ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

Multiscale Spatio-Temporal Data Aggregation and Mapping for Urban Data Exploration

Etienne Côme[1] and Anaïs Remy[2] *

1- Université Paris-Est, COSYS, GRETTIA, IFSTTAR,
F-77447 Marne-la-Vallée, France

2- SNCF, Innovation & Recherche,
40 avenue des terroirs de France, 75611 Paris, France

Abstract. Maps seem the most intuitive way to visualize massive urban data but they also raise some well-known graphical problems (such as visual clutter, etc.). This paper focuses on processing massive spatio-temporal data in order to ease multi-scale exploration. To this end, we describe a preprocessing tool that enables the automatic creation of a multi-resolution grid from a high resolution grid of spatio-temporal data in a format compatible with webmapping applications (vector tiles). The use of this tool is exemplified through a prototype that offers the possibility to navigate into a massive itinerary request dataset collected in the Ile-de-France region.

1 Interactive Maps and Massive Urban Data

In this work, we focus on processing massive mobility data in order to provide multi-scale information while tacking into account temporal dynamics. To address these issues, we developed a code called "pypixgrid" for vector tile generation from grid data. Tile systems are now commonly used for geographical multi-scale data management. We propose here a generic tool that automatically generates vector tiles from any massive grid data and allows:

- A precise aggregation control process from a zoom level to another;
- The management of temporal aspects additionally to geographic features.

In a second step, we developed an interactive visualization prototype, called "living mobility generators" based on itinerary requests data and directly connected to the previously generated vector tiles. This case study is a first illustration of "pypixgrid" code usage. Our goal is to visually explore mobility dynamics and their links with territorial features. A key issue is to provide a simple interface allowing an intuitively navigation through spatial and temporal features.

This work is a proof of concept developed in collaboration between SNCF Innovation & Research department and IFSTTAR. Data processing and visualization have been implemented on a set of historical data and on different French areas at urban and regional levels.

*The authors thank SNCF Transilien and Keolis Lyon for supporting this research project.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

2 Spatio-Temporal Multi-Scale Data Aggregation for Interactive Exploration

One of the main challenge in providing a fluid navigation in massive urban datasets come from the size of the data and their spatio-temporal nature. A classical way to solve this question is to perform some sort of aggregation (binning) in order to reduce the size. This technique may also be used to avoid visual clutter and even enhanced using a data-dependent binning [1]. Such lossy compression techniques could solve the problem of data weight but open the question of defining an appropriate bin size. Depending on the chosen bin size, it is well known (in statistics and geography [2]) that different aspects of the underlying data could be observed (or hidden). A natural extension of this concept is therefore to propose several scale of analysis to the viewer by computing several aggregations with different bin sizes. This approach has another advantage: if the aggregation scale is proportional to the View scale (using coarse binning for large scale views and finer binning for zoomed views) the size of the data needed to draw the map can in fact be controlled. At fine scales where the binning is small (and where therefore the data weight more) only a small part of the whole dataset is visible to the user and we can provide only the relevant part of the dataset to the user. This technique is commonly used in web-mapping where network bandwidth limits the amount of data that can be rapidly transfered to the viewer. Building on this well known technique for spatial data we build an automatic multi-scale aggregation tool for statistical data aggregation that we call "pypixgrid". Taking the finest grid as input, this tool builds a serie of grids by iteratively aggregating the previous grid by a specified factor using a user defined aggregating function. With such an approach the series of grids has typically the following form:

$$(30m \times 30m) \rightarrow (60m \times 60m) \dots (240m \times 240m) \rightarrow (480m \times 480m)$$

The first grid uses cells of 30m by 30m, the second's are four time bigger and so on. Eventually, each of these grids is sliced into vector tiles using a format commonly used in web-mapping applications in order to enable a fast access to the portion of the data needed to display the current view to the user (i.e. which depends on the current position and zoom level). The definition of the relationship between the grid pyramid and the tiles pyramid (which defines the available zoom levels) is also parameterizable using a simple scheme.

This solution enables the exploration of massive spatial datasets and can be extended to include temporal information and this was one of the motivations for this work.

In the same spirit as for space, a nested binning scheme can be used for time. This will increase the complexity of the interaction with the user during exploration since two kinds of zoom must be handled in this case (spatial and temporal). In this first prototype, we limited ourselves to a unique aggregation period for the time dimension of the data to avoid this issue. To conclude,

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

this dataprocessing tool can handle a spatial grid over which several statistical variables are observed. These variables may vary with respect to time or not. The tool produces web-map tiles ready to be visualized with web-mapping tools enabling an easy exploration of spatio-temporal data at different scales.

3 Case Study: Itinerary Request Data

Trip planners are used to find multimodal itineraries in response to "door to door" travel demands. This service is available through various supports: website, smartphone applications, etc. These tools are commonly used on many French regions. For example, there are about 80 thousand itinerary requests per hour in the Ile-de-France region (Paris and its suburbs) through SNCF Transilien tools. So, why not examine these massive digital activities in order to discover new insights on mobility demand on territories?

For technical reasons, trip planning tools generate automatically log files with full itinerary requests made by users and calculated responses. Figure 1 presents the data collected on the user query and the service response in these logs. Although this information can contain bias (demand expressed by only a part of transit users) [3, 4, 5], this kind of data are interesting because they provide a new source about transit information needs and "door to door" mobility demand on territories.

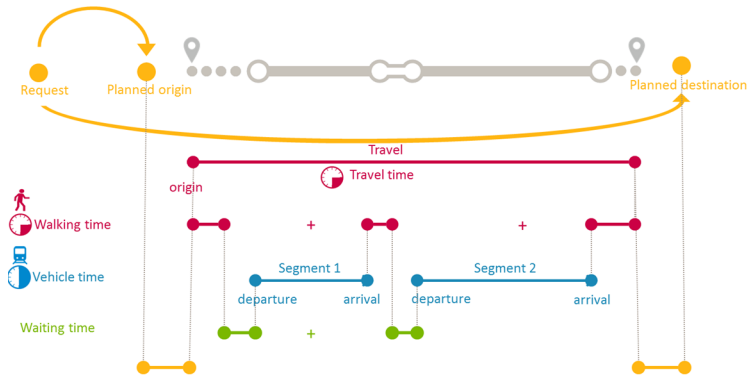

Fig. 1: Itinerary request and response data collection.

It's noteworthy that theses log files record each server call and response and don't necessarily reflect the users' actual behaviors. For example, if a user requests the same itinerary three times, we count three itineraries although there is only one demand. It is one of the actual limits of the data and it can cause some bias.

However, the data also provide very detailed information with precise locations of origin and destination that can be an address, a station or a point of interest. We also know the moment when the user requests an itinerary and

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

when he wants to start or finish his travel. The response gives back information about transportation supply through several indicators like walking time, vehicle time and waiting time [6].

The scope of the study is on both the Ile-de-France region and Lyon urban agglomeration in France. Both territories are characterized by strong population densification and mass transit issues. We collected all itineraries requests made by users and paths calculated in response during few months in 2015. These dataset like other automated data collection systems became quickly large, about 1 hundred millions of itineraries requests.

We also collected others territorial data such as employment, facilities and special events (football matches, concerts, etc.) thanks to INSEE's open data and some available API. Then, we defined a synthetic "contextual" indicator based on these data: Each contextual feature has a relative weight by geographical zone (with the sum of the weights equal to one). By combining these contextual data with mobility demand (through itinerary requests) we want to visually establish a link between them in order to improve our understanding of mobility generators.

From an applicative perspective, the high temporal resolution of such a dataset is of main interest compared to traditional surveys which can't be used to answer questions regarding demand variability or the impact of specific temporal events since they only provide an average vision of the mobility demand. An interactive visual exploration of such a sizable dataset can be a first step towards a better understanding of such questions, and is therefore of great interest. However providing a fluid environment for spatio-temporal data exploration is not a trivial task, building on existing technologies we propose a proof of concept for this dataset based on multi-scale data aggregation that could be used for other types of spatio-temporal data. With respect to the implementation of this tool it's a python module which works with a postgis database commonly used to store spatio-temporal data. The aggregation process is defined in a simple text file and the outputs of the program fulfill the vector tile specifications[1].

4 Interface Design

New technologies and possibilities of data visualization are now very impressive in order to make tangible some complex phenomena from large data sets. This is a growing research area that requires various skills in data processing, IT development, and also generative design.

In this case, we want to visually highlight mobility dynamics from itinerary requests data in order to:

- Identify attractors / sprayers of mobility during the day;
- Visualize links with contextual environment features (employment, housing, facilities, cultural and social events).

[1] https://www.mapbox.com/vector-tiles/specification/

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

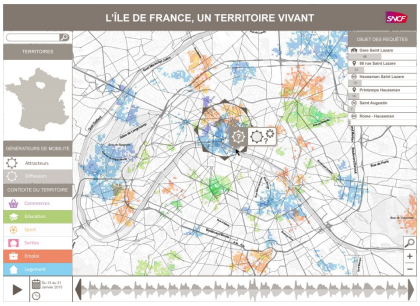

Fig. 2: First sketch of the prototype.

As a first step, we present a sketch of data visualization made by a designer [7] in Figure 2. The middle area is a map showing origins and destinations of planned trips during the day, and different color layers represent contextual data like employment, housing, facilities, cultural and social events.

Several users' interactions are available:

- **Territories module**: To select the geographic scope of study;
- **Attractors / Sprayers module**: To select only origins or destinations of planned trips;
- **Contextual module**: To filter one or more color layers (employment, housing, facilities, events, etc.).

In addition, the user can explore the whole dataset by choosing the map zoom level (by scroll interaction) and the time step thanks to an hourly timeline with playback and control capability. The main difficulty in the prototype was to develop a graphical encoding to combine simultaneously different contextual data. To solve this visualization issue, some inspirations were derived from printing process. The final solution represents the weight of the different context variables for each of the grid cells together with the total volume of itinerary requests with origin (or destination). A screen-shot of the final live prototype with real data is given in Figure 3.

5 Conclusion

In this paper we have described a pre-processing tool for spatio-temporal datasets that enables multi-scale data aggregation and produces vector tiles for an easy integration with web-mapping libraries. One application of such an approach on itinerary request data is also described in this paper but the general approach is easily expendable to another massive spatio-temporal datasets with several scales of interest. With respect to our current and further works, we are working on automatic tools to ease the definition of the graphical encoding used to display such multi-scale spatio-temporal dataset.

ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017, i6doc.com publ., ISBN 978-287587039-1.
Available from http://www.i6doc.com/en/.

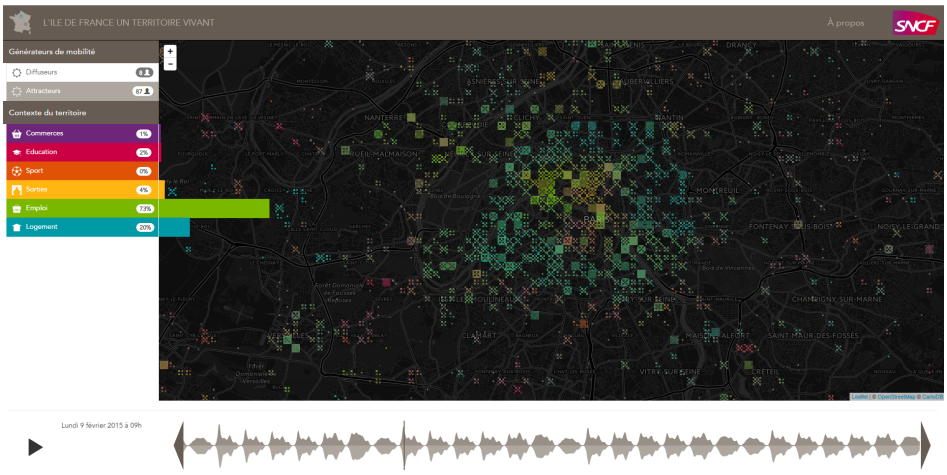

Fig. 3: Final prototype on a large scale view.

References

[1] A. Chua and A. Vande Moere. Binsq: visualizing geographic dot density patterns with gridded maps. *Cartography and Geographic Information Science*, pages 1–20, 2016.

[2] S. Openshaw and S. Openshaw. The modifiable areal unit problem. *Geo Abstracts University of East Anglia*, 1984.

[3] M. Trépanier, R. Chapleau, and B. Allard. Can trip planner log files analysis help in transit service planning? *Journal of Public Transportation*, 8:79–103, 2005.

[4] P. Colpaert, A. Chua, R. Verborgh, E. Mannens, R. Van de Walle, and A. Vande Moere. What public transit api logs tell us about travel flows. In *Proceedings of the 25th International Conference Companion on World Wide Web.*, WWW '16 Companion, pages 873–878, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[5] J. Jones, C. Cloquet, A. Adam, A. Decuyper, and Isabelle Thomas. Belgium through the lens of rail travel requests: Does geography still matter? *ISPRS International journal of geo-information*, 5(11):216–238, 2016.

[6] A. Remy, M. Chandesris, S. Mastalerz, A. Hyenne, and E. Bousquie. Multimodal travel demand based on itinerary requests. In *Proceedings of the World Conference on Transport Research*, 2016.

[7] C. Guérin, M. Chandesris, and A. Rémy. Territoire vivant. *Sciences du Design*, 3:30–33, 2016.