A Seriation Based Framework to Visualize Multiple Aspects of Road Transport from GPS Trajectories

Alexandre Dubray[1], Siegfried Nijssen[1], Isabelle Thomas[2] and Pierre Schaus[1]

***Abstract*— Heavy good vehicles GPS trajectories can tell us a lot about the structure of the road network and the spatial organization of a territory. In this paper we introduce a framework for visualizing multiple aspects of road transport on a single map. The final result allows unveiling at one glance areas with similar properties and lets the user gain insight in the geography of a studied area. More concretely, the map is divided into basic spatial units and for each spatial unit relevant features are computed from the GPS trajectories, such as the time spent in congestion, or resting. A seriation algorithm is then used to discover a linear order of the spatial units where successive units of the order share similar features. A rainbow color scale is used to assign a color to each spatial units to visually reveals similarities with respect to the extracted features. As a case study, we analyze the Belgian territory using a real-world data set. We compare our approach with clustering based methods, demonstrating that it reveals more of the spatial structure. The source code of our implementation is available under an open-source license for anyone interested in analyzing GPS trajectories.**

I. Introduction

Road freight transport has constantly increased over the last decades. Estimates suggest that freight transport will increase by 60% by 2050 in Europe [1], creating new logistic challenges. In parallel, the abundance of GPS-derived spatial data available is also constantly growing. Studies aiming to understand and analyze GPS trajectories are fairly recent [2, 3, 4, 5]. Such studies could provide insights to decision makers, allowing for decisions on organizational or infrastructure changes. An important aspect here is how to communicate the information hidden within data to decision makers. Flexible tools are needed to create useful visualizations.

An important challenge in communicating data about geographical territories is that territories may have large numbers of different attributes associated to them, related to both economic activities and infrastructure. Existing methods for visualizing data on geographical areas can be put into two categories: (1) methods that create one separate map per attribute [3, 6] or (2) methods that create one map in which colors correspond to clusters of similar areas [7, 8, 9]. The disadvantage of category (1) is that a large number of maps need to be generated; the disadvantage of category (2) is that detail is lost in the clustering process. Moreover, since a cluster might contain areas with different values for some attributes, they are less easy to interpret.

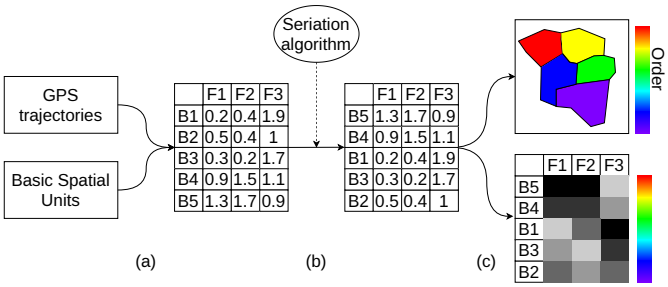


Fig. 1: Example of the proposed approach. a) Starting from a set of GPS trajectories, three numerical features (F1, F2 and F3) are computed on five BSUs (B1, B2, etc.). b) A seriation algorithm is then used to find an order of the BSUs so that successive BSUs are similar c) Two visualizations are generated. A map that show the BSU and their color (top) and a heatmap (higher values in each features darker) of the feature space (bottom).

To solve these issues, we propose the approach illustrated in Figure 1 on a small example. It starts from a set of trajectories and some basic spatial units (BSU) (e.g. cells of a grid, sub-national boundaries). Then for each spatial unit, a set of numerical features is derived from the trajectories (Figure 1a). Examples of such features are the average speed of the trucks, the proportion of time the trucks are stuck in congestion, the fuel consumption, etc.

Subsequently, we use a novel approach to visualize these features for the BSUs. The key idea is to create a visualization that consists of two components that are linked to each other: (1) a heatmap of the numerical features, in which each BSU corresponds to a row and every row has an identifying color, and (2) a map in which each BSU is identified using the color chosen in the heatmap. By showing these two components side by side, it is in principle possible to identify the features for each BSU shown on the map.

The challenges in this approach are (1) how to order the rows of the heatmaps in an insightful manner and (2) how to pick the colors used to identify the BSUs. Here, we propose to use the results of the literature on *seriation*. Seriation is an area that studies how to order the rows (or columns) of a data matrix in a manner that provides insight in the data. In our context, we will use this process to order the rows of the heatmap as well as to choose the colors of the BSUs used on the map.

[1] are with the Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM) at Université catholique de Louvain.

[2] is with the Louvain Institute of Data Analysis and Modeling in economics and statistics (LIDAM).

[1,2] E-mails: {first.second}@uclouvain.be

The advantages of this approach include that (1) the local variations between the BSUs are kept, while still also overall groups in the data are shown, and (2) it is still possible to understand how colors translate into features.

We use our approach to analyze the Belgian territory using a real-world data set of heavy goods vehicles' trajectories. The extracted features related to the proportion of time they pass in different states (e.g. driving, congestion) are used with the aforementioned method to explore the spatial structure of the country. We compare a number of different seriation methods and clustering methods for assigning colors to BSUs [7, 8, 9]. We show that seriation allows to unveil more of the spatial structure of the country.

Although in this work, we focus our analysis on the case of how trucks spend their time on the Belgian territory, the method is generic enough to be applied to other data sets and other forms of analysis. The implementation of our method is available under an open-source license and can be easily extended with new features or integrated in existing visualization tools.

The rest of this paper is organized as follows. Section II gives a short overview of existing analyses of truck trajectories. In Section III the workflow of the framework is explained in more details and the application to a real data set is reported in Section IV. In Section V seriation and clustering methods are compared. Section VI concludes and gives future work directions.

II. Related work

In [6], useful indicators on a studied region (e.g. number of trucks, entry/exit time, distance driven) are derived from a large data set of GPS trajectories. The purpose of the indicators is to provide insight about the use of the road network and its environment to city authorities, municipalities, etc. Each indicator is visualized individually and, when appropriate, on a map. In [5] the authors proposed to use community detection algorithms to find the interdependencies between spatial units. They divide the studied area using a grid of rectangular cells and then link them using an origin-destination matrix, derived from a large database of GPS sequences, before applying a community detection algorithm. In [2], the authors analyze the efficiency of a road network using truck trajectories. Starting from a set of GPS trajectories, they assign a color to each segment of the road network and then analyze how the mean speed evolves with the structure of the network. Finally, in [3], the authors analyzed international trips of freight vehicles between Canada and the United States. They show on a map the intensity of international trips starting in subnational entities and analyze the resulting patterns.

III. Methodology

The data-flow pipeline of the visualization framework starting from a set of raw trajectories is illustrated in Figure 1. First for each BSU, a set of numerical features is derived. The tool comes with some predefined features, but can be extended with features for specific applications. An example of how to characterize BSUs starting from GPS trajectories, in terms of numerical features, is given in Section III-A. Finally, the algorithmic details on how the color of each BSU is chosen and the heatmap visualization are given in Section III-B.

A. Characterizing the Basic Spatial Units

The first step of the methodology is to characterize each BSU with a fixed number of numerical features computed from the trajectories. The computed features depend on the data set as well as the goal of the analysis that needs to be performed. A feature can be as straightforward as the average speed of the trajectories crossing a BSU, or more complex, such as the estimated time spent in congestion.

This paper aims to be as general as possible; therefore we only use features computed from the raw GPS trajectories; more advanced features could be used starting from annotated trajectories. A raw trajectory is defined as

$$T = \langle T_1, \ldots, T_n \rangle = \langle (lat_1, lon_1, t_1), \ldots (lat_n, lon_n, t_n) \rangle$$

where lat_i, lon_i and t_i are respectively the ith latitude, longitude and timestamp.

In order to analyze how trucks spend their time, we introduce a non-exhaustive list of four simple states derived for each T_i ($1 \leq i \leq n$): *Driving*, *Congestion*, *Rest* and *Work related*. These estimate the actions[1] performed by a truck driver at the corresponding time.

a) Driving: This category is used when the truck is driving at normal speed between two GPS points T_i and T_{i+1}. All other points will be considered as *Stop points*. In this paper we decided to use a velocity threshold: if the average velocity between T_i and T_{i+1} is more than 15 km/h, then T_i is considered as a driving point. Every other point is considered as a Stop point and falls into one of the remaining categories.

b) Congestion: Following the analysis performed in [5] (which uses the same data set), every stop point of less than 10 minutes will be considered as congestion. This rule does not prevent to wrongly consider a short stop as a congestion stop, but it is accurate in practice despite its simplicity. Moreover, the aggregation made for each BSU, explained later, is not impacted significantly by a small number of errors.

c) Rest: This category represents the stops in a rest area located alongside the main highways. These can be identified by clustering [10] or can be provided in an additional data set.

d) Work related: Finally, this category represents all the other stops. These are stops of at least 10 minutes that are not located in rest areas and thus are assumed to be work-related, such as loading or unloading a truck.

Then, each trajectory can be rewritten as follows;

$$T = \langle (T_1, L_1), \ldots, (T_n, L_n) \rangle,$$

where T_i is defined as before and L_i is the category of T_i.

[1]Notice that more advanced techniques such as [4] exist to assign semantic labels to GPS points that could be reused in the proposed framework.

A sub-trajectory of T, T', is defined as a subsequence, not necessarily contiguous, of T. For such a sub-trajectory, the proportion of time in a category L can be computed as follows:

$$prop(T', L) = \frac{\sum_{(T_i, L_i) \in T' | L_i = L} t_{i+1} - t_i}{\sum_{(T_i, L_i) \in T'} t_{i+1} - t_i} \quad (1)$$

Note that Equation (1) can be used with any categorical feature. For numerical features, such as the average speed, the mean over the sub-trajectory can be taken instead.

Each BSU has one numerical value per category L defined above. The general idea is to compute these values based on the sub-trajectories that pass in the neighborhood of the BSUs. For instance, in the case of the congestion type of points, that value represents the proportion of time the trucks are in congestion in the area of the BSUs. More formally, given a set of BSUs $\mathcal{B}$, the neighborhood of $B \in \mathcal{B}$ is given by

$$N(B, \delta) = \{B' \in \mathcal{B} \mid d(B, B') \leq \delta\}$$

where δ is a user-defined radius and d is the distance between two BSUs. For a trajectory T, its sub-trajectory in the neighborhood of B is given by

$$T_{|_B} = \{(T_i, L_i) \in T \mid \exists B' \in N(B, \delta) \ : T_i \in B'\}$$

Finally, the numerical values for each category of GPS points L can be computed by taking the average of the values from Equation 1. That is, the average of the $prop(T_{|_B}, L)$ for every trajectory T such that $T_{|_B}$ is not empty.

B. Visualizing the features

The second step of the methodology is to apply a seriation algorithm that finds a linear ordering of the BSUs. In our framework, we use two of the most popular methods to perform seriation: one based on the Traveling Salesman Problem (TSP) [11] (solved using the Concorde solver [12]) and the Optimal Leaf Ordering method [13] (proposed in the scipy Python library [14])[2]. Intuitively, seriation is the problem of finding a linear arrangement of a given set of objects such that consecutive objects are similar. Thus, a good ordering is such that consecutive BSUs have similar features according to a similarity measure such as the Euclidean distance. If the features are not by nature comparable (e.g. distance driven with average speed), then a normalization ensures an equal weight for each feature.

Figure 2 shows an example with two features. On the left, the BSUs are plotted in the feature space and their order, found by a seriation algorithm, is shown. Let us denote p_i the ith BSU in the ordering. A color is assigned to each BSU depending on its order, using a rainbow color scale. We chose to use a rainbow color scale instead of a uniform one, as in the output of the seriation, there is no information on the similarity between BSUs far away. As an example, let us consider p_0 and p_{11} which are the furthest away in the ordering. It can be seen that p_0 is as much similar to p_{11} than p_5, even though p_5 is closer in the ordering.

[2]Those are standard approaches but other seriation methods have been proposed.

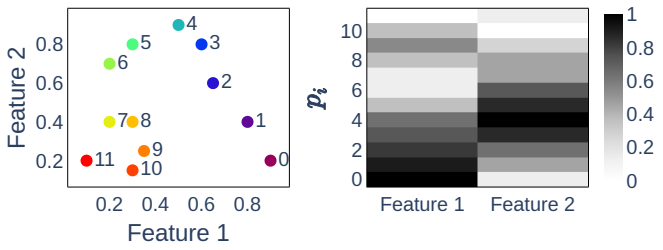


Fig. 2: Example of heatmap visualization with two features. On the left, the initial data points are shown as well as the order found by the seriation algorithm. The ith BSU in the ordering is denoted p_i. The points are colored using a rainbow color scale, following the order found by the seriation. On the right, the heatmap of features reordered.

In our approach we choose to make the similarity in color between adjacent BSUs dependent on the similarity of the rows in the heatmap. Hence, if in the order of the heatmap we have a number of rows that are very similar, we believe it helps for the interpretability of the results if their colors are also more similar. As an example, in Figure 2, p_7 and p_8 are more similar than p_0 and p_1; thus we wish to give these points more similar colors.

We achieve this by first recognizing that a rainbow color scale can be seen a smooth transition between 5 anchor colors, in order: pure purple, blue, green, yellow and red, as can be seen in Figure 2. The other colors (e.g. orange) will be interpolated between these anchor colors based on the total similarity of an order, defined by

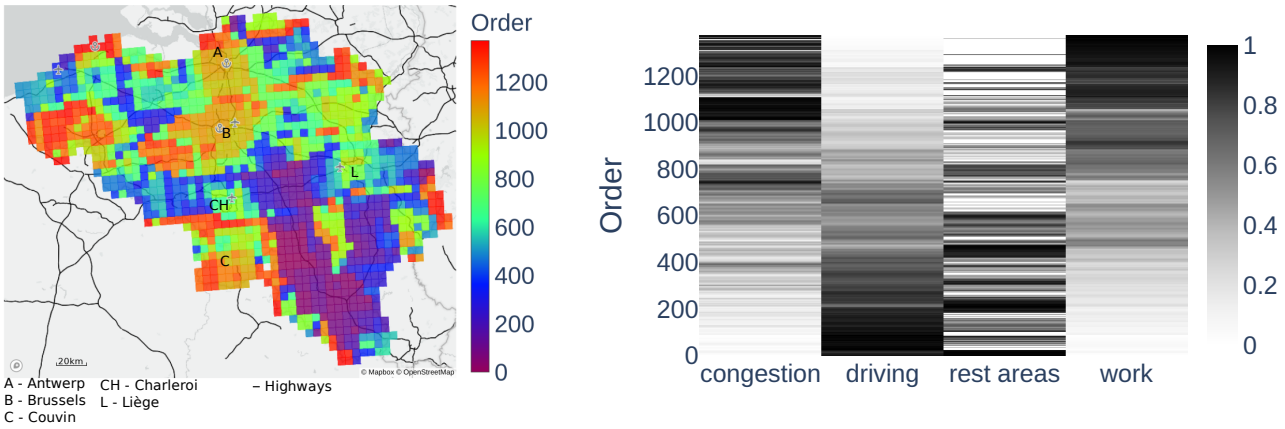
$$D = \sum_{i=0}^{|\mathcal{B}|-1} s(p_i, p_{i+1}),$$

with s a similarity metric. Based on this, we can calculate a position $0 \leq D_{p_i} \leq 1$ for every BSU:

$$D_{p_i} = \frac{\sum_{j=0}^{i-1} s(p_j, p_{j+1})}{D}$$

The interpolation is based on the D_{p_i} values such that there is a direct correspondence between a D_{p_i} value and a color. The anchor colors are evenly spaced between 0 and 1. In Figure 2, the purple is located at 0, the blue at 0.25, the green at 0.5, etc. The color of a BSU p_i is determined by its D_{p_i}. For example, if $D_{p_i} = 0.125$, then the color assigned to the BSU is halfway between purple and blue. As a result, some ranges of colors are more present than others in the final visualization. Indeed, if successive BSU of the ordering are very similar, the increase in the D_{p_i} values will be small, and the resulting colors will be more similar.

In order to interpret the colors in terms of features, a heatmap of the feature space, reusing the order of the



(a) (b)

Fig. 3: Visualization of how the trucks spend their time on the Belgian territory where the BSU are cells of 5 kilometers side. The order of the cells is found using the Optimal Leaf Ordering method.

seriation algorithm, is shown (right of Figure 2), revealing patterns in terms of features. In such visualization, there is one row for each BSU, one column per feature and the cells represent the values of the feature for the BSUs (darker cells represent higher values). In most cases, a scaling of the features such that they follow the same distribution is needed to better visualize the patterns. In the framework, the feature values shown in the heatmap are scaled to a uniform distribution. Thus 1 represents the highest value of a feature, 0.75 the third quartile, 0.5 the median, etc.

IV. Case Study

In this section, the framework is used to analyze how and where trucks in Belgium spend their time during their trip. First, the data set used is described followed by the analysis of the figures produced by the framework[3].

A. The data

Since the 1st of April 2016, all trucks that use the Belgian road network must be equipped with an On-Board Unit (OBU) that records their timestamped latitude-longitude point in order to pay a per-kilometer toll. For more details on the data set and its characteristics, see [5]. Our data set is composed of the recordings of these pings for all trucks that used the Belgian road network on the 15th of November 2016. The trucks are identified by a unique identifier that changes every night between 1:50a.m. and 2a.m. To avoid spurious changes in the trajectories, we decided to filter out the ones that started before 2a.m. Moreover, as done in [5], trajectories of less than 10 points are dropped from the data set. Overall the data set contains almost $90,000$ trucks that generated roughly 30 million GPS points. In this section as well as for Section V, the BSUs are 5×5 km grid cells provided by Eurostat, the European office for statistics [4]. The neighborhood radius varies depending on the BSUs used and we set it to 10 kilometers in this paper. This value was chosen as we found it is the smallest value that visually reveal the structure of the country. Using larger values will smooth the values in the BSU and variations in smaller spatial areas would be less visible.

B. How do trucks spend their time?

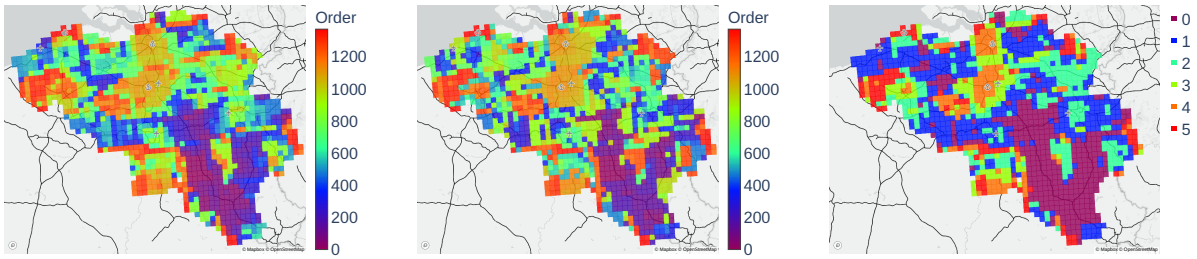
Figure 3 shows the result of our methodology. In order to better understand the results, let us first remind the reader of some general aspets of the geography of Belgium. The two major cities are Brussels and Antwerp, in which most of the economic activities are concentrated. The Northern part of the country is more densely populated and the road and city networks are tight. The South of the country (below the Charleroi-Liège axis) is less densely populated; the city network is looser and hence the intercity distances much larger and the economic activities more spread out. Still, there are two cities with higher economic and transport activities in Wallonia: Liège and Charleroi. We refer to Adam et al. for further analysis on traffic or stop intensities [5].

The ordering and analysis of the cells reveal the following. The first cells in the ordering, in purple and blue, are characterized by the highest values for driving while congestion and work activities are the lowest. They are mainly located on the axis going from the South of the Brussels agglomeration to Luxembourg, which, as explained, is an area where trucks drive more. The same is observed along the highway going from Liège to Luxembourg.

Next in the ordering, there are cells in light blue and green with median values for all features. These cells are spread all over the country, and cover a large part of Flandres as well

[3]Our source code is available at `https://github.com/AlexandreDubray/seriation-map`

[4]`https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/grids`



(a) Seriation (Optimal Leaf Ordering) (b) Seriation (TSP) (c) Self-organizing maps

Fig. 4: Comparison of multiple methods

as the areas of Liège and Charleroi. In these areas where the economical and commercial activities are more present, a larger proportion of the trucks do work-related actions, increasing the work and congestion features. However the driving feature is higher than in the densest part of the country due to the highways (e.g. in Liège or Charleroi) or to more rural areas where speeds are higher.

Finally, the last group of cells is colored in orange-red and corresponds to the cities with the densest economic activities such as Brussels and Antwerp. Since most of the economical and industrial activities are located in these areas [5], the proportion of time doing work-related actions is the highest. The case of Couvin is an exception that shows how temporary events can influence the results. Indeed, from 2011 to the end of 2017, the portion of the E420 highway between Couvin and the French border was under construction. The gigantic road construction increased temporarily the proportion of work time in these cells. Without such construction works, the level of activity in that area would have been much lower.

Overall it can be seen that the framework allows unveiling the urban and economic structure of Belgium. The North/South division of the country as well as the large imprint of Brussels and Antwerp are clearly visible. The exception of Couvin is worth mentioning as it corresponds to a huge construction site, making the area resemble major cities of the country. Let us remind here that we do not consider trucks of less than 3.5 tons (not tolled). Also note that there is a border effect present on the map due to the simple state assignment process defined in Section III. Indeed if a truck leaves the country and comes back later, the point is wrongly considered as a work-related point because the OBU only emits GPS points on the Belgian territory.

V. Comparisons

Let us now compare the colored assigned by seriation and clustering methods. We chose to compare the seriation methods available in the framework with Self-Organizing Maps (SOM) [15]. They have been well studied to visualize georeferenced data [7, 8, 9] because they try to preserve the topology of the input space and thus are similar to seriation. We decided to use 1-dimensional SOM implementation provided by the Minisom Python library [16], because it is closer to what is done by Seriation. We set the number of neurons in the SOM to 20, but the size of the SOM did not impact significantly the results.

First let us compare seriation based methods (Figure 4b-4a) with SOM clustering (Figure 4c). In SOM, every neuron is linked to its neighbors and thus there is a natural notion of distance between the resulting clusters (i.e. the cluster 0 is closer to cluster 1 than 4). In Figure 4c, the structure explained in Section IV-B can also be seen, but with less granularity. For example, the Charleroi and Liège areas are not as visible as in Figure 4a. This is due to the fact that in SOM, every BSU is assigned to the closest neuron. Thus, BSUs that are midway between two neurons are assigned to one of them and this notion of in between is lost. This problem is not present with seriation because there are no groups and the continuous color scale allows to have intermediate colors, which allows a finer interpretation.

Finally, let us look at the output of the seriation done with the TSP, shown in Figure 4b. The structure explained in Section IV-B can be distinguished, but it is less structured. To further analyze this, Figure 5 shows the heatmap of the feature space ordered using the output of the TSP based method. In Figure 3b, there was a clear transition from high driving to low and the congestion and work features were going smoothly from low to high. However in Figure 4b the changes from low to high, in each feature are more present and the spatial structures are less visible. Thus in our particular use case, the output of the TSP based seriation is less easy to interpret.

Overall, seriation allows to better retrieve the spatial structure of the country. Because there is no strict assignation to a given cluster, the natural groups in the BSUs are visible as well as the in between ones. Moreover the interpretation of the clusters in terms of features is easily done using the heatmap, as shown in Figure 3b, which is lacking with the SOM method. The choice of the seriation algorithm also plays an important role as it produces different results, with some more easily interpreted than others.

VI. Conclusion

Analyzing GPS trajectories can give good insights in the spatial structure of a country and can help make logistical

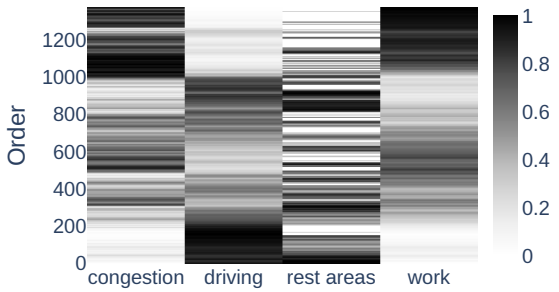


Fig. 5: Heatmap of the feature space when ordering using TSP bases seriation

decisions. In this work we introduce a framework that allows visualizing multiple aspects of road transport on a single map. To do so, the country is divided into small basic spatial units and characteristics are computed, based on the trajectories in their neighborhood. Then the BSUs are ordered using a seriation algorithm such that successive BSUs have similar features. A color is assigned to each unit from a rainbow color scale and visualized on a map alongside a heatmap of the feature space. This framework was used to analyze how trucks spend their time on the Belgian territory and it was able to unveil the spatial structure of the country. We compare seriation and clustering for the assignment of colors to the BSUs and showed that seriation methods allow a finer interpretation of the results. The proposed framework is very flexible and can integrate more advanced features and data sets. We think that it can be used to analyze a large variety of GPS trajectories, not only trucks, and thus the code is available under an open-source license and easily extensible.

As future work it would be interesting to see how the seriation process can be enhanced by using the geographical information (e.g. land use, industry localization). Moreover we did not visualize the impact of the time of the day in this analysis which impact the features we computed. For example, the congestion level is often maximized during peak hours and almost absent otherwise. A dynamic visualization which shows the evolution of the map according to the time of the day might unveil such time-dependent relationships.

References

[1] E. Commission, "Transport in the european union current trends and issues," 2019.

[2] M. Flaskou, M. A. Dulebenets, M. M. Golias, S. Mishra, and R. M. Rock, "Analysis of freight corridors using gps data on trucks," *Transportation Research Record*, 2015.

[3] K. Gingerich, H. Maoh, and W. Anderson, "Characterization of international origin–destination truck movements across two major us–canadian border crossings," *Transportation Research Record*, 2016.

[4] M. Taghavi, E. Irannezhad, and C. G. Prato, "Identifying truck stops from a large stream of gps data via a hidden markov chain model," in *ITSC 2019*, 2019.

[5] A. Adam, O. Finance, and I. Thomas, "Monitoring trucks to reveal belgian geographical structures and dynamics: From gps traces to spatial interactions," *Journal of Transport Geography*, 2021.

[6] S. Hadavi, S. Verlinde, W. Verbeke, C. Macharis, and T. Guns, "Monitoring urban freight transport based on gps traces of heavy-goods vehicles," *IEEE Transactions on Intelligent Transportation Systems*, 2018.

[7] J. Gorricha and V. Lobo, "Improvements on the visualization of clusters in geo-referenced data using self-organizing maps," *Computers & Geosciences*, 2012.

[8] A. C.-D. Lee and C. Rinner, "Visualizing urban social change with self-organizing maps: Toronto neighbourhoods, 1996–2006," *Habitat International*, 2015.

[9] V. Moosavi, "Contextual mapping: Visualization of high-dimensional spatial patterns in a single geo-map," *Computers, Environment and Urban Systems*, 2017.

[10] R. Aziz, M. Kedia, S. Dan, S. Basu, S. Sarkar, S. Mitra, and P. Mitra, "Identifying and characterizing truck stops from gps data," in *Industrial Conference on Data Mining*, 2016.

[11] G. Laporte, "The seriation problem and the travelling salesman problem," *Journal of Computational and Applied Mathematics*, 1978.

[12] D. Applegate, R. Bixby, and W. C. V. Chvátal. Concorde. [Online]. Available: https://www.math.uwaterloo.ca/tsp/concorde/index.html

[13] Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola, "Fast optimal leaf ordering for hierarchical clustering," *Bioinformatics*, 2001.

[14] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright *et al.*, "Scipy 1.0: fundamental algorithms for scientific computing in python," *Nature methods*, 2020.

[15] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological cybernetics*, 1982.

[16] G. Vettigli, "Minisom: minimalistic and numpy-based implementation of the self organizing map," 2018. [Online]. Available: https://github.com/JustGlowing/minisom/