T-MAPS: A spatiotemporal model to describe traffic conditions based on Twitter

Bruno P. Santos[†], Paulo H. L. Rettore[†], Heitor S. Ramos[‡], Luiz F. M. Vieira[†], and Antonio A. F. Loureiro[†]
[†]Computer Science Department – Universidade Federal de Minas Gerais (UFMG) – Brazil
[‡]Computer Institute – Universidade Federal de Alagoas (UFAL) – Brazil
Email: [†]{bruno.ps, rettore, lfvieira, loureiro}@dcc.ufmg.br, [‡]heitor@ic.ufal.br

***Abstract*—The understanding of urban mobility has gained more attention as the number of people living in urban areas is increasing and technology is becoming more sophisticated. Similarly, it is noticed the increase of accessibility and shared real-time content by Location Based Social Media (LBSM) platforms, which becomes a data source for recent studies of urban mobility, particularly on traffic subjects. Such studies can promote impacts on our daily activities by changing the way we move around. In this work, we argue that LBSM feeds may offer a new layer to increase our traffic and transit comprehension, even if it remains to overcome some LSBM data aspects such as data imprecision and users' bias. Initially, we show the significant positive correlation between Twitter's feed and the traditional traffic sensors. Then, we present the Twitter MAPS (T-MAPS) a low-cost spatiotemporal model to explain the traffic events through tweets. T-MAPS uses instantaneous, historical and sentimental data into its model. This strategy allows T-MAPS to enhance traditional traffic sensors by bringing the users' point of views into the urban mobility comprehension. In our evaluation, T-MAPS and Google Maps performed their route recommendation service, which presented an average similarity of 62% among the routes, and for one-fourth of the evaluated trajectories, we observed a high similarity varying from 75% to 100%. We also presented a service of route sentiment and description, which aims to increase the information about the route, performing the tweets' text analysis.**

I. Introduction

The transport infrastructure might be able to promote people's movement efficiently, but it also implies in the constant need for planning and management of the transportation system. In this sense, understanding urban mobility (traffic and transit) has been the focus of governments, researchers, and industries [1]. Usually, traffic and transit specialists use traditional raw data sources (e.g., data from inductive loops, traffic cameras, and origin-destination matrix) to perform their analyzes. Unfortunately, access to these data sources is, in general, limited to those who are connect to governmental entities or large corporations. This become a barrier to better understand urban mobility that asks for other solutions.

In that way, the Location Based Social Media (LBSM) (e.g., Twitter, Instagram, and Foursquare) becomes an alternative data source to study urban mobility. These platforms allow users to share their thoughts, viewpoints, and activities related to their feelings about almost everything, which include traffic conditions. There are different research issues which benefit from using LBSM as a low-cost data source [1], [2], [3], [4].

According to the *Twitter*, about 313 million users are active every month in their network (June 2016 data). This huge attendance may open up several opportunities. In this work, we investigate the traffic scenario, which affects our daily activities profoundly. However, data from LBSM also brings several issues that can lead to other challenges such as data imprecision, users' bias, and spatiotemporal assignment or inconsistency. Therefore, we need to overcome those data issues before making complete use of LBSM's data.

In this work, we performed a study to understand better the relationship between the real traffic scenario and the data provided by Twitter, a very well-known and largely used LBSMs platform. The first step focused on the characterization of the collected data and its properties. After that, we proposed Twitter MAPS (T-MAPS), a low-cost spatiotemporal model to explain traffic condition events based on tweets. T-MAPS intends to enrich the current navigation context by connecting LBSM's data in different ways, for example, by evaluating tweets frequency and LBSM's users perspective of the region of interest. The model is flexible enough to incorporate instantaneous, historical and even sentimental data. In this work, we present a case study where we collected a dataset from New York City (NYC for short) and demonstrated the T-MAPS significant strengths and shortcomings. Then, we evaluated T-MAPS route recommendation against Google Directions routes through similarity, and we also discussed T-MAPS route description service. We highlight two main contributions of in this work: 1) The characterization of LBSM data, especially from Twitter, as a data source to better understand and describe the traffic conditions. 2) The proposition of T-MAPS, a low-cost spatiotemporal model, to explain traffic condition events through tweets.

An important question emerges based on the inherent subjectivity of enriching the traffic description and suggesting routes, as we proposed. To the best of our knowledge, there is no ground truth for the best route, even when there is a complete description of the traffic. For that reason, there are many tools that aim to offer their traffic viewpoint like Google Maps, Here Wego, and TomTom maps. The main reason that motivated us to develop T-MAPS is to the desire to demonstrate the potential of using LBSM data, in particular traffic data. Also, we aim to encourage the design of applications, models and analysis of urban mobility using LBSM.

This work is organized as follows. Section II details the

TABLE I
EXAMPLE OF USERS

Account name	# *Tweets*
@511NYC	126925
@TotalTrafficNYC	20267
@WazeTrafficNYC	7850
@Traffic4NY	3789
...	...
@NYC_DOT	3680
Total:	≈ 354K

Fig. 1. Tweet example in the dataset.

process of data collection from Twitter and its correlation with data from traditional traffic sensors. Section III describes the issues related to the LBSM's data. Section IV shows the modeling of T-MAPS. We ran the T-MAPS and Google Maps route recommendation services, and then evaluated their similarities in Section V, followed by related works in Section VI. Finally, Section VII presents the final remarks and future directions.

II. DATASET

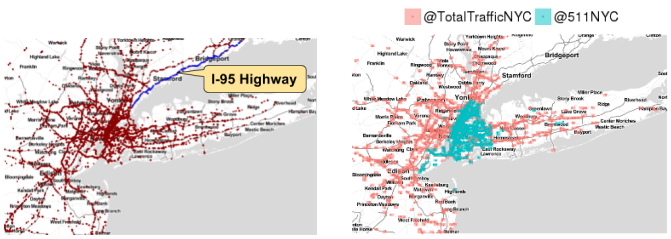
One of the most significant challenge to study urban mobility is the absence of open data in such context. Therefore, most of the work in this field lies on theoretical field or has a large private data provider (e.g., government agencies, Google, Tomtom among others.). Fortunately, the growing LBSMs adoption allows people share on online platforms their thoughts, viewpoints, or activities. These content are related to users feelings about the world, including the traffic and transit conditions. With the right tools and code, we were able to collect data from Twitter, where many users periodically share information about traffic and transit events. For this end, we use Twitter APIs respecting the restriction terms[1].

The dataset consists of 353.807 tweets from twenty-one manually selected users accounts. Those accounts are maintained by departments of transport, specialists on traffic and transit reports such as news channels or dedicated companies. Table I shows some accounts and tweeting frequency in the dataset. The number of tweets with geotagging is 307.020, most of them in NYC. We explored Manhattan where has 38.112 tweets. The dataset was collected on last three months of 2016.

Figure 1 displays an example of a tweet in the dataset. The tweet consists of a rich explanation of the traffic event (textual address, the cause of the event and even delays). Also, the tweet has metadatas with the hour of the post, a geotag signature, counters (e.g., retweets and likes) among others. Note, the dataset does not contain regular users due to high user bias in their tweets regarding traffic feelings, we return to this issue in Section III-B. Next, we analyze the dataset in spatiotemporal context, which gave for us initial insights.

Figure 2(a) shows tweets with geotagging in the dataset. The majority of tweets are over the road network, for instance, if we zoom in on the image, it is possible to see the I-95 highway with tweets over its extension. Also, the central region has higher tweets density than non-central ones, which can indicate a tendency of the user's preferred region to report information.

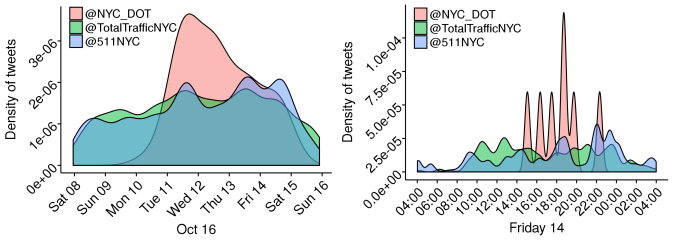
[1]https://developer.twitter.com/en/docs



(a) Tweets on New York City map (b) Spatial coverage of two users.

Fig. 2. Dataset coverage in New York City and neighborhood.

Figure 2(b) shows the filtered dataset which has only tweets from @TotalTrafficNYC and @511NYC. As expected, we note that different accounts have contrasting coverage. While @511NYC focused on reporting traffic information within NYC boundaries, @TotalTrafficNYC had broader coverage. Aware of this characteristic, one might use many spatial complementary accounts to cover a region.



(a) Weekly tweeting density (b) Hourly tweeting density

Fig. 3. Temporal coverage of three accounts.

Figure 3 shows a week of data from @NYC_DOT, @TotalTrafficNYC, and @511NYC. On the left, Figure 3(a) the density of tweets along the week. As expected, different accounts have disparate behavior in their posting rate. Although the department of transportation account (@NYC_DOT) posts mainly on working day, @TotalTrafficNYC and @511NYC do posts everyday. However, they still have different tweeting rate behavior. On the right, Figure 3(b) displays hourly tweeting density. The most of the @NYC_DOT posts occur during business hours, while @TotalTrafficNYC and @511NYC do posts all over the day. Note that some peaks of tweets appear during rush times, e.g., the @TotalTrafficNYC presents a high volume of posts from 7:00 to 10:00, this can suggest that traffic events occur while people are starting their daily activities. The peaks also appear in @511NYC's curve, one from 12:30 to 15:00, another from 17:30 to 20:00, and from 21:00 to 00:00. These peaks suggest a high posting rate at lunch time, when people are finish business day, and when people are starting their nightlife respectively. Aware of this, one might use complementary users data to increase the temporal coverage.

A. Twitter as a traffic sensor

Are tweets related to traditional traffic sensor data? Understand how well related they are is fundamental to reveal the potential of such content to enhance and complement

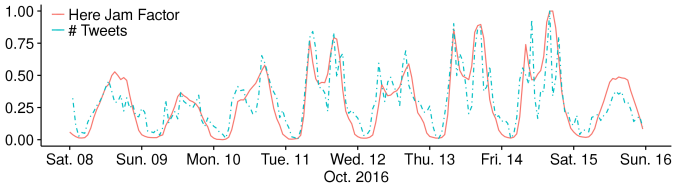


Fig. 4. Tweets frequency and Here Jam Factor time series.

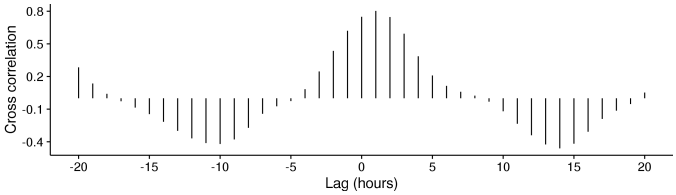


Fig. 5. Cross-correlation between Jam Factor and # tweets time series.

traditional ways to see traffic and transit. For example, if a traditional traffic sensor detects an anomalous event, can tweets explain such atypical event?. This section intends to present directions to answers these questions.

First, it is required to get access to classic traffic measurement data, such as inductive loop detector count, traffic cameras, vehicle GPS traces on road network, or origin-destination matrices, among others. With these data, traffic specialists are able to study demand and supply aspects. Demand can be seen as vehicles and pedestrians while supply is related to streets, highways, sensors and control devices [1]. Thus, it is possible to study the interactions between demand and supply, and eventually develop efficient transportation systems, which optimizes urban mobility and decrease transit congestion.

Unfortunately, access raw traditional sensor data is a challenge for the regular community. Actually, raw traffic data are kept by government entities or large companies, and it is not freely accessible. Usually, the traditional sensors sense three variables of interest: velocity, density, and flow. These quantities relate to each other allowing traffic behavior analyses and visualizations [1], [5]. On the other hand, LBSMs data is more accessible, which it allows urban mobility studies [1], [2]. Also, it is common that users share their thoughts, viewpoints, and activities on LBSM platforms. It expands the sensing capacity by capturing the users perspective about the situation.

Naturally, raw data holders perform some data fusion process and present the result in their services or statistics. For example, the Google gather heterogeneous data like GPS traces, cameras, and inductive loops, then it makes a data fusion and presents the result in colors over the map. In that way, companies like Google, Here, and TomTom allow access to the resulting data fusion process. In this article, we use the Jam Factor (JF) from HERE API[2] as aggregated traditional traffic sensor data. According to Here documentation, the JF is a fused representation from traditional heterogeneous data. JF ranges from 0 to 10, where 0 means the way is free and 0 congested. We choose Here JF due to no other company provides such kind of data.

Figure 4 shows the correlation between Here Jam Factor and tweets in the dataset along a week in October 2016. The time series in blue is the aggregated Here Jam Factor, and the orange time series corresponds to the number of tweets. Due to differences in time series scales, we put both on a scale between 0 and 1, and we aggregated each series hourly. Then, we observe that the curves are similar one each other. We compute the Spearman's rank (ρ) a nonparametric correlation coefficient to assay relationship between two variables. The ρ has a value between -1 and +1, where -1 mean the observations are fully dissimilar and +1 meaning a similar or identical. We apply Spearman's rank in the time series resulting in $\rho = +0.81$. When ρ is positive and near to +1, one can interpret that the # tweets tend to increase when Jam Factor increases.

One can ask about the delay between those time series. By applying the cross-correlation technique, it is possible to figure out where time series match [6]. Figure 5 shows on the y-axis the cross-correlation between JF and # tweets, and on the x-axis the lag between the time series, we use JF as the test waveform. The highest correlation (0.8) appear when the lag is +1 meaning that # tweets curve is 1 hour ahead from JF. One can interpret that tweets appear on the platform before JF increases, but note that the time series were hourly aggregated. Therefore this is not necessarily a fact, but a indication.

III. Data Aspects

Often data from Twitter has aspects that lead to issues in its use on traffic context. In this section, we classify the data aspects into four classes *Data imprecision, User bias, Spatiotemporal assignment, and Inconsistencies*. Our taxonomy is in agreement, in parts, with the literature [7], [8]. Then discussed the potential proposed solutions for each one aspect.

A. Data Imprecision

LBSMs' data, in general, comes with a certain degree of imprecision. Frequently, the data imprecision present at least one of the characteristics: *incomplete data, vagueness, granularity effects*. The inherent heterogeneity of the data sources, and "freedom" of data input on LBSM's platforms may promote the imprecision. We used the following tweet to illustrate those data characteristics:

"Now 8:00AM an accident at 100 W 33rd St #NYC #BadTraffic #creepedOut".

It is possible to obtain rich information about the event, in the tweet above, like the user sentiment, traffic condition, and hour. However, the *Incomplete* aspect appears when the tweet lacks some information that might be essential such as geotagging or the event severity. Some researchers have proposed techniques to mitigate data incompleteness, e.g., in [9] proposed a record linkage approach to enrich incomplete data. In [10], [11], the authors used possibility theory and the probability of fuzzy events to handle imperfect data.

[2]https://developer.here.com/rest-apis/documentation/traffic/topics/quick-start.html

The *Vagueness* corresponds to the no clear description or context of the data. The tweet sample shows vagueness because we are not able to precisely define the extension, position, its cause or even those involved of the accident. Usually, a way to deal with vagueness is matching and fusing additional data from different sources to understand the situation. There are research works handling data vagueness by using Fuzzy logic and probabilistic approaches [8].

The *Granularity* ranges from fine-grained and coarse-grained. When fine-grained, the data contains enough information to accurately describe the following items: event location, direction, the severity of accidents, and other information. Otherwise, coarse-grained consider a macro view of events with a broad description.

B. User Bias

On traffic and transit context, the LBSM users can interpret the traffic congestion in different ways, and use their freedom to post any information. For instance, suppose a scenario where a person, from a small city, is in the traffic of a metropolis, and then he/she can interpret the situation as a mess and then doing posts on the LBSM platforms, while a metropolis resident may understand as a typical situation. Consequently, the user's perception may lead to bias introduction on data traffic and transit collected from LBSMs.

The dedicated users (accounts which professionally reports traffic scenario) on reporting traffic and transit information also can introduce bias. Such users can, for instance, feed information for a specific audience or even they may have greater or lesser intention to publish data to a specific place over another one. In this work, we choose manually dedicated users accounts to overcome regular users bias, but despite the diverse nature of users in the dataset (department of transport, news specialists, dedicated companies, so on), the data may follow inherent bias of users interests and intentions.

C. Spatiotemporal Assignment

The spatiotemporal assignment is one of the most critical aspects of data collected from LBSMs regarding traffic and transit context. The geolocation and temporal tagging allow traffic specialists to study and characterize a region at an instant or interval of time. Following, we discuss the general issues to extract the LBSM's data spatiotemporal information.

- **Spatial:** It is fundamental to assigning a location to the data, aiming to understand the context surrounding the information. However, deriving this information, even when it is present it is not always a trivial task. Suppose a *tweet* containing the spatial location in written form instead geotag, therefore it requires a way to textual address location extraction. Although such techniques already exist, the inherent unstructured form and freedom of writing (e.g., abbreviations, only 140 characters, so on) on LBSMs make challenging the task of spatial textual extraction. In the case of tweets, its particularities often result in information subjectivity or misinterpretation. There are research efforts to overcome these issues, for example, in [12], [13], the authors use Natural Language Processing (NLP) techniques to obtain parts of speech and entity extraction to label sequences of words that are the name of things. In [14], Li optimizes NLP techniques to tweets text.
  The availability of information is another issue that affects the spatial data assignment. It is expected that some regions will have more significant spatial coverage than others due to several factors. For example, large cities tend to have higher spatial coverage than smaller towns, the cause of this may simply be due to the more substantial number of users, companies and the traffic of information, or a complex social matter.
- **Temporal:** associating a temporal tag (timestamping) to the shared data has a key role to understand the past, present, and possibly the future scenario of the transport networks. LBSM platforms usually assign a timestamp tag when users input data the system. However, this markup may not represent the same moment as when the event occurred. Some open questions arise about temporal assignment aspect such as *What is the validity of data published by a user of LBSM?*, *How does characterize the delay between the event and the data input on LBSM platforms?*.

D. Inconsistencies

In this work, we discussed only inconsistencies like conflicts and out of order data, others ones aspects are covered in the literature [8], [7].

- **Conflict:** The conflicting data from LBSMs appear when two or more data sources diverge about a specific event. For instance, suppose that two users Alice and Bob shared their feelings about the same traffic event incident. Alice reports that nothing serious happened and the traffic flow is good, while Bob reports that a severe accident happened, which promote a substantial negative impact on the traffic. In this situation, based only on these two points of view is difficult to determine what happened about the event. In the literature, works on Dempster-Shafer evidence theory have gained notoriety to reducing data source divergences [15], [16]. Also, it is possible to give a reputation weighting to users accounts, e.g., a dedicated account may have high reliability than a regular user when they post traffic information, then rules can be applied to decide the right information.
- **Out of sequence:** The freedom offered by LBSMs platforms allow users to enter traffic and transit information out of sequence into the system. These data appear as inconsistent to the systems which will use it. In traffic and transit context, out of sequence data is often related to the temporal dimension of the data, for example, one user may share information about a past traffic event. One may think how to use such data properly. Usually, the trivial solution is to discard the out of sequence data. However, if the data was identified correctly and then

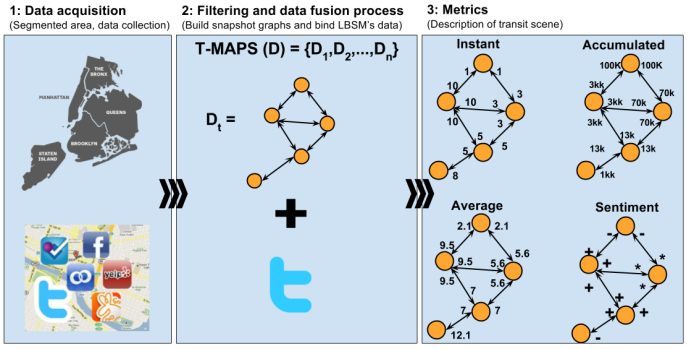


Fig. 6. T-MAPS's modeling process.

sorted, it may be used as a feedback data at the cost of more processing and storage resources.

IV. Modeling Process of Twitter Maps (T-MAPS)

We presented a low-cost spatiotemporal model to explain traffic events through tweets. The model allows the representation of the traffic scenario in different ways. The T-MAPS uses instantaneous, historical, or even users sentiment into its model. Following, we presented the steps of the T-MAPS's modeling process.

1) ***Data acquisition:*** This step consists of segmenting the area of interest and retrieving data from the LBSMs platforms. It is possible to segment the region of interest in several resolutions, ranging from micro (at the level of roadways and streets segments) to macro resolutions (those at the level of entire roads, boroughs, city). The segmentation resolution may be adjusted to fit the spatial coverage of the data.
2) ***Filtering and data fusion process:*** This step aims to filter and bind LBSM's data to the segmented region. We proposed the use of a weighted time-varying digraph as a model to map theses areas and data. The time-varying digraph is represented as a series of static networks, one for each time step. Formally, let R be the set of segments of the region, then a snapshot digraph is defined as $D_t = (V, E, m)$, where $V = \{r | r \in R\}$ denoting the segmented region, and $E = \{(u, v) \in V | u$ is adjacent to v in R segmentation$\}$ denoting directed edges between physically connected regions. The T-MAPS's time-varying digraph is a sequence of snapshot digraphs, thus T-MAPS$(D) = \{D_{t=t_{min}}, D_{t+\Delta}, ..., D_{t_{max}}\}$, where t_{min} and t_{max} are the start and end time of the available dataset, and Δ can be adjusted conveniently.
3) ***Metrics:*** it consists in assigning cost weights to the directed edges. Formally $m(u, w) : E \rightarrow value$, where $m(u, w)$ is a function mapping the directed edges to a metric value cost. The metrics function represents the analyzed traffic and transit scenario using the LBSM data.

Figure 6 illustrates a toy example of the T-MAPS modeling process. First, we segment the NYC map into five regions of interest, then we collect LBSM available data. Next, we obtain the digraph $G = (V, E, m)$, where V are the regions, and E are directed edges between adjacent regions. Then, we bind Twitter's traffic data to the resulting regions graph. Finally, the weights are assigned to the edges using different metrics functions. The resulting time-varying digraph allows us to make analysis of the traffic scenario condition and description. We present some metrics functions below:

- **Instant:** This metric function considers all tweets in a given t time on a day by fusing and filtering them properly. This strategy corresponds to a snapshot view of the traffic at that moment. The smallest t must agree with the configured Δ of T-MAPS model.
- **Accumulated:** This approach considers all previous available data given a time. It requires two parameters, t_{start} and $t_{reference}$, where $t_{start} < t_{reference}$ and it must respect the temporal dataset availability. Then, it accumulates all data between t_{start} and $t_{reference}$. One can interpret this metric as a historical metric looking to the past until the reference time point. In our experiments $t_{start} = t_{min}$.
- **Average:** It uses the same approach of *Accumulated*. However, the value assigned to the edges are the average of tweets occurrences over a time, day, week, year so on. It must respect the data availability. This information must be passed as a parameter to the metric function. One can interpret this metric as a typical traffic condition metric, that put into the account historical information.
- **Sentiment:** This is a description metric where each tweet passes through a text analyzer, and then it is associated with a sentiment score. Next, for each region, the metric function ranks the sentiment scores and assign the most common sentiment to the region. One can interpret this metric as a Twitter's users feelings about the traffic condition for each region.

On top of T-MAPS model a variety of services related to urban mobility can be designed such as analysis of the traffic and transit scenario, route recommendation service or a route events description. We discussed and evaluated these services in the next section.

V. Evaluation

The evaluation of the T-MAPS services was conducted by comparing its recommended routes against Google Directions[3] routes. Afterwards, we discussed the route description service, demonstrating an opportunity to enriches the description of the traffic scenario. The results correspond to the Manhattan region segmentation in NYC and the dataset presented in Section II. The Manhattan region was segmented into 29 official neighborhoods[4], consequently a T-MAPS digraph snapshot contains 29 vertices. We removed all tweets outside the Manhattan region. Besides, the minimum time interval between two consecutive T-MAPS graphs corresponds to a $\Delta = 1$ hour.

[3]We consider Google Directions as the most accurate representation of the traffic scenario, although it is not necessarily a fact.

[4]www1.nyc.gov/site/planning/data-maps/open-data/districts-download-metadata.page

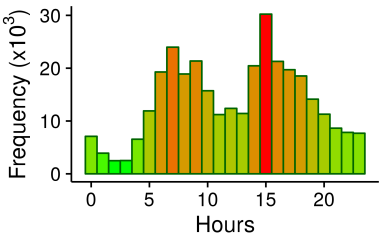


Fig. 7. The frequency of tweets per hours in the dataset.

A. Route Services

We evaluated the routing service by comparing the similarity between T-MAPS and Google Directions recommendations. In this work, the T-MAPS route suggestion service considers a macro resolution of the regions on the map, but our model is flexible enough to encompass fine-grained resolution as well (see Section IV). On macro resolution, T-MAPS aims to tell the users that the recommended regions have the best conditions regarding the metrics applied.

We query the T-MAPS and Google Direction recommenders for 812 routes in Manhattan neighborhoods. The routes were derived from the combination $2 \times C_k^n$, where $n = 29$ (Manhattan neighborhoods) and $k = 2$ (origins and destinations). Note that we are considering routes $A \rightarrow B$ and $B \rightarrow A$. The routes start and end at the center of the region. Also, we rule out routes that start and end in the same region. We query the routes in three different moments (7:00, 15:00, and 19:00) of the day along of the week. Those moments were chosen purposely; first, they represent rush hours, second, they have a higher volume of tweets in the dataset, third, the Here Jam Factor increases on that moments. Figure 7 shows the frequency of tweets per hour in the dataset.

The similarity technique measures the matched areas where the recommended routes by T-MAPS [5] and Google Directions passed through. Figure 8 displays the similarity between routes along of eight days in the dataset considering three metric functions. The box-plots summarize 58.464 routes analyzed. T-MAPS with *Instant* metric shows a high variation of similarity rate, its median range from 50% up to 66.7%, while *Accumulated* metric shows 60% to 70% and *Average* metric 60% to 66.7%. It means that half of the evaluated routes present at least 50% with Google Directions. We expected that *Instant* metric poses the lower similarity due to its intrinsic disparity with others metrics and it does not take into the account historical data. As a global evaluation, we averaged all route similarities resulting that, in average, T-MAPS has 62% routes similarity with Google Direction for half of the routes evaluated. The upper quartile (1/4 of the routes) until the maximum value presented a similarity between 75% and 100%.

Discussion: we argue that T-MAPS showed up a high degree of route similarity with Google Directions. T-MAPS exclusively employed Twitter's dedicated users feeds, while Google Directions uses several data sources (not fully known). Again, there is not a ground truth for what is the best route and we claim that LBSM data can be used as data source for traffic scenario description.

[5]We use Dijkstra's algorithm on T-MAPS model to recommend routes.

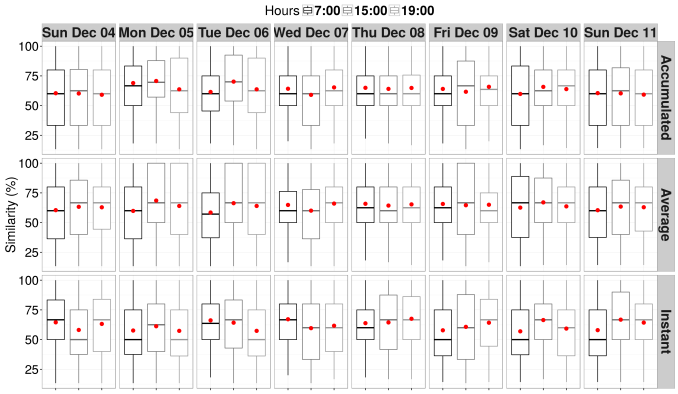
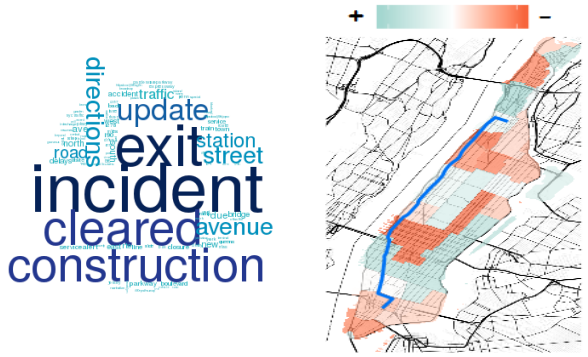


Fig. 8. Route recommendation similarity between T-MAPS and Google Directions. Dots represent the mean.

B. Route description

We conducted a pre-processing step on the tweet's texts, aiming to provide a route description service on top of the T-MAPS by sensing users feelings through their tweets. Initially, we performed the cleaning phase which converts the text to lowercase, removes accents, tokens extracting, filter stops words, links, special characters, and punctuation. Then, we performed the *stemming* processing on the text, which removes suffixes from words such as jam**ming**, jam**med** resulting in the word *jam*. As a result, we generated a word cloud of the route, Figure 9(a) where the word size indicates high-frequency of use on the tweets over the route.



(a) Word cloud of the route. The size indicates word frequency.

(b) Route suggestion in Manhattan on Oct. 14, 2016.

Fig. 9. Route sentiment based on the tweets text analysis.

we used the open source *syuzhet* package available for R tool to extract the sentiment of the tweets text. Syuzhet contains algorithms to extract sentiment of texts by using a dictionary of words and its associated emotions/feelings. The sentiment of the text depends on the number of occurrences of

these words/feelings in the text, and then a score is associated with the positive or negative sentiment.

We put all together on top of T-MAPS and provide the route sentiment service. Figure 9(b) shows the resulting of sentiment analysis and a route recommended by T-MAPS and Google Directions with high similarity. With this information (path and sentiment), the T-MAPS can enrich the recommended route, indicating to the users the sentiment of the way or even providing routing based on these feelings.

VI. Related Works

Many studies exist in the literature on event detection and diagnostics by using LBSMs [17], [18], [19]. Much of them are focused on detect general events and language patterns recognition to understand events. In [3], the authors propose a technique to detect traffic events and display them in near real-time on the web. Our study also focuses on traffic and transit events, besides that, we propose a model that handles LBSM's data to provide a traffic scenario conditions description.

In [20], [21], the authors study sentiment analysis by using data from LBSMs. In [20], Bertrand studies how to measure high-resolution sentiment in NYC from spatiotemporal perspective. Our work uses similar sentiment techniques, for example, we are interested in understanding the sentiment of a route. Although we use sentiment analysis techniques, our work propose a model to traffic conditions description. Also, T-MAPS can recommend routes.

Kim proposes SocRoutes a safe route recommendator based on Twitter data [4]. Our work goes beyond by providing a model to clarify the traffic condition, and it is flexible enough to take into account instantaneous, historical, sentiment data. Giridhar et al., in [22], focuses on explain unusual traffic events using social media feeds, but their work does not provide ways to recommend routes. Finally, we highlighted data aspects in previous work [7] where a taxonomy was proposed, and discussed the state of the art solutions.

VII. Conclusions

In this work, we presented a characterization and relationship between collected data from Twitter (a LBSM platform) and real traffic and transit data. Based on that, we developed a time-varying digraph-based model, T-MAPS. With T-MAPS, we can bind LBSM data and map aiming to enrich the navigation context into the city. We evaluated two T-MAPS' services: a route recommendation and route description. The results showed that the T-MAPS suggested routes have, in average, 62% of similarity to those suggested by *Google Directions*, also 25% of the evaluated routes presented similarity degree between 75% and 100%. This indicates that the T-MAPS provides a meaningful correspondence with *Google Directions* routes, even using tweets exclusively. In the route description service, we discuss that a route can be enriched with LBSM data by putting the LBSM's users viewpoint and feelings into the account.

Finally, we intended as future work to extend the T-MAPS for regions with larger dimensions, besides to explore the tweets texts to extract semantic information about the events reported. Also, we aim to use regular, and dedicated user accounts by plugging a reputation model on T-MAPS to handle conflicting information.

References

[1] A. L. Bazzan and F. Klügl, "Introduction to intelligent systems in traffic and transportation," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 7, no. 3, pp. 1–137, 2013.

[2] J. Yin and Z. Du, "Exploring multi-scale spatiotemporal twitter user mobility patterns with a visual-analytics approach," *ISPRS International Journal of Geo-Information*, vol. 5, no. 10, p. 187, 2016.

[3] S. S. Ribeiro Jr, C. A. Davis Jr, D. R. R. Oliveira, and W. Meira Jr, "Traffic observatory: a system to detect and locate traffic events and conditions using twitter," in *5th ACM SIGSPATIAL*, 2012.

[4] J. Kim, M. Cha, and T. Sandholm, "Socroutes: safe routes based on tweet sentiments," in *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014, pp. 179–182.

[5] M. Keyvan-Ekbatani, A. Kouvelas, I. Papamichail, and M. Papageorgiou, "Exploiting the fundamental diagram of urban networks for feedback-based gating," *Transportation Research Part B: Methodological*, vol. 46, no. 10, pp. 1393–1403, 2012.

[6] R. N. Bracewell and R. N. Bracewell, *The Fourier transform and its applications*. McGraw-Hill New York, 1986, vol. 31999.

[7] P. H. Rettore, B. P. Santos, A. B. Campolina, L. A. Villas, and A. A. Loureiro, "Towards intra-vehicular sensor data fusion," *19th International Conference on ITS*, 2016.

[8] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information Fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[9] C. Pinto, R. Pita, G. Barbosa, J. Bertoldo, S. Sena, S. Reis, R. Fiaccone, L. Amorim, M. Y. Ichihara, M. Barreto, M. Barreto, and S. Denaxas, "Probabilistic integration of large brazilian socioeconomic and clinical databases," *30th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2017)*, vol. 1, no. 1, 2017.

[10] D. Dubois and H. Prade, "Possibility theory and data fusion in poorly informed environments," *Control Engineering Practice*, vol. 2, no. 5, pp. 811–823, 1994.

[11] R. R. Yager, "Generalized probabilities of fuzzy events from fuzzy belief structures," *Information sciences*, vol. 28, no. 1, pp. 45–62, 1982.

[12] X. Liu, S. Zhang, F. Wei, and M. Zhou, "Recognizing named entities in tweets," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011.

[13] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2005.

[14] C. Li and A. Sun, "Fine-grained location extraction from tweets with temporal awareness," in *37th ACM SIGIR*. ACM, 2014, pp. 43–52.

[15] L. A. Zadeh, "Review of a mathematical theory of evidence," *AI Magazine*, 1984.

[16] M. C. Florea, A.-L. Jousselme, É. Bossé, and D. Grenier, "Robust combination rules for evidence theory," *Information Fusion*, vol. 10, no. 2, pp. 183–197, 2009.

[17] Crooks, Andrew and Croitoru, Arie and Stefanidis, Anthony and Radzikowski, Jacek, "# Earthquake: Twitter as a distributed sensor system," *Transactions in GIS*, vol. 17, no. 1, pp. 124–147, 2013.

[18] Weng, Jianshu and Lee, Bu-Sung, "Event detection in twitter." *ICWSM*, vol. 11, pp. 401–408, 2011.

[19] Hasan, Mahmud and Orgun, Mehmet A and Schwitter, Rolf, "A survey on real-time event detection from the Twitter data stream," *Journal of Information Science*, p. 0165551517698564, 2017.

[20] K. Z. Bertrand, M. Bialik, K. Virdee, A. Gros, and Y. Bar-Yam, "Sentiment in new york city: A high resolution spatial and temporal view," *arXiv preprint arXiv:1308.5010*, 2013.

[21] A. Giachanou and F. Crestani, "Like it or not: A survey of twitter sentiment analysis methods," *ACM Computing Surveys (CSUR)*, vol. 49, no. 2, p. 28, 2016.

[22] Giridhar, Prasanna and Amin, Md Tanvir and Abdelzaher, Tarek and Wang, Dong and Kaplan, Lance and George, Jemin and Ganti, Raghu, "ClariSense+: An enhanced traffic anomaly explanation service using social network feeds," *Pervasive and Mobile Computing*, 2017.