Multi-perspective Analysis of Movement Data with Visual Analytics

Gennady Andrienko
Natalia Andrienko
http://geoanalytics.net
Visual Analytics

Enabling synergetic work of humans and computers
Types of spatio-temporal data

Spatial events

Trajectories

Spatial time series

Local time series

Spatial distributions

(id, location, time, attributes)
Transformations of spatio-temporal data structures

Spatial events → Spatial time series

Trajectories → Spatial time series

Integration

Extraction, disintegration

Aggregation

Extraction

Spatial time series

Local time series → Spatial time series

Spatial distributions → Spatial time series

Projection
Transformations enable multi-perspective analysis of movement data
Running example dataset: trajectories of cars in Milan

GPS-tracks of 17,241 cars in Milan, Italy

Time period: from Sunday, the 1st of April, to Saturday, the 7th of April, 2007

Received from Octo Telematics
www.octotelematics.com
special thanks to T.Martino

Data structure:
- Anonymized car identifier
- Date and time
- Geographic coordinates
- Speed

The trajectories from one day are drawn on a map with 5% opacity
Perspective 1: Movement data in the form of trajectories
Density-based clustering of trajectories: What distance measure to use?

- Trajectories are time series of spatial positions and other movement attributes.
- Trajectories are complex objects with heterogeneous properties: positions in space and in time, shape, dynamics of speed, ...
- A single distance measure accounting for all properties would be hard to implement and results would be hard to interpret.
- It is more feasible to create a library of simple distance measures (a.k.a. distance functions) addressing different properties. For example,
  - spatial distance between origins and/or between destinations,
  - average spatial distance between corresponding points along the routes,
  - average spatial distance between points reached at the same times, ...
- Different aspects of trajectories are studied using different distance functions.
DB clusters of trajectories (example 1)

Distance function: the average spatial distance between the origins and between the destinations;

\[ R=750m, \; N=5 \]

Only 18 largest clusters are shown.
Summarised representation of clusters of trajectories

Minor flows are omitted for a clearer view.
DB clusters of trajectories (example 2)

Distance function: “route similarity”, i.e., the average spatial distance between the corresponding points along the route; $R=750m$, $N=5$

Only 18 largest clusters are shown.
The same clusters represented in a summarised form
Interactive progressive clustering

Applying different distance measures at different steps

Data: trajectories of cars in Milan

Step 1: clustering according to the spatial proximity of the end points
Distance function: “common ends”

Question: what are the most frequent destinations of car trips?
Interactive progressive clustering

Applying different distance measures (2)

Data: one (or more) selected cluster(s) from the previous step

Step 2: clustering according to the similarity of the routes (shapes)

Distance function: “route similarity”

Question: what routes are usually taken to get to the selected destination?
Clustering of very large sets of trajectories

Problem: clustering of complex objects (such as trajectories) involving non-trivial distance functions (such as “route similarity”) can only be done in RAM, i.e. for a relatively small dataset.

Our approach:

1. Take a subset (sample) of the objects suitable for processing in RAM.
2. Discover clusters in the subset.
3. Load the remaining objects into RAM by portions. 
   *Classify* each object = identify to which of the discovered clusters the object belongs.
   Store the result of the classification in the database.
4. Take the objects that remained unclassified and apply steps 1 to 3 to them.
   Repeat the procedure until no meaningful new clusters can be discovered.

Question: how to identify the cluster where an object belongs?
Classifier, the main idea

- From each cluster $C_i$ select one or more representative objects (prototypes) and respective distance thresholds: 
  \[
  \{ (p_{t_1}, d_1), \ldots, (p_{t_n}, d_n) \} \text{ such that } \forall o \in C_i \exists k, 1 \leq k \leq n: \text{distance}(o, p_{t_k}) < d_k
  \]
- The set of all cluster prototypes with the respective distance thresholds defines the classifier.

- A new object $o'$ may be ascribed to the cluster if the same condition holds for it.
  \[\Rightarrow\text{For each object from a large database:}\]
  - measure the distances to all prototypes;
  - take the closest prototype among those with the distances below the thresholds and ascribe the object to the respective cluster;
  - if no such prototypes found, label the object as unclassified.

- To select prototypes:
  - Divide the cluster into “round” subclusters
  - Take the medoid of each subcluster as one of the prototypes
  - Take the maximum of the distances from the subcluster medoid to the subcluster members as the distance threshold for this prototype
Dividing a cluster into round sub-clusters: an illustration using points

This can be done by a variant of the K-medoids clustering algorithm where the desired maximum radius of a subcluster is a parameter.
Divison of a cluster of trajectories into "round" subclusters

<table>
<thead>
<tr>
<th>prototype ID</th>
<th>Distance threshold</th>
<th>Original subcluster size</th>
<th>N neighbours found in the test</th>
<th>Mean distance to the original neighbours</th>
<th>Mean distance to the found neighbours</th>
</tr>
</thead>
<tbody>
<tr>
<td>89133</td>
<td>438.2</td>
<td>4</td>
<td>0</td>
<td>240.5</td>
<td>0</td>
</tr>
<tr>
<td>96013</td>
<td>200.0</td>
<td>8</td>
<td>0</td>
<td>96.9</td>
<td>0</td>
</tr>
<tr>
<td>6548</td>
<td>526.5</td>
<td>29</td>
<td>0</td>
<td>161.1</td>
<td>0</td>
</tr>
<tr>
<td>43285</td>
<td>200.0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>34239</td>
<td>414.3</td>
<td>19</td>
<td>0</td>
<td>186.7</td>
<td>0</td>
</tr>
<tr>
<td>32809</td>
<td>368.2</td>
<td>15</td>
<td>0</td>
<td>121.2</td>
<td>0</td>
</tr>
<tr>
<td>141138</td>
<td>485.0</td>
<td>10</td>
<td>0</td>
<td>271.2</td>
<td>0</td>
</tr>
<tr>
<td>109120</td>
<td>200.0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Maximum subcluster radius

To select appropriate cluster prototypes, the density-based clusters will be divided into "round" subclusters.

Maximum radius of a subcluster? 1000.0
To obtain meaningful results, the analyst may need to review and, possibly, edit the classifier.

<table>
<thead>
<tr>
<th>prototype ID</th>
<th>Distance threshold</th>
<th>Original subcluster size</th>
<th>N neighbours found in the test</th>
<th>Mean distance to the original neighbours</th>
<th>Mean distance to the found neighbours</th>
</tr>
</thead>
<tbody>
<tr>
<td>89133</td>
<td>436.2</td>
<td>44</td>
<td>0</td>
<td>240.5</td>
<td>0</td>
</tr>
<tr>
<td>96013</td>
<td>200.0</td>
<td>82</td>
<td>0</td>
<td>94.9</td>
<td>0</td>
</tr>
<tr>
<td>6548</td>
<td>526.5</td>
<td>29</td>
<td>0</td>
<td>161.1</td>
<td>0</td>
</tr>
<tr>
<td>43285</td>
<td>200.0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>34239</td>
<td>414.3</td>
<td>19</td>
<td>0</td>
<td>186.7</td>
<td>0</td>
</tr>
<tr>
<td>32809</td>
<td>368.2</td>
<td>15</td>
<td>0</td>
<td>121.2</td>
<td>0</td>
</tr>
<tr>
<td>141138</td>
<td>485.0</td>
<td>10</td>
<td>0</td>
<td>271.3</td>
<td>0</td>
</tr>
<tr>
<td>109120</td>
<td>200.0</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Should I keep the three branches in one cluster?

Or should I divide the cluster into two or three clusters?

Some trajectories are not very similar to the others. Should such trajectories be in the cluster?

Is it good to have this prototype? This is not a core trajectory of the cluster.
Example of interactive editing
What are the most frequent routes on Wednesday?

Result of clustering of single-day trajectories by route similarity
How frequent are these routes during the whole week?

Result of building a classifier and applying it to the whole set of trajectories
Further analysis of the trajectories

- The analysis is continued by loading a subset of the unclassified trajectories ("noise") to RAM, applying clustering to it, building a new classifier, and applying the classifier to the whole set of unclassified trajectories.

- Empirical experience:
  - With each new iteration step, the number and the sizes of discovered clusters substantially decrease in comparison to the previous step.
  - After 4-5 steps of the procedure, only very small clusters can be discovered.
  - The analyst’s effort needed for editing of the classifier also decreases.
  - The editing effort is high for big clusters with high internal variation, which mostly appear in the first step; the following clusters are smaller and “cleaner”.

- Unfortunately, no formal criterion for terminating the procedure.
Where to read more


Interactive Visual Clustering of Large Collections of Trajectories

*IEEE Visual Analytics Science and Technology (VAST 2009)*

Proceedings, IEEE Computer Society Press, 2009, pp. 3-10
Analysis of movement attributes

Investigate speed variation along a selected route: single day
Analysis of movement attributes

Investigate speed variation along a selected route: whole week

Speed (km/h)
- < 5.00
- 5.00 - 10.00
- 10.00 - 15.00
- 15.00 - 30.00
- 30.00 - 50.00
- 50.00 - 75.00
- 75.00 - 100.00
- > 100.00
- NaN
Perspective 2: Movement data in the form of spatial events
Example of analysis focusing on movement events

- Data: trajectories of cars in Milan
- Task: find places of traffic congestions and determine their characteristics (times of the congestions, durations, numbers of cars involved, ...)
- Traffic congestion \( \approx \) dense spatio-temporal cluster of low speed movement events
  - Movement direction must be taken into account
- Places of interest: areas where at least one traffic congestion occurred \( \approx \) areas containing the clusters
- Characteristics of places: time series of event counts, vehicle counts, ...
- Data transformations:
  Trajectories \( \rightarrow \) Events \( \rightarrow \) Places \( \rightarrow \) Spatial time series
Step 1: extract low speed events from the trajectories

Low speed := speed ≤ 10 km/h
Step 2: density-based clustering of events
by spatio-temporal positions and directions

Distance function:

\[
d = \begin{cases} 
\infty, & \text{if } (d_s > D_s) \text{ or } \exists i \mid (d_i > D_i), \quad i = 0..n \\
D_s \cdot \max \left( \frac{d_s}{D_s}, \frac{d_0}{D_0}, \ldots, \frac{d_n}{D_n} \right), & \text{if } (a) \quad \text{– neighbourhood defined as a cube} \\
D_s \cdot \sqrt{\frac{(d_s}{D_s})^2 + \sum_{i=0}^{n} \left( \frac{d_i}{D_i} \right)^2}, & \text{if } (b) \quad \text{– neighbourhood defined as a sphere}
\end{cases}
\]

\(D_s\) – spatial distance threshold; \(D_0, D_1, \ldots, D_N\) - distance thresholds for other attributes
\(d_s, d_0, d_1, \ldots, d_N\) – distances; \(d_s\) – distance in space

Distance in time \((t_1, t_2)\) are intervals:

\[
d_t(t_1, t_2) = \begin{cases} 
t_2^{\text{end}} - t_1^{\text{start}}, & \text{if } t_1^{\text{end}} < t_2^{\text{start}} \\
t_1^{\text{start}} - t_2^{\text{end}}, & \text{if } t_1^{\text{start}} > t_2^{\text{end}} \\
0, & \text{otherwise}
\end{cases}
\]

Distance for a cyclic attribute \((V\) is the cycle length):

\[
d(v_1, v_2, V) = \begin{cases} 
|v_1 - v_2|, & \text{if } |v_1 - v_2| < V/2 \\
V - |v_1 - v_2|, & \text{otherwise}
\end{cases}
\]

E.g., direction: \(V = 360^\circ\); \(d(5^\circ, 355^\circ, 360^\circ) = 10^\circ\)
The result of the density-based clustering of the slow movement events by their spatial positions, temporal positions, and movement directions (STD) with the distance thresholds 100 meters, 10 minutes, and 20 degrees and the minimum number of neighbors 5.
The STD-clusters, noise hidden
Step 3: unite STD-clusters in SD-clusters

Cluster the events from the STD-clusters by the spatial positions and directions.

The result of the density-based clustering with the spatial distance threshold of 100 m and direction distance threshold of 20°.
Events that occurred in same or close places but in different times were formerly in different clusters, but now they are in the same clusters.
One SD-cluster includes one or several STD-clusters.
Step 4: outline the places of interest

Build spatial buffers around the SD-clusters of events

The places are painted according to the prevailing movement directions of the respective events.
Belt road north-south on the east of the city (A50)

Extended areas of congested traffic directed to the south and southeast

Smaller areas of obstructed movement directed to the north and northwest

Belt road west-east on the north of the city (A4)

Very long area of congested traffic directed to the east

Long area of congested movements directed to the west
Step 5: aggregate data by the places and by suitable time intervals, e.g., hourly

Place-referenced time series of the counts of slow movement events
The temporal diagrams show the variation of the attribute value (vertical dimension) over time (horizontal dimension).
Map fragment (northwest) enlarged

Congested traffic in the afternoon in the direction out of the city (northwest)

Congested traffic in the morning in the direction to the south
Other map fragments enlarged

Northeast

morning and midday

morning

morning

morning and afternoon

afternoon

morning and afternoon

afternoon
Where to read more

- IEEE VAST 2011 paper (best paper award)

  G. Andrienko, N. Andrienko, C. Hurter, S. Rinzivillo, S. Wrobel
  From Movement Tracks through Events to Places: Extracting and Characterizing Significant Places from Mobility Data

- Extended version, covering also scalable clustering of events

  G. Andrienko, N. Andrienko, C. Hurter, S. Rinzivillo, S. Wrobel
  Scalable Analysis of Movement Data for Extracting and Exploring Significant Places
  *IEEE Transactions on Visualization and Computer Graphics*, 2013, 19(7), 1078-1094
Perspective 3: Movement data in the form of spatial situations
Spatio-temporal aggregation of trajectories
Division of the territory

Characteristic points from the trajectories  Spatial clusters of characteristic points

Cluster centres → seeds for Voronoi tessellation

Details:
Natalia Andrienko, Gennady Andrienko

Spatial Generalization and Aggregation of Massive Movement Data
http://doi.ieeecomputersociety.org/10.1109/TVCG.2010.44
Spatial situations: presence

Circle area is proportional to value:

- → 186.0
- → 0.0
Spatial situations: flows

...03-04h

04-05h

05-06h

N moves by time intervals

0.00 91.00
Clustering of spatial (flow) situations by similarity
Comparison of clusters of spatial situations

Values for cluster 9 have been subtracted from values for all other clusters
Where to read more

N.Andrienko, G.Andrienko, H.Stange, T.Liebig, D.Hecker

Visual Analytics for Understanding Spatial Situations from Episodic Movement Data


http://dx.doi.org/10.1007/s13218-012-0177-4
Perspective 4: Movement data in the form of local time series
An alternative view of spatial time series: a set of local time series
An alternative view of spatial time series: a set of local time series

We wish to represent the essential characteristics of the ST-variation explicitly by a formal model or a set of models.
Methods for spatio-temporal modelling (e.g. STARIMA)

- Account for spatial and temporal dependencies
- Require prior specification of multiple weight matrices expressing impacts among locations for different temporal lags
  - may be difficult (the impacts are not easy to quantify)
- Build a single global model of the entire spatio-temporal variation
  - It does not necessarily perform better than a set of local temporal models
- Assume spatial smoothness of the modelled phenomenon, i.e., closer places are more similar than more distant ones
  - May be not very suitable for spatially abrupt phenomena
Existing techniques for time series modelling

+ Widely available in numerous statistical packages and libraries → can be applied to spatially referenced time series

- The modelling methods are designed to deal with singular time series → hard to use for a large number of time series

- Separate consideration of each time series ignores the phenomenon of spatial dependence (relatedness and similarities among spatial locations or objects)

- Separate consideration of each time series does not allow data abstraction and generalisation over space
Combination of spatial and temporal modelling

- **Approach 1:**
  1. Model the temporal variation independently for each location
  2. Model the spatial variation of the parameters of the temporal models, e.g., as a random field
  - Assumes that the character of the temporal variation is the same everywhere and only the parameters differ

- **Approach 2:**
  - Model the spatial variation independently for each time step, e.g., as a random field
  - Model the temporal variation of the parameters of the spatial models at each location

- Both approaches assume spatial smoothness of the phenomenon
Our approach

Step 0: Prepare data

Step 1: Group time series
- Clustering
- Progressive clustering
- Interactive re-grouping

Step 2: Analyse and model
- Select method
- Set parameters
- Compute and view model

Step 3: Evaluate model
- Predict values and compute residuals
- Visualise and analyse residuals

Step 4: Store model set

For each group:

Details:
Natalia Andrienko, Gennady Andrienko
A Visual Analytics Framework for Spatio-temporal Analysis and Modelling
Data Mining and Knowledge Discovery, 27(1), 55-83, 2013
http://dx.doi.org/10.1007/s10618-012-0285-7
Step 1: Clustering of local TS

- Here: k-means (Weka) but may be another partition-based method
- Tried different k from 5 to 15
- Immediate visual response facilitates choosing the most suitable k
Step 1: Re-grouping by progressive clustering
Step 2: Analysis and modelling

A) Check automatically detected time cycles in the data.
B) Select the current class (cluster) for the analysis and modelling.
C) Build the representative TS.
D) Select the modelling method.
E) View and modify model parameters (this section changes depending on the selected modelling method).
Step 2: Analysis and modelling
Step 2: Analysis and modelling

Time series analysis

Check the presence of cyclic variation of the data:
- Cycle: daily
- Cycle: weekly

N moves by hours

Perform modelling based on the
- percentile 50
- mean
- excluding 5%

alpha (overall smoothing) = 0.8
beta (trend smoothing) = 0.03
gamma (seasonal smoothing) = 0.1
Step 2: Analysis and modelling
Step 3: Model evaluation (analysis of residuals)

- The goal is not to minimise the residuals
  - The model should not reproduce all fluctuations and outliers present in the data
  - This should be an abstraction capturing the characteristic features of the temporal variation
  - High values of the residuals do not mean low model quality

- The goal is to have the residuals randomly distributed in space and time (no detectable patterns)
  - This means that the model correctly captures the characteristic, non-random features of the temporal variation
No systematic bias: approximately equal numbers of positive and negative errors in each time step

No periodic increases and decreases at the level of the whole group

However, we are not sure about individual objects
More detailed analysis by subgroups

It may be reasonable to consider this subgroup separately -> back to re-grouping
Use of a model for prediction

- We obtain a common model for a group (cluster) of time series
  - Predicts the same values for all objects/places of the group
  - The statistical properties of the distribution of the predicted values in each place differ from the distribution of the original values

- Adjustment of the prediction for individual objects/places:
  - Compute and store the basic statistics (quartiles) of the original values for each object/place i: $Q_{1i}, M_i, Q_{3i}$
  - Compute the statistics of the model-predicted values for the same time steps as the original values: $Q_1, M, Q_3$ (common for the cluster)
  - Shift (level adjustment): $S = M_i - M$
  - Scale factors (amplitude adjustment): $F_{low} = \frac{M_i - Q_{1i}}{M - Q_1}$, $F_{high} = \frac{Q_{3i} - M_i}{Q_3 - M}$
  - Let $v^t_i$ be the model-predicted value for an arbitrary time step $t$ and $v^t_i$ the individually adjusted value for the place/object $i$
    
    $v^t_i = \begin{cases} 
    M + F_{low} \cdot (v^t - M) + S, & \text{if } v^t < M \\
    M + F_{high} \cdot (v^t - M) + S, & \text{otherwise}
    \end{cases}$
Use of a model for prediction: example

Common prediction for a cluster:

Set of individually adjusted predictions for this cluster:
Prediction based on the models

[Diagram showing time intervals and data analysis graphs]
Comparison of actual values with predicted (e.g., in monitoring)

Absolute differences

Normalized differences

Actual N moves by hours - predicted

Actual N moves by hours - predicted divided by variance of predicted
Analysis and modelling of relationships between two time-variant attributes
Data transformation and clustering

- Dependency of attribute $A(t)$ on attribute $B(t)$:
  - Divide the value range of $B$ into intervals
  - For each interval, collect all values of $A$ that co-occur with the values of $B$ from this interval
  - Compute statistics of the values of $A$: minimum, maximum, median, mean, percentiles …
  - For each of these, there is a series $B \rightarrow A$, or $A(B)$
Dependency modelling: flow → maximal mean speed
Dependency modelling: mean speed $\rightarrow$ maximal flow
Graphical representation of the models built

9th decile of Average speed (km/h) by hours

Max of N moves*50 depending on Average speed (km/h)
Use of the models:

*simulation of extraordinary traffic from given places*
The bottlenecks can be revealed even before the simulation.
Simulated trajectories

Some traffic re-routed to the south:
The speeds on the northern motorway
Animation of simulation results
Presence and flows for selected time intervals

20:00 – 20:10
21:00 – 21:10
21:30 – 21:40
22:00 – 22:10
23:00 – 23:10
00:00 – 00:10

Circle area is proportional to value.
N moves by time intervals
Multi-perspective analysis of movement

Trip destinations, routes...

Movement data

- Moving objects
- Trajectories
- Spatial event data
- Spatial time series
- Locations
- Local time series
- Times
- Spatial distributions

Low speed events → traffic jams

Periodic variation of flow volumes; Dependencies volume vs. speed

Periodic (daily and weekly) variation of spatial situations
Ch.1. Introduction
Ch.2. Conceptual framework
Ch.3. Transformations of movement data
Ch.4. Visual analytics infrastructure
Ch.5. Visual analytics focusing on movers
Ch.6. Visual analytics focusing on spatial events
Ch.7. Visual analytics focusing on space
Ch.8. Visual analytics focusing on time
Ch.9. Discussion and outlook