# MAT-Builder: a System to Build Semantically Enriched Trajectories

Chiara Pugliese
*ISTI-CNR, Pisa, Italy*
*University of Pisa, Italy*
chiara.pugliese@isti.cnr.it

Francesco Lettich
*ISTI-CNR, Pisa, Italy*
francesco.lettich@isti.cnr.it

Chiara Renso
*ISTI-CNR, Pisa, Italy*
chiara.renso@isti.cnr.it

Fabio Pinelli
*IMT Lucca, Italy*
fabio.pinelli@imtlucca.it

2022 23rd IEEE International Conference on Mobile Data Management (MDM) | 978-1-6654-5176-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/MDM55031.2022.00058

***Abstract*—The notion of *multiple aspect trajectory* (MAT) has been recently introduced in the literature to represent movement data that is heavily semantically enriched with dimensions (*aspects*) representing various types of semantic information (e.g., stops, moves, weather, traffic, events, and points of interest). Aspects may be large in number, heterogeneous, or structurally complex. Although there is a growing volume of literature addressing the modelling and analysis of multiple aspect trajectories, the community suffers from a general lack of publicly available datasets. This is due to privacy concerns that make it difficult to publish such type of data, and to the lack of tools that are capable of linking raw spatio-temporal data to different types of semantic contextual data. In this work we aim to address this last issue by presenting MAT-BUILDER, a system that not only supports users during the whole semantic enrichment process, but also allows the use of a variety of external data sources. Furthermore, MAT-BUILDER has been designed with modularity and extensibility in mind, thus enabling practitioners to easily add new functionalities to the system and set up their own semantic enrichment process. The demonstration scenario, which will be showcased during the demo session, highlights how MAT-BUILDER's main features allow users to easily generate multiple aspect trajectories, hence benefiting the mobility data analysis community.**

***Index Terms*—trajectory enrichment, semantic enrichment, trajectory analysis, multiple aspect trajectory**

## I. INTRODUCTION AND MOTIVATIONS

The notion of *multiple aspect trajectory* (MAT) has been recently introduced in the literature [1] to represent movement data (i.e., a moving object trajectory) that is *heavily semantically enriched*. These trajectories can be seen as positioning data augmented with different dimensions, or *aspects*, representing various types of semantic information that are relevant or contextual to the data they are associated with. A few examples of aspects can be stops, moves, weather, POIs, transportation means, activities performed, social media posts, and so on. The aspects associated with a trajectory may be large in number, heterogeneous, and structurally complex.

While there are already contributions to the modelling and analysis of semantic trajectories and multiple aspects trajectories [2] – for instance, MATs are already used in tasks dealing with next point prediction, performance of recommendation systems, understanding human behavior, and traffic prediction –, the amount of datasets available in the community is still *scarce*. This is mainly caused by privacy concerns and regulations that make the publication of multiple aspect trajectory datasets difficult. Moreover, even when the data to be enriched is not strictly privacy-sensitive, we observe a general lack of tools that can support users during the complex process of creating these datasets. Generally speaking, this process requires to identify (1) the parts of the trajectory to be enriched, (2) the various types and sources of semantic data to be used for the enrichment, and (3) the most suitable approaches to properly associate spatio-temporal data with semantic information. For what concerns existing solutions, there are several libraries [3], [4] (e.g., Geopandas[1]), dashboards (e.g., [5], [6]), and ontology-based approaches (e.g., [7]) that are able to process and extract insights from trajectory data, while others propose trajectory enrichment with specific data sources (e.g., [8]). Unfortunately, these solutions either do not perform semantic enrichment or, when they do, they are limited to a fixed number of aspects, are not extensible, or they do not support the use of external data sources.

In this work we aim to address the above challenges and issues by introducing MAT-BUILDER, an interactive system that supports practitioners during the whole trajectory semantic enrichment process and that enables users to generate datasets of multiple aspects trajectories. The main contributions of this work are as follows: (1) we provide a system that supports the user in the complex process of building multiple aspect trajectories from heterogeneous data sources and with different semantic aspects, (2) this system is able to extract, combine, and build enriching information from a variety of external data sources (e.g., OpenStreetMap[2]), and (3) it is designed with modularity and extensibility in mind, thus allowing developers to add new aspects, external data sources, and functionalities. It is important to point out that we designed MAT-BUILDER to reuse functionalities provided by existing mobility data management libraries, including some of those that we already mentioned and referred above, i.e., GeoPandas, Scikit-mobility [3], and PTrail [4].

The rest of the paper is organized as follows. In Section II we present some fundamental notions which will be then used in the subsequent sections. Section III presents the modular architecture of MAT-BUILDER. Section IV presents a running example of the demonstration scenario we intend to showcase

[1]https://geopandas.org/
[2]https://www.openstreetmap.org/

2375-0324/22/$31.00 ©2022 IEEE
DOI 10.1109/MDM55031.2022.00058

during the demo session. Finally, Section V draws some final conclusions.

## II. Preliminaries

In this section we introduce some fundamental notions that will be used throughout the rest of the paper. We define a *raw trajectory* to be a sequence of time-stamped spatial coordinates representing the movement of an object. We define an *enriched trajectory* to be a raw trajectory enriched with semantic information. A *multiple aspect trajectory* is an enriched trajectory with multiple types, or *aspects*, of semantic information (e.g., weather, transportation means, POIs). From here on we will use the two terms above interchangeably to refer to multiple aspects trajectories. A *segmented trajectory* is a raw trajectory partitioned into sub-trajectories, or *segments*, according to some criteria. A common example of segmented trajectory is that obtained with the *stop* (i.e., the object is not moving) and *move* (the converse) segmentation [9]. The *segment enrichment* task associates to each segment one or more semantic aspects. For instance, a stop segment can be enriched with the information of a nearby geographical object (e.g., the point of interest aspect). Stops falling within the same area more than a given number of times $\tau$ are considered a *systematic stop*. Common examples are a person staying at home or at their workplace. *Occasional stops* can be then defined as stops that are not systematic.

## III. System Architecture

The MAT-Builder system[3] is written in Python and is made up of a *user interface* (UI) and, the core component of the system, the *modular backend*. The UI exposes the MAT-Builder's functionalities to the users, and translates the users' needs into queries that are then processed by the backend. We postpone the UI illustration to the demonstration scenario (Section IV). The backend is the core component of our system as it represents the MAT-Builder's query processing engine. Following the process described in [10], we designed the backend to include three main *modules*: trajectory pre-processing, trajectory segmentation, and segment enrichment. Each module provides a subset of the system functionalities. Figure 1 reports an overview of the modules (and the underlying information flow) that the system currently provides and that will be showcased during the demo session.

It is important to highlight that the backend is designed to be *extensible* with new modules, thus allowing developers to provide additional functionalities (e.g., additional aspects or management of further external sources) to the system. Indeed, every module in the system must extend a *common interface* that specifies the methods that every module should implement in order to be integrated and used within the system. More precisely, the interface requires every module to implement the following methods: (1) one that sets in the UI its *input parameters*, (2) one that *executes* the module operations according to the input parameters, and (3) one that specifies how the results should be provided via the UI. Note that any new module may also be built on top of preexisting modules via subclassing. Finally, our system provides *data workflow management capabilities* that let the user specify which modules shall be used among those available, and the order in which these shall be executed.

In the following we discuss the general goals of each backend module, and provide some details concerning their current implementation. Furthermore, we explain how the modules are connected to each other through intermediate saved results.

The **pre-processing** module (blue block in Figure 1) takes in input a set of raw trajectories and filters out noisy or unusable data to facilitate the activities of the other modules. One of the functions of the pre-processing module is to discard trajectories that have an insufficient sampling rate. The module also filters out the outliers by analysing their spatio-temporal characteristics. Another functionality is the trajectory compression adopted from the scikit-mobility library [3]. Note that trajectory compression can be critical, since it may drastically reduce the computation time of other modules by feeding trajectories with fewer samples.

The **segmentation** module (green block in Figure 1) takes in input a set of pre-processed trajectories and partitions every trajectory into sub-trajectories (or *segments*). A well known and widely used segmentation criterion is the *stop* and *move* [9]. Accordingly, this module outputs a set of segmented trajectories which can be then further processed by other modules. In the present version, the module makes use of the stop-move detection algorithm provided by the scikit-mobility library [3].

The **segment enrichment** module (yellow block in Figure 1) takes the output of the previous module and identifies the different segments to enrich, the aspects to consider, the datasets to be used to enrich the segments with different aspects, and the enrichment criteria. In MAT-Builder's current implementation this module has been divided into two sub-modules, one dealing with the enrichment of stops and the other with the enrichment of moves.

The **stop enrichment** sub-module enriches the stop segments with the *regularity* aspect, i.e., it distinguishes between systematic and occasional stops. The geographical area of interest is discretized into a spatial grid by means of a Geohash function[4], which is then used to assign each stop segment to a cell. Stops that fall within the same cell more than a given number of times $\tau$ are considered to belong to the same systematic stop. Conversely, stops that do not satisfy this criterion fall within the *occasional* category.

Systematic stops are then further enriched with the *activity* aspect by inferring the activity performed in it. The module is currently tailored on people's movements, therefore classifying systematic stops as *home*, *work*, and *others*. The *home* activity enriches a systematic stop that occurs in the night. *Work*, on

[3]The code is available at: https://github.com/chiarap2/MAT_Builder

[4]https://github.com/vinsci/geohash

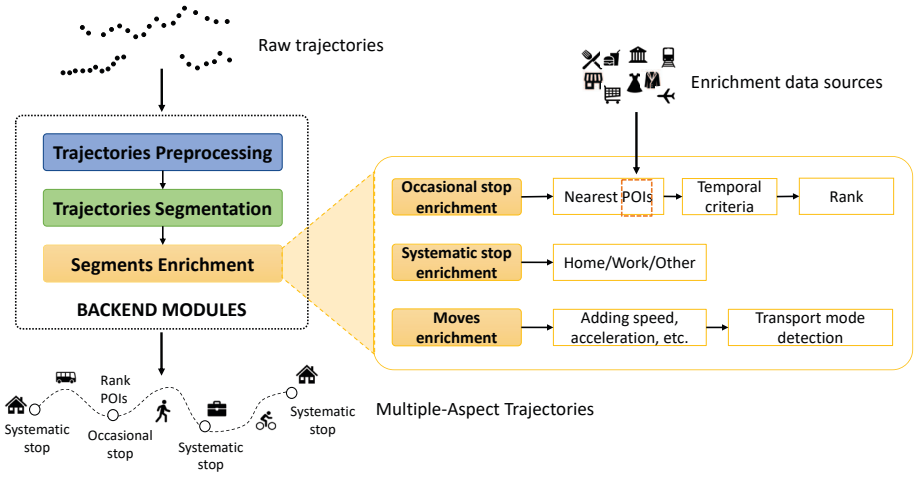


Fig. 1. MAT-BUILDER backend.

the other hand, is an activity that enriches a systematic stop that occurs on weekdays during working hours (i.e., a user defined time range). Systematic stops that do not satisfy any of the above criteria are enriched as *other*.

Occasional stops are harder to characterize than systematic stops, as they do not exhibit any substantial spatio-temporal regularity. The module presently enriches such stops with POIs that may be relevant to them. More precisely, for each stop the module first ranks the POIs according to some given order of preference, and then associates to the stop the top k ones. In the module current implementation, POIs are ranked according to two criteria, i.e., their distance from the stop and the overlap between their opening hours and the stop starting and ending times – the latter criterion allows us to filter out POIs that are closed when the stop occurs. We report that our system can associate POIs having any kind of geometric shape, e.g., points, lines, and polygons. Finally, note that the module can retrieve POIs either from Open Street Map or from other external data sources, assuming that minimum information is provided (e.g., identifier, latitude, and longitude) and that they comply with the required formatting criteria.

The **move enrichment** sub-module focuses on enriching the move segments. In the present version, the module enriches with two aspects: (1) quantitative numerical measures and (2) transportation mean. The first aspect associates to each move quantitative numerical information extracted from the underlying sub-trajectory, i.e., maximum and average speed, acceleration, bearing rate, and total length. These information are computed via the PTrail library [4]. The second aspect is enriched by inferring the transportation mean used during each move. To this end, the module leverages a random-forest classifier (using the scikit-learn library [11]) trained on the GeoLife dataset [12] to recognize the following transportation means: *walk*, *car*, *bike*, *bus*, *subway*, and *train*.

## IV. DEMONSTRATION SCENARIO

The MAT-BUILDER system targets users who want to semantically enrich movement data, possibly with many different aspects, and who would like to have control on how the various enrichment tasks are conducted. Indeed, thanks to MAT-BUILDER's highly flexible modular design, users can decide which and how many aspects they want to consider to enrich their trajectories. Furthermore, the operations conducted by the various MAT-BUILDER's modules are highly customizable, as a large set of modifiable parameters enable users to precisely control and fine-tune the whole enrichment process. To show our system capabilities we introduce a running example of the demonstration scenario that participants will experience during the demo session.

Demo attendees will be able to interact with the system through the MAT-BUILDER's UI at the different steps of the enrichment process. The UI presents three tabs corresponding to the three backend modules. In the *pre-processing tab* (Figure 2) the attendee will be able to select and input the raw trajectory dataset they intend to enrich. Here the pre-processing tab lets the user customize some of the pre-


Fig. 2. MAT-BUILDER UI: *pre-processing* step

processing operations, i.e., they will be able to indicate the *minimum number of points* a trajectory should have and a *km/h threshold* between two consecutive points the module uses to filter out outliers. Once the raw trajectories have been pre-processed, the UI will show the results of this step.

The attendee will then proceed to the *trajectory segmentation tab*. Here the interface lets the user specify the *minimum duration* and the *spatial radius* the system will use to identify the stop segments (and, indirectly, the move ones). Once the trajectories have been segmented, the UI will again show a few statistics about this step.

The user will finally proceed to the *segment enrichment tab* (Figure 3), where they will be able to enrich the stop and

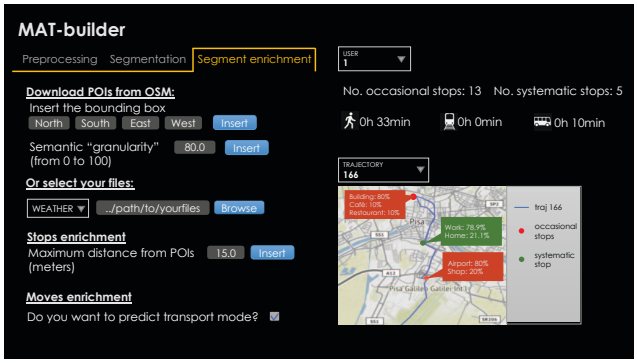


Fig. 3. MAT-BUILDER UI: *segment enrichment* step

move segments with different aspects. The first aspect added to the stop segments concerns their regularity, i.e., whether they belong to some type of systematic stop or they are an occasional one. Next, the attendee will be able to select the data sources to be used to enrich the occasional stops. Here, the user can provide a file for each semantic aspect they want to add – the system currently supports enrichment with POIs, weather, and social media posts. Moreover, MAT-BUILDER is able to gather POIs from OpenStreetMap in the eventuality a POI file cannot be provided: in this case, the user must first specify a bounding box of geographical coordinates from which the POIs should be retrieved. POIs might have an extremely large number of attributes, and many of these may have lots of missing values. To deal with this issue the UI lets the user specify a value to discard attributes with too many missing values. With the POIs in place, the module decides which POIs should be used to enrich the occasional stops by ranking them according to the criteria described in Section III (i.e., distance and temporal overlap). Finally, the attendee will be able to choose whether to enrich the moves with the transportation mean. This aspect is estimated with a random-forest classifier as described in Section III.

During the demo session we will finally show how developers can add new modules to the system, or modify existing ones. We will also show how the MAT-BUILDER's data workflow management capabilities allow to select modules among those available and manage their execution.

## V. CONCLUSIONS AND FUTURE WORK

In this paper we proposed MAT-BUILDER, a new system that supports users in creating multiple aspect trajectory datasets starting from raw trajectories and external data sources. The semantic enrichment process offered by MAT-BUILDER includes trajectory pre-processing, trajectory segmentation, and segment enrichment. The backend, which is the core component of MAT-BUILDER, implements said process and it is currently instantiated with stop and move segmentation, enrichment with systematic and occasional stops, activity inference, and transportation mean estimation. We highlight that key characteristics of MAT-BUILDER are its modularity and extensibility which, in conjunction with MAT-BUILDER's data workflow management capabilities, enable users to easily set up their own semantic enrichment processes as well as add further functionalities to the system.

## ACKNOWLEDGEMENTS

This work has been supported by the EC H2020 projects MOBIDATALAB (GA 101006879) and MASTER (GA 777695).

## REFERENCES

[1] R. Mello, V. Bogorny, L. O. Alvares, L. H. Z. Santana, C. A. Ferrero, A. A. Frozza, G. A. Schreiner, and C. Renso. MASTER: A multiple aspect view on trajectories. *Trans. GIS*, 23(4):805–822, 2019.

[2] C. Renso, V. Bogorny, K. Tserpes, S. Matwin, and J. A. F. de Macêdo. Multiple-aspect analysis of semantic trajectories(master). *Int. J. Geogr. Inf. Sci.*, 35(4):763–766, 2021.

[3] L. Pappalardo, F. Simini, G. Barlacchi, and R. Pellungrini. scikit-mobility: A python library for the analysis, generation and risk assessment of mobility data. *arXiv preprint arXiv:1907.07062*, 2019.

[4] S. Haidri, Y. J. Haranwala, V. Bogorny, C. Renso, V. P. da Fonseca, and A. Soares. Ptrail – a python package for parallel trajectory data preprocessing. *CoRR*, 2021.

[5] M. Berlingerio, F. Calabrese, G. Di Lorenzo, R. Nair, F. Pinelli, and M. L. Sbodio. Allaboard: A system for exploring urban mobility and optimizing public transport using cellphone data. In *Machine Learning and Knowledge Discovery in Databases*, pages 663–666, 2013.

[6] F. Calabrese, E. Cobelli, V. Ferraiuolo, G. Misseri, F. Pinelli, and D. Rodriguez. Using vodafone mobile phone network data to provide insights into citizens mobility in italy during the coronavirus outbreak. *Data & Policy*, 3:e22, 2021.

[7] T. P. Nogueira, R. B. Braga, C. T. de Oliveira, and H. Martin. Framestep: A framework for annotating semantic trajectories based on episodes. *Expert Systems with Applications*, 92:533–545, 2018.

[8] Nikolaos Koutroumanis, Georgios M. Santipantakis, Apostolos Glenis, Christos Doulkeridis, and George A. Vouros. Scalable enrichment of mobility data with weather information. *GeoInformatica*, 25(2):291–309, 2021.

[9] S. Spaccapietra, C. Parent, M. L. Damiani, J. A. de Macedo, F. Porto, and C. Vangenot. A conceptual view on trajectories. *Data & knowledge engineering*, 65(1):126–146, 2008.

[10] A. Ibrahim, H. Zhang, S. Clinch, and S. Harper. From GPS to semantic data: how and why - a framework for enriching smartphone trajectories. *Computing*, 103(12):2763–2787, 2021.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[12] Y. Zheng, X. Xie, W. Ma, et al. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33(2):32–39, 2010.