City Research Online

City St George's, University of London

Citation: Andrienko, G., Andrienko, N., Fuchs, G. & Wood, J. (2017). Revealing Patterns and Trends of Mass Mobility through Spatial and Temporal Abstraction of Origin-Destination Movement Data. IEEE Transactions on Visualization and Computer Graphics, 23(9), pp. 2120-2136. doi: 10.1109/tvcg.2016.2616404

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/15488/

Link to published version: https://doi.org/10.1109/tvcg.2016.2616404

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

Revealing Patterns and Trends of Mass Mobility through Spatial and Temporal Abstraction of Origin-Destination Movement Data

Gennady Andrienko, Natalia Andrienko, Georg Fuchs, and Jo Wood

Abstract— Origin-destination (OD) movement data describe moves or trips between spatial locations by specifying the origins, destinations, start, and end times, but not the routes travelled. For studying the spatio-temporal patterns and trends of mass mobility, individual OD moves of many people are aggregated into flows (collective moves) by time intervals. Time-variant flow data pose two difficult challenges for visualization and analysis. First, flows may connect arbitrary locations (not only neighbors), thus making a graph with numerous edge intersections, which is hard to visualize in a comprehensible way. Even a single spatial situation consisting of flows in one time step is hard to explore. The second challenge is the need to analyze long time series consisting of numerous spatial situations. We present an approach facilitating exploration of long-term flow data by means of spatial and temporal abstraction. It involves a special way of data aggregation, which allows representing spatial situations by diagram maps instead of flow maps, thus reducing the intersections and occlusions pertaining to flow maps. The aggregated data are used for clustering of time intervals by similarity of the spatial situations. Temporal and spatial displays of the clustering results facilitate the discovery of periodic patterns and longer-term trends in the mass mobility behavior.

Index Terms—Movement data, mobility behavior, spatial flow situation, flow map

1 Introduction

Spatio-temporal patterns and long-term trends of mass mobility behavior are difficult to analyze and understand due to the complexity of the phenomenon itself as well as the high dimensionality and large volumes of data reflecting it. One of the possible forms of data on mass mobility is a large number of records about individual moves or trips including origin, destination, start time, and end time. Such records can be aggregated into flows (collective moves) by origin-destination pairs and time intervals. The flows are characterized by the counts of the people that moved and, possibly, further attributes, such as mean or median move duration, speed, or characteristics of the people that moved. Data reflecting mass mobility may be originally available in this aggregated form, for example, population migration data. This form of data will henceforth be referred to as *OD (origin-destination) flow data*, or, shortly, *flow data*. The size and complexity of flow data is a product of the number of possible origin-destination pairs, which is quadratic with respect to the number of distinguishable locations in the data, and the number of time intervals. When the number of locations and/or time intervals is large, comprehending characteristic features of mass mobility behavior and its variation across space and time requires an appropriate abstraction of the flow data.

There are two complementary views on time-dependent flow data [1][2]: as multiple time series of scalar attribute values (move counts and others) associated with the OD pairs, called *links*, and as a single time series of *spatial flow situations*, which are in the focus of our paper. One spatial flow situation corresponds to one time interval. It is a directed weighted graph with the nodes anchored in geographic space, that is, the nodes are discrete spatial locations. A directed edge (link) exists between two nodes A and B if there were any moves from A to B. The number of these moves, called flow magnitude, is taken as the weight of the link. The flow magnitudes and, consequently, the link weights change from one time interval to another. The links may also have other time-dependent attributes.

- *Gennady Andrienko and Natalia Andrienko are with Fraunhofer Institute IAIS and City University London. E-mail: {gennady|natalia }.andrienko@iais.fraunhofer.de.*
- *Georg Fuchs is with Fraunhofer Institute IAIS. E-Mail: georg.fuchs@iais.fraunhofer.de*
- *Jo Wood is with City University London. E-mail: J.D.Wood@city.ac.uk.*

For any location, there may be many outgoing and many incoming links; in the extreme, a flow situation is a complete graph where each location is linked with every other location. Numerous intersecting links make flow situations very difficult to visualize and to comprehend, but it is even more difficult to deal with long time series of flow situations.

Our research goal was to develop a general procedure for analyzing the temporal dynamics of spatial flow situations in order to understand characteristic spatio-temporal features and trends in mass mobility behavior over long time periods. We found that spatial and temporal abstraction are necessary for dealing with the intrinsic complexities of OD flow data. We propose an approach in which the spatial abstraction aggregates OD flows with a common origin or a common destination by direction and distance ranges. This reduces the dimensionality of the data and permits visual representation of flow situations by diagram maps, thus avoiding line intersections, which are inevitable in flow maps with links represented by lines. The temporal abstraction is based on clustering of time intervals by similarity of the flow situations. The aggregations resulting from the spatial abstraction are used as feature vectors describing the flow situations. Clusters of time intervals represented visually in a calendar-like display show the overall temporal patterns and trends, and also exhibit disruptions and outliers.

The overall analytical procedure, which is the main contribution of our paper, includes the following components:

- Spatial simplification and analysis:
  - A data abstraction technique that simplifies a flow situation by aggregating OD flows by direction and distance ranges.
  - A cartographic visualization method for representing an abstracted flow situation on a map by a set of radial diagrams.
  - An approach to supporting visual comparisons between flow situations.
  - An approach to combining an overall abstracted view of a spatial situation with "details on demand" [3][4].
- Temporal simplification and analysis:
  - Data abstraction by means of interactive clustering of time intervals according to similarity of the flow situations.
  - A calendar-like visualization of the overall temporal distribution of the clusters of flow situations.
  - Summarization of flow situations by time clusters.

These components are organized in a workflow presented in Section 3. The analytical procedure is targeted on revealing and exploring the following types of temporal and spatial patterns in flow data:

- Temporal patterns: time intervals of similar flow situations, small and large changes of flow situations, disruptions of continuity,

periodic repetition of similar flow situations with regard to temporal cycles (daily, weekly, and seasonal), temporal outliers.

- Spatial patterns: major hubs (locations with many outgoing or incoming flows), spatial flow trends (prevalence of flows in certain directions and/or to certain distance ranges in different parts of the territory), regions of attraction and repulsion (where incoming or outgoing flows prevail).
- Spatial patterns of changes: locations and regions of increase or decrease of outgoing or incoming flows, changes with regard to flow directions and/or distances.

These pattern types are specializations of the highly general pattern types 'association', 'differentiation', and 'arrangement' ([5], p. 91).

We would like to stress that the goal of the paper is to present a general analytical procedure but not a particular system that we used for developing and testing the procedure and for making illustrations. The specifics of the software implementation are thus irrelevant.

2 Related Works

2.1 Spatial simplification of flow data

Spatial simplification can be achieved by grouping the places of the flow origin and destination into larger regions and aggregating the original flows into flows between the regions. Guo [6] proposes a spatially constrained graph partitioning method that groups spatially neighboring places so that there are more connections within the groups than between the groups. For data representing individual trips, regions can be defined by spatial clustering of the points of the trip origins and destinations [7]. Gao et al. [8] apply place clustering for simplification of time-dependent flow situations; however, for each time step, the places and flows are aggregated separately, which complicates tracking changes between time steps. Von Landesberger et al. [9] use density-based clustering for aggregation of strongly connected neighboring places into regions and then apply graph drawing techniques to represent flows between the regions in a more abstracted manner. Flow data can also be simplified by grouping and aggregating spatially close OD flows using hierarchical clustering [10] or kernel-based density estimation [11]. After the simplification, the most important flows are visualized on a flow map, i.e., minor flows are hidden for reducing the display clutter. Our approach to spatial simplification preserves the original set of places (i.e., does not unite them) but aggregates OD flows with common origins or destinations by direction and distance classes, which allows flow maps to be transformed into diagram maps. This approach can also be applied after a previous simplification of OD flow data by any of the earlier mentioned methods.

2.2 Temporal simplification of time-variant flow data

To analyze changes of flow situations over time, it was proposed to group similar situations corresponding to different time intervals by means of clustering [1][2][9]. The situations are specified by large feature vectors consisting of the flow magnitudes for all existing links. The result is clusters of time intervals, which are shown on temporal displays, such as a calendar display [12]. For these time clusters, average flow situations are computed and represented on a small multiple flow map display. However, the problem of over-plotting of the flow lines remains unsolved. Another problem is the length of the feature vectors describing the situations. With increasing the number of distinct places, the number of links and, hence, the length of the feature vectors, grow quadratically. The data become too heavy and problematic for clustering tools [13]. Simplification of the data is required to reduce the dimensionality and decrease the impacts of occasional fluctuations [9].

Alternatively to time clustering, the vectors representing spatial situations at different time steps can be projected onto a plane, as proposed recently for generic (non-spatial) graphs [14] and other types of time-variant data [15]. Each situation ("snapshot" [14]) is represented by a point on the plane. Points corresponding to similar situations tend to form clusters. By interactively selecting points, the user can see the corresponding situations. All points are sequentially connected by lines in chronological order, which creates a "time curve" [15]. The shape of the curve gives an idea about the character of the temporal variation. However, this representation is not optimal for data with periodic variation, especially in presence of two nested temporal cycles, such as daily and weekly.

2.3 Visualization of flow data

The recent reviews of visualization methods for OD flows [16][17] discuss three main classes of techniques: OD matrix [18], flow map [19], and a hybrid of a matrix and a map called OD map [20][21]. In an OD matrix, the rows and columns correspond to locations and the cells contain flow magnitudes represented by color shades. The rows and columns can be automatically or interactively reordered for uncovering connectivity patterns. Disadvantages of the matrix display are the lack of spatial context and the limited number of different locations that can be represented.

In flow maps, links between locations are represented by straight or curved linear symbols analogously to node-link diagrams. Various possible representations of directed links are discussed and evaluated by Holten et al. [22]. Flow magnitudes are shown by proportional line widths or by color shades. Proposed approaches to dealing with display clutter and occlusion rely on reducing or simplifying the data. These include the spatial simplification methods discussed in section 2.1. Visual simplification can be achieved by varying the opacity according to the flow magnitudes [21]. Filtering is used to show only flows with magnitudes above a chosen threshold [6][10][8][19][23] or only flows between selected locations [24]. Edge bundling [25][26][27] simplifies the display by merging or grouping spatially close flows. Edge bundling methods are a popular research topic in the graph drawing community [28]. In flow maps, unlike general non-spatial graphs, edge bundling works well only for showing flows from one or two locations or in special cases, e.g., when radial flows from/to one location prevail over all others [27]. Besides, edge bundling on a geographic map introduces undesired geographic artifacts, such as arterial roads that do not exist in reality.

Clutter on a flow map can also be reduced by removing the middle parts of the lines connecting origins and destinations and showing the start and end parts in two different colors [29]. FlowStrates [17][30] shows flow origins in one map and destinations in another. Between the two maps, there is a table display of time series of flow magnitudes. The origins and destinations in the two maps are connected by lines with the corresponding table rows. This technique is suitable for tracing individual links and viewing the associated time series, but the spatial patterns of the flows are lost.

OD maps [20][21] are based on space transformation in which the locations are arranged in a rectangular layout (i.e., a matrix) so as to minimize the distortions of their relative spatial positions with respect to each other. As a result, each location is represented by a matrix cell, which is filled with a small matrix of the same structure as the overall matrix representing the flows from/to this location to/from all other locations. This display is free from occlusion, but the space distortion complicates the perception, and the overall spatial pattern of flows is broken into multiple location-specific patterns. Besides, the method implies the user to view multiple small matrices, which may be too difficult for a large number of locations.

A straightforward approach to showing time-variant flows is to use multiple maps arranged either temporally in map animation or spatially in a small multiple maps display [31]. Map animation is not effective [32] because the user cannot memorize and mentally compare multiple spatial situations. In small multiples, a limited number of spatial situations can be shown simultaneously; hence, this approach is not suitable for long time series. Clustering of spatial situations [1][2][9] can be used to reduce the number of distinct situations that need to be shown. A completely different approach is to show the time series of flow magnitudes separately from maps, for instance, as it is done in FlowStrates [17][30]; however, the spatial situations and their changes over time cannot be seen.

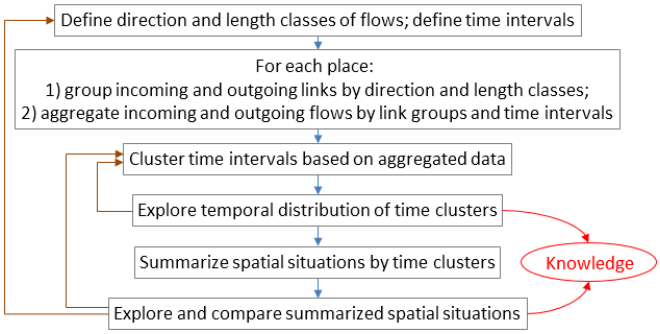

Fig. 1. The proposed analytical workflow.

In our approach, similarly to [9], a calendar-like display [12] shows how groups of similar spatial situations are distributed over time, and multiple maps show representative spatial situations for the groups. To alleviate the problem of overplotting on a flow map, we propose a diagram map providing an overview of a spatial situation.

3 Analytical Workflow

Figure 1 represents the proposed analytical workflow composed of data transformations, clustering, and knowledge discovery with the help of interactive visual displays. The blue arrows in the diagram show the ordering of the analysis steps, such that steps appearing later in the sequence use results of the preceding steps. The brown arrows represent possible returns to earlier stages of the analysis for refinement of observed patterns or for testing their sensitivity to the parameter settings of the methods involved in the analysis. The oval labelled "Knowledge" represents the knowledge gained through the analysis, i.e., the understanding of the spatial and temporal aspects of the mass mobility on the studied territory. The steps of the workflow are presented in the following sections. Section 4 deals with the data transformations, i.e., the first two boxes of the diagram. Clustering of time intervals and exploration of their temporal distribution are described in section 5, and section 6 focuses on the visualization, exploration, and comparison of spatial situations.

The approach has been tested on several examples of real flow data. In the paper, we shall use data concerning journeys by public bicycles in London as a running example for illustrating the methods.

The London Cycle Hire Scheme (LCHS) allows people to pick up a bicycle from one of several hundred automated docking stations around central London, and after riding, return it to any of the docking stations in the network. Each bike removal or docking is logged in a database along with a timestamp, bicycle ID, and docking station ID. The data have been made publicly available by the governmental transport authority Transport for London (http://www.tfl.gov.uk/info-for/open-data-users/our-feeds).

The LCHS dataset that we used consists of 5,177,679 records about individual bike trips made during the period of 28 weeks from Sunday, July 22, 2012 till Saturday, February 2, 2013. The origins and destinations of the bike trips are 569 bike stations distributed over the central part of London. Aggregation of the individual trips into flows between the stations results in a graph with 218,187 links. LCHS data from earlier time periods were explored previously [21][33]. It was interesting to check whether the earlier detected regular patterns preserved over time. We successfully detected these patterns, which signifies that our approach is feasible and valid.

To demonstrate the applicability of the proposed analytical procedure to other data, we shall also use trajectories of 17,241 cars in Milan (Italy) collected by means of GPS tracking over one week. The dataset was first introduced in paper [34]. The data have been earlier analyzed from various perspectives (e.g., [35][2]), but the spatio-temporal patterns of the car journeys in terms of the origins and destinations have not been explored yet. Knowing that car positions were only recorded during movement, we extracted the journeys by dividing the entire position sequence of each car into subsequences separated by time gaps of at least 30 minutes length. We thus obtained 51,498 trips, from which we took only the beginning and ending positions and times. For trip aggregation, we used 63 territory compartments with approximate radii of 3 km resulting from data-driven tessellation [36] of the Milan territory.

4 Spatial Abstraction

We propose a data abstraction technique based on aggregating links with a common origin or a common destination by direction and distance intervals. Except links with coinciding origins and destinations, each link has a certain spatial direction and length, i.e., the distance between the origin and the destination. Let L_{max} be the maximal link length in the dataset. The analyst divides the range of distances $(0, L_{max}]$ into k intervals by choosing suitable breaks L_1, …, L_{k-1} based on domain knowledge or on the statistical distribution of the link or trip lengths. The length 0 is considered separately. The analyst also divides the range of spatial directions from 0° to 360° into m intervals by choosing breaks D_1, …, D_m. Since the range of directions is cyclic, i.e., 0°=360°, the breaks D_1, …, D_m determine the set of direction intervals $\{(D_i, D_{i+1}] \cup (D_m, D_1] \mid 1 \le i \le m\text{-}1\}$. For convenience of human perception, the direction breaks may be chosen so that the resulting intervals correspond to the four cardinal directions or to the eight principal compass rose directions.

These divisions define $k \times m+1$ possible classes of links, where $k \times m$ stands for all possible combinations of length and direction intervals and 1 stands for the class of links representing round trips. These links have coinciding origins and destinations, zero length, and no direction. For each location, the links starting and/or ending at it are grouped according to these classes. For each time interval, the flow magnitudes on the links are summarized by these classes, i.e., each spatial location S has an associated vector of flow magnitudes $(M_0, \{M_{ij}^{in}\}, \{M_{ij}^{out}\})$. M_0 is the magnitude of the flow from S to S. M_{ij}^{in} and M_{ij}^{out}, $1 \le i \le m$, $1 \le j \le k$, are the cumulated magnitudes of the incoming and outgoing flows to/from S from/to within the direction interval i and the distance interval j. A flow situation in each time interval is represented by a combination of these vectors for all locations, which can also be considered as a single feature vector consisting of $n \times (2 \times k \times m+1)$ components, where n stands for the number of the spatial locations.

When the number of distinct locations is large, the transformation notably reduces the data size. With n distinct locations, there are $n \times n$ possible origin-destination pairs; hence, one situation is represented by a vector of $n \times n$ flow magnitudes. The vector length is quadratic with respect to the number of locations. The vector length for transformed data is $n \times (2 \times k \times m+1)$, where $2 \times k \times m+1$ is constant, i.e., the length is linear with respect to the number of locations. When $n >> 2 \times k \times m+1$, the data reduction is substantial.

For the LCHS dataset, we divide the range of directions [0, 360) degrees into the eight compass rose directions North, Northeast, East, and so on using breaks {22.5, 67.5, 112.5, 157.5, 202.5, 247.5, 292.5, 337.5}. Based on the statistical distribution of the bike trip lengths, we divide the links into three distance classes: short (0, 2 km], medium (2 km, 5 km], and long (5 km, ∞). According to these divisions, k=3, m=8, and the flows to and from one location are represented by a vector composed of 2×3×8+1=49 summarized flow magnitudes. One flow situation is represented by a vector with 49×569=27,881 components, i.e., 8.61% of the 569×569=323,761 values for all possible origin-destination pairs or 12.78% of the 218,187 actually existing links.

For the Milan car trips, we use the same set of directions. The distances between the trip start and end positions are taken as the trip lengths, irrespectively of the paths followed. The trips to distances not more than 500 m are treated as round trips. Based on the statistical distribution of the lengths of the remaining trips, we take the following distance classes: (500 m, 8 km] treated as short trips, (8 km, 13.5 km] as medium, and (13.5 km, ∞) as long. The maximal trip length in the dataset is 27.1 km. The proportions of the round, short, medium, and long trips are, respectively, 5.7, 32.6, 31.1, and 30.5%. We divide the territory of Milan into 63 spatial compartments

(further called cells) and aggregate the trips originating and ending in these cells into out- and in-flows by the direction and distance ranges. One spatial situation is represented by a vector with 49×63=3,087 components. In this case, the data reduction is low (7.3%, compared to the existing 3,300 origin-destination pairs). The main benefit of the aggregation is obtaining clearer views of spatial situations than it could be achieved with traditional flow maps.

5 Temporal Abstraction

5.1 Defining time intervals

For the temporal abstraction, the time range of the data needs to be divided into intervals. The choice depends on the length of the time period under study, the relevance of particular time cycles, and the spatio-temporal density of the movements. Let us explain these criteria by example of the LCHS data.

Movements of people typically adhere to the daily and weekly cycles. The time period length of our data is 196 days, or 28 weeks. To analyze both weekly and daily patterns, we need to divide the daily cycle into such intervals that can capture the natural differences in the mass mobility behaviors throughout a day: morning rush hours, business hours, afternoon-evening rush hours, and late evening-night quiet time. The spatio-temporal density of our data is not sufficient for a division into hour-long intervals or shorter. Since only a few links were actually used in each interval, aggregation by short time intervals would result in a great number of zero flow magnitudes and high fluctuations in the flow time series. Therefore, we divide a day into four longer intervals: morning [6:00, 10:00), midday [10:00, 16:00), afternoon-early evening [16:00, 20:00), and late evening-night 20:00, 6:00). We have chosen these breaks based on a histogram of the distribution of the trips by hours of the day. The whole time range is thus divided into 4×196=784 time intervals.

For the Milan example, where the time span of the data is short (only one week), we divide it into 168 hourly intervals.

5.2 Suitable clustering methods

After the data transformation, a spatial situation in one time interval is represented by a feature vector consisting of $n\times(2\times k\times m+1)$ summarized magnitudes of outgoing and incoming flows, where n is the number of distinct locations, k is the number of distance intervals, and m is the number of direction intervals. Clustering is applied to the set of feature vectors corresponding to different time intervals. It groups the time intervals based on the similarity of the spatial situations. Thereby, temporal abstraction is gained.

According to our approach, it is appropriate to use partition-based clustering methods, such as k-means [37] (used in this paper). Partition-based methods divide items into groups so that items within a group are similar and items from different groups are less similar. Density-based methods, such as DBScan [38] and OPTICS [39], are not suitable since their goal is to find dense groups of similar or close items; the remaining items are treated as "noise". An item is included in a group when it is similar to at least a given minimal number of other group members, but it does not need to be similar to all group members. Hence, a density-based algorithm can construct a cluster in which two arbitrary members may be very dissimilar. Therefore, the density-based clustering concept is not appropriate for the proposed kind of analysis, which requires within-cluster consistency.

Self-organizing map (SOM) [40] is a kind of partition-based method that builds a network of prototype vectors (a.k.a. neurons or cells) and associates each data item with the nearest (i.e., the most similar) prototype. Not every single cell necessarily represents a meaningful cluster. It may be useful to take a combination of nearby cells as one cluster. The u-matrix [41] showing pair wise distances between neighboring cells helps analysts to see what cells are similar and thus can be interactively joined in one cluster [42].

As a measure of similarity between feature vectors, partition-based clustering methods typically use Euclidean or Manhattan distance. In case of high dimensional data, Euclidean distance may

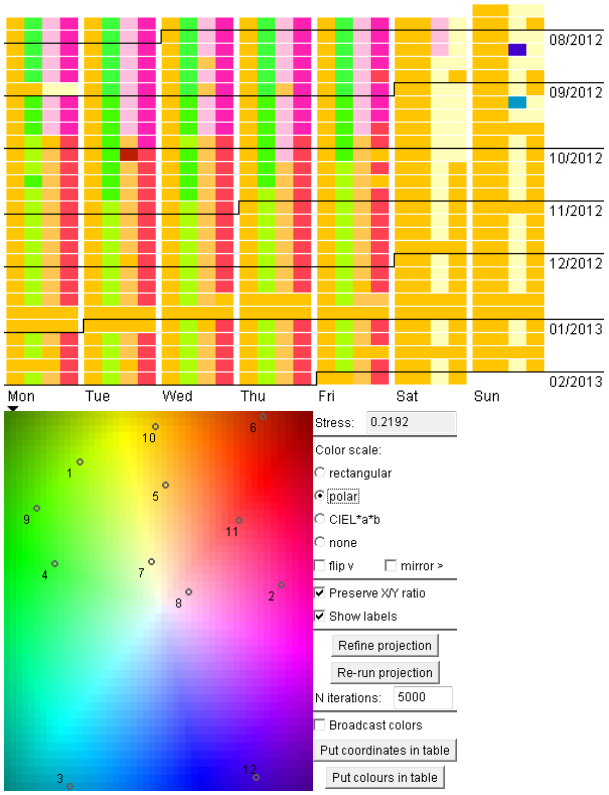

Fig. 2. Time clusters are visualized in a time arranger display (top). Colors are assigned to clusters by projecting cluster centers onto a color plane (bottom).

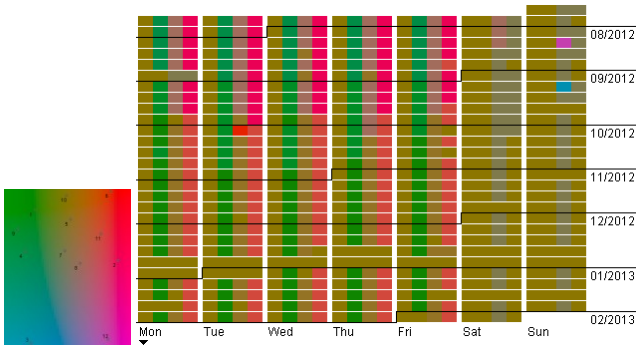

Fig. 3. When the CIELAB color space is used for filling the projection background, the number of well distinguishable colors decreases.

not be a good choice [43]. Manhattan distance gives better results, but fractional distances (i.e., Minkowski distance of order $p<1$) can work even better [43][44]. However, the choice of optimal p is application-dependent. It is also worth noting that the triangle inequality is violated for $p<1$, that is, the distance is not a metric.

5.3 Temporal display of clustering results

The resulting clusters of time intervals need to be visualized on a temporal display. Detection and investigation of periodic temporal patterns can be supported by a visualization method that may be called "time arranger". Chronologically ordered time intervals are represented by rectangular blocks arranged in rows, similarly to the calendar display [12], but the time arranger is more generic, being applicable to time cycles of any length. The row length is set to be equal to the number of time intervals in a relevant time cycle. If the time span of the data does not start from the beginning of the time cycle, an appropriate offset is made in the first row. To represent a shorter time cycle within the cycle represented by the row length, the rows may be divided into sections corresponding to the shorter cycle.

An example can be seen in Fig. 2 (top), which represents the 784 time intervals of the aggregated LCHS data. The intervals are arranged in rows according to the weekly time cycle, i.e., each row represents one week starting from Monday. The row length is set to 28 = 7 days of the week × 4 time intervals per day. Since the time span of the data starts from Sunday, the first row is offset by 24 = 6 days × 4 intervals. The rows are divided by spaces into sections corresponding to the daily cycle. The section length is 4 blocks, as there are 4 time intervals per day. The first block in each section represents the night time interval.

It is a usual practice to represent clustering results in visualization by color-coding. A unique color is assigned to each cluster, and display elements representing cluster members are painted in these colors. In a time arranger, the blocks representing time intervals are colored according to the cluster membership of the intervals. For meaningful assignment of colors to clusters, the following approach can be applied. The vectors representing cluster centers are projected onto a two-dimensional color plane [1][45] (Fig. 2, bottom) using multidimensional scaling [46], Sammon's mapping [47], or another method. A two-dimensional continuous color space is used as a projection background. Clusters that are close in the projection space receive similar colors and distant clusters receive dissimilar colors.

The combination of coloring and arrangement of the blocks in a time arranger reveals various temporal patterns and trends. Periodic temporal patterns manifest themselves through vertical alignments of identically or similarly colored blocks. Temporal trends are detected from the color shades gradually changing in the horizontal or vertical direction. Temporal outliers are manifested by blocks colored in high contrast to their neighbors in both horizontal and vertical directions.

For illustrations in this paper, we use a color space that gives well distinguishable colors for the clusters. However, it is not perceptually optimal since the ratio between the human-perceived difference in colors and their spatial distance is not uniform throughout the space. CIELUV and CIELAB color spaces [48] are perceptually more uniform, but give quite a limited number of well distinguishable colors (Fig. 3). When cluster colors are intended to be mainly used as cluster labels in various displays, perceptual uniformity can be compromised for getting a larger number of distinct colors.

5.4 Interactive visually supported clustering

Typically, partition-based clustering algorithms require the user to specify the number of clusters in which the data must be divided; for SOM, the maximal number of clusters is determined by the user-chosen dimensions for the map layout. The suitable number of clusters is often not known in advance. We propose to perform clustering iteratively, starting with a small number of clusters and gradually increasing it. This process is supported by the projection display of the cluster centers (Fig. 2). When the distances between the cluster centers in the projection space are large, it is reasonable to try a larger number of clusters. If the next clustering step results in two or more cluster centers located very closely, it makes sense to return to the previous step with a smaller number of clusters.

Additionally, the quality of the clusters is assessed based on the distances of the cluster members from the cluster centers. The distances can be represented in a time arranger view by block sizes. In Fig. 4, the block sizes are inversely proportional to the distances, that is, they show the closeness of the time intervals to their cluster centers, i.e., the larger, the closer. Hence, large blocks represent core cluster members and small blocks represent possible outliers. A cluster with high internal variation is recognized from presence of many small blocks. Details for the blocks, including the distances to the cluster centers, can be accessed by mouse-pointing.

Cluster quality can also be judged from the distance statistics. A large difference between the mean and median distances indicates that the cluster includes outliers and should be refined. However, it is not guaranteed that re-running of the clustering method with increasing the desired number of clusters will refine this particular cluster. We suggest progressive clustering [1], i.e., application of the clustering algorithm only to the clusters needing refinement.

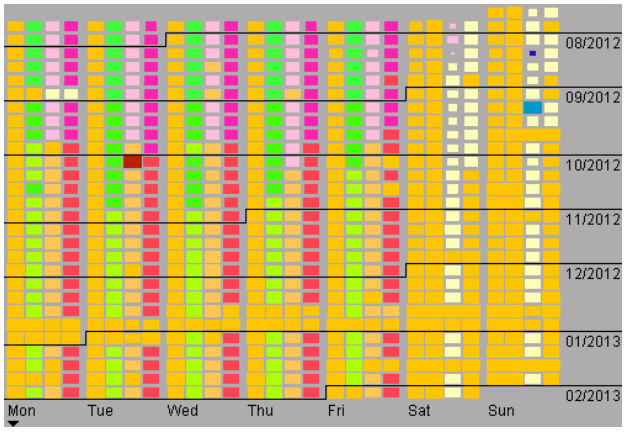

Fig. 4. The closeness of the cluster members to the cluster centers is represented by proportional sizes of the blocks.

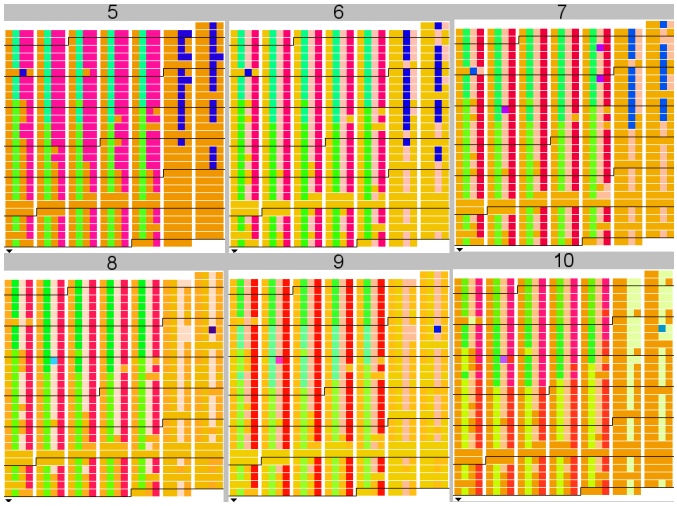

Fig. 5. Clustering results for different number of clusters.

Please note that it is not the ultimate goal of the analysis to obtain perfect time clusters. The goal is to uncover the temporal patterns and trends existing in the data. This is achieved through repeated re-clustering of time intervals with different clustering parameters and observing the color patterns emerging in the time arranger. When increasing the number of clusters does not uncover additional patterns in comparison to previous results but only highlights outliers, the process may be finished. Even when the analyst is interested in finding outliers, there is no need to strive at separating them into individual clusters (singletons). When outliers are included into larger clusters, their distances to the cluster centers are large, and this can be seen from the block sizes in a time arranger display.

To facilitate comparisons of results of different clustering runs, it would be good to preserve the consistency of cluster colors between the runs. We apply the following approach. For each cluster resulting from the latest run, the interactive clustering tool finds the closest cluster from the previous run, i.e., such that the distance between the vectors of the cluster centers is minimal. More formally, let $\{C_i^N, 1 \le i \le k(N)\}$ be the set of clusters obtained in N-th run of the clustering algorithm; k(N) denotes the number of the clusters obtained. Let cv_i^N be a vector of flow values representing the center of cluster C_i^N. Let $projection(cv_i^N) = (x_i^N, y_i^N)$ be the projection point of the cluster center onto the color plane.

When $N>1$, for each C_i^N, the tool finds a matching cluster from the previous run $M_i^{N-1}=C_m^{N-1}$, $1 \le m \le k(N-1)$, such that $data_distance(cv_i^N, cv_m^{N-1}) < data_distance(cv_i^N, cv_j^{N-1})$ for any $1 \le j \le k(N-1)$, $j \ne m$. Here *data_distance* is the distance between two vectors of flow values computed by the same distance function as has been used for the clustering, i.e., Manhattan distance, Euclidean distance, or fractional Minkowski distance.

Then the tool runs the projection method multiple times for the set of cluster centers $\{cv_i^N, 1 \le i \le k(N)\}$ to obtain different projection variants. Each run of the projection method may arrange the vectors in a different way. Only the relative distances between the vectors are preserved but not their absolute positions. From the different projection variants, the tool selects the one with the smallest sum of weighted distances between the positions of the new cluster centers and the positions of their matching cluster centers in the projection used for the previous clustering results. The distances are weighted by the cluster sizes. Formally, the tool minimizes the sum

$$\sum_{i=1}^{k(N)} |C_i^N| \times distance(projection(cv_i^N), projection(mv_i^{N-1})),$$

where $|C_i^N|$ is the cardinality of cluster C_i^N, cv_i^N is the center of C_i^N, mv_i^{N-1} is the center of the matching cluster M_i^{N-1} for cluster C_i^N, and $distance(projection(cv_i^N), projection(mv_i^{N-1}))$ is the spatial distance between the projection points of the cluster centers cv_i^N and mv_i^{N-1} in the projection space.

In this way, new clusters receive colors similar to the colors of the corresponding old clusters, with giving higher priority to larger clusters. This precludes radical changes of the color patterns in the time arranger. As an example, Fig. 5 includes 6 images of the time arranger representing the results of applying the k-means clustering method with the Manhattan distance function to the flow situations obtained from the LCHS data. The images correspond to different values of the parameter k (the number of clusters) from 5 to 10. We see that the colors are mostly consistent across the different images.

5.5 Temporal analysis by means of clustering

Clustering of time intervals by the similarity of the flow situations in combination with coloring and block arrangement in a temporal display reveals existing periodic patterns and longer-term trends in the evolution of the mobility behavior; however, several runs of clustering with different parameters are needed for uncovering all relevant patterns and gaining high confidence in their validity. Thus, in our LCHS example, the following temporal patterns and trends exist (these are known from previous studies [21][33] but also expected based on the common knowledge of human behavior):

- periodic daily pattern repeated in 5 weekdays;
- distinct daily patterns occurring on Saturdays and Sundays;
- repetition of the weekly pattern composed of these daily patterns;
- seasonal changes of the mobility behavior while keeping the periodicity of the daily and weekly patterns;
- deviations from the periodic patterns due to holidays, such as the Summer Bank Holiday and the Christmas – New Year period.

All these patterns can be seen in the time arranger view already for k=6. The result for k=5 is worse as it puts together the midday and evening time intervals of the weekdays. Still, already with k=5 we see seasonal changes, specifically, changes of the weekly pattern that occurred in mid-October and are manifested by the changes of the cluster membership of the morning intervals of the weekdays and of the midday intervals of Saturdays and Sundays. The results for k=7 and k=8 expose a couple of outliers but do not refine the previously uncovered patterns. With k=9, we additionally see a change at the beginning of September: similarly to October, the mornings of the weekdays changed their cluster membership. For k=10, we see that the mobility behavior in the weekday evenings also changed in October. Further increasing of the number of clusters puts more outliers into singletons, but does not reveal new patterns. Hence, we may conclude that the relevant temporal patterns have been captured.

Fig. 6 shows the time clusters obtained for the Milan car trips. The uppermost row in the time arranger corresponds to Sunday and the lowest row to Saturday; the columns correspond to the hours of the day. The columns are numbered from 0 to 23, where number x means the time interval $[x, x+1)$. Clustering with different values of k reveals differences between the night and day hours, between the week days and the weekend, and between the morning hours (starting from hour 5) of the week days and the remaining times of

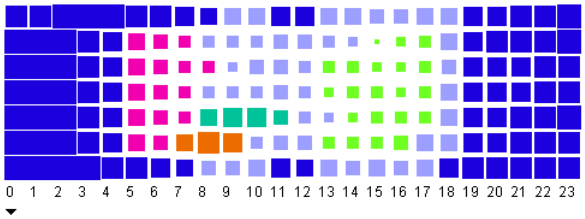

Fig. 6. Time clusters for the aggregated Milan car trips.

the day. The result for k=6 (Fig. 6) also separates afternoon hours of the week days. There are two small time clusters composed of late morning hours of Thursday and Friday, which, evidently, differ from corresponding hours in the other week days.

Besides revealing temporal patterns, clustering of time intervals greatly reduces the number of situations that need to be studied. Long time series of flow situations can be substituted by a relatively small number of representative situations for the time clusters. In this sense, clustering of time intervals is a means of temporal abstraction and simplification of long time series. In the LCHS example, we replace the original 784 time intervals by 10 time clusters, which is less than 1.3% of the original size. For Milan, 168 hourly intervals can be replaced by 6 time clusters (3.57% of the original length). Besides representative situations for time clusters, analysts may also need to investigate detected outliers. This will not increase the workload dramatically. When regular temporal patterns exist, outliers are few, whereas presence of numerous outliers rather indicates the absence of regular patterns. In such a case, investigating every single outlier is hardly useful.

6 Visualization and Analysis of Spatial Situations

6.1 Flow diagrams

6.1.1 The design space

The data transformation described in section 4 allows representing a flow situation on a map using diagrams that show the magnitudes of the flows in different directions and different distance ranges. A diagram can represent either incoming or outgoing flows. To view both, a display with two coordinated diagram maps can be used.

Figure 7 shows different variants of flow diagrams organized in a table with the columns corresponding to different shapes of diagrams or their components and rows corresponding to possible ways of combining components in a diagram. The columns have labels P for polygons, B for bars, R for rose diagrams [49], A for angular components, and C for a circular diagram. The labels of the rows are O for overlay, J for juxtaposition, and S for segmentation. The designs can thus be referred to using two-character codes, e.g., PO means a diagram with overlaid polygons.

In all designs except CJ, colors are used to distinguish the distance ranges. Dark gray represents round trips, and colors from yellow through orange to red correspond to the short, medium, and long distances. This color scale is based on one of the Color Brewer [50] sequential multi-hued scales. For a better contrast between components within a diagram and better visibility of diagrams on a cartographic background (Fig. 8), we, first, picked non-neighboring colors and, second, increased the brightness of the chosen colors.

Representing distance categories by color hues can support an overall view (i.e., holistic perception) of a map with diagrams. The visual variable 'color' (hue) is both selective and associative [51], that is, a map viewer can perceptually select and associate all marks of the same color, i.e., see them all at once. We invite the readers to test this with the example maps in Fig. 8 and others. The visual variable 'value' (i.e., color lightness) is not associative and, hence, cannot support holistic perception of a diagram map.

All diagram variants in Fig. 7 encode the same combination of flow magnitude values. In design groups P, B, and R, the magnitude for the round trips is represented by the radii of the dark grey circles drawn using a dashed stroke. To represent flows in different spatial

directions, we use polar coordinates in PO and radial layout in the remaining designs. In PO, the flow magnitudes are represented by the distances of the polygon vertices from the polar coordinates origin. In B and R, the flow magnitudes are represented by the lengths of the bars and sectors, respectively. RO2 differs from RO1 by applying Flannery perceptual scaling [52] of the sector lengths. Perceptual scaling is also applied in RS2. In BS and RS, the full lengths of the bars and sectors encode the sums of the flows to all distance ranges in the respective directions. The bars and sectors are proportionally divided into segments representing the flows to the different distance ranges. In BS and RS1, the division is applied to the lengths of the bars and sectors. In RS2, the sectors are proportionally divided into sub-sectors, such that the flows to the different distance ranges are represented by the angle sizes. A problem of RS1 is that the segments corresponding to longer distance ranges are wider, which leads to substantial over-estimation of the flow magnitudes. We also tried to represent flow magnitudes by sector areas rather than lengths; however, this impedes or even disables comparisons of flow magnitudes for different distance ranges and different directions.

The two variants of design AJ correspond to a suggestion of reviewers of an earlier version of this paper to represent the distance ranges by lengths of diagram components and the flow magnitudes by widths, which may be more intuitively understandable to users. In AJ1, flow magnitudes are encoded into sector widths, i.e., the lengths of their ending arcs. In AJ2, we applied a slightly different idea. The distance ranges are represented by the distances of the sectors to the diagram center. All sectors have the same length, and flow magnitudes are represented by angle sizes.

A drawback of AJ is that, unlike in the other designs, round trips cannot be represented in a way allowing direct comparison to the other flows. We represent round trips using small circles in the diagram centers. Their full area corresponds to the highest magnitude attained in the data set. The magnitude for round trips is represented by the proportion of the circle area painted in dark gray. In AJ1, the value is encoded by the area of a dark gray circle and in AJ2 by the angle size of a dark gray circle sector, which is more consistent with representing the other flow magnitudes also by angle sizes. However, for the other flows, the maximal angle size is 90° while for the round trips it is 360°; hence, comparisons are hindered.

In a recently proposed design CJ [53], different distance ranges are represented by concentric rings while the inner circle corresponds to round trips. The rings are divided into sectors corresponding to the flow directions. The flow magnitudes are represented by shades of gray, so that darker shades encode higher values. To distinguish zero values (absence of trips) from low magnitudes, the sectors corresponding to zero values are not filled.

6.1.2 Difference diagrams

In analysis of flows, it may be necessary to compare values at all locations (further referred to as "local values") with (a) the values from a selected reference location or (b) the values at the same locations for another time interval or time cluster. In both cases, two vectors are compared for each location: the vector of local values and a vector of reference values, which is common for all locations in case (a) and specific to each location in case (b). To facilitate comparisons, we propose to apply explicit encoding of differences [54]. The components of the reference vector are subtracted from the corresponding components of the local vector, and the differences are shown using *difference diagrams*. We considered two design variants for difference diagrams shown in the upper and lower rows in Fig. 9. The variant in the upper row corresponds to diagrams where colors are used for representing distance ranges, and the variant in the lower row corresponds to circular diagrams CJ. The diagram examples from left to right represent different combinations of difference values. The corresponding diagrams in the upper and lower rows represent the same combinations.

In the upper row, the diagram components corresponding to different combinations of direction and distance ranges each consist

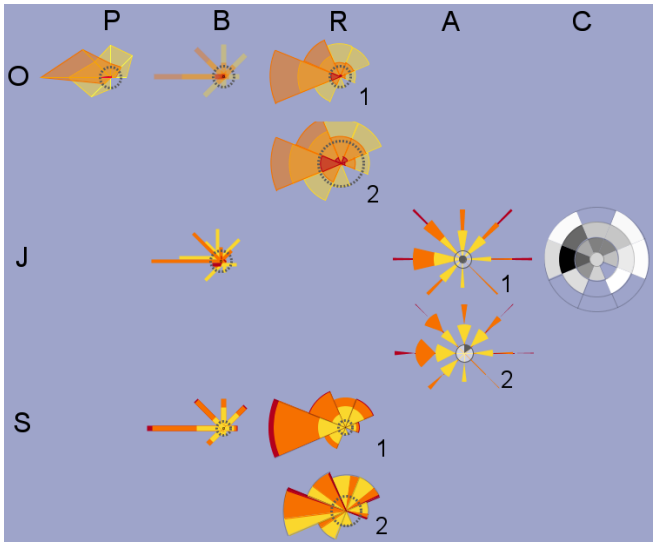

Fig. 7. Variants of flow diagram designs. Column labels denote shapes: P – polygon, B – bar, R – rose, A – angle, C – circle. Row labels denote methods of combining components: O – overlay, J – juxtaposition, S – segmentation. All diagrams represent the same combination of values.

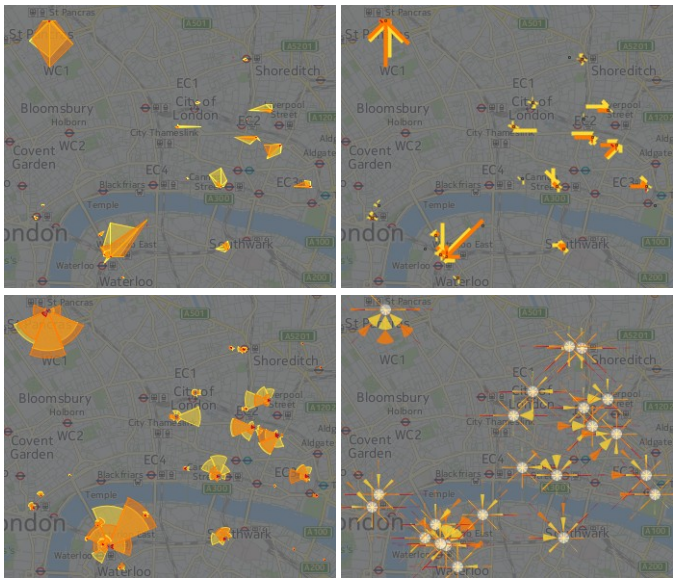

Fig. 8. Map fragments representing the same data using different diagram designs.

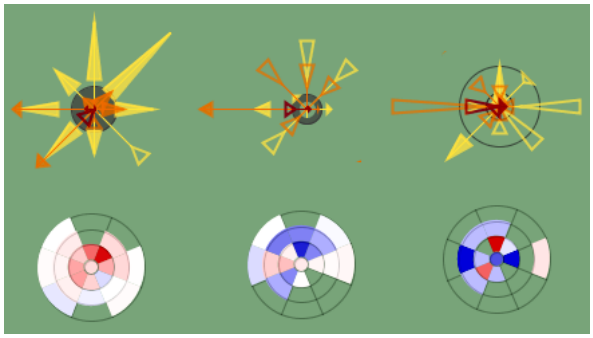

Fig. 9. Examples of difference diagrams.

of two primitives: an axial line and a triangle. The lengths of the axial lines represent the local values. The triangles, which are drawn at the ends of the axial lines, show the differences between the local values and the reference values. The heights of the triangles are proportional to the absolute values of the differences. Positive

differences (i.e., when a local value is larger than the reference value) are represented by filled triangles oriented outwards from the diagram center. Negative differences are represented by hollow triangles oriented towards the diagram centers. The bases of the triangles are positioned on the axes proportionally to the reference values and the base-opposite points coincide with the ends of the axial lines, that is, their positions are proportional to the local values. As a result, filled "positive" triangles lie on top of the axial lines and hollow "negative" triangles hang at the ends of the axial lines.

The component representing movements with coinciding starts and ends consists of two concentric circles representing the local and reference values. The circle for the local value is thicker and darker than the one for the reference value. If the reference value is smaller, the space between the circles is filled in gray; in the opposite case, it remains unfilled. Hence, the component has an appearance of a filled or hollow wheel with the section height being proportional to the absolute difference between the local and reference values.

In circular diagrams (lower row in Fig. 9), positive and negative differences are encoded by the saturation of shades of red and blue, respectively. A disadvantage of this design is that, unlike the first one, it shows only differences but not the values that are compared. However, this design is much simpler and easier to understand.

6.1.3 Evaluation of diagram designs

Each of the designs in Fig. 7 has its strengths and weaknesses. Rose diagrams are visually more prominent than bar diagrams, but a viewer may not be sure whether the sector lengths or areas represent the flow magnitudes and, hence, have interpretation difficulties. The bar diagram design does not have this problem and permits not only semi-transparent overlay of bars corresponding to different distance ranges (BO) but also juxtaposition of bars (BJ), in which the bars are more salient and easier to discern. PO can be holistically perceived as a single graphical object with a particular shape rather than a combination of objects, as may happen with a diagram consisting of several components. A weakness is that the polygon area has no meaning but a viewer may try to interpret it. Another weakness is that a polygon representing only movement in one direction or in two opposite directions collapses to a single line, which has low salience.

In the analytical procedure, the role of maps is to represent spatial flow situations so that spatial patterns (hubs, trends, attraction and repulsion areas; see section 1) can be readily perceived. To check how well maps with different variants of flow diagrams can fulfil this role, we conducted an informal study with participation of our partners from two EU-funded research projects on mobility analysis topics, 24 persons in total, including 7 experts in maritime or air transportation domains, 10 professional data analysts, and 7 computer scientists specializing in machine learning and databases. We designed the study so as to use at most 15 minutes of the valuable time of the professionals. We prepared printouts of maps in which the same data were represented using different types of diagrams, as in map fragments in Fig. 8. To avoid overwhelming the participants with too many diagram variants, we have chosen six variants: PO, BJ, RO2, AJ1, AJ2, and CJ. The maps were printed on separate sheets of paper together with brief (at most two lines) texts explaining the encoding and two questions to be answered by the participants. The same two questions were repeated on each sheet:

1. What are the major hubs?
2. What are the overall trends of the flows over the territory?

It is intentional that both questions require the overall reading level [51], i.e., holistic grasp of the represented situation.

The participants were given the printouts and asked to, first, read the questions, second, choose the map they find the most suitable for answering the questions, third, write brief answers on the sheet with the chosen map and, fourth, rank the designs from 1 (the most suitable) to 6 (the least suitable). The participants were also encouraged to write their comments on any of the designs. This way of conducting the evaluation had the following rationale.

Rather than engaging casual test participants, we wanted to involve professionals who understand well the OD data type and are capable to use and interpret maps. Since the professionals that we involved could not devote much time to the study, we could not do the evaluation in a more formal way, with asking the participants to fulfil similar tasks using each of the designs. At the same time, we could rely on the capability of our participants to compare the designs, make a rational choice, and explain it. In this way, we expected to receive qualified comments on all designs. We were not so much interested in obtaining the design ranks as in understanding the ranking criteria, i.e., what features of the designs make them more or less suitable for holistic perception of spatial patterns.

TABLE 1
RANKS OF THE DIAGRAM DESIGNS

Measure	AJ1	RO2	AJ2	BJ	CJ	PO
Count of rank 1	7	8	1	5	2	1
Count of rank 6	1	1	0	1	10	10
First quartile	1	1	2	2	2	4
Median	2	3	3	3	5	5
Third quartile	4	4	4	4	6	6
Mean	2.67	2.83	3.04	3.08	4.42	4.83

Of course, it was important to check whether the participants could properly fulfil the tasks using the chosen maps. It was so in all but one cases. One participant, who used AJ1, misinterpreted the sector-shaped diagram components as arrows indicating movement directions. A possibility for this misinterpretation of AJ1 and AJ2 was mentioned in a comment from another participant. However, two people found AJ very intuitive and requiring little explanation.

Table 1 shows statistics of the ranking. The columns are ordered based on the median and mean ranks of the corresponding designs. The first two rows show the counts of the participants who gave ranks 1 (most suitable) and 6 (least suitable) to the diagram variants. Each design was preferred by at least one participant, and for each design except AJ2 there was at least one participant who found it the least suitable. The designs RO2 and AJ1 were top ranked by the largest numbers of participants, and the designs CJ and PO got the highest number of the lowest rank. From the participants' comments, we elicited the following ranking criteria:

C1. Negative: clutter and overlapping of diagrams. The equal-sized diagrams AJ1, AJ2, and CJ were criticized the most.
C2. Positive: variation of the overall diagram sizes (RO2, BJ, PO). It makes major hubs and spatial trends (particularly, attraction and repulsion areas) easily identifiable. RO2 was especially praised for this. AJ1, AJ2, and CJ were judged as not supportive.
C3. Positive: good visibility of all diagrams and their components (equal-sized diagrams CJ, AJ1, AJ2).
C4. Negative: overlapping of colors in the diagrams (RO2, PO).
C5. Negative: unclear flow directionality (BJ, PO).
C6. Positive: easy comparison of flow magnitudes for different distance ranges (BJ).
C7. Positive: support of detailed readings (CJ).
C8. Negative: insufficient contrast between the colors of the diagram components (all designs).

Given these criteria, none of the tested designs is perfect. Moreover, creating a perfect design is hardly possible since some criteria are conflicting (e.g., C2 against C1 and C3).

Some participants' comments contained suggestions for improving the maps. For reducing display clutter, it was suggested that small flows should be hidden. One participant proposed to apply spatial aggregation, i.e., put together flows from near places and represent them by a single diagram.

None of the designs tested in this study supports perception of the total flows in different directions including all distance ranges. The diagrams with component segmentation (BS, RS1, and RS2) were designed to support this task. To evaluate also these designs, we specifically addressed those participants who preferred bar or rose diagrams and gave detailed explanatory comments. We got responses from 5 out of 7 participants asked. All of them judged the segmented bar diagrams (BS) as the best suitable and RS2 as the least suitable,

because the division of the sectors into sub-sectors complicated the interpretation of the directions. For RS1, the participants noted that volumes of more distant flows can be overestimated due to larger widths of the segments representing them, which conforms to our own judgement.

6.1.4 Accounting for study results

Since there is no single design that is preferred by everyone and can suit all purposes, users should be able to choose from several design variants depending on the task to be performed. The selection can be facilitated by a task-oriented interactive interface. Based on the ranks and comments obtained, the following options can be proposed:

- Identify major hubs and flow trends (RO2).
- Compare flows to different distance ranges (BJ).
- Compare total flows in different directions (BS).
- See complete flow compositions (AJ1).
- See all details for individual places (CJ).

RO2 can appear as the default representation. When the user selects another option, the map immediately changes.

The comments of the study participants indicate that it is reasonable to allow users to change the color scale from sequential to contrast. Thus, we constructed a contrast color scale consisting of red, yellow, and blue colors selected by means of Adobe Color Wheel [55] (triad model). We order the colors from red (hot) to blue (cold) to represent the distance ranges from near to far. The contrast color scale is applied in the illustrative maps in Figs. 11-14.

To reduce display clutter by hiding small flows, our experimental software allows interactive focusing, which is described in the next section. As suggested, display clutter can also be reduced by spatial aggregation, i.e., uniting flows to and from neighboring locations. For the aggregation, the territory can be divided into compartments by means of data-driven tessellation [36], which is based on clustering of spatially close locations. Figure 10 demonstrates the division of the central area of London based on the spatial distribution of the LCHS docking stations. It can be seen that near docking stations tend to be included in the same compartment, as, for example, the docking stations at the Waterloo railway station, which is pointed at by an arrow. For a given territory division, data referring to locations fitting in the same compartment are summarized and represented by a single diagram. Such spatial aggregation is applied in the illustrative maps in Figs. 11-14.

The use of spatial aggregation should not exclude the possibility to see the detailed information at the level of individual locations. In our prototype, it is possible to have aggregated and detailed data shown in two map displays or to put them in the same map display as two information layers, the visibility of which can be interactively switched on and off when more detail or more abstraction is needed.

6.1.5 Interactive focusing

The maximal radius of flow diagrams, including difference diagrams, corresponds to the maximum value of flow magnitude available in the represented data. The sizes of the diagram components in the bar and rose diagrams are proportional to the values they represent. Thus, some diagrams may be very small and hardly visible. By interactive focusing, the user can limit the represented range of values. The result of decreasing the upper limit is that the maximum diagram radius is used for a smaller data value; hence, previously small diagrams increase in size. The diagrams in which the maximal value exceeds the upper limit remain still visible, but their components are drawn without filling and the sizes do not proportionally increase beyond the maximum radius. It is also possible to increase the lower limit of the represented value range. The diagrams where the maximal values are below the lower limit are hidden from the view. In this way, the user may disregard locations with small flows and consider only those with major flows.

For difference diagrams, interactive focusing can also be done according to the distances between the local and reference vectors. The user can select one of three methods to compute the distances: maximum of the absolute component-wise differences, Manhattan distance, and Euclidean distance. The range of computed distances is shown to the user, and the user can limit the range to be represented. The diagrams where the distances are beyond the limits are hidden.

Both types of focusing can be done using sliders or by direct setting of the upper and/or lower limits. The diagrams on a map are dynamically updated in response to user's manipulations.

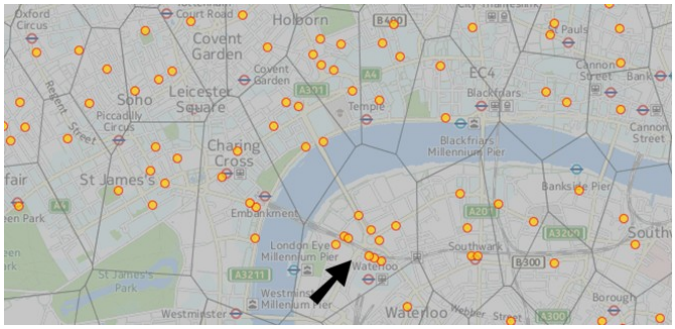

Fig. 10. The territory covered by the LCHS services has been divided into areas for reducing map clutter by means of spatial aggregation. The arrow points at a spatial cluster of docking stations located near the Waterloo railway station.

6.2 Interactive maps of spatial flow situations

A spatial flow situation is represented by a map with flow diagrams drawn at the positions of the flow origins and/or destinations. A map can represent a flow situation for a time interval or a representative flow situation for a time cluster (Figs. 11 and 13). Multiple maps corresponding to different time intervals/clusters can be arranged in a temporal sequence (map animation) or a spatial layout (small multiples). The latter approach is practically possible only for a small number of selected time intervals/clusters, because diagram maps need to be quite large to be legible. In an animated display, the user selects the time interval/cluster t to be currently presented. The map shows the respective spatial situation. The current t can be changed step wise, or by dragging a slider, or by directly setting the value.

In the mode of comparison between locations, the reference location is chosen by clicking in the map. In the mode of temporal comparison, the reference interval or cluster is set explicitly.

One diagram map can show incoming flows, or outgoing flows, or their differences. To view incoming and outgoing flows simultaneously, a display with two maps is used (Figs. 11-14), where the maps are coordinated in several ways. They are simultaneously animated and always represent the same time interval or time cluster. Zooming and panning operations are applied to both maps in parallel, so that they always show the same territory. Selection of different diagram designs, switching to comparison mode, interactive focusing, and selection of reference locations, time intervals, or time clusters in the comparison mode are applied to both maps simultaneously. As an example, the maps of out- and in-flows in Fig. 14 represent the changes between time cluster 2 (summer weekday mornings) and cluster 6 (autumn and winter weekday mornings).

Flow diagram maps are used for getting overviews of spatial situations instead of traditional flow maps, which are illegible due to heavy over-plotting. However, the transformation of the OD flow data into the direction-and-distance vectors involves quite large information loss. As a compensation, the user can obtain details on demand [3][4] by interactively selecting subsets of OD flows to be shown on the maps in addition to the diagrams (Fig. 12).

Analogously to the transformed data, the original time series of OD flows can also be summarized by time clusters. When a diagram map shows a situation for a time cluster, the OD flow magnitudes for the same time cluster can be seen, as in Fig. 12. Subsets of OD flows to view can be selected by means of several interactive filters: (1) attribute-based filter, which can select links based on their flow magnitudes, directions, or lengths, (2) spatial filter, which selects links fitting in a user-drawn spatial window, and (3) filter by link origins and/or destinations. For the latter filter, the origin and/or

destination places are selected by clicking on the map. Four filtering modes are possible: for selected locations, the map can show (a) the outgoing flows, (b) the incoming flows, (c) both incoming and outgoing flows, or (d) the flows with both origins and destinations being selected. Fig. 12 shows the OD flows originating or ending in two selected places, which have high aggregated out-flows.

Furthermore, the comparison mode can also be applied to OD flows. Thus, Fig. 14 demonstrates the comparison between time clusters 2 and 6 applied both to the flow diagrams and to the OD flow lines for a selected subset of links (the same as in Fig. 12). The flow magnitude values for cluster 2 have been subtracted from the values for cluster 6. The cyan color, the same as for the original flow lines in Fig. 12, represents the positive differences, i.e., the increased flows from time cluster 2 to 6. The red color, which is opposite to cyan, represents the negative differences, i.e., the decreased flows.

In our experimental implementation, the maps are zoomable, the information layers can be switched on and off, and the maximal sizes of the diagrams can be increased or decreased. It is also possible to choose between background maps from several map servers. Excessive prominence of the map background can be reduced by covering it with a semi-transparent gray rectangle, the degree of darkness and transparency of which can be interactively controlled.

7 Guidelines for executing the procedure

The procedure begins with defining the direction and distance classes and time intervals. The direction classes can be defined according to the eight principal compass rose directions, which are commonly known. To define the distance classes, the range from the minimal to the maximal trip lengths is divided into two or three intervals. A larger number of intervals may complicate the visual perception and analysis of spatial situations with the use of flow diagram maps, because the complexity of the diagrams would increase.

The class breaks can be chosen using a histogram showing the statistical distribution of the trip distances. When the distribution is close to uniform, it is reasonable to divide the value range into intervals of equal length. For a distribution having one or more prominent peaks, the breaks are chosen so that each peak with its neighborhood is included in one class. If there is a long "tail", it can be divided into parts containing approximately equal number of trips.

The time span of the data can be divided into equal intervals, the length of which depends on the total length of the time period and the desired level of detail in the analysis. Domain knowledge and/or data properties may suggest a division into unequal intervals, as we showed by the LCHS example (subsection 5.1).

When all divisions are defined, the data are automatically transformed as described in section 4. The next step in the workflow is clustering of the time intervals based on the similarity of the spatial situations; section 5 gives detailed guidelines. Representative spatial situations for the time clusters are examined and compared using flow diagram maps (subsection 6.2). By observing the temporal distribution of the time clusters and the spatial distributions of the flows corresponding to the time clusters, the analyst gains knowledge of the spatio-temporal patterns of mass mobility.

The flow chart in Fig. 1 shows that the analysis procedure may be performed iteratively, with returns to previous steps and modifying the previously made choices. In sections 5.4-5.5, we have described iterative time clustering: after exploring clustering results by means of temporal displays, the clustering is repeated for other parameter settings. A return to the stage of time clustering may occur after the investigation of the representative spatial situations for the time clusters, to check whether cluster refinement may affect the spatial patterns seen in the maps of the representative spatial situations. To gain even more confidence in the analysis results, the analyst may also return to the stage of data transformation, in which the analyst can modify the direction and distance classes or the time intervals. After re-aggregating the OD flow data, the analysis is repeated, and the consistency of the new results with the previous ones is checked.

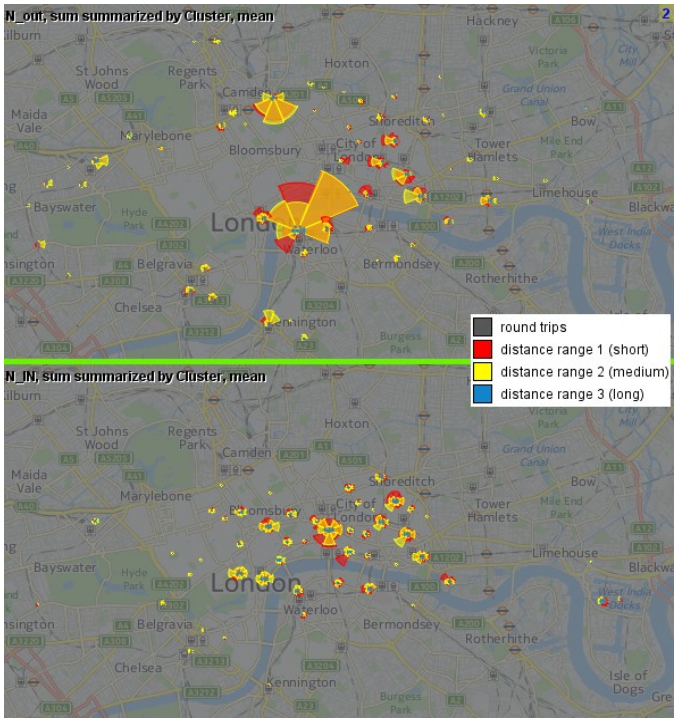

Fig. 11. Out- and in-flows for time cluster 2 (summer weekday mornings) are shown on two coordinated maps.

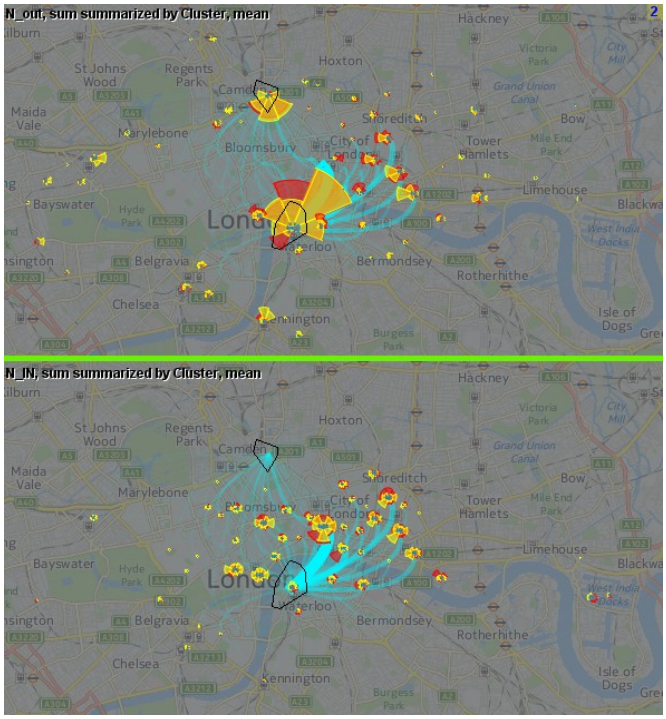

Fig. 12. A subset of OD flows (here: incoming and outgoing flows for two selected places) can be shown on demand by flow lines.

8 Case studies

8.1.1 Bicycle trips in London

Section 5.5 lists the temporal patterns that we discovered by means of the visually supported time clustering and justifies the choice of the clustering result with 10 time clusters for the further analysis. We summarized the spatial situations in these 10 time clusters and explored the summarized situations using interactive maps.

As could be expected, in the night times, there were almost no movements. In the weekday mornings (Figs. 12-13), there were two major hubs with strong outgoing flows in multiple directions (Fig. 12, top), one at the railway station Waterloo and another, smaller, at the train stations King's Cross and St. Pancras. These hubs represent multimodal journeys with large numbers of commuters using rail travel in the morning to get to London and then taking bikes to get to their final destinations or to other public transportation facilities [33]. The flow destinations were distributed over and around the City of London (Fig. 12, bottom). In the weekday afternoons, the situation was opposite: the morning hubs turned to the major receivers, and the morning receivers had large outgoing flows. The prevailing distance ranges were short and medium.

The afternoon situation differs from the morning also by the presence of round trips at the corners of Hyde Park. These may be leisure bicycle trips in or around the park. Round trips in this area occurred also in the middays of the week days. Apart from that, the flow magnitudes in the midday times were low over the whole territory. Much more round trips took place in the weekend midday times in the summer and early autumn (Fig. 13). This is consistent with the interpretation of the round trips as leisure bike rides, which are done more often on weekends when the weather conditions are good. An exceptionally large number of such trips occurred in the midday of Sunday, September 9, which was put in a separate cluster.

The seasonal changes, i.e., the differences between the same intervals of the weekly cycle that fall in different time clusters, are not as obvious as the differences within the weekly cycle. We use the comparison mode, as in Fig. 14 (by focusing, we have hidden the diagrams where the maximal flow differences are less than 5). In this example, we see an overall decrease of incoming flows in the City area (Fig. 14, bottom). The outgoing flows (Fig. 14, top) also mostly decreased. Yet, in the Waterloo area, there was a small increase of the short-range outgoing flows to the north and a slightly larger increase of the mid-range outgoing flows to the northeast while the short-range flows in this direction preserved. The OD flow lines in Fig. 14 show that the main receiver of the increased northeastern flow was the area of Moorgate, for which the flow difference diagram in the lower map indicates a small increase of incoming mid-range flows from the southwestern direction. In the King's Cross – St. Pancras area, the mid-range flows to the south also slightly increased and the short-range flows to the south preserved.

In this study, we detected all previously known or expected spatio-temporal patterns in the use of the public bicycles, which confirms the validity of our approach. Besides, we could reveal and investigate the seasonal changes better than it was possible before. It was known that the use of bicycles decreased in late autumn and winter. We additionally found that the weekday morning flows from the areas of King's Cross and Waterloo towards the City increased despite the overall decreasing trend. So, our approach allowed us not only to detect the expected but also to uncover the unexpected [52].

8.1.2 Car trips in Milan

The time clusters for the Milan case study are shown in Fig. 6. Cluster 1 (dark blue) consists of night hours from 0 till 4, i.e., till interval [04:00, 05:00), and evening hours starting from 19. As could be expected, it is characterized by low movement activity throughout the territory. On the weekend, also morning hours till 8 on Sunday and 7 on Saturday (and midday hours 11 and 12 belong to cluster 1.

A common feature of the spatial situations in all time clusters, even the quiet cluster 1, is large volumes of outgoing and incoming long distance flows in three areas on the northwest, northeast, and southeast of the territory. As an example, Figure 15 presents the flow patterns of time cluster 2 (bright pink). Here we use the diagram design with juxtaposed bars (BJ), because many places in Milan have almost equal flows to different distance ranges in the same directions. Such flows are hard to distinguish and compare using RO.

In Fig. 15, three corners of the maps have diagrams with long dark red bars oriented towards the remaining corners. We have interactively selected the areas containing these diagrams for seeing

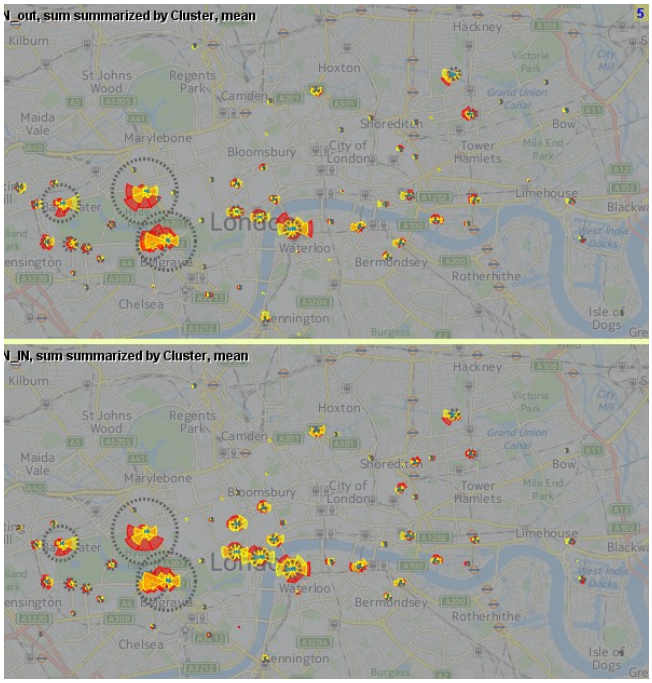

Fig. 13. Out- and in-flows for time cluster 5 (weekend midday and afternoon times till beginning of October and weekend midday times in later autumn and winter).

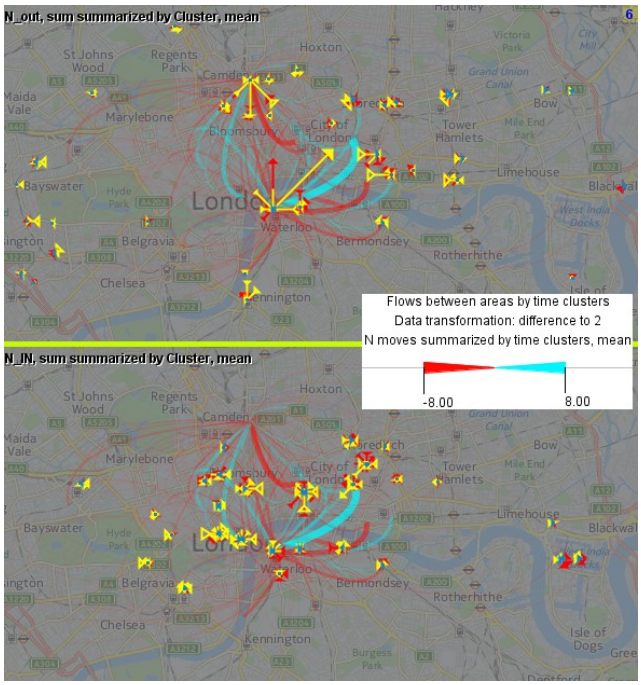

Fig. 14. The difference diagrams show the changes from time cluster 2 (summer weekday mornings) to 6 (autumn and winter weekday mornings). Additionally, the line symbols in cyan and red show the changes of the OD flows for two selected places.

their OD links, which are shown by flow lines in purple. The links with the highest flow magnitudes (i.e., the thickest lines) connect the northwest to the northeast and southeast, indicating that many car trips started in one of the corners and ended in one of the two others. Hence, many cars just passed the city using the belt motorway. There were many such transit car trips at all times. The magnitudes of the long distance out- and in-flows in the corner areas are always much higher than the values in the remaining areas; therefore, the diagrams over the whole territory, except for the corners, are too small for clearly showing the flow patterns over the territory.

To alleviate this problem, we have applied interactive focusing as described in section 6.1.5. The result is shown in Fig. 16. It is the same time cluster 2 as in Fig. 15. Now we can better see several secondary hub areas around the city. In the map of out-flows (left), we observe notable volumes of mid- and short-distance flows from the outskirts towards the inner city, in addition to the long-distance flows made by transit cars. In the map of in-flows (right), we see that there were many car trips from various directions that ended in the central part of the city, especially on the east of the center. We also see that district Corsico southwest from the center was a strong attractor, especially for short trips from the southwest. On the east of the city, many trips in all distance ranges ended at the Linate airport.

Time cluster 3 (light green, composed of weekday afternoon hours) differs from cluster 2 (weekday mornings) by increased out-flows and decreased in-flows in the central part, as is exhibited by the difference diagram maps in Fig. 17. Here we use the circular design of the difference diagrams, with shades of blue and red showing negative and positive differences. In this example, the selective and associative power of the visual variable 'color' [51] groups multiple charts with prevailing red or blue shades into regions of increase or decrease. Thus, on the left, we see that out-flows increased not only in the center but also on the east and south and decreased on the north and west. On the right, we see that in-flows decreased on the east and south and increased on the north.

Most of the remaining times are in time cluster 6 (light blue), including midday times of the weekdays and weekend afternoons. This time cluster is characterized by notably lower flow volumes over the whole territory (except the hub areas) than in the weekday mornings and afternoons. Small time clusters 4 (bluish green) and 5 (orange) occurred in late mornings of Thursday and Friday. The corresponding situations are characterized by higher flow volumes than in the same hours of the other weekdays. In comparison to the morning flows, cluster 4 and 5 mostly differ in terms of transit flows. Besides, the flows from the center slightly increased and the flows to the center decreased. In comparison to cluster 4, cluster 5 has lower out-flows on the northwest and northeast.

To summarize, we found in this case study that a large proportion of car trips are transit trips that enter the territory from one of the corners on the northwest, northeast, or southeast and leave it in another corner. The magnitudes of the flows derived from these trips are high at all times and do not significantly vary over a day. For the non-transit trips, the prevailing flows in the weekday mornings are to the city and Linate airport. The pattern reverses in the weekday afternoons. In the remaining times, the flows significantly decrease.

9 Discussion and Conclusion

Our approach involves data abstraction reducing the dimensionality of the data. The reduction factor is proportional to the number of distinct locations in the data. The reduction can make originally very large data suitable for interactive clustering. It also permits a drastic decrease of clutter and occlusion in cartographic visualization of the data compared to traditional flow maps. The abstraction decreases the level of detail with regard to the destinations (for out-flows) or origins (for in-flows). Thereby, it reveals spatial directional trends and allows disregarding minor fluctuations in movement destinations or origins, e.g., between two neighboring docking stations.

Unlike the other methods for spatial simplification [6]-[11], our approach does not change the original set of locations present in the data, but it can be applied to OD flow data that have been previously simplified using any of these methods. The latter reduce the number of locations and links in the data but preserve the graph-like data structure, whereas our approach transforms the graph into a set of multidimensional vectors associated with the locations.

Our method for spatial simplification has disadvantages due to the transformation of the flow directions and distances into discrete classes. As with any discretization, close values may fit in distinct classes while more different values may fit in the same class and become indistinguishable. Still, discretization is commonly used in data analysis. To be used properly, it requires careful definition of classes. Guidelines are provided in section 7. Another disadvantage of our method is the loss of links between locations. To compensate for this, selected subsets of OD links are shown on diagram maps.

Besides the spatial abstraction, our approach involves temporal abstraction by clustering of time intervals according to the similarity of the respective spatial flow situations. This supports analysis of data that cover long time periods. A very large number of time intervals can be reduced to a much smaller number of time clusters to be studied. The temporal distribution of the clusters is visualized in a way enabling detection of periodic temporal patterns and long-term trends. The flow situations corresponding to the clusters can be

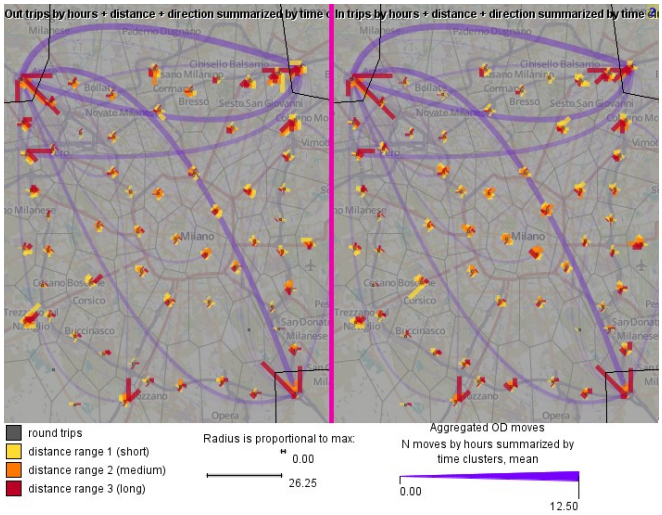

Fig. 15. Spatial patterns of the out- and in-flows for time cluster 2 (weekday mornings) and OD flows for three selected areas.

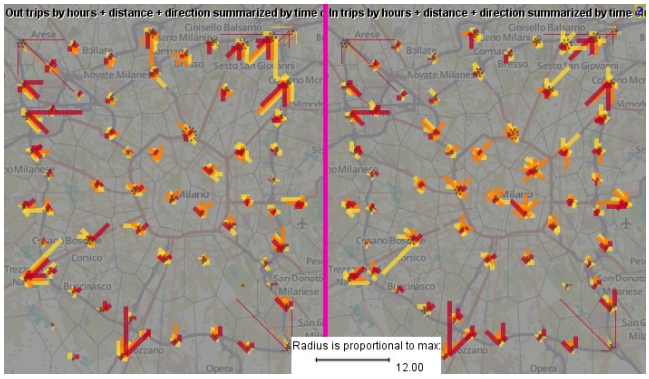

Fig. 16. Interactive focusing has decreased the visual prominence of the major hubs and made the overall patterns better visible.

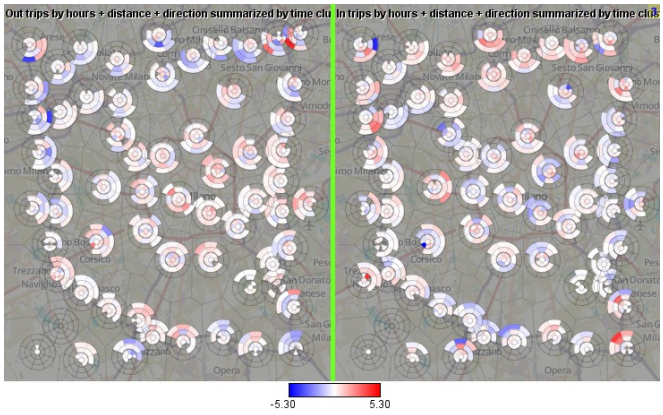

Fig. 17. Differences of time cluster 3 (weekday afternoons) from time cluster 2 (weekday mornings).

examined by viewing cluster representatives. Study of differences between time clusters is supported by interactive difference maps.

Hence, the approach handles the specific complexities of origin-destination flow data: the spatial complexity due to the high connectivity of the flow graph and the temporal complexity due to the length of the flow time series and the complex character of the temporal variation, which is governed by interplaying time cycles. Using the proposed methods, a complex mobility behavior represented by long-term flow data can be effectively studied.

Our approach is comparable to another systematic approach to analyzing long time series of OD data that was published recently [9]. The latter also involves spatial and temporal abstraction. The temporal abstraction is also achieved by means of partition-based clustering of time intervals, whereas the spatial abstraction is done through density-based clustering of neighboring strongly connected locations. Furthermore, spatial situations are visually represented using graphs (node-link diagrams) rather than maps. This alternative approach achieves high spatial simplification, which facilitates the following visual analysis. However, details in dense areas are lost, which may be a disadvantage. Besides, since the graph-like data structure is preserved, link intersections may be a problem even in simplified graphs. The abstracted representation of spatial situations as graphs has both pluses and minuses. On the one hand, certain mobility patterns (such as flows between the center and the periphery) can be more readily perceived; on the other hand, it is not immediately clear what areas in space are represented by the graph nodes. Our approach eliminates flow intersections and presents spatial situations by maps, where locations are easily recognizable.

To test the feasibility and effectiveness of our methods, we applied them to two datasets of differing sizes and complexity. The results of the testing are positive: we see that the approach works and both confirms existing knowledge of transport experts and generates new understanding about spatial and temporal patterns of mobility. The approach is generally applicable to any OD flow data reflecting movements in geographical space. Geographical space is an essential applicability condition for our approach since it deals with spatial directions and assumes that spatial locations are fixed. It is thus not applicable to abstract graphs, where nodes can be arbitrarily moved, nor abstract spaces where spatial directions may have little meaning. It is applicable, for example, to data on population migration, transportation of people and goods, and movements retrieved from georeferenced posts in Twitter, Flickr, and other social media [2][9].

ACKNOWLEDGEMENT

This research was supported by EU within projects VaVeL (grant agreement 688380), SoBigData (grant agreement 654024) and BigData4ATM (grant agreement 699260).

REFERENCES

[1] G. Andrienko, N. Andrienko, H. Stange, T. Liebig, and D. Hecker. Visual analytics for understanding spatial situations from episodic movement data. *Künstliche Intelligenz*, 26(3): 241-251, 2012.

[2] G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel. *Visual Analytics of Movement*. Springer, 2013.

[3] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages*, pp 336-343, 1996.

[4] D.A. Keim, F. Mansmann, and J. Thomas. Visual analytics: how much visualization and how much analytics? *ACM SIGKDD Explorations Newsletter*, 11(2): 5-8, 2009.

[5] N. Andrienko and G. Andrienko. *Exploratory analysis of spatial and temporal data: a systematic approach*. Springer, 2006.

[6] D. Guo. Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics*, 15(6): 1041–1048, 2009.

[7] D. Guo, X. Zhu, H. Jin, P. Gao, and C. Andris. Discovering Spatial Patterns in Origin-Destination Mobility Data, *Transactions in GIS*, 16(3):411-429, June 2012.

[8] S. Gao, Y. Liu, Y. Wang, and X. Ma. Discovering spatial interaction communities from mobile phone data. *Transactions in GIS*, 17(3):463–481, 2013.

[9] T. von Landesberger, F. Brodkorb, P. Roskosch, N. Andrienko, G. Andrienko, and A. Kerren. MobilityGraphs: Visual Analysis of Mass Mobility Dynamics via Spatio-Temporal Graphs and Clustering. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):11-20, 2016.

[10] X. Zhu and D. Guo. Mapping Large Spatial Flow Data with Hierarchical Clustering. *Transactions in GIS*, 18(3):421-435, 2014.

[11] D. Guo and X. Zhu. Origin-Destination Flow Data Smoothing and Mapping. *IEEE Transactions on Visualization and Computer Graphics*, 20(12): 2043-2052, December 2014.

[12] J.J. van Wijk and E. van Selow. Cluster and Calendar-based Visualization of Time Series Data. In: G. Wills and D. Keim (eds.), *Proc. IEEE InfoVis'99*, pp. 4-9, 1999.

[13] H.-P. Kriegel, P. Kröger, and A. Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(1): Article No 1, March 2009.

[14] S. van den Elzen, J. Blaas, D. Holten, and J.J. van Wijk. Reducing Snapshots to Points: A Visual Analytics Approach to Dynamic Network Exploration. *IEEE Transactions on Visualization and Computer Graphics*, 22(1): 1-10, 2016.

[15] B. Bach, C. Shi, N. Heulot, T. Madhyastha, T. Grabowski, and P. Dragicevic. Time Curves: Folding Time to Visualize Patterns of Temporal Evolution in Data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1): 559-568, 2016.

[16] N. Andrienko and G. Andrienko. Visual analytics of movement: An overview of methods, tools and procedures. *Information Visualization*, 12(1): 3-24, 2013

[17] I. Boyandin. Visualization of Temporal Origin-Destination data. PhD thesis N 1786, University of Fribourg (Switzerland), 2013; URL: ilya.boyandin.me/assets/thesis.pdf; last accessed 18 March 2014.

[18] D. Guo. Visual analytics of spatial interaction patterns for pandemic decision support. *International Journal of Geographical Information Science*, 21(8): 859-877, 2007.

[19] W. Tobler. Experiments in migration mapping by computer. *The American Cartographer*, 14(2): 155-163, 1987.

[20] J. Wood, J. Dykes, and A. Slingsby. Visualisation of origins, destinations and flows with OD maps. *The Cartographic Journal*, 47(2): 117 – 129, 2010.

[21] J. Wood, A. Slingsby, and J. Dykes. Visualizing the dynamics of London's bicycle hire scheme. *Cartographica*, 46(4): 239 – 251, 2011.

[22] D. Holten, P. Isenberg, J.J. van Wijk, and J.-D. Fekete. An Extended Evaluation of the Readability of Tapered, Animated, and Textured Directed-Edge Representations in Node-Link Graphs. In *2011 Pacific Visualization Symposium (PacificVis)*, pp. 195-202, 2011.

[23] A. Rae. From spatial interaction data to spatial interaction information? Geovisualisation and spatial structures of migration from the 2001 UK census. *Computers, Environment and Urban Systems*, 33(3): 161–178, 2009.

[24] S. van den Elzen and J. van Wijk. Multivariate network exploration and presentation: From detail to overview via selections and aggregations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12): 2310-2319, Dec 2014.

[25] D. Phan, L. Xiao, R. Yeh, P. Hanrahan, and T. Winograd. Flow map layout. In Proc. *IEEE Symp. Information Visualization (InfoVis 2005)*, pp. 219-224, 2005.

[26] K. Verbeek, K. Buchin, and B. Speckmann. Flow map layout via spiral trees. *IEEE Transactions on Visualization and Computer Graphics*, 17(12): 2536–2544, 2011.

[27] O. Ersoy, C. Hurter, F. Paulovich, G. Cantareiro, and A. Telea. Skeleton-based edge bundling for graph visualization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12): 2364-2373, 2011.

[28] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. van Wijk, J.-D. Fekete, and D. Fellner. Visual analysis of large graphs:

state-of-the-art and future research challenges. *Computer Graphics Forum*, 30(6): 1719–1749, 2011.
[29] I. Boyandin, E. Bertini, and D. Lalanne. Visualizing the World's refugee data with JFlowMap. Poster at *Eurographics/IEEE Symp. Visualization EuroVis 2010*.
[30] I. Boyandin, E. Bertini, P. Bak, and D. Lalanne. Flowstrates: An approach for visual exploration of temporal origin-destination data. *Computer Graphics Forum*. 30(3): 971-980, 2011.
[31] G. Andrienko, N. Andrienko, P. Bak, D. Keim, S. Kisilevich, and S. Wrobel. A conceptual framework and taxonomy of techniques for analyzing movement. *Journal of Visual Languages and Computing*, 22(3): 213-232, 2011.
[32] B. Tversky, J.B. Morrison, and M. Betrancourt. Animation: can it facilitate? *International Journal of Human-Computer Studies*, 57(4): 247-262, 2002.
[33] R. Beecham, J. Wood, and A. Bowerman. Studying commuting behaviours using collaborative visual analytics. Computers, Environment and Urban Systems, doi: 10.1016/j.compenvurbsys.2013.10.007, 2013.
[34] G. Andrienko and N. Andrienko. Spatio-temporal aggregation for visual analysis of movements. In *IEEE Visual Analytics Science and Technology (VAST 2008)*, pp.51-58, 2008.
[35] G. Andrienko, N. Andrienko, C. Hurter, S. Rinzivillo, and S. Wrobel. Scalable Analysis of Movement Data for Extracting and Exploring Significant Places. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 19(7): 1078-1094, 2013.
[36] N. Andrienko and G. Andrienko. Spatial Generalization and Aggregation of Massive Movement Data. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 17(2):205-219, 2011.
[37] I.H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition), Morgan Kaufmann, San Francisco, CA, 2005.
[38] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. ACM KDD* 1996, 226-231.
[39] M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander. OPTICS: Ordering points to identify the clustering structure. In *Proc. ACM SIGMOD* 1999, pp. 49–60, 1999.
[40] T. Kohonen. *Self-Organizing Maps (Third Edition)*. Springer, Berlin, 2001.
[41] A. Ultsch. Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series. In *Kohonen Maps*, Elsevier, pp. 33–46, 1999.
[42] G. Andrienko, N. Andrienko, S. Bremm, T. Schreck, T. von Landesberger, P. Bak, and D. Keim. Space-in-Time and Time-in-Space Self-Organizing Maps for Exploring Spatiotemporal Patterns. *Computer Graphics Forum*, 29 (3): 913-922, 2010.
[43] C.C. Aggarwal, A. Hinneburg, and D. Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Spaces. *In ICDT '01 Proceedings of the 8th International Conference on Database Theory*, Springer-Verlag London, UK, PP. 420-434, 2001.
[44] D. François, V. Wertz, and M. Verleysen. The concentration of fractional distances. *IEEE Trans. on Knowledge and Data Engineering*, 19(7): 873-886, 2007.
[45] N. Andrienko and G. Andrienko. A visual analytics framework for spatio-temporal analysis and modelling. *Data Mining and Knowledge Discovery*, 27(1): 55-83, 2013.
[46] T.F. Cox, M.A.A. Cox. *Multidimensional Scaling*. Chapman and Hall, 2001.
[47] J.W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18(5): 401–409, May 1969.
[48] J. Schanda. *Colorimetry*. Wiley-Interscience, 2007.
[49] L. Brasseur. Florence Nightingale's visual rhetoric in the rose diagrams. *Technical Communication Quarterly*, 14(2): 161-182, 2005.
[50] C.A. Brewer. http://colorbrewer2.org/; last accessed 20 May 2016.
[51] J. Bertin. *Semiology of Graphics. Diagrams, Networks, Maps*. University of Wisconsin Press, Madison, 1983. Translated from J. Bertin: *Sémiologie graphique*, Gauthier-Villars, Paris, 1967.
[52] J.J. Flannery. The relative effectiveness of some common graduated point symbols in the presentation of quantitative data. *The Canadian Cartographer*, 8(2): 96-109, 1971.
[53] X. Jiang, C. Zheng, Y. Tian, and R. Liang. Large-scale taxi O/D visual analytics for understanding metropolitan human movement patterns. *Journal of Visualization*, 18(2): 185-200, 2015.
[54] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C.D. Hansen, and J.C. Roberts. Visual comparison for information visualization. *Information Visualization* 10(4): 289–309, 2011.
[55] Adobe Color CC. https://color.adobe.com/create/color-wheel/
[56] J.J. Thomas and K.A. Cook, eds. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, IEEE CS Press, 2005.

Gennady Andrienko is a lead scientist responsible for the visual analytics research at the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) and professor (part-time) at City University London. He co-authored two monographs 'Exploratory Analysis of Spatial and Temporal Data' (Springer, 2006) and "Visual Analytics of Movement" (2013) and 70+ peer-reviewed journal papers. Since 2007, Gennady Andrienko is chairing the ICA Commission on GeoVisualization. He co-organized scientific events on visual analytics, geovisualization and visual data mining, and co-edited 11 special issues of journals.

Natalia Andrienko has been working at GMD, now Fraunhofer IAIS, since 1997. Since 2007, she is a lead scientist responsible for the visual analytics research. Since 2013 she is professor (part-time) at City University Londond. She co-authored the monographs 'Exploratory Analysis of Spatial and Temporal Data' (Springer, 2006) and "Visual Analytics of Movement" (2013) and over 70 peer-reviewed journal papers. She received best paper awards at AGILE 2006, EuroVis 2015 and IEEE VAST 2011 and 2012 conferences, best poster awards at AGILE 2007, ACM GIS 2011 and IEEE VAST 2016, and VAST challenge awards 2008 and 2014.

Dr. Georg Fuchs is a senior research scientist and project manager at Fraunhofer IAIS working in the field of visual analytics with a strong emphasis on spatio-temporal data analysis. His research interests include information visualization in general and visualization of spatio-temporal data in particular, visual analytics methodologies, task-driven adaptation of visual representations and Smart Visual Interfaces, as well as computer graphics and rendering. Georg Fuchs has co-authored 38+ peer-reviewed research papers and journal articles, received a best short paper award at Smart Graphics 2008 and a VAST challenge award 2014.

Jo Wood is a Professor of Visual Analytics at the giCentre, City University London, where he designs, builds and applies data visualization software. He has particular interests in data visualization, narrative in visual analytic design and the intersection between visualisation design and art. His background is in Geographic Information Science and terrain analysis, and has, since 1990 been developing new methods and software that bridge the GI Science and Data Visualization domains. He has over 50 peer-reviewed articles in this area including best paper and honourable mention awards from the EuroVis, PacificVis and Infovis conferences. He is head of the Department of Computer Science at City University London and has been a member of the organising and programme committees of a number of international conference series in GI Science and visualization including IEEE Infovis and VAST, Eurovis, GIScience, Spatial Accuracy and Geomorphometry.