Mixed data modelling of transportation and incidents

Veneta Markovska[1, a)] and Stanimir Kabaivanov[2, b)]

[1]*University of Food Technology - Plovdiv*
26 Maritza Blvd, 4000 Plovdiv, Bulgaria
[2] *Plovdiv University "Paisii Hilendarski"*
24 Tzar Assen Str., 4000 Plovdiv, Bulgaria

Author Emails
[a)] Corresponding author: stanimir.kabaivanov@uni-plovdiv.bg
[b)] v_markovska@uft-plovdiv.bg

Abstract. To create and maintain a well-functioning transportation system is a very complex task, that requires knowledge from different domains. As the number of different means of transportation grows new dependencies and dynamic behavior patterns emerge. In this paper we analyze different options for modelling transportation systems with the use of heterogeneous data inputs and sources. Our study builds on use of machine learning and big data analysis for extracting information on characteristics of transportation systems and reducing the number of incidents. We aim at developing a flexible and extensible approach, that can be adapted to transportation problems of different complexity.

INTRODUCTION

Automated processing of heterogeneous data has been subject to various research papers as in [1], [2]. Due to the convenience and speed it offers, it is possible to process and analyze large number of inputs and integrate different sources. Yet there are situations, where application of a fully automated solution may not be optimal due to the high "price" in terms of loss of material and even human lives in case of errors. We believe that transportation, as a critical system, that if not handled properly can trigger substantial economic losses and loss of human life, should be treated in a way, that focuses on the impact of errors.

Due to the fact that information regarding transportation systems and incidents related to them includes different input types, it is necessary to use them together in order to get a full picture of the undergoing processes. Working with a number of different data types, that has to be used simultaneously suggests the following:

- created models are domain and context dependent;

This means, that models will be harder to transfer without further re-calibration into different context. While this may not be equally valid for all models, it certainly is an issue for those of them, that require special inputs like images from traffic camera or special sensor inputs (like for example air quality and concentration of certain substances). Further arguments on context-dependency can be found in [3].

- data pre-processing is crucial for the overall success rate of the models;

Preliminary data processing is well known and widely used. Some algorithms actually impose, that inputs should be scaled and normalized, while in other cases it may just contribute to improving the model performance ([4], [5]).

With regard to special inputs like images and video streams, pre-processing may be a cumbersome task on its own, as it requires use of customized algorithms to extract features of interest and store them in an easy to use format.

- scalability and flexibility of created models is limited by the special inputs used.

As Table 1 indicates, in addition to numeric and categorical values, transportation system may provide special inputs like video feeds, images, vehicle identification data. Models built to depend on them would be tied to the respective sources and may require extra efforts in order to transfer them to a new context or to solve different kind of problems.

TABLE 1. Types of data commonly use in modelling

Common cases of mixed data types	Usage remarks
Combination of discrete and continuous inputs	Depending on the goals of the model, it is possible to transform continuous inputs into categories/groups doing it in a way that minimizes information loss.
Combination of numerical, categorical and nominal input variables	In this case nominal inputs can be used as a mean to separate the available data set and conduct the analysis separately on each group. If this is not feasible, then a specialized model can be selected - [6].
Use of special inputs like images, meta-data, geo location	Special data inputs can be handled with appropriate pre-processing (for example for feature extraction in processing video stream), or by integrating them into the analysis through new calculated variables (for example in case of geo-spatial data new numeric input can be created for distance or speed).
Use of inputs with different or varying over time precision/accuracy	In addition to the obvious solution of rounding, different precision/accuracy can be handled by transforming the inputs in a way to form ranges that would minimize the impact of varying precision.

If we view the examples from Table 1 as special cases of mixing structured, unstructured and semi-structured data, the general idea of using combined data set can be expressed in the way shown on Figure 1.

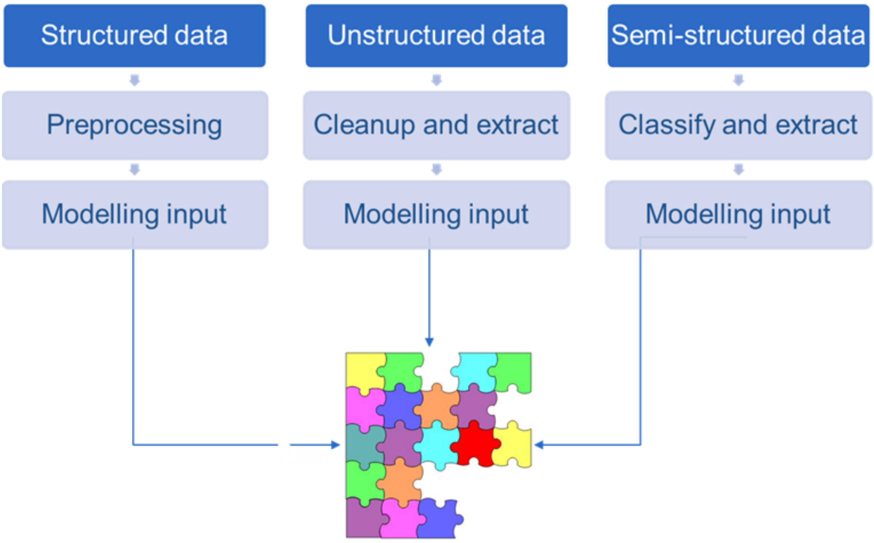

FIGURE 1. Overview of mixed data modelling of transportation

There are several important benefits, that justify going an extra mile and mixing different inputs:

- as there are multiple factors, that influence the performance and quality of public transportation systems models need to take all of them into account in order to provide adequate results;
- combination of different data sources makes it possible to gain a better understanding and prepare more accurate forecasts on the examined transportation problems;
- model findings can be validated through surveys and qualitative sources;
- experts from different domains can be involved in solving issues and improving transportation systems.

With regard to incident analysis, as a special class of problems, the benefits from mixing different inputs are even more obvious. They can not only improve the accuracy of predicting potential incidents, but also provide different point of view on incident causes, that may not be possible to obtain only from numerical data. It should also be noted, that mixed inputs can support another important part of incident analysis – estimation of the total impact (which is both economic and social) and comparing the costs and benefits for reducing the incident rates.

MIXED TYPE INPUTS AND DEEP LEARNING ALGORITHMS

There are various models, that can be used with data of different dimensions and types like linear ([7], [8]) and logistic regression ([9]). We use a neural network approach for modelling transportation and incidents. There are several reasons that justify such a decision, in particular:

- neural networks offer a more flexible and easily adaptable approach, compared to other algorithms;

Considering the fact, that internal structure and type of the network can evolve and be modified, this offers a significant advantage, when adding new inputs or when migrating the model to new environment and new goals. This comes at the expense of harder to explain networks and results, that need extra care in interpretation. Methods like defensive distillation ([10]), probabilistic robustness ([11]) and adversarial retraining ([12]) can be used to improve robustness of neural network models and reduce the risks associated with use of solutions, that are not stable enough.

- processing data in multiple layers is a possibility to include some inputs at different stages of the analysis.

Following the idea of ensemble learning ([13], [14]), this option allows to combine different algorithms and use the output of each of them in an attempt to improve model accuracy.

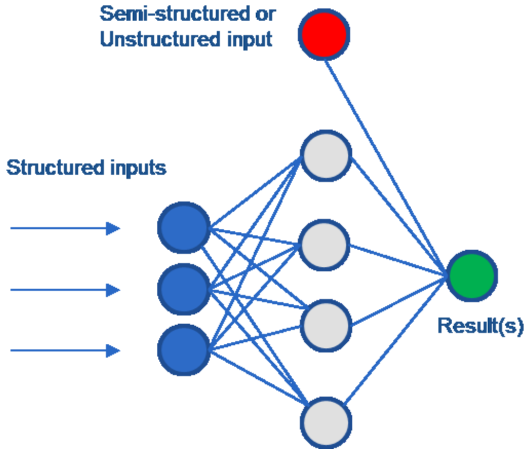

FIGURE 2. Combined use of neural networks with different inputs and models

In contrast to combining similar models to reduce variance of the outputs, the benefit of using neural networks in transportation and incident analysis is related to application of different input types at different processing steps, as shown on Figure 2. Unstructured data can be first pre-processed to create usable metrics, that can be provided as part of the analysis. For example, unstructured data as texts describing transportation issues and/or accidents, can be studied with recurrent neural networks (RNN [15]) in order to better assess sentiment embedded in the inputs.

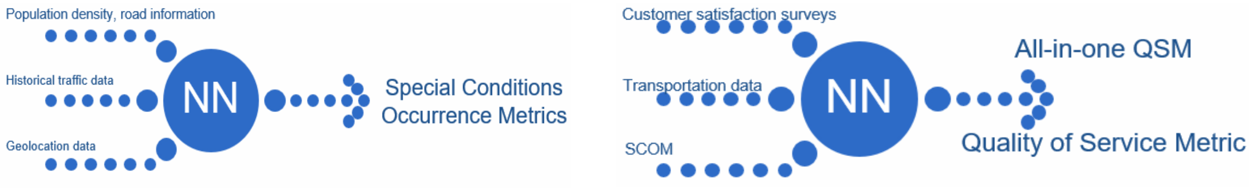
Population density, road information
Historical traffic data
Geolocation data
NN
Special Conditions Occurrence Metrics
Customer satisfaction surveys
Transportation data
SCOM
NN
All-in-one QSM
Quality of Service Metric

FIGURE 3. Use of special conditions occurrence metrics (SCOM)

Figure 3 provides an example of mixing different data types in order to improve public transportation system analysis and eventually reduce the number of traffic incidents. Special conditions occurrence metrics (SCOM) may have different root causes, thus presenting the output of distinct complementary models, that aim at analyzing such reasons.

INPUT DATA AND MODEL ASSESSMENT

Numerical experiment has been conducted on publicly available data regarding U.S. accidents in the period of 2016-2020 ([16]), where only information about traffic delays and impact on traffic is used as a feature of primary interest. In addition to numeric inputs, the data set consist of geo-spatial, categorical and textual data. A reduced number of inputs has been selected, as shown in Table 2.

TABLE 2. Data used as input

Feature	Description
Number of observations	Out of 2906610 entries in the original data set, we have selected to use only data for the second half of 2020, which reduces the number of elements to 684312. Out of this sample, we have selected only cases, where duration is larger than 360 minutes and there are no missing field of interest, which further reduces the number of observations to 36357 (split into 29085 observations in the training set and 7271 in the test one).
Number of variables	Original data set contains 49 variables, out of which 10 are used for conducting the base experiment (numeric inputs like temperature, humidity, visibility in meters, duration of the incident, wind speed, precipitation). One of the base inputs is a categorical variable – severity of the incident on a scale from 1 to 4. In the secondary experiment 2 more variables are used – textual description which is pre-processed and geo-location data of the incident.
Target variables	Incident duration (calculated as difference in minutes between the recorder start and end times) and severity are used a target variable, that model should be able to estimate and classify.

Figure 4 represents the distribution of target variable values, where there is additional limit to show cases with incident duration of less than 4000 minutes. All cases with recorded difference between end and start time, that is bigger are considered as one group.

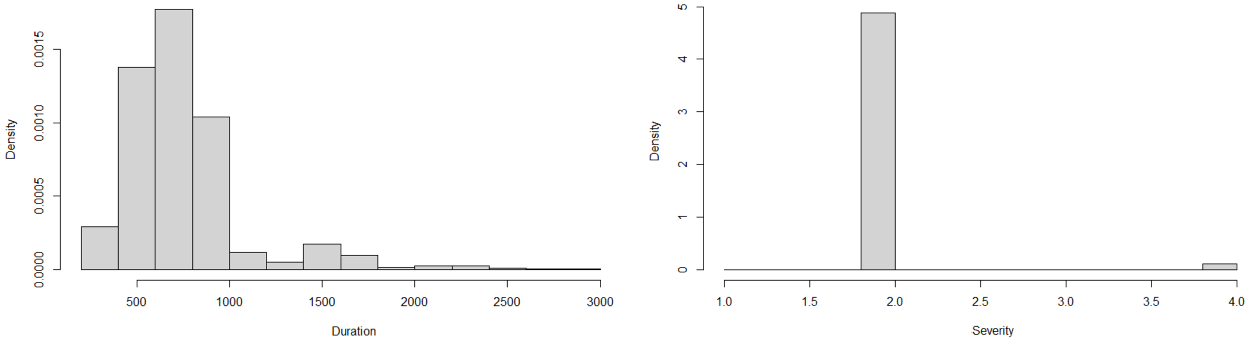

FIGURE 4. Density of target variables used in example models

The analysis is carried out as follows:

1) Numerical data is first cleaned up and scaled, before being processed by a very simple neural network without hidden layers and using resilient backpropagation algorithm ([17]). The total number of training steps is limited to 15000.
2) A GLM is used on the same inputs and data set in order to serve as a benchmark of the neural network accuracy. Considering the limited number of features used, as well as the use of only one type of neural network, we expected to have similar values for the MSE on the GLM and network models.
3) Textual description of each case is processed to represent the unstructured data associated with each case. As there are many ways, that text can be processed, we went for the most common one, using simple sentiment analysis.
4) The sentiment information (in terms of categories associated with each description) is used to extend the initial model and compare the change in MSE with neural network and GLM approaches.

Only the MSE metrics are shown in Table 3, as it is important to track down how it changes with both approaches. As expected, both neural network and GLM have similar MSE metrics in the initial model. When sentiment information is used, the MSE is reduced, with the values still being close.

TABLE 3. Numerical results for example model MSE

Case	MSE value
Initial model – neural network	0.9674065
Initial model - GLM	0.9790668
Extended model – neural network	0.9552502
Extended model – GLM	0.9552501

MSE calculations support the notion, that non-structured data can be valuable in getting better understanding on the events and incidents being studied. Taking into consideration, that target variable was the duration of the event, application of similar models can be use in practice to better assess the severity of special cases and incidents and improve allocation of available resources in order to reduce negative impact and avoid extended times, where transportation system may be blocked or functioning with limited capacity. Yet another important effect in this case would be better understanding of reasons for special events, which may fall beyond the information conveyed by simple numeric and categoric inputs. While sentiment analysis on its own cannot state such reasons, it can help in pinpointing some of the cases and segment them for further analysis.

As Table 3 indicates, there is no huge difference in the MSE for neural network and GLM. While this may be due to the fact, that we have a limited amount of data (compared to the initial data set size) and a reduced set of inputs, it also reminds, that simple models may be also very useful for well-defined and simple scenarios. Extended model demonstrated here only uses part of the available textual inputs and does not benefit from other types of data, like geo-location elements. By adding these elements and further optimizing the network structure (as well as testing different network types and learning algorithms) we can increase the accuracy of the neural network model.

CONCLUSION

Our study builds on the assumption, that both structured and unstructured data can be used to model special conditions in transportation and improve our understanding on problems and root causes of incidents. To be able to handle different data types, we use artificial neural networks and pre-processing for inputs, that need special handling before they can be used as part of the machine learning process. In the example model, this has been demonstrated with the use of geolocation and textual data.

Numerical results support the idea, that neural networks provide a flexible and powerful approach for integrating data of different types, meaning and sources. In contrast to simple pre-processing or encoding of inputs, steps followed in the experiment have one important advantage – special input types can be included at a predefined step of the modelling process. This allows to create ensemble-like solutions, where non-numeric data, like geo-location, image processing inputs and text can be used in parallel with available numerical characteristics, that describe transportation system and incidents.

REFERENCES

[1] G. Pal, K. Atkinson and G. Li, "Managing Heterogeneous Data on a Big Data Platform: A Multi-criteria Decision Making Model for Data-Intensive Science," in *IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2020.

[2] E. Trunzer, I. Kirchen, J. Folmer, G. Koltun and B. Vogel-Heuser, "A flexible architecture for data mining from heterogeneous data sources in automated production systems," in *IEEE International Conference on Industrial Technology (ICIT)*, 2017.

[3] B. Edmonds, "Complexity and Context-Dependency," *Foundations of Science,* vol. 18, no. 1, p. 745–755, 2013.

[4] A. Adeyemo, H. Wimmer and L. M. Loreen Marie Powell, "Effects of Normalization Techniques on Logistic Regression in Data Science," *Journal of Information Systems Applied Research ,* vol. 12, no. 2, pp. 37-44, 2019.

[5] J. Pan, Y. Zhuang and S. Fong, "The Impact of Data Normalization on Stock Market Prediction: Using SVM and Technical Indicators," in *International Conference on Soft Computing in Data Science: SCDS 2016: Soft Computing in Data Science*, 2016.

[6] D. Hedeker, "Multilevel Models for Ordinal and NominalVariables," in *Handbook of Multilevel Analysis*, New York, Springer, 2008, pp. 239-270.

[7] C. M. Cuadras and C. Arenas, "A distance based regression model for prediction with mixed data," *Communications in Statistics - Theory and Methods,* no. 6, pp. 2261-2279, 1990.

[8] N. Alghanmi and X.-J. Zeng, "A Hybrid Regression Model for Mixed Numerical and Categorical Data.," in *UK Workshop on Computational Intelligence*, Cham, Springer, 2019, pp. 369-376.

[9] B. Levin and P. E. Shrout, "On extending bock's model of logistic regression in the analysis of categorical data," *Communications in Statistics - Theory and Methods,* vol. 10, no. 2, pp. 125-17, 1981.

[10] N. Papernot, P. McDaniel, X. Wu, S. Jha and A. Swami, "Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks," in *IEEE Symposium on Security and Privacy*, San Jose, CA, 2016.

[11] M. Ravi, A. V. Nori and A. Orso, "Robustness of neural networks: A probabilistic and practical approach.," in

IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results, IEEE, 2019.

[12] M. Nag, M. Moh and T.-S. Moh, "Defending deep learning models against adversarial attacks.," *International Journal of Software Science and Computational Intelligence,* vol. 13, no. 1, pp. 72-89, 2021.

[13] D. Xibin, Z. Yu, W. Cao, Y. Shi and Q. Ma, "A survey on ensemble learning," *Frontiers of Computer Science,* vol. 14, no. 2, pp. 241-258, 2020.

[14] B. E. Rosen, "Ensemble learning using decorrelated neural networks," *Connection science 8,* Vols. 3-4, pp. 373-384, 1996.

[15] W. Xingyou, W. Jiang and Z. Luo, "Combination of convolutional and recurrent neural network for sentiment analysis of short texts," in *The 26th international conference on computational linguistics: Technical papers*, 2016.

[16] Kaggle, "A Countrywide Traffic Accident Dataset," 21 05 2019. [Online]. Available: https://www.kaggle.com/sobhanmoosavi/us-accidents/metadata. [Accessed 21 05 2021].

[17] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," in *1993*, Proceedings of the IEEE International Conference on Neural Networks.

ACKNOWLEDGMENTS

Authors wish to acknowledge the assistance and funding of this research as part of project UIA05-202 "INNOAIR - Innovative demand responsive green public transportation for cleaner air in urban environment", funded by the European Union initiative - Urban Innovative Actions (UIA).