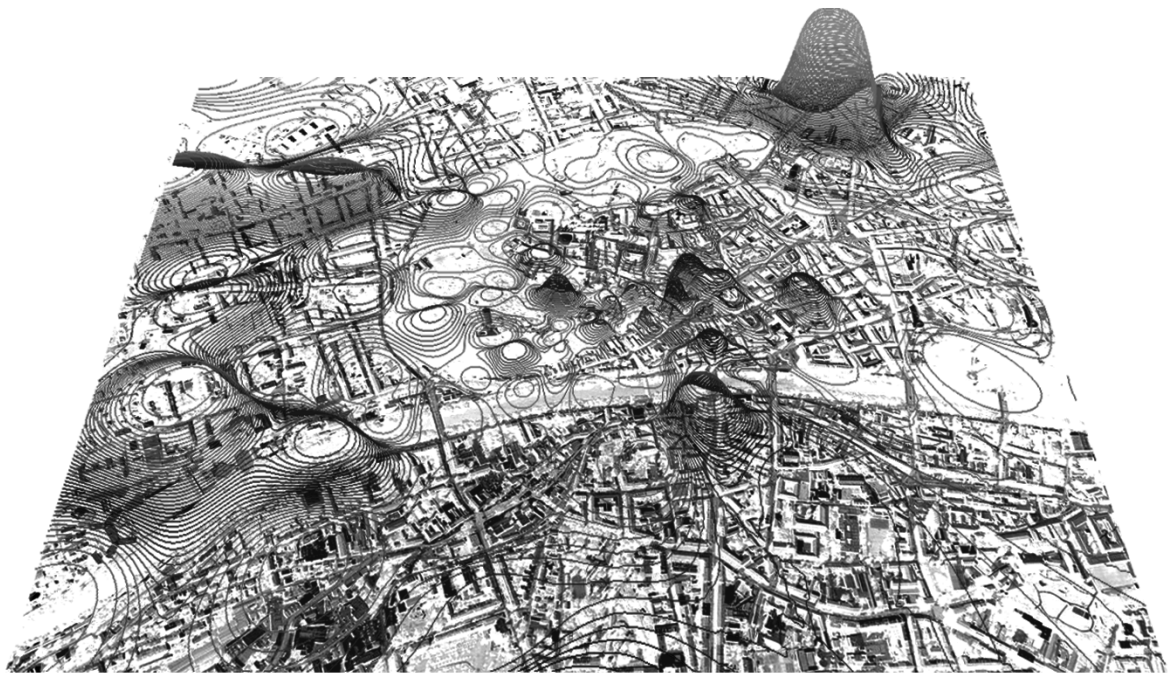
We have also mentioned another method of division, namely, division of the set of entities into groups according to similarity of their IMBs by means of clustering. After the application of clustering, it is useful to look at various statistics for the resulting clusters in order to judge, for example, whether there are any links between the properties and/or activities of the entities and the features of their IMBs.

There are also purely visual methods for searching for links. Thus, overlaying several information layers in maps may support detecting links between the movement characteristics and various properties of the underlying territory as well as spatial and spatio-temporal phenomena and spatially located events. Fig.4.11 provides an example of overlaying streets on a density surface represented by colouring. Fig.4.12 illustrates how a smooth density surface may be shown in a 3D view on top of a map or an aerial or satellite image of the territory. Movement data may be represented as flows or vectors on top of a layer representing weather or land cover information by area painting and a layer representing rivers and other waters by lines or shapes of a specific colour. Another option is encoding values of several spatially referenced characteristics in different properties of graphical elements such as size, colour, brightness, shape, orientation, and texture. For example, in a vector map like in Fig.4.7, the vector orientation may show the movement direction, the size may be proportional to the number of moving entities, and the colour of the vectors may encode the air temperature or land cover type. It is also possible to establish links by comparing two or more map displays presenting different information related to the same territory; however, it may be more difficult to detect correspondences than in a case when all information is present in the same display.

For other display types, ways to incorporate additional information can sometimes be found while there are no general approaches. For detecting links with various temporal events, these events can be indicated on displays of movement data according to the times of their occurrences. For example, "small multiples" or animation frames corresponding to these times may be specially labelled or marked. On displays representing the time by means of one of the display dimensions, the times of the events can be marked at the corresponding positions within this dimension. Thus, when a user looks for connections between the temporal variation of movement speeds and some temporally distributed events, the times of the events can be marked on a time series display representing the variation of the speeds.



**Fig. 4.11.** A density surface, which is represented by colouring, is overlaid with a street map (Graz 2005, Ratti et al. 2005).



**Fig. 4.12.** A perspective view of a density surface overlaid upon an aerial or satellite image of the territory (Graz 2005, Ratti et al. 2005).

In order to detect links between the movements and the temporal cycles, it is useful to look at “small multiples” representing movements at different times and arranged according to the temporal cycles. For example, displays of city traffic data aggregated by days may be arranged on the screen into a matrix with seven columns corresponding to the days of the week and the rows corresponding to different weeks. This arrangement facilitates noticing commonalities and differences within and between the cycles.

A classical visualisation technique that supports looking for correlations between numeric or ordinal variables is scatterplot. For massive data, binned scatterplot, as in Fig.4.9, may be used. In particular, one of the axes in such a scatterplots may represent absolute or relative (transformed) time or positions within a temporal cycle.

#### 4.5 Summary of techniques for supporting pattern detection

As can be seen, quite a number of tools are needed for detecting patterns of various types in movement data as well as relating movement characteristics and behaviours to other phenomena. Table 4.3 summarises what has been suggested. This choice of techniques results from a theoretical analysis and, certainly, needs a practical verification.

**Table 4.3.** Computational and visualisation techniques for detecting various types of patterns.

Pattern types	Computational or database techniques	What is visualised	Visualisation techniques
Similarity and difference between IMBs	Clustering on the basis of various distance functions Data aggregation with various temporal granularity	Statistics of the movements within the clusters	Density map Tiered maps representing flows Histograms; temporally arranged segmented bars (non-spatial characteristics)
		Individual behaviours included in a cluster	Map with trajectory lines Animated map Space-time cube
Constancy, changes, and various arrangements in the development of the MCB over time	Methods for generalisation of spatial distribution of points: kriging, ...	Density surfaces for different time moments	Animated displays or “small multiples”: Density map Perspective view
	Aggregation by spatial compartments	Various statistics for the pairs compartment + time moment: number of entities, averaged characteristics, variance indicators, ...	Animated displays or “small multiples”: Choropleth map Map with graduated symbols Map with diagrams Map with vectors

	Statistical aggregation over the whole set of entities by time moments or intervals	Various overall statistics for the time moments or intervals	Sequence of histograms, bar charts, or star diagrams (“small multiples”) Temporally arranged segmented bars
	Computing changes by spatial compartments	Differences or ratios for the pairs compartment + time moment	Animated maps or “small multiples” using a diverging colour scale to distinguish between increase and decrease
Various connectional patterns	Database queries involving movement data and other types of data	Subsets of the movement data related in the specified manner to the other data	Special marking (highlighting) of graphical elements corresponding to the selected data, depending on the type of display
	Dividing the movement data and computing statistics for the subsets	Statistics of the characteristics by the subsets	Multiple histograms, bar charts, or star diagrams Multiple maps showing aggregated positions
Links between IMBs and static properties or activities of the entities	Clustering of the IMBs (see above)	Statistics of the static properties or activities of the entities within the clusters	Histograms (numeric properties) Bar charts; pie charts (qualitative properties)
Links between movements and characteristics of the space or spatial phenomena	Spatial generalisation or aggregation (see above)	Aggregated or generalised movement data together with other spatial data	Overlaying two or more information layers in a map or perspective view (animated display or “small multiples”) Presentation of different information in separate maps
Links between movements and temporal cycles	Spatial or statistical generalisation or aggregation (see above)	Movements by time moments or intervals	Arrangement of “small multiples” according to the temporal cycles
Links between movements and events		Times and, possibly, spatial positions of the events	Including information about the events in various displays as labels, symbols, marks, etc.
Links between two numeric (ordered) attributes or between one such attribute and linear or cyclic time	Data aggregation by intervals of attribute values or time	Counts of occurrences of value combinations for each pair of intervals	Binned scatterplot

## 4.6 Visualisation of patterns

In this section, we have discussed a number of visual and interactive techniques that can help users in detecting various types of patterns in movement data. A different approach is to design algorithms and software tools to perform automated search, or “mining”, for patterns. The visual and automatic methods of pattern detection can be used complementarily. The use of automatic methods requires the results to be presented to the user in a way allowing the user to interpret and evaluate them. In other words, the patterns need to be made perceptible to human mind. Symmetrically, the use of visual methods requires that the user be able to represent the patterns discovered so that they were perceptible to other people and to this user after some time. In both cases, adequate methods are needed for pattern *visualisation* (let us recall that the word “visualise” is defined in a dictionary as “to make perceptible to the mind or imagination”).

To our knowledge, the research on the visualisation of patterns extracted from data is currently in its infancy and consists mainly of a few ad hoc methods devised for the visualisation of particular types of data mining results. In the area of data visualisation, the researchers are focused on the task of enabling users to detect patterns and do not consider the problem of how these patterns can be documented. There is a more general research on knowledge visualisation (Tergan & Keller 2005), which is mainly conducted in the fields of knowledge management and education. Most methods suggested for knowledge visualisation are based on the use of node-link structures, or graphs. This includes semantic networks, also known as concept maps or cognitive maps, mind maps, argumentation maps, storyboards, etc. These visualisations may also incorporate multimedia displays or interactive links to such displays (Alpert 2005, Cañas et al. 2005).

Generally, graphs as quite powerful instruments for representing various relationships are widely used. In particular, graphs are used in the visualisation of some types of data mining results such as association rules and decision trees. Note that arrow symbols are paramount for visual communication; they are used multi-purposely to represent directions, movements, orders, relations, interactions, and so forth (Kurata & Egenhofer 2005). This makes node-link structures quite suitable for the visualisation of some types of patterns that can be extracted from movement data, in particular, temporally annotated sequential patterns (Giannotti et al. 2006), which may be related to locations or regions in space. An example of such a pattern is a frequently appearing sequence of places A, B, C with the transition time from A to B being  $t_1$  and the transition time from B to C being  $t_2$  ( $t_1$  and  $t_2$  may be average times or intervals). Nodes may be used to represent the places and links (arrows) may indicate the temporal order in which these places were visited. The arrow symbols may differ in width, colour, brightness, and/or texture to represent the transition times and other characteristics. Additionally, text labels may be used. The graph may be drawn on top of a map to allow the user to recognise the places, see their relative positions, and relate the patterns to various geographical information. Alternatively, the graph may contain interactive links to appropriate map displays where nodes refer to specific geographical locations or regions.

However, simultaneous visualisation of all sequential patterns extracted from movement data may be impracticable not only because of their potentially great number but also because of possible overlaps between the patterns when one and the same place appears in two or more patterns. Therefore, additional tools are required for navigation through the set of patterns and selection of patterns for more detailed examination and comparison. The user should be able to select subsets of the patterns according to various criteria, in particular, spatial (e.g. patterns involving place A or patterns where the movement direction is outwards from the centre) and temporal (e.g. patterns occurring in the morning).

Besides sequential patterns, node-link drawings superimposed on a map can also represent rules referring to specific places, for example:

- traffic\_jam (Pisa, 7:30AM)  $\Rightarrow$  traffic\_jam (Lucca, 8:30AM) or
- traffic\_jam (Pisa,  $t$ )  $\Rightarrow$  traffic\_jam (Lucca,  $t+1h$ )

However, this approach will not work for rules involving more general spatial concepts such as “city centre” and “outskirts”, “pedestrian area” and “major thoroughfare”, etc., which have no precise localisation and/or crisp boundaries, and hence cannot be adequately represented on a map display. Instead, such concepts can be represented verbally or symbolically, for example, with the use of the system of signs, the so-called “chorèmes”, suggested by Roger Brunet for the representation of spatial objects and relations (cited in Elzakker 2004). It seems reasonable to study which form, symbolic or verbal, is more effective and convenient for users.

While specific research on visualisation of patterns that can be found in movement data is yet to be done, the following general considerations may provide some guidance:

1. Node-link structures are widely used and are therefore familiar to users and easily understood. Such structures are well suited to representation of patterns involving various kinds of relationships. In particular, they can represent sequence patterns in movement data as well as correlation and dependency patterns.
2. Node-link structures may incorporate various media, in particular, maps.
3. The use of maps is reasonable when patterns refer to specific geographical locations or regions.
4. Besides techniques for pattern visualisation, it is necessary to design tools for navigation through the set of patterns and for management of the patterns, which includes filtering, re-arrangement, and establishing of links.

## 4.7 Conclusion

Current state-of-the-art methods and tools for visual and interactive exploration of movement data have significant limitations regarding the volumes of data they can be applied to. In this section, we have outlined a road map to developing methods for visual analysis of massive datasets, with numerous moving entities and long time series of measurements. The methods are based on data aggregation, which is performed prior to the visualisation. A number of technical problems need to be solved; in particular, effective linking between several displays presenting differently aggregated data.

The main goal of data exploration is detecting patterns and relationships in the data. We have considered the possible types of patterns an analyst may seek for in movement data. The role of interactive visual techniques is to allow the user to detect these patterns. We have also pointed out the need in tools for recording discovered patterns and in methods for the visualisation of patterns. Visualisation is necessary for a joint analysis of all detected patterns in order to gain an overall understanding of the data. This applies both to patterns detected by a human analyst and to patterns derived automatically, e.g. using data mining algorithms. Visualisation of patterns is also required when the analyst wishes to communicate his/her discoveries to others. Currently, the problem of pattern visualisation, in particular, visualisation of movement patterns, is far from being solved and requires further research efforts. Another unsolved problem is the problem of supporting knowledge synthesis, which has been indicated in section 3.7.

## References

1. Ahlberg, C., Williamson, C., and Shneiderman, B. (1992): Dynamic queries for information exploration: an implementation and evaluation. In: *Proceedings of ACM CHI'92*, ACM Press, New York, pp. 619–626
1. Alpert, S.R (2005): Comprehensive Mapping of Knowledge and Information Resources: The Case of Webster, in Tergan, S.-O., and Keller, T., editors: *Knowledge and Information Visualization: Searching for Synergies*, Springer-Verlag Berlin Heidelberg 2005, Germany, pp. 220-237
2. Andrienko, N. and Andrienko, G. (2006): *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach* (Berlin: Springer)

3. Brewer, C.A. (1994): Color use guidelines for mapping and visualization. In: *Visualization in Modern Cartography*, ed. by MacEachren, A.M., Fraser Taylor, D.R. (Elsevier, New York) pp 123–147
4. Cañas, A.J., Carff, R., Hill, G., Carvalho, M., Arguedas, M., Eskridge, T.C., Lott, J., and Carvajal, R. (2005): Concept Maps: Integrating Knowledge and Information Visualization, in Tergan, S.-O., and Keller, T., editors: *Knowledge and Information Visualization: Searching for Synergies*, Springer-Verlag Berlin Heidelberg 2005, Germany, pp. 205-219
5. Drecki, I., and Forer, P. (2000): Tourism in New Zealand - International Visitors on the Move (A1 Cartographic Plate); Tourism, Recreation Research and Education Centre (TRREC): Lincoln University, Lincoln.
6. Elzakker, van, C.P.J.M. (2004): *The use of maps in the exploration of geographic data*, doctor's dissertation, Utrecht : University of Utrecht (Netherlands Geographical Studies 326), pp. 202
7. Giannotti, F., Nanni, M., and Pedreschi, D. (2006): Efficient mining of temporally annotated sequences. In *Proc. of the SIAM Conference on Data Mining (SDM06)*, 2006.
8. Graz (2005): Mobile Landscapes: Graz in Real Time, a Web page by SENSEable City Laboratory of the Massachusetts Institute of Technology, <http://senseable.mit.edu/projects/graz/graz.htm>
9. Imfeld, S. (2000). Time, points and space: Analysis of wildlife data in GIS. Unpublished Dissertation, University of Zürich, Department of Geography, Zürich (<http://www.geo.unizh.ch/~imfeld/diss>)
10. Inselberg, A. (1985): “The plane with parallel coordinates”, *The Visual Computer*, 1 (1), 1985, pp.69-91
11. Kurata, Y. and Egenhofer, M. (2005): Semantics of simple arrow diagrams, In *AAAI Spring Symposium on Reasoning with Mental and External Diagram: Computational Modeling and Spatial Assistance*, March 2005, Stanford, CA (Menlo Park: AAAI Press), pp. 101-104.
12. Ratti, C., Sevtsuk, A., Huang, S., and Pailer, R. (2005): Mobile Landscapes: Graz in Real Time, *Proceedings of the 3rd Symposium on LBS & TeleCartography*, 28-30 November, Vienna, Austria, also available at the URL <http://senseable.mit.edu/papers/pdf/RattiSevtsukHuangPailer2005LBSVienna.pdf>
13. Tergan, S.-O., and Keller, T., editors (2005): *Knowledge and Information Visualization: Searching for Synergies*, Springer-Verlag Berlin Heidelberg 2005, Germany, pp. 385
14. Tobler, W. (2005): Display and Analysis of Migration Tables, [http://www.geog.ucsb.edu/~tobler/presentations/shows/A\\_Flow\\_talk.htm](http://www.geog.ucsb.edu/~tobler/presentations/shows/A_Flow_talk.htm)
15. Tufte, E.R. (1983): *The Visual Display of Quantitative Information*, Graphics Press, Cheshire CT
16. Tukey, J.W. (1977): *Exploratory Data Analysis*, Addison-Wesley, Reading MA, 1977

## 5. SPATIO-TEMPORAL REASONING

In this section we report the role of reasoning in the GeoPKDD process. We will show how the reasoning steps are pervasive in the discovering process. Furthermore, we will analyse the types of reasoning (deduction, induction, abduction) that occur at each step in order to create usable knowledge (insight).

### 5.1 The role of metaphors in reasoning to discover patterns

Tiezzi (2004) recently stated that knowledge of nature must come from a global and systemic view of patterns as well as from a study of the network of information joining the various forms of metaphors in time and space. Currently, a geographic knowledge discovery process is supported by computer-based environments that provide a global and systemic view of patterns by allowing users to be interactive and iterative, involving their visual thinking (perceptual-cognitive process) and automated information processing (computer-analytical process) with many decisions made by the users about how to fit models to, or how to determine patterns from data.

However, patterns do not explain a metaphor (Gendlin 1995). But on the contrary, it is the metaphor, and only after it makes sense, that an unknown set of patterns from a GKDD process can be interpreted and understood by an expert of an application domain. Why are metaphors important? Metaphors are artefacts of understanding, specifically understanding one kind of conceptual domain in terms of another. They are not just a pattern or a logical form. Johnson (1987) proposes metaphors as a “concrete and dynamic, embodied imaginative schemata”, which are surely not just logical patterns, images, or diagrams. Moreover, Lakoff (1987) argues that metaphors are something “non-propositional”, which should not be thought of as if they were commonalities, classes, structures, or image schemata, although we might be interested to formulate those.

In a GKDD context, metaphors will help the comprehension of what makes one pattern structurally and meaningfully different from another. Ideally, metaphors would be constructed in the domain of the expert, having a high-level or abstract reason that makes sense within a specific problem context. They will lead to the “discovery” of higher level entities, relationships or processes within some application domain of interest. Poore and Chrisman (2005) drawn attention to the fact that information metaphors do not relate directly to reality, but instead, they are more successful when they can have the effect of structuring reality to fit. For example, in GIS, the landscape-as-layer metaphor has structured the landscape into a set of layers, and nowadays, software packages encourage organisations to collect their data according to layers. Although researchers have proposed new information metaphors such as objects (Wachowicz 1999, Peuquet 2002) and agents (Deadman 1999, Ligtenberg et al. 2004), numerous practitioners are locked into the layer form of reasoning.

What will be the information metaphors of a GKDD process? In particular, the movement-as-trajectories metaphor is already being used to structure the history of the past and current locations of mobile entities. Pfoer and Jensen (2001) employ the metaphor of trajectory as polylines consisting of connected line segments, which can be grouped accordingly to two movement scenarios, termed as unconstrained movement (vessels at sea) and constrained movement (cars and pedestrians). Another example is given by the account of the movement-as-balance metaphor that provides an interpretive artefact of a balance scale for analysing the traffic flow of cars in a presence of transportation problems (Richmond 1998). A transportation system that operates under the conditions of free-flow will be in balance. On the contrary, if the components such as road and rail are in wrong proportions, they are out of balance, having as a result a traffic that is unbearable with a need to remove the load from the roads.

In a GKDD process, the challenge relies of mapping the discovered patterns with metaphors such as movement-as-trajectories or movement-as-balance. For example, how discovered patterns, such as clusters or association rules, can be understood as representing those patterns occurred in

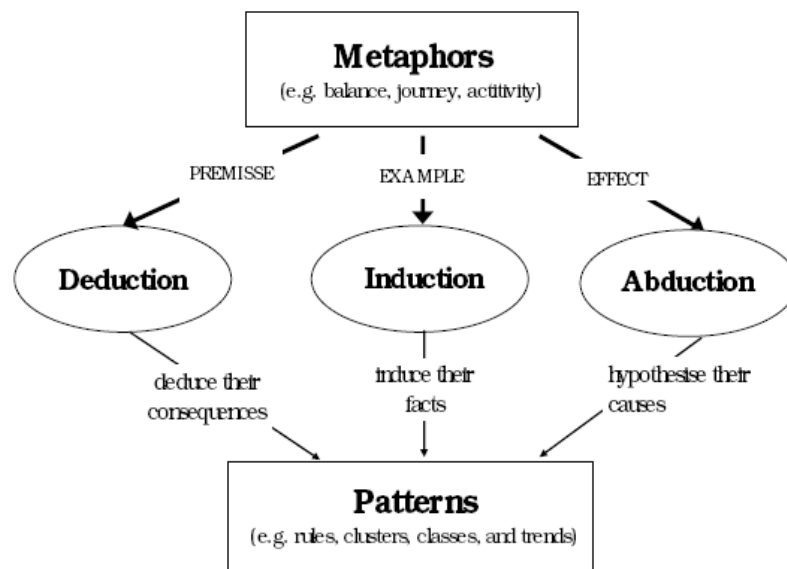
low density fringe growth in urban developments that can show the reduced effectiveness of public transport and increased reliance on the private car. It is still to be proven that information metaphors will enhance the likelihood that an expert will “see” not only the movement patterns, but will understand their meaning as well. However, it is already clear that information metaphors can generate a chain of commonalities and differences, not a single pattern. A better account of the role of information metaphors in a GKDD process would allow the experts to form and operate on concepts, not on GKDD steps.

The complex relationship between information metaphors and GKDD must remain a topic for further research. In this chapter, we outline our first effort on understanding such a relationship by looking at the reasoning paradigm. Reasoning is the ability of experts to form and operate on concepts in abstractions (i.e. metaphors). In our research, reasoning constitutes the “logic of discovery” as already proposed by the philosopher and logician C.S. Peirce (1878). Therefore, three different approaches have been distinguished according to the type of reasoning task. They are:

- **Deduction:** a reasoning task by which one infers a consequence from a set of patterns. The consequences are drawn from the general (patterns) down to the specific (metaphor). In this case, the metaphor is already known by the experts, and it usually forms the empirical basis of a GKDD process, because the relationship between the metaphor and the patterns can be verified straightforwardly. The metaphor C is known, if the findings from a GKDD process reveal pattern A and pattern B have occurred. One example is the movement-as-journey metaphor, where movement is conceptualised as a journey that begins and ends at home, and can include one or more stops. The GKDD process might reveal cluster A showing that people make more journeys and spend more time travelling on weekdays, rather than weekends. Other cluster B might reveal the patterns of a large neighbourhood shopping centre, where people spend less time travelling and travel less kilometers on weekdays. The journey metaphor underlies the expert’s understanding of people’s behaviour based on properties such as travel distance (from home to destination) and cyclic time (e.g. weekdays). Movement metaphors are needed for deductive reasoning since they underlie an assumption that represents the expert’s understanding of the patterns revealed by a GKDD process.
- **Induction:** a reasoning task by which one infers a generalisation from a set of patterns. It implies reasoning from detailed facts (examples) to general principles (conclusions). This approach of reasoning supports “learning by example”, where the example is the metaphor which contains more information than what was contained in the patterns themselves. The challenge is to uncover what metaphors can explain the causes for the observed patterns. Movement metaphors are needed for inductive reasoning since they rise from generalisation. For example, the movement-as-activity metaphor explains how people organise their movement in a geographic environment by defining a sequence of activities that comprise a person’s existence at any temporal scale (daily, monthly, lifetime) and social extent. For example, after the discovery of distinct linear patterns of a set of trajectories between mornings, afternoons and evenings, it would be possible to infer that leisure and work activities are the most common activities conducted outside the home, closely followed by grocery shopping and bring/get activities such as bringing/getting a child to/from school. If activity is used to explain the linear patterns extracted via inductive reasoning, then a generalised form of “activities” metaphors needs to be a priori known in order to explain the discovered patterns in the forms of rules, clusters, or classes.
- **Abduction:** a reasoning task by which one infers to the best cause for the occurrence of a set of patterns. An explanation is a relation between one or more hypothesis and the patterns they account for. It is flexible because it is not restricted to using existing metaphors of an application domain, but is instead free to create new metaphors that help to explain the patterns presented. If some theory states that if pattern A causes pattern B, and the GKDD process reveal the occurrence of pattern B, then by abduction an expert can infer A. However, data

mining methods do not operate in this way, most either attempt to locate pre-defined patterns (deduction) or else learn from examples that are presented or selected (induction). Ideally, the new metaphor would be unravelled by the expert, mapping the discovered patterns into a new hypothesis in an application domain.

Figure 5.1 illustrates the relationship between metaphors and the reasoning tasks of a GKDD process. It is important to emphasise the role of metaphors in clarifying, naming, and structuring what might otherwise be vague and inapplicable patterns within the context of an application domain. Therefore, reasoning is an integral part of the discovery process, and we propose that discovery and reasoning should be studied together. This will facilitate not only the extraction of patterns from very large databases, but also to infer knowledge from these patterns.



**Fig. 5.1.** The role of metaphors in different reasoning tasks

Previous research on spatio-temporal reasoning has primarily dealt with hierarchical metaphors based on static and well-defined, closed environments, and unfortunately, without having them associated to a geographic knowledge discovery process. Some examples include the spatio-temporal granularity description of spatial regions (Stell 2003), and the concept of perceptual hierarchical spatial units for representing people behaviour in urban environments (Reginster and Edwards 2001). The dominant view has been that these representations are hierarchically organised (McNamara, Hardy and Hirtle 1989), and the locations, objects, circumstances, and factors may be perceived and understood in separate representations which are required accordingly to a particular situation or task (Huttenlocher et al. 1991). Most of the studies have been conducted at a specific scale level by building scenarios on the variations in urban forms characteristics such as urban morphology, transportation network, availability of facilities, and density of a city and the relative location of neighbourhoods. The reasoning task involved has been of deriving the most likely explanations of the known facts and assumptions about urban form characteristics, and their influence on travel behaviour. Such explanations have usually pointed out to four major factors that have explained such an influence on a specific scale. They are: density of development, land use mix, transport networks, and layout development (Stead and Marshall 2001).

## 5.2 The multi-tier ontological framework

A GKDD process constructed from a multi-tier ontological perspective aims to integrate different reasoning tasks in a unified system by mapping the complex relationship between movement metaphors and patterns. Knowledge discovery is not a trivial process and it requires the examination of metaphors of characteristics, similarities and differences, interrelations, behaviour and evolution of what experts believe the world is like. This will lead to uncovering new and innovative hypothesis of distributions, patterns and structures across very large databases. Therefore, these metaphors will not rely on similar reasoning backgrounds but will be derived from the integration of different inference modes (i.e. abduction, induction and deduction).

This section describes the multi-tier ontological framework which has been developed from two previous fundamental research works. First, the work on a set of tiers of ontology previously proposed by Frank (2003) for defining consistency constraints, data interoperability, and more recently data quality in Geographical Information Systems (Navratil and Frank 2006). Second, our multi-tier framework has been largely based on the three “spaces” paradigm that has been proposed by Ernst Cassirer (1874-1945), a philosopher of the Marburg school, who describes a learning process as a truly dynamic activity of the mind of the human experience of spaces and time. The spaces are from an observed space through sensors and senses (interpretation), to an abstract model of space (guide), to a higher level of concepts incorporated in an internal and cognitive space (synthesis).

Our aim is to describe a GKDD process using ontological tiers that will provide the common base for the organisation of different nature and sources of knowledge of the movement metaphors used by experts of application domains. The tiers also establish the movement metaphors for the integration of different reasoning tasks in a unified system. This can only be achieved since each Tier instantiates the metaphors of the previous tier, enabling the understanding of interesting, meaningful, and previously unknown patterns. Figure 5.2 illustrates the proposed multi-tier ontological framework, in which a successive set of tiers refine the steps of GKDD process, which are named as sampling, relating to a geographic context, discovering patterns, generating new insights, and confronting them with previous background knowledge. Therefore, five Tiers have been defined as Reality Space; Positioning Space; Geographic Space; Social Space; Cognitive Space.

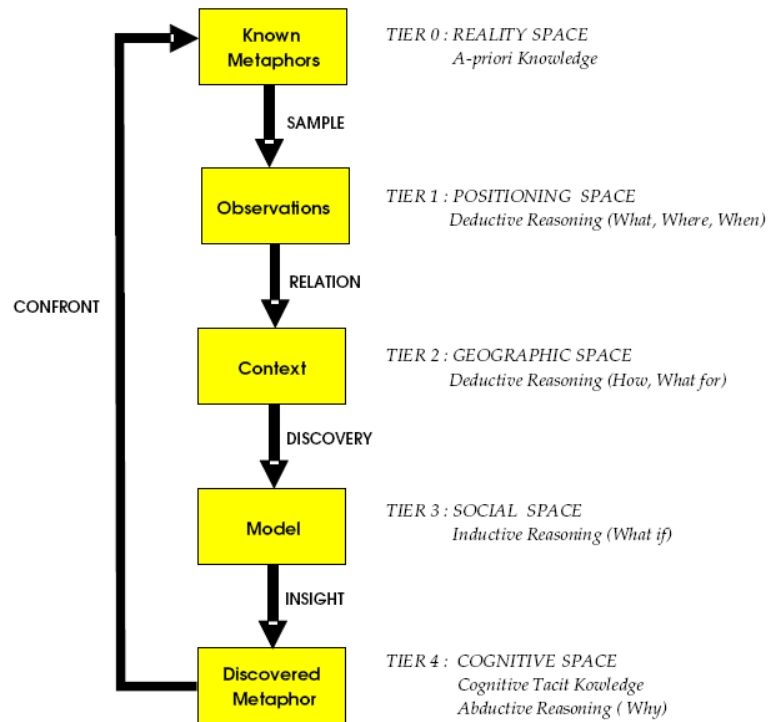


Fig. 5.2. The multi-tier privacy-aware geographic knowledge discovery process.

From a privacy perspective, the multi-tier framework allows a number of legal frameworks, often specific to application domains, to be adhered to throughout the GKDD process. Therefore, one of the initial steps of this process requires understanding which of these laws or regulations apply to each one of the Tiers. Often such frameworks require the sensor carriers and domain experts to get the consent of those whose data is being collected to the primary or secondary use of that data. This is particularly challenging in a GKDD process, since the metaphors and the context of their use may not yet be determined during the Tier 1 when the collection of data is carried out. Furthermore, the other Tiers may also have privacy constraints on the results of the GKDD process that are unknown by the data miner as well as the expert of an application domain.

### 5.2.1 Tier 0 – The Reality Space

Tier 0 of the ontology represents the “Reality Space”, which recognises the existence of a known-world as a four dimensional continuous field in space and time. Usually, natural language is used to formalise the background knowledge that is derived from metaphors formulated by experts within an application domain. The process of geographic knowledge discovery may use this type of knowledge for generating a-priori knowledge as predetermined hypothesis, training examples or rules. Several known movement metaphors are currently being used, including the movement-as-journey metaphor already mentioned in the previous section. The existing a-priori knowledge might formulate that one or two journeys on a day are most common. Over three quarters of all journeys is usually a single-stop tour, while combining more than 3 stops in a journey is very rare. In contrast, the movement-as activity metaphor also mentioned in the previous section can generate a-priori knowledge statements such as most out-of-home activities have a considerable duration. More than 50% of all-out-of home activities take more than an hour and over 30% take even more than two hours (Axhausen and Gärling 1992).

In Tier 0, it is important to establish whether there are any privacy concerns from any of the sensor carriers and the experts of an application domain. This means that it is necessary to define a level of privacy according to who are the sensor carriers from whom data will be collected and

who are the involved experts who will define the purposes of collecting these data. In the case of applications for transportation management, the sensor carriers might be those traveling from home to work, and the experts might be the company managers who have privacy goals towards the collected data. Company managers may not want it to be known where their employees travel during work hours, since this could point out to information about the activities of that company and those who are interested in using the data, for example, the supermarkets in the area may be interested in the trajectories relevant for better advertising. Once the stakeholders are identified in Tier 0, it is also necessary to identify what their privacy requirements are, which could be stated in terms of hierarchical levels of privacy.

### 5.2.2 Tier 1 – The Positioning Space

The Tier 1 describes the “Positioning Space” that contains the observations of the four dimensional continuous field in space over time. Observations are measurement values at every point in space and time, based on some measurement scale, which may be quantitative or qualitative. Besides, observations are always marked by some degree of uncertainty, which depends on the type of sensors being used for collecting the location and movement information of mobile entities, such as X,Y,Z coordinates, speed, and time. They can be navigation sensors (e.g. GPS, INS, MEMS sensors, digital compasses, etc.), remote sensing sensors (e.g. frame-based cameras, thermal cameras, laser scanners, etc.), and wireless technologies.

In this Tier, the movement metaphors can be used to infer some empirical knowledge from the discovered patterns, such as density clusters of points in space. It is important to point out that we have only a set of observations of a finite sequence of time-referenced locations, represented by points where the movement of an entity starts at  $t_0$  and ends at  $t_{end}$  in space. We do not have a trajectory representation of these points yet. However, it is possible to distinguish the observations according to four point representations. We distinguish among:

- **Stop:** a cluster of points that represent stops with a very short duration of some minutes due to traffic light or stop signs.
- **Stop over:** a cluster of points that represent a change of speed. For example, a road accident.
- **Short stay:** a cluster of points that represent stays with a short duration of some hours due to an activity such as working, shopping or leisure.
- **Long stay:** as cluster of points that represent stays with a long duration of several hours that will correspond to the sensor device being switch off or being at home.

The reasoning task consists of allowing one to infer a consequence from a set of point patterns. The consequences are drawn from these point patterns down to a specific metaphor such as, for example, the movement-as-urban forms metaphor. Human behaviour is constraint by urban forms, such as urban morphology, transportation network, availability of facilities, and density of a city and the relative location of neighbourhoods. The movement-as urban forms metaphor is needed to provide the premises for inferring the knowledge about the point patterns generated at this Tier, according for example, the shape characteristics of urban morphologies such as radial, linear, concentric, and grid. Figure 5.3 illustrates some examples of these urban forms, and their associated shape characteristics.



Fig. 5.3. Examples of possible urban forms as a movement metaphor:

- (1) Ring network in a concentric city: concentric pattern
- (2) Radial network in a lob city: radial pattern
- (3) Linear poly-nuclear city: linear concentric pattern
- (4) Concentric poly-nuclear city: circular concentric pattern
- (5) Linear network in a linear city: linear pattern
- (6) Grid city: square concentric pattern

In the Real-Time Graz experiment in Austria, observations of cell phone usage have been collected through the city based on a location system where the movement of the cell phones was recorded and tracked with the agreement of the customers (Ratti 2004). Figures 4.5 and 4.12 illustrate the visual density clusters found after a 24 hour experiment in the city of Graz. It is already possible to realise the important role of a metaphor such as movement-as-urban forms in order to infer some knowledge about the clusters found of this experiment. In this example, it is possible to visually identify the circular concentric patterns representing possibly a concentric poly-nuclear city, with some linear poly-nuclear patterns as well. It is also important to point out that in this example; the trajectory representation does not exist yet.

In terms of privacy, the main issue is to make sure the granularity of data collection is in according to the respect to the privacy requirements of the sensor carriers. The domain experts can also make sure that the granularity of positioning data set is appropriate for the needs of the application domain and it complies with the privacy requirements of the different sensor carriers.

### **5.2.3 Tier 2 – The Geographic Space**

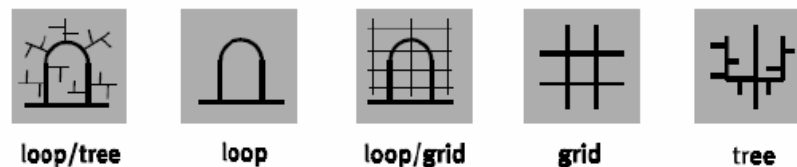
Tier 2 represents the “Geographic Space” where our cognitive system from an array of properties values, is capable of forming trajectories and reasoning about them. In geometrical terms, the movement of an entity is termed as a trajectory (we will use “movement” and “trajectory” interchangeably as already proposed by Pfoser and Jensen (2004)). From the movement metaphors and representations already defined in the previous Tier 1, a trajectory is therefore defined in this Tier as any polyline between stops, stop overs, short stays, and/or long stays. Moreover, in this Tier is where the privacy requirements of the sensor carriers need to be implemented using security constraints. If there are sensor carriers who want their trajectories to be unobservable or who want to remain anonymous, then the necessary steps need to be taken here by applying different trajectory privacy preserving methods such as cloaking and mixes.

The deductive reasoning task is characterised by inferring descriptive knowledge such as the trajectory characteristics (e.g. space and time), the geographic environment where the trajectory occurs (i.e. landscape), the topological relations between the trajectories, and the association between the trajectories and specific features of a landscape. Such information can be a set of properties which are associated to individual trajectories themselves, or a pre-defined group of trajectories. The overall goal is to help the experts to deduce the consequences for the existence of linear patterns of the movement of the trajectories. A set of movement metaphors is necessary to be defined by using some kind of classification scheme, set of association rules or clustering. For example, the aim might be to discover linear clusters which may be understood by which of these categories; allowing the experts to generate one or more internal theories to explain the discovered linear clusters of trajectories.

Currently, the main metaphor being used at this ontological level is accessibility, which can be obtained by the calculation of trajectory distances (lengths). This is carried out by using the average distances between zone centroides (regional scale) or using the distance between the origin and destination zone centroides (city scale). Depending on the size of zones, the actual trajectory distance may be significantly different to the distance calculated using average centroid distances (Spence and Frost 1995, Banister et al. 1997, Troy 1992). The calculations also do not account for the configuration of the transport network in order to establish the actual route distances. In fact,

they are only based on straight-line distances between origin and destination zones (Stead and Marshall 2001).

However, the GKDD process is entirely propositional and different movement metaphors need to be taken under consideration at this ontological level. For example, the metaphors of movement-as-urban form and movement-as-accessibility can be used to deduce the consequences of the linear patterns from the trajectories based on the generalisation of the trajectories using the transportation networks such as the types of streets in an urban area (Figure 5.4). Accessibility is constraint by urban forms such as the transportation network. For example, the tree street patterns usually impede the movement of people on reaching a destination; meanwhile grid street patterns facilitate the fast reaching to a particular destination. A GKDD process might provide, for example, the data mining query mechanism necessary to discover an anomaly in the trajectories corresponding to different types of local streets; the similarities and dissimilarities among the trajectories accordingly to the different characteristics of street types; and finally, discover point clusters of non-movement among trajectories and their association with the type of local street.



**Fig. 5.4.** Examples of possible types of local streets in an urban area.

Previous initiatives on gathering information about the trajectories of people at the street level include the experiment undertaken by the Waag Society in which a mobile device was provided to citizens willing to participate in generating information about their movement behaviour in the city of Amsterdam, the Netherlands. Some drawn conclusions from this experiment point out the transport, modality, the location of home, and street patterns as the main accessibility factors on determining the trajectory taken by each individual.

### 5.2.4 Tier 3 – The Social Space

The Tier 3 encompasses the model that underlies our daily trajectories and their fundamental relations with human activities. Traditional spatial planning theory usually considers the geographic environment as a space where human activities take place and represent the geographic environment according to the goal of a spatial planning. For example, if the national government develops its spatial policies for a country it requires a representation of the geographic environment that contains for example the cities, main infrastructure, population densities, nature areas etc, flows of people and goods. Municipalities on the other hand, developing their detail spatial policy for their cities use less abstract representations of the geographic environment. They need detailed, high level information about the individual functions of the buildings, detail infrastructure, and social compositions of different neighbourhoods. The information is usually described as in terms of transportation modalities (e.g. car or public transport), commuting time or distance, spatial distribution of jobs and housing locations, total vehicle miles travelled, average trip lengths, and congestion on links and intersections. Finally, from socio-economic statistics, information can be obtained about the geographic environment, such as income and education of a neighbourhood.

The above examples show that the planners use various metaphors for the geographic environment, depending on the context of the spatial planning. These metaphors are, however, currently based on mostly static models of activities. Relations between representations and activities are based on assumptions, spatial-analysis or activity-based models. However, the same

geographic environment should also be considered as the result of movement patterns of people represented by their invisible footprints of trajectories on the landscape. Pulselli (Pulselli et al. 2005) has already pointed out that although positioning data sets of mobile entities are becoming increasingly available, surprisingly enough, they have not been used to describe the social and spatial systems.

A social system consists of individuals, groups and organisations that maintain relations through intentional (cooperative) activities based upon a more or less common set of rules, norms and values and act within the boundaries of the institutions that are derived from it (Kleefmann, 1984). The spatial system is composed of biotic and a-biotic components, processes that alter these components and relations between them. An important difference between social and spatial systems is that the latter has been mostly described in geographic terms while the first has not. One exception is found in the work of Hägerstrand (1970), where three types of constraints have been formalised to represent the location of trajectories according to the human activities in both spatial and social systems. They are capability constraints (limit of activities of individuals due to their physical capabilities and or resources), coupling constraints (constrain where, when, and for how long and individual can join others to produce, transact, and consume), and authority constraints (impose certain conditions of an individual's accessibility).

Therefore, it is important to realise that a trajectory takes place in a social-spatial system. The main metaphor used at this Tier level is movement-as-activity. The socio-spatial organisation concept defines social-activities in a spatial perspective and can be used to analyse the interactions between social developments and the spatial system (Wisserhof, 1996). The spatial system and the social system are strongly intertwined and should not be analysed separately. There is a structural coupling between the geographic environment and the social system that acts upon it. Processes in the social system such as the economic, political or cultural subsystems have spatial consequences and vice versa.

In this Tier, *t* requires that some sort of target must already have been identified, and the task becomes one of uncovering “what if” scenarios to explain the trajectory patterns within a social context. For example, instead of finding the best location for a supermarket based on the proximity of objects on the landscape, the problem becomes about finding the best location based on the patterns of the trajectories of people, which in turn, suggest that the human activity on a landscape is potentially more complex and probabilistic.

Consequently, induction is most commonly applied at the Tier 3 ontological level in order to determine a model that can best “fit” the trajectories. Some examples are given below:

1. Discover patterns that explain the occurrence of a certain activity (e.g. shopping, recreation).
2. Discover the dependencies between different characteristics of activities.
3. Discover the activities subsets and time periods with the corresponding patterns.
4. Detect the occurrence of an unexpected activity.

The Tier 3 is where the “classification anonymity” or the “categorisation anonymity” requirements of the sensor carriers need to be guaranteed. The data miner and the domain expert need to make sure that exact rules which contain the complete data population do not breach any of the “group privacy” requirements. Further, there may be privacy requirements in the form of:

- If the size of an identified group is less than 10% of the complete population, then a sensor carrier wants to be unobservable, or it should not be inferred that the sensor carrier belong to this group.

Or:

- If an unexpected activity (example number 4 above) is detected, and this contains information about clearly identifiable locations, small set of trajectories or small groups of people, then this information should either not be released, or only accessible to trusted parties.

In Tier 3, the experts need to be aware of sensor carrier's privacy requirements about inferred information and the context in which this information may be used. Inferences about the movement on trucks on highways and city centres may be meaningful for traffic balancing, but may be a threat to companies whose weaknesses on product distribution may then become inferable by other companies.

### **5.2.5 Tier 4 – The Cognitive Space**

The Tier 4 represents the "Cognitive Space" where the goal of a geographic knowledge discovery process is to gain knowledge through abductive reasoning that can function in the absence of pre-determined hypotheses, training examples, or rules. Abduction is flexible because it is not restricted to using existing knowledge of pre-defined patterns (deduction) or else learns from examples that are presented or selected (induction), but instead free to create new structures that help to explain the patterns of a data mining process. Cognitive tacit knowledge is a non-linguistic non-numerical form of knowledge that is highly personal and context specific, and deeply rooted in individual experiences, ideas, values, and emotions. It refers to ingrained schema, beliefs, and mental models that are taken for granted (Nonaka and Takeuchi 1995).

Therefore, it is important to point out the difference between tacit and implicit knowledge. Implicit knowledge is something experts might know, but not wish to express while tacit knowledge is something that experts know but cannot express, it is personal, difficult to convey, and which does not easily express itself in the formality of language. Searle (1995) argues that cognitive tacit knowledge is not a form of knowledge (such as beliefs, theories, and empirical hypothesis) but rather the preconditions of forming an individual's background knowledge. This raises the possibility that at least some metaphors of background knowledge can be confronted with the ones of cognitive tacit knowledge, which implicates that the features of the world are not independent of the mind.

## **5.3 Conclusions**

This section introduces a geographic knowledge discovery process; in which the primary goal of identifying, associating, and understanding patterns is used to infer the location, identity, and relationships among mobile entities, and their respective trajectories in a spatial environment. In this case, the different types of inferences play a different role accordingly to what a domain expert want to infer, that is, the location, changes, properties, identity, or relationship among the appropriate metaphors. It is the metaphor, and only after it makes sense that an unknown set of patterns can be interpreted and understood by a domain expert. Basically, three modes of reasoning are presented using a multi-tier ontological framework. They are deductive, inductive, and abductive modes of reasoning.

In the deductive mode of reasoning, the geographic knowledge discovery process involves the search for common attributes among a set of mobile trajectories, and then the arrangement of these trajectories into classes, clusters, or patterns according to a meaningful metaphor. The focus is on applying statistical approaches (probability distributions, hypothesis generation, model estimation and scoring) for exploring classes, clusters or patterns from a data set. In the inductive mode of reasoning, the geographic knowledge discovery process is based on learning as the reduction of uncertainty in knowledge. Several techniques have been developed, such as rule induction, neural networks, genetic algorithms, case-based learning and analytical learning (theorem proving). Many techniques partition the target data set into as many regions as there are classes by using a function, for example, a posterior probability or linear discriminate functions. These techniques provide a data fit, in the sense that the main goal is to generate derived knowledge describing the data, often called concept hierarchies.

In the abductive mode of reasoning, the importance of cognitive tacit knowledge needs to be considered. Will the information have the same meaning and weight (in terms of privacy) if the

patterns are used in contexts other than it was meant to be? In this case, the value of the discovered knowledge is judged and the decision is taken on its role in making decisions for the application domains such as transport management, spatial planning and geomarketing. It might turn out that final decisions made are not in line with patterns suggested by the knowledge discovery process. The political, economic, or social realities of the decision-making process are sometimes prevalent above the rational knowledge inferred from a geographic knowledge discovery process. Questions like, why do people choose certain modality of transportations at certain times of the day, or why are certain transportation modalities more present in area a than in area b; are some examples where new metaphors could explain the relations between certain movement behaviour and the characteristics of a geographic environment.

The inevitable challenge facing the research community at the moment is directed toward a more complete integration of these modes of reasoning and their association to movement metaphors within a geographic knowledge discovery process.

As we have seen, the role of reasoning is pervasive in the spatio-temporal discovery process, driving each step in producing new knowledge. It is interesting to highlight how this vision meets the Data Mining Query Language definition (See Report on DMQL). Indeed, an interesting direction for further studies is to define the DMQL task inside this framework.

A first proposal is defining a DMQL as capable of expressing some tiers and relative reasoning steps. An example is shown in the following picture (Fig.5.5), where the red box highlights a possible role of DMQL in the ontological framework.

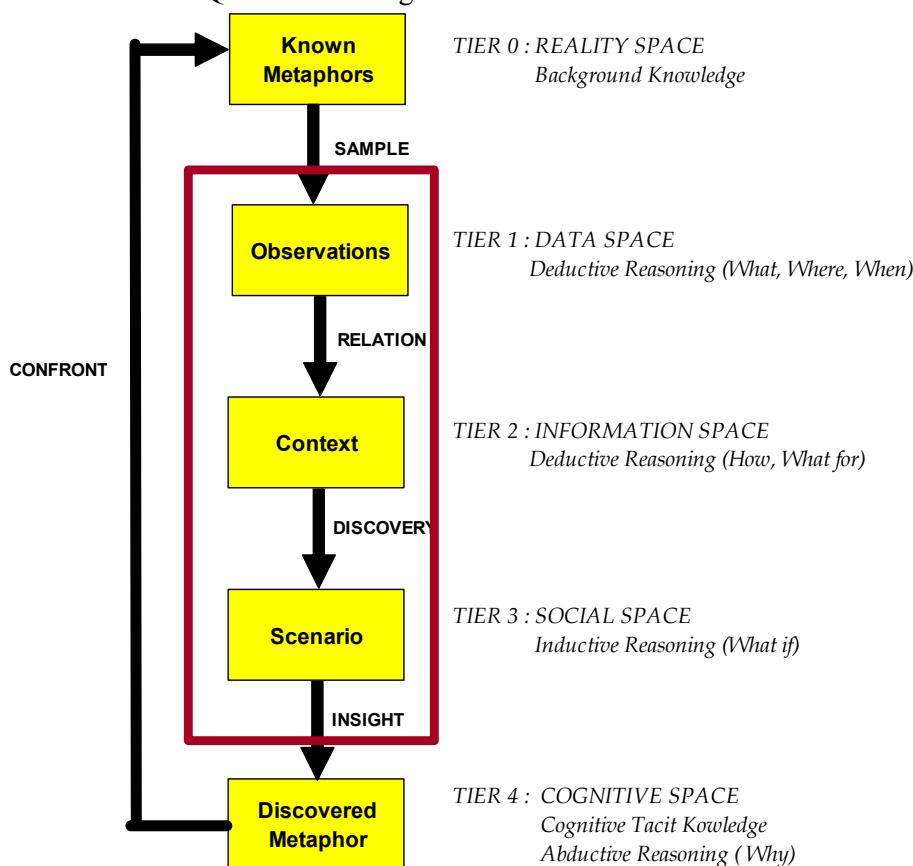


Fig. 5.5. A possible role of DMQL in the ontological framework is outlined by the red box.

## References

1. Axhausen K.W. and Gärling, T. (1992). Activity-Based Approaches to Travel Analysis: Conceptual Frameworks, Models and Research Problems. *Transport Reviews*, 12, pp.324-341.
2. Banister, D.; Watson, S. and Wood C. (1997) Sustainable cities, transport, energy, and urban form. *Environment and Planning B: Planning and Design*, 24(1), pp. 125-143.
3. Deadman, P. (1999). Modelling individual behaviour and group performance in an intelligent agent-based simulation of the tragedy of the commons. *Journal of Environment Management*, 56,159–172.
4. Frank, A. U. (2003). Ontology for spatio-temporal databases. In: *Spatio-temporal databases: The Chorochronos approach*. (Manolis Koubarakis and et al. eds). Berlin: Springer-Verlag, pp. 9-78.
5. Gendlin, E.T. (1995). Crossing and dipping: some terms for approaching the interface between natural understanding and logical formulation. *Minds and Machines*, 5, pp.547-560.
6. Hägerstrand, T., 1970, What about people in regional science? *Papers of the Regional Science Association*, 24, 7–21.
7. Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: prototype effects in estimating spatial location. *Psychological Review*, 98, pp.352–376.
8. Johnson, M. (1987). *The body in the mind*. U. Chicago Press.
9. Kleefman F. (1984). *Planning als zoekinstrument*. VUGA, 's Gravenhage.
10. Lakoff, G. (1987). *Women, fire, and dangerous things*. U. Chicago Press.
11. Ligtenberg, A., Wachowicz, M. Bregt, A., Beulens, A. and Kettenis, D.L. (2004). A design and application of a multi-agent system for simulation of multi-actor spatial planning. *Journal of Environment Management*, 72(2004), pp.43-55.
12. McNamara, T. P., Hardy, J. K., & Hirtle, S. C. (1989). Subjective hierarchies in spatial memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, pp.211–227.
13. Navratil, G. and Frank, A.U. (2006). Data Quality for Spatial Planning - An Ontological View. *CORP 2006 & Geomultimedia05*, Vienna, Austria.
14. Nonaka, I. and Takeuchi, H. (1995). *The knowledge creating company*. Oxford University Press, Inc.
15. Peuquet, D.J. (2002). *Representations of Space and Time*. The Guilford Press.
16. Pfoser, D. and Jensen, C.S. (2001). Querying the trajectories of on-line mobile objects.
17. Poore, B.S. and Chrisman, N.R. (2006). Order from Noise: Toward a social theory of geographic information. *Annals of the Association of American Geographers*, 96(3), pp. 508-523.
18. Pulselli, R.M., Pulselli, F.M., Ratti, C. and Tizzi, E. (2005). Dissipative structures for understanding cities: Resource flows and mobility patterns. *Proceedings of the First International Conference on built Environment Complexity, BECON 2005*, Liverpool, England, pp. 271- 279.
19. Ratti, C. (2004). Space syntax: some inconsistencies. *Environmental and Planning B: Planning and Design*, 31, pp.487-499.
20. Reginster, I. and Edwards, G. (2001). The concept and implementation of perceptual regions as hierarchical spatial units for evaluating environmental sensitivity. *URISA Journal*, 13(1), pp.5-16.
21. Richmond, J.E.D. (1998). Simplicity and complexity in design for transportation systems and urban forms. *Journal of Planning Education and Research*, 17, pp.220-230
22. Searle, J. (1995). *The construction of social reality*. New York, Free Press.

23. Spence, N. and Frost, M. (1995) Work travel responses to changing workplaces and changing residences. In: *Cities in Competition. Productive and sustainable cities for the 21st century*, Brotchie, J.; Batty, M.; Blakely, E.; Hall, P. and Newton, P. (eds.) Longman Australia Pty Ltd., Melbourne. pp. 359-381.
24. Stead, D. and Marshall, S. (2001). The relationships between urban form and travel patterns: An international review and evaluation. *European Journal of Transport and Infrastructure Research*, 1(2), pp.113-141.
25. Stell, J.G. (2003). Qualitative extents for spatio-temporal granularity. *Spatial Cognition and Computation*, 3(2&3), pp. 199-136.
26. Tiezzi, E. (2004). *Beauty and science*, WIT Press, Southampton, UK
27. Troy, P.N. (1992) Let's look at that again. *Urban Policy and Research*, 10(1), pp.41-49.
28. Wachowicz, M. (1999). *Object-Oriented Design for Temporal GIS*. Taylor and Francis.
29. Wissershof, J.: 1996, *Landelijk gebied in onderzoek : ontwikkeling en toepassing van een interdisciplinair conceptueel kader*, KU Nijmegen, Nijmegen.

## 6. CONCLUSION

This document presents the current status of an ongoing research, which is conducted in workpackage WP3 of the project GeoPKDD. The major progress that has been achieved thus far is a systematic design of a toolkit that could support visual exploration and analysis of massive collections of movement data.

When data are massive, it is insufficient to use only visual displays but it is necessary to involve the database technologies and computational methods of data processing and analysis. Still, the visualisation plays the central role since it allows the innate perceptual and cognitive capabilities and background knowledge of a human analyst to be utilised in the process of data exploration and analysis. These capabilities and knowledge cannot be replaced by purely machine processing. Hence, the combination of the visualisation with computer operations makes a ground for a truly synergetic work of human and computer.

In order to find out what set of methods and techniques could appropriately support the work of an analyst with a large set of movement data, we first considered the general structure of movement data. On this basis, we defined the types of patterns that could be detected in movement data and between movement data and data about other phenomena. Then, we reasoned out what kinds of data transformations, computations, and visualisations could allow the analyst to detect these pattern types. We did not try to invent any absolutely new visualisation or data processing techniques but referred to existing approaches, techniques, and technologies, which can be quite serviceable if properly integrated and made accessible to analysts. However, at the present moment, we are not aware of any existing toolkits that could comprehensively support visual exploration and analysis of massive movement data. We hope that this study can provide useful guidelines for developers of such toolkits. Prototype implementation of some of the suggested tools will be done within GeoPKDD.

Some challenging research problems related to WP3 are yet far from being solved. One of them is visualisation of patterns extracted from movement data either through visual analysis or through application of automatic methods (data mining). We have outlined some approaches, which we deem to be promising. Visualisation of patterns is a part of a more general problem of supporting knowledge synthesis. Another part of this problem is supporting analytical reasoning, when the analyst establishes essential links between various phenomena and between different aspects of the same phenomenon. We have considered three modes of reasoning, deductive, inductive, and abductive, using a multi-tier ontological framework. We have also analysed the role of metaphors, which help the comprehension of what makes one pattern structurally and meaningfully different from another. These modes of reasoning and their association to movement metaphors need to be effectively supported in the geographic knowledge discovery process. Further research is required to find ways to achieve this.