Knowledge Gap Identification - Initial Report

Big Data and Data Analytics Committee

1. Introduction

The Big data and data analytics committee is a horizontal committee within the framework of the nine professional committees of the Israeli Smart Transportation Research Center (Figure***1***). Being a horizontal committee, it aims to identify knowledge gaps that address all areas of smart transport as well as the various methods and techniques included in big data analytics and machine learning (ML). This broad view poses challenges in implementing the traditional manner for identifying knowledge gaps, i.e., reviewing individual articles as a basis for drawing the necessary conclusions. Moreover, the rapid growth in the availability of types and quantities of relevant data as well as in the state-of-the-art machine-learning methodologies require constant updating of knowledge gaps.

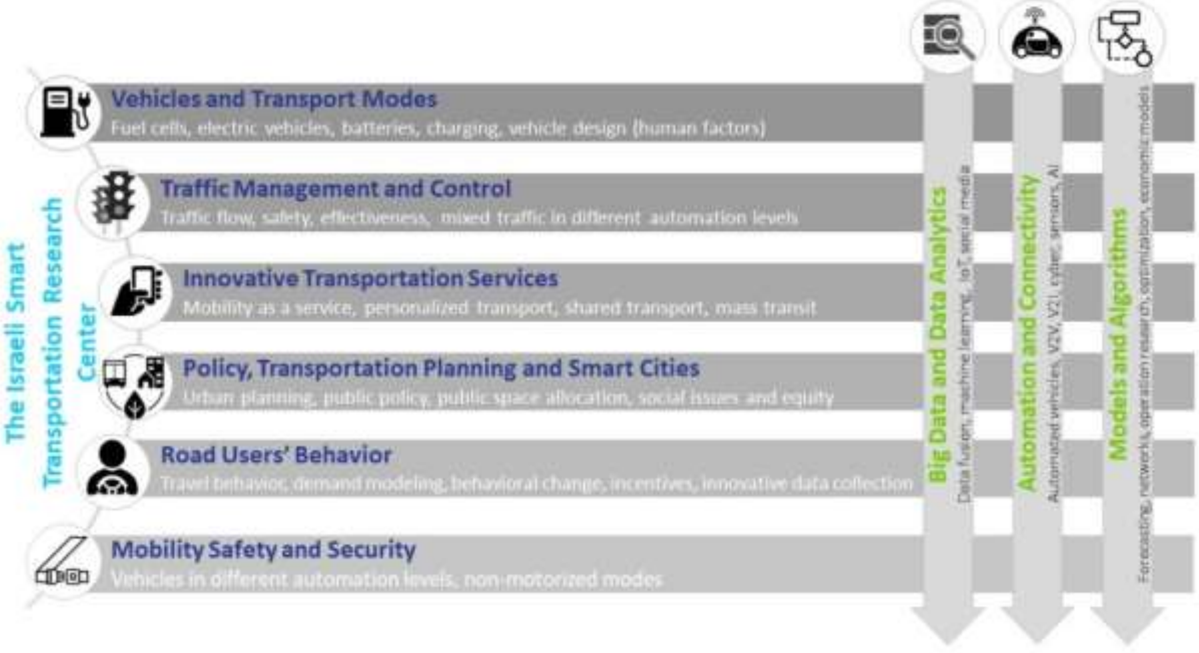

The Israeli Smart Transportation Research Center
Vehicles and Transport Modes
Fuel cells, electric vehicles, batteries, charging, vehicle design (human factors)
Traffic Management and Control
Traffic flow, safety, effectiveness, mixed traffic in different automation levels
Innovative Transportation Services
Mobility as a service, personalized transport, shared transport, mass transit
Policy, Transportation Planning and Smart Cities
Urban planning, public policy, public space allocation, social issues and equity
Road Users' Behavior
Travel behavior, demand modeling, behavioral change, incentives, innovative data collection
Mobility Safety and Security
Vehicles in different automation levels, non-motorized modes
Big Data and Data Analytics
Data fusion, machine learning, IoT, social media
Automation and Connectivity
Automated vehicles, V2V, V2I, cyber, sensors, AI
Models and Algorithms
Forecasting, networks, operation research, optimization, economic models

Figure1 : Professional committees of the Israeli Smart Transport Center

It was therefore decided to pursue two approaches for knowledge gap identification. The first one, described in section 2 of this document, is based on surveying recent review papers that focus on the use of big data and ML in the transport field in general and in some specific transport-related domains in particular. This process revealed some interesting knowledge gaps, however it also revealed that some of the transport-related domains are not covered by recent review papers. This finding doesn't lead to the conclusion that there is no literature addressing the application of ML techniques to these domains, however it clarified that identifying knowledge gaps using merely review papers is not adequate.

It was therefore decided to use an additional approach. This approach aims to develop a tool for quantitative assessment of knowledge gaps, i.e., the extent to which each ML technique was applied to tackle problems in each transport domain. The assessment will be based on the number of publications in each category of domain-technique combination, obtained by automatic classification existing literature. Visual representation of the results, including drill-down abilities into sub-categories, will provide the basis for knowledge-gap identification.

2. Knowledge gaps and recommendations for further research based on recent review papers

Main knowledge gaps and recommendation for future research are hereby presented based on eight recent review papers focusing on big data and ML in the transport field. One paper was published in 2018, three in 2019 and four in 2020.

Two of the review papers have not provided definitive conclusions regarding knowledge gaps or specific recommendations for future research (Amin et. al, 2019; Mounica & Lavanya, 2020), however others highlighted interesting insights and ideas for research themes that are still to be explored. Twelve recommended research topics have emerged from synthesizing the insights presented in the review papers.

Welch & Widita (2019) reviewed studies addressing various public transport-related problems while using big data sources and techniques. The authors recommended three research directions in this specific area:

1) Studies focusing on multimodality and the first/last mile, and specifically the relationship between public transport and other transport modes (both motorized and non-motorized).
2) Studies on resilience and health/safety in the public transit realm as well as studies addressing equity and the impact of transit on public health.
3) Using social media data for enhancing the understanding of travelers' perceptions of public transport.

Koushik et al. (2020) reviewed ML techniques in activity-travel behavior and propose the following:

4) Address the spatiotemporal transferability of ML-based models by focusing more on interpretability rather than increasing accuracy in order to aligned with the main aim of activity-travel behaviour models.

Several research gaps associated with international freight transportation management were identified by Barua et al. 2020:

5) Exploiting the advantages of the sequential nature of ensemble more for problems in the field of international freight transportation management.
6) Investigation across different ML methods in order to improve the understanding of the applicability of different ML methods to diverse international freight transportation management problems, and the sensitivity of these methods to data types, problem sizes, and problem types.

General research topics, that are relevant for all aspects of the transport systems, have been recommended by various authors:

7) Research exploiting the potential of hybrid methodologies, combining operation research and ML, for solving problems in which actions should be taken (Barua et. Al, 2020).
8) Enhancing the use of heterogeneous data sources relevant for the transport domain (Neilson et al., 2019; Welch & Widita, 2019).
9) Studies proposing ways to access new data sources and methods to easily process data from multiple sources and provide user-friendly output (Welch & Widita, 2019, Zhu at al., 2018).
10) Managing veracity inherent in transportation data (Neilson et al., 2019, Zhu at al., 2018).
11) Exploring ways for ensuring that privacy and data security are managed properly across different networks (Neilson et al., 2019, Zhu at al., 2018).
12) Research addressing transport modes beyond road transport (Kaffash at al., 2020).

2.1. Further insights identified by the committee’s members

Surprisingly, no thorough updated review paper focusing on the use of ML and data mining for traffic management and control was found in the literature search. However, many papers describing various implementations of these techniques to problems in the field of traffic management and control have been published in recent years. Some of them address the prediction of specific traffic parameters (volume, speed etc.) and others deal with classification of traffic states in various types of networks. Identifying specific research gaps in this field will hopefully be available through the automatic classification of articles that will be developed by the committee next year.

Yet, looking at this type of papers in a non-exhaustive manner raises a general impression that the potential of both data fusion and model fusion methodologies has not been realized in this area.

On top of the specific topics numbered 1-6 in the previous section, all the general research topic identified in the literature (numbered 7-12) are relevant to all sub-areas in transportation research. Moreover, personalization is an intriguing direction that big data can set the basis for. Personalization is relevant for many elements of the transportation system, such as planning transport services (especially multi-modal journeys), information related to transport services, logistic-related services, driving behavior (especially towards the era of autonomous and semi-autonomous vehicles) etc.

3. Quantitative assessment of knowledge gaps

A tool for quantitative assessment of knowledge gaps is based on the concept that the number of publications in professional journals that address a specific transport-related domain by specific means of ML provides a good measure that can serve as a starting point for researchers to identify areas/techniques that have not yet been intensively explored.

In order to achieve these goals, the following steps should be taken, some of which have already begun and others that will be carried out in the next months:

Step 1 – Defining the categories of transport domains and ML techniques into which the articles will be classified

Step 2 – Defining the features of the knowledge-gap identification tool in terms of content

Step 3 – Defining the features of the knowledge-gap identification tool in terms of user interface

Step 4 – Developing a classification model

Step 5 – Developing a visualization tool

The four first steps are briefly described in the following sections. The fifth one is more technical and is therefore out of the scope of this document.

3.1. Defining the categories into which the articles will be classified (step 1)

The initial categories of the transport domains are based on seven of the nine committees included in Figure1 :

1. Vehicles and Transportation modes
2. Traffic Management and Control
3. Emerges transportation services
4. Transportation management policy and Smart Cities
5. Travel Behaviour
6. Safe and secure transport
7. Automation and connectivity

Given the classification to these domains is successful, each one will be furthered divided into su-domains such as the type of transport mode addressed in a safety-related article or a distinction between urban and interurban traffic management. The ML-related categories were defined based on the know-how of the Big data committee members with the relevant expertise:

1. Supervised machine learning for structured data
2. Unsupervised machine learning for structured data
3. Supervised machine learning for unstructured data
4. Unsupervised machine learning for unstructured data
5. Deep Learning
6. Image Processing
7. Speech Processing
8. Text/Natural Language Processing
9. Audio / Signal Processing
10. Reinforcement Learning
11. Classic AI - Agents, Planning, Search
12. Data Visualization
13. Anomaly Detection
14. GIS and Spatial Data Mining
15. Ethics, Privacy and other Legal issues of AI

An additional hierarchy of the ML categories is considered, which will refer to the data source, i.e., end user-based data, service operators-based data or authorities-based data. Additionally, the top layer categories might be further refined during the classifier-building process based on the insights gained.

3.2. Defining the features of the knowledge-gap identification tool in terms of content (step 2).

Each publication will be double-classified, i.e., into a transport domain (a horizontal category) and into an ML technique (a vertical category). Although when defining the categories, it was clear that clear-cut distinction among the horizontal categories as well as among the vertical categories is not always possible, preliminary human annotation of articles into transport-domain categories provided additional evidence to this phenomenon.

It was decided to limit the classification into no more than two horizontal and two vertical categories, i.e., the two that are most representative. The overlaps among categories will be taken into account when defining target values for the performance indicators of the classifier to be developed. The performance indicators will address both False positive and False negative aspects, and the target values will be finalized based on the results of a cross-annotation process carried out by transport experts. As an initial value, an accuracy of about 80% is aimed for.

3.3. Defining the features of the knowledge-gap identification tool in terms of UI (step 3)

Given the visualization and drill-down capabilities of business intelligence (BI) tools, one of the commercially available software services, such as QlikSense, Tableau, will be used. The two figures below demonstrate a mock-up of the interactive user interface of the tool to be developed. Figure 2 demonstrates a visualization of the highest level of the classifier's results.

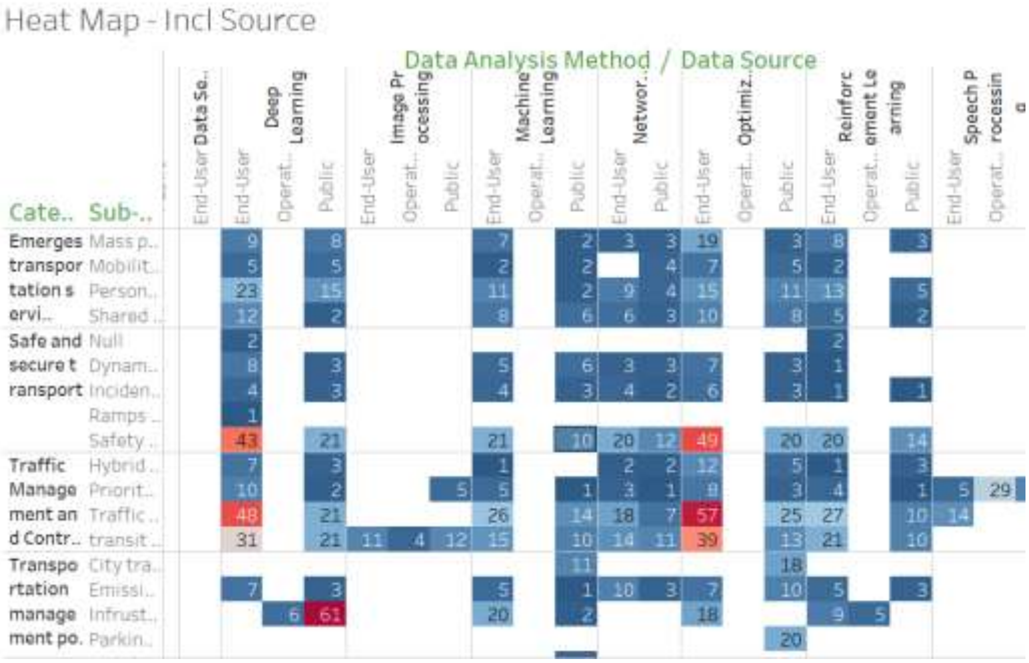
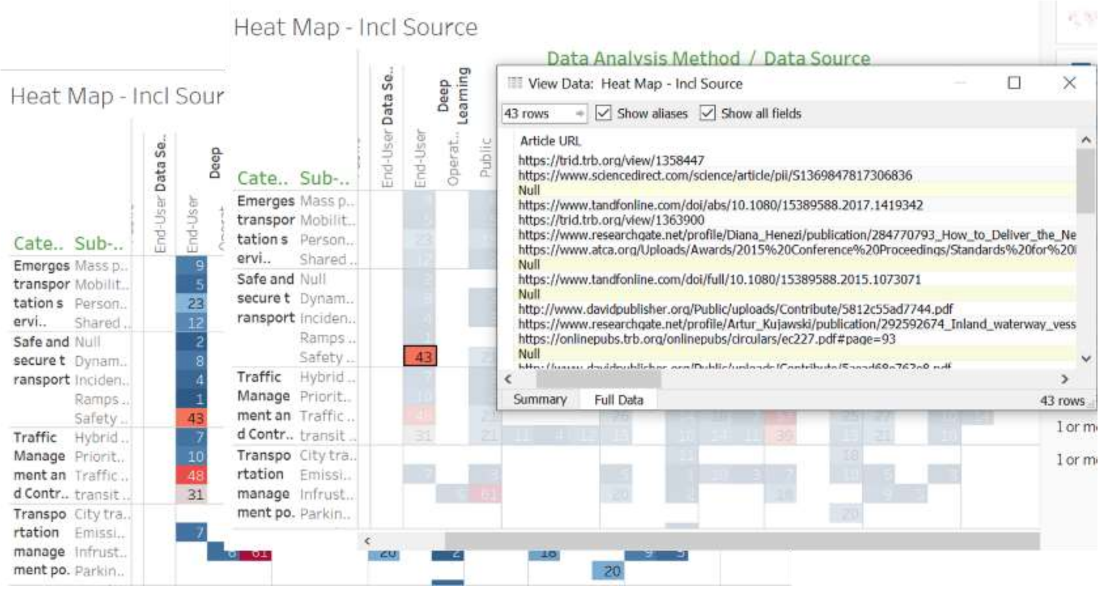


Figure 2 : Visual View using BI tool of the highest level of classifier's results

Figure 3 presents the ability to browse from the BI tool into the area of interest, including a link to the relevant papers.



View Data: Heat Map - Incl Source

43 rows ☑ Show aliases ☑ Show all fields

```
Article URL
https://trid.trb.org/view/1358447
https://www.sciencedirect.com/science/article/pii/S1369847817306836
Null
https://www.tandfonline.com/doi/abs/10.1080/15389588.2017.1419342
https://trid.trb.org/view/1363900
https://www.researchgate.net/profile/Diana_Henezi/publication/284770793_How_to_Deliver_the_Ne
https://www.atca.org/Uploads/Awards/2015%20Conference%20Proceedings/Standards%20for%20
Null
https://www.tandfonline.com/doi/full/10.1080/15389588.2015.1073071
Null
http://www.davidpublisher.org/Public/uploads/Contribute/5812c55ad7744.pdf
https://www.researchgate.net/profile/Artur_Kujawski/publication/292592674_Inland_waterway_vess
https://onlinepubs.trb.org/onlinepubs/circulars/ec227.pdf#page=93
Null
```

43 rows

Figure 3 : Ability to zoom into areas of interest and link to relevant papers

3.4. Developing a classification model (Step 4)

Data source

The four main features characterizing professional publications that were considered as an input for the to the classifier to be developed were: Title, Keywords, Abstract, Full paper. Human annotation of several dozens of articles led to the conclusion that the abstract is essential, but the full paper might more often decrease the quality of classification rather than improve it.

Several free-of-charge search engines exist for retrieving publications, three of which were considered as the source for obtaining the articles' required features: google scholar, Scopus and web of science. Scopus was selected as it enables efficient definition of search conditions and provides a convenient interface for exporting user-defined in an easy-to-handle CSV format.

Classification technique

Classifying articles into categories, based on content analysis, can be achieved by implementing various ML techniques.

The two prominent approaches of ML are unsupervised and supervised learning. Clustering, that relies on unsupervised ML techniques, can be used for grouping articles, aiming to obtain groups that reflect similar topics and/or ML techniques within each group and distinct attributes between groups. There are numerous clustering algorithms, such as K- and hierarchical clustering, and the selection of most appropriate one depends on the characteristics of the problem at hand. Regardless of the clustering algorithm used, there is a need to identify the meaningful commonalities within each cluster, a task that is sometimes challenging and the explainability of results may be difficult. The supervised learning approach, realized through classification algorithms, uses labeled items as input, i.e., each paper is labeled by the transportation domain/s it addresses and the ML technique/s implemented. The algorithm uses the labeled examples to learn rules that will be used to label a new publication. ML classification models, such as decision trees, are much more intuitive and results are easier to interpret. However, the labeling itself often reflects preconceptions of the researchers and might overlook meaningful similarities and differences among articles.

Additional possible direction to be implemented is creating a classification rule base by implementing Phrase-Based Classification (PBC) (Bekkerman and Gavish, 2011). PBC is very natural for multilabel text classification, which is the case at hand, given a single publication might refer to more than one transportation domain and ML technique.

It was decided to explore both the unsupervised and supervised approaches, and optionally also the rule-base methodology.

Human annotation of articles

As supervised and optionally also rule-based approaches will be implemented, human annotation of a sufficient amount of papers is required. As an initial step, approximately 900 papers were extracted from Scopus using the following search rules:

a. Data related articles, the search included the following words:

- **learning** - for capturing "machine learning ", "deep learning" or "deep reinforcement learning"
- **mining – for data mining**
- **processing – for "image processing", "signal processing" and similar**
- **big data**

b. Articles, were limited to the following types: article (ar); Review (re), Book (bk), Book Chapter(ch)

c. Language: English

d. Years: 2015-2020

e. Source titles: transportation related magazines. We chose this feature to increase the probability of retrieving articles with high relevancy to transportation.

300 papers were manually annotated by a chair of the Big data and data analytics committee and approximately 50 papers were re-annotated by the chairs of the six committees dedicated to the various transport domains. At this stage, it was assumed that considering these domains as the appropriate transport-related classes is sufficient. It was discovered along the way that Automation and connectivity (a horizontal committee) is an additional class needed to considered, and it will serve as a category by itself in further steps. Another decision that was made is to focus on ground transportation (with the exception of drones). 255 papers out of the 300 annotated were found to be relevant for ground transportation.

Each paper was classified into one or more of the transport domains. If the annotator considered an article as clearly associated with one of the domains, it was categorized solely into this class.

In case the annotators identified more than two domains the article addressed, the two most apparent ones were chosen for classification. Table 1 depicts the annotation results of the 255 articles.

	Vehicles and Transportation modes	Traffic Management and Control	Emerging transportation services	Policy, transportation planning and Smart Cities	Travel Behavior	Safe and secure mobility
Vehicles and Transportation modes	11	2	0	1	0	0
Traffic Management and Control		43	2	20	0	3
Emerging transportation services			2	18	8	0
Policy, transportation planning and Smart Cities				49	25	1
Travel Behavior					9	6
Safe and secure mobility						53
Total papers per topic	14	70	30	114	48	63

Table 1: Papers distribution among transport domains based on human annotation

As the distribution among topics was not balanced, the number of papers sent to the each of the committees' chairs for re-annotation was not identical. Still, an attempt was made to include papers from various journals and each set of examples for re-annotation included:

a. Papers that were classified as belonging solely to the committee's domain.

b. Papers that were classified as belonging to the committee's topic and to another domain.

c. Papers that were classified as belonging to topics other than the committee's domain.

The cross-agreement ratio also varied from one domain to the other, and ranged from 70% to 85%. Discussions addressing the disagreements led to better clarification of each committee's scope of topics, however it should be noted that precise definition of this scope is difficult to obtain and some blur boundaries will probably always remain.

4. Summary

This document describes the two approaches taken for knowledge-gap analysis related to the implementation of machine learning methodologies for addressing challenges in the various domains of transportation. The attempt to identify these gaps based on recent review papers revealed that, although providing some important insights, this approach is insufficient.

An additional approach that will therefore be taken aims to develop a tool for quantitative assessment of knowledge gaps, based on the number of publications in each category of transport domain-ML technique combination. The main steps for developing such a tool were formulated and the main requirements of each step were defined. A classification model for associating articles into the various categories is at the core of this tool, and preliminary steps towards the construction of an appropriate classifier are described.

In the following months this process will continue and its results will be disseminated among the professional community.

References

Amin, M.A., Hadouej, S. and Darwish, T.S., 2019, February. Big Data Role in Improving Intelligent Transportation Systems Safety: A Survey. In *International Conference on Emerging Internetworking, Data & Web Technologies* (pp. 187-199). Springer, Cham.

Bekkerman, R. and Gavish, M., 2011, "High-precision phrase-based document classification on a modern scale." In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 231–239). ACM.

Barua, L., Zou, B. and Zhou, Y., 2020. Machine learning for international freight transportation management: a comprehensive review. *Research in Transportation Business & Management*, *34*, p.100453.

Kaffash, S., Nguyen, A.T. and Zhu, J., 2020. Big data algorithms and applications in intelligent transportation system: A review and bibliometric analysis. International Journal of Production Economics, p.107868.

Koushik, A.N., Manoj, M. and Nezamuddin, N., 2020. Machine learning applications in activity-travel behaviour research: a review. *Transport reviews*, *40*(3), pp.288-311

Mounica, B. and Lavanya, K., 2020, February. Social Media Data Analysis for Intelligent Transportation Systems. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-8). IEEE.

Neilson, A., Daniel, B. and Tjandra, S., 2019. Systematic review of the literature on big data in the transportation domain: Concepts and applications. *Big Data Research*, *17*, pp.35-44.

Welch, T.F. and Widita, A., 2019. Big data in public transportation: a review of sources and methods. Transp*ort reviews*, *39*(6), pp.795-818.

Zhu, L., Yu, F.R., Wang, Y., Ning, B. and Tang, T., 2018. Big data analytics in intelligent transportation systems: A survey. IEEE Transactions on Intelligent Transportation Systems, 20(1), pp.383-39.