DATA-DRIVEN METHODS FOR ACTIVITY SCHEDULE MODELING

ANIL NP KOUSHIK

DEPARTMENT OF CIVIL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

AUGUST 2024

DATA-DRIVEN METHODS FOR ACTIVITY SCHEDULE MODELING

by

ANIL NP KOUSHIK

DEPARTMENT OF CIVIL ENGINEERING

Submitted

In fulfilment of the requirements of the degree of doctor of philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

AUGUST 2024

CERTIFICATE

This is to certify that the thesis entitled "**Data-Driven Methods for Activity Schedule Modeling**" is being submitted by **Mr. Anil NP Koushik** to the Indian Institute of Technology, Delhi, India, for the award of the degree of Doctor of Philosophy. This is a record of the original bonafide research work carried out by him under our guidance and supervision. The doctoral research work presented in this thesis has not been submitted, in part or in full, to any other University or Institute for the award of any degree or diploma.

Dr. M. Manoj
Associate Professor
Department of Civil Engineering
Indian Institute of Technology Delhi
Hauz Khas, New Delhi - 110016,
India

Dr. Nezamuddin
Associate Professor
Department of Civil Engineering
Indian Institute of Technology Delhi
Hauz Khas, New Delhi - 110016,
India

ACKNOWLEDGEMENTS

I wish to extend my profound gratitude to the numerous individuals and organizations who have played an indispensable role in guiding and supporting me on my research journey.

First and foremost I'm immensely thankful for my thesis supervisor Dr Manoj M for his persistent guidance and support. His mentorship has been unwavering, guiding me through the intricacies of research, from reviewing articles to writing them. He encouraged me to participate in seminars, conferences and workshops constantly urging me to upgrade my knowledge and skills. I immensely appreciate his approach where he allowed me to pursue my own ideas and explore my research while providing crucial and useful guidance whenever needed. In addition to being a supervisor, he has played the role of a mentor and a friend providing emotional and financial support which has been invaluable through this journey. I will always remember how he would draw upon his own research experiences to guide me through mine. The belief he has shown in my abilities got me through thick and thin of research.

Dr. Nezamuddin, my other esteemed supervisor, has provided unwavering support and guidance throughout this research journey. His valuable contributions to my research and paper writing, along with his constant encouragement, have played a pivotal role in advancing this work. I am genuinely thankful for his unwavering encouragement, guidance, and financial support, especially during challenging times.

I acknowledge the contributions of the Student Research Committee members, including Prof. Dhanya C. T., Prof. Aravind Krishna Swamy, and Dr. Deepak U Patil, and for their valuable input into this research. My gratitude also extends to former SRC members, Prof. Geetam Tewari and Prof. Rakesh Khosa.

Furthermore, I appreciate the valuable insights provided by Prof. Geetam Tewari, Prof. Kalaga Ramachandra Rao, and Prof. Aravind Krishna Swamy during the semester progress presentations. I would like to express my gratitude for the invaluable support and guidance provided by Dr. Pratosh AP, a member of the Electrical Engineering Department. His involvement has been instrumental in the progress of my research, and I am genuinely thankful for his contributions to this work. I extend my gratitude to Prof. Pushkin Kachroo for his encouragement and for generously sharing exciting ideas and concepts with me.

I wish to acknowledge the Indian Institute of Technology (IIT) Delhi for affording me the opportunity, providing the essential resources, and creating a conducive environment for my research. The faculty and staff of the Civil Engineering Department and Transportation Research and Injury Prevention Center (TRIP-C) have provided administrative support for my research, and I am thankful for their contributions.

I wish to express my gratitude to Centraal Bureau voor de Statistiek (CBS) and Rijkswaterstaat (RWS-WVL) in the Netherlands for granting access to the dataset used in this study. Their swift response to any queries and their support in accessing this data were essential for the successful completion of this research. I also acknowledge the Python coding and Machine Learning community for their open science practices and readily available support, which significantly facilitated coding and model development challenges.

I would like to express my appreciation to Chetan, who has been not only a close personal friend but also a dedicated lab mate. His friendship and companionship have been a source of support and encouragement that I hold in high regard. I am deeply thankful to Punyabeet for the assistance provided during my visit to Bhubaneswar and the invaluable help regarding choice modeling. I would like to extend my gratitude to Poonam, Vamshi, and Mustansir for their invaluable help and support especially during the times when I was working away from the lab. To my other lab mates, Dr. Hari Krishna, Dr. Lakshmi Devi, Dr. Mohit Kumar Singh, Dr. Abhary, Dr. Deotima, Dr. Shayesta, Dr. Tarapada Mandal, Aishwarya, Pranav, Sourav, and Krishnan, I extend my heartfelt gratitude for your friendship and unwavering support throughout this journey.

My family, including my father, Padmanabha NR, and mother, Rukmini, have provided continuous support and stood by me through life's challenges. Their unwavering presence has been a source of strength. My wife, Shruthi, and children, Aneesh and Sadhvi, have been constant sources of support and relaxation during the demanding phases of my research. I am also thankful to my in-laws for their support during challenging times. Special thanks to my sister's family, who provided invaluable help and support, particularly during my relocation to Delhi.

There are countless others who have contributed in various ways to this research, and while I cannot thank them all individually here, their support and encouragement will forever be etched in my heart.

Each individual mentioned above, and countless others who have contributed in diverse ways, has played an indispensable role in making this research a reality. I extend my heartfelt appreciation to each and every one of you for your support, encouragement, and contributions to this work.

Anil NP Koushik

ABSTRACT

Activity schedule modeling is recently exploring data-driven methods, which offer advantages over traditional approaches by deriving relationships directly from the data and requiring fewer assumptions from the modeler. However, the existing data-driven models for activity schedules typically utilize activity episode index as a reference, which does not provide meaningful implications for the schedule and limits their ability to capture useful patterns. To address this limitation, this thesis proposes treating activity schedule as a time series for modelling, allowing time to serve as the reference for analysis. This approach enables the model to capture important patterns in activity schedules, while facilitating simultaneous prediction of different schedule components. This avoids imposition of any assumptions from the modeler's side and helps account for all possible interdependencies among schedule components. It also ensures time budget constraints are inherently satisfied and avoids episode overlaps in the schedule. To model the resulting long time series and address long-term dependencies, the thesis employs a data-driven tool - long short-term memory (LSTM) network. The proposed framework is implemented using activity-travel diary data from respondents in the Netherlands, specifically the OViN 2016 dataset. To assess the model's robustness, it is also tested on a time use dataset encompassing 23 different activity types. Furthermore, different approaches to address class imbalance problems arising from infrequently performed activities in the schedule are explored and compared.

Two prominent issues stand out when the literature on Artificial Neural Networks (ANNs) is reviewed: the lack of model transferability and interpretability (often referred to as the black box issue). To address these challenges, the spatial transferability of the developed activity schedule model is tested by initially developing it for the city of Rotterdam and then transferring it to Amsterdam. Transfer learning, a technique that enhances transferability using a small amount of local data, is employed to improve the model's transferability. The effectiveness of transfer learning is first evaluated on various mode choice models before being extended to the proposed activity schedule model to enhance its transferability. To tackle the black box issue, the study utilizes a post hoc, model-agnostic feature attribution method called SHapley Additive exPlanations (SHAP). SHAP is employed to explain the contributions of individual features towards the model's output. Initially, SHAP is demonstrated on a simple mode choice model

using feed-forward neural networks and then extended to explain LSTM-based activity schedule models, providing insights into the factors driving the model's predictions. Lastly, this thesis introduces a novel ANN-based formulation called ANNProbit, designed for classifying activity patterns. This formulation allows for the incorporation of correlated error terms and enables realistic substitutions between similar activity patterns. The effectiveness of ANNProbit is evaluated on both synthetic datasets and the publicly available Swissmetro dataset. In the end, ANNProbit is employed to model activity patterns and analyze behavioral responses under various scenarios.

The findings of this study suggest that the proposed activity schedule model exhibited accurate predictions for the frequency of activity participation, within 10% closeness. The bidirectional LSTM model successfully estimated activity start-time distributions for most activities, as confirmed by the Kruskal Wallis test. The impact of class imbalance on model performance was significant, but the use of a weighted loss function effectively mitigated the issue. In addition, the model's performance remained consistent across different subgroups of data, aligning with the results at the aggregate level. Further, the model demonstrated efficient handling of diverse activity types when validated using time use data. When the proposed model was transferred spatially from Rotterdam to Amsterdam, initially poor performance was observed. However, employing transfer learning resulted in significant improvement. Even with only 50% of the local data, the transfer learning-enhanced model yielded comparable results to a locally developed model. While addressing the black box issue, SHapley Additive exPlanations (SHAP) was found to provide plausible explanations for the model's predictions. It was found that day of week, social participation and month of the year had significant impact on activity schedule. Household composition and urban class were also identified as key factors influencing activity participation. The variable attributions derived from SHAP were consistent with existing literature and intuitive understanding, and this induces better confidence to the modeler in model results. The proposed ANNProbit framework outperformed other models, including ANN with Softmax activation, Multinomial Probit (MNP), and Multinomial Logit (MNL), in terms of prediction ability for both simulated and real-world datasets. The incorporation of correlated error terms in ANNProbit enhanced the flexibility of substitution patterns. Further, applying ANNProbit to model activity patterns resulted in behavioral results consistent with prior research.

Keywords: Activity schedule modeling, Machine Learning, Artificial Neural Networks, Spatial Transferability, Explainable AI

सार

गतिविधि शेड्यूल मॉडलिंग तेजी से डेटा-संचालित तरीकों की ओर रुख कर रही है, जो सीधे डेटा से संबंध प्राप्त करके और मॉडलर से कम मान्यताओं की आवश्यकता के द्वारा पारंपरिक दृष्टिकोण पर लाभ प्रदान करती है। हालाँकि, गतिविधि शेड्यूल के लिए मौजूदा डेटा-संचालित मॉडल आमतौर पर गतिविधि एपिसोड इंडेक्स को संदर्भ के रूप में उपयोग करते हैं, जो शेड्यूल के लिए सार्थक निहितार्थ प्रदान नहीं करता है और उपयोगी पैटर्न को पकड़ने की क्षमता को सीमित करता है। इस सीमा को संबोधित करने के लिए, यह थीसिस गतिविधि अनुसूची को एक समय श्रृंखला के रूप में मानने का प्रस्ताव करती है, जिससे समय विश्लेषण के लिए संदर्भ के रूप में काम कर सके। यह दृष्टिकोण मॉडलर को गतिविधि शेड्यूल में महत्वपूर्ण पैटर्न को पकड़ने में सक्षम बनाता है, जबकि विभिन्न शेड्यूल घटकों की एक साथ भविष्यवाणी की सुविधा प्रदान करता है। यह मॉडलर की ओर से किसी भी धारणा को थोपने से बचता है और शेड्यूल घटकों के बीच सभी संभावित अन्योन्याश्रितताओं को ध्यान में रखने में मदद करता है। यह यह भी सुनिश्चित करता है कि समय बजट की कमी स्वाभाविक रूप से संतुष्ट हो और शेड्यूल में एपिसोड ओवरलैप से बचा जाए। परिणामी लंबी समय श्रृंखला को मॉडल करने और दीर्घकालिक निर्भरता को संबोधित करने के लिए, थीसिस विभिन्न मशीन लर्निंग तकनीकों की व्यापक समीक्षा के बाद लंबी अवधि की मेमोरी (एलएसटीएम) नेटवर्क, एक तंत्रिका नेटवर्क-आधारित वास्तुकला को नियोजित करती है।

प्रस्तावित रूपरेखा नीदरलैंड में उत्तरदाताओं से गतिविधि-यात्रा डायरी डेटा, विशेष रूप से ओवीआईएन 2016 डेटासेट का उपयोग करके कार्यान्वित की जाती है। मॉडल की मजबूती का आकलन करने के लिए, इसे 23 विभिन्न गतिविधि प्रकारों वाले समय बजट डेटासेट पर भी परीक्षण किया जाता है। मॉडल की मजबूती का आकलन करने के लिए, इसे 23 विभिन्न गतिविधि प्रकारों को शामिल करने वाले समय बजट डेटासेट पर भी परीक्षण किया जाता है। थीसिस अनुसूची में कभी-कभार की जाने वाली गतिविधियों के कारण उत्पन्न होने वाले वर्ग असंतुलन के मुद्दों को संभालने के लिए विभिन्न तरीकों की खोज और तुलना करती है। इसके अलावा, अनुसूची में कभी-कभार की जाने वाली गतिविधियों से उत्पन्न होने वाली वर्ग असंतुलन की समस्याओं के समाधान के लिए विभिन्न तरीकों का पता लगाया जाता है और उनकी तुलना की जाती है।

कृत्रिम तंत्रिका नेटवर्क (एएनएन) पर साहित्य की समीक्षा करते समय, दो प्रमुख मुद्दे सामने आते हैं: मॉडल हस्तांतरणीयता और व्याख्यात्मकता की कमी (अक्सर ब्लैक बॉक्स मुद्दे के रूप में जाना जाता है)। इन

चुनौतियों का समाधान करने के लिए, विकसित गतिविधि अनुसूची मॉडल की स्थानिक हस्तांतरणीयता का परीक्षण शुरू में इसे रॉटरडैम के लिए विकसित करके और फिर इसे एम्स्टर्डम में स्थानांतरित करके किया जाता है। ट्रांसफ़र लर्निंग, एक ऐसी तकनीक जो स्थानीय डेटा की थोड़ी मात्रा का उपयोग करके ट्रांसफ़रेबिलिटी को बढ़ाती है, मॉडल की ट्रांसफ़रेबिलिटी को बेहतर बनाने के लिए नियोजित की जाती है। स्थानांतरण सीखने की प्रभावशीलता का मूल्यांकन पहले इसकी हस्तांतरणीयता को बढ़ाने के लिए प्रस्तावित गतिविधि अनुसूची मॉडल तक विस्तारित करने से पहले विभिन्न मोड विकल्प मॉडल पर किया जाता है। ब्लैक बॉक्स मुद्दे से निपटने के लिए, अध्ययन पोस्ट हॉक, मॉडल-अज्ञेयवादी फीचर एट्रिब्यूशन विधि का उपयोग करता है जिसे SHapley Additive exPlanations (SHAP) कहा जाता है। मॉडल में व्यक्तिगत विशेषताओं के योगदान को समझाने के लिए SHAP का उपयोग किया जाता है। प्रारंभ में, SHAP को फ़ीड-फ़ॉरवर्ड न्यूरल नेटवर्क का उपयोग करके एक सरल मोड विकल्प मॉडल पर प्रदर्शित किया जाता है और फिर LSTM-आधारित गतिविधि शेड्यूल मॉडल को समझाने के लिए विस्तारित किया जाता है, जो मॉडल की भविष्यवाणियों को चलाने वाले कारकों में अंतर्दृष्टि प्रदान करता है। यह पाया गया कि सप्ताह के दिन, सामाजिक भागीदारी और वर्ष के महीने का गतिविधि कार्यक्रम पर महत्वपूर्ण प्रभाव पड़ा। घरेलू संरचना और शहरी वर्ग को भी गतिविधि भागीदारी को प्रभावित करने वाले प्रमुख कारकों के रूप में पहचाना गया। अंत में, यह थीसिस एएनएनप्रोबिट नामक एक नए एएनएन-आधारित फॉर्मूलेशन का परिचय देती है, जिसे गतिविधि पैटर्न को वर्गीकृत करने के लिए डिज़ाइन किया गया है। यह सूत्रीकरण सहसंबद्ध त्रुटि शर्तों को शामिल करने की अनुमति देता है और समान गतिविधि पैटर्न के बीच यथार्थवादी प्रतिस्थापन को सक्षम बनाता है। ANNProbit की प्रभावशीलता का मूल्यांकन सिंथेटिक डेटासेट और सार्वजनिक रूप से उपलब्ध स्विसमेट्रो डेटासेट दोनों पर किया जाता है। अंत में, ANNProbit को गतिविधि पैटर्न को मॉडल करने और विभिन्न परिदृश्यों के तहत व्यवहारिक प्रतिक्रियाओं का विश्लेषण करने के लिए नियोजित किया जाता है।

इस अध्ययन के निष्कर्षों से पता चलता है कि प्रस्तावित गतिविधि अनुसूची मॉडल ने 10% निकटता के साथ गतिविधि भागीदारी की आवृत्ति के लिए सटीक भविष्यवाणियां प्रदर्शित कीं। जैसा कि क्रुस्कल वालिस परीक्षण द्वारा पुष्टि की गई है, द्विदिशात्मक एलएसटीएम मॉडल ने अधिकांश गतिविधियों के लिए गतिविधि प्रारंभ-समय वितरण का सफलतापूर्वक अनुमान लगाया है। मॉडल प्रदर्शन पर वर्ग असंतुलन का प्रभाव महत्वपूर्ण था, लेकिन भारित हानि फ़ंक्शन के उपयोग ने इस मुद्दे को प्रभावी ढंग से कम कर दिया। मॉडल का प्रदर्शन डेटा के विभिन्न उपसमूहों में सुसंगत रहा, जो समग्र स्तर पर परिणामों के अनुरूप था। इसके

अतिरिक्त, समय उपयोग डेटा का उपयोग करके सत्यापित होने पर मॉडल ने विभिन्न गतिविधि प्रकारों के कुशल संचालन का प्रदर्शन किया। जब मॉडल को रॉटरडैम से एम्स्टर्डम में स्थानांतरित किया गया, तो शुरू में खराब प्रदर्शन देखा गया। हालाँकि, ट्रांसफर लर्निंग को नियोजित करने से महत्वपूर्ण सुधार हुआ। यहां तक कि केवल 50% स्थानीय डेटा के साथ, ट्रांसफर लर्निंग-एन्हांस्ड मॉडल ने स्थानीय रूप से विकसित मॉडल के तुलनीय परिणाम दिए। ब्लैक बॉक्स मुद्दे को संबोधित करने के लिए, अध्ययन ने शेप्ले एडिटिव एक्सप्लानेशंस (एसएचएपी) का उपयोग किया, जिसने मॉडल की भविष्यवाणियों के लिए प्रशंसनीय स्पष्टीकरण प्रदान किया। SHAP से प्राप्त परिवर्तनीय विशेषताएँ मौजूदा साहित्य और सहज समझ के अनुरूप थीं। इसके अलावा, प्रस्तावित एएनएनप्रोबिट फ्रेमवर्क ने सिम्युलेटेड और वास्तविक दुनिया डेटासेट दोनों के लिए भविष्यवाणी क्षमता के मामले में सॉफ्टमैक्स सक्रियण, मल्टीनोमियल प्रोबिट (एमएनपी), और मल्टीनोमियल लॉगिट (एमएनएल) के साथ एएनएन सहित अन्य मॉडलों से बेहतर प्रदर्शन किया। ANNProbit में सहसंबद्ध त्रुटि शब्दों के समावेश ने प्रतिस्थापन पैटर्न के लचीलेपन को बढ़ाया। मॉडल गतिविधि पैटर्न में ANNProbit को लागू करने से पूर्व शोध के अनुरूप व्यवहारिक परिणाम प्राप्त हुए।

Table of Contents

List of figures

List of Tables

List of abbreviations

ABM	Activity Based Model
ADAM	ADAptive Moment
AI	Artificial Intelligence
AMOS	Activity MObility Simulator
ANN	Artificial Neural Network
CART	Classification And Regression Trees
CHAID	Chi-squared Automatic Interaction Detector
CPM	Computational Process Model
DCM	Discrete Choice Model
DT	Decision Trees
HMM	Hidden Markov Model
IIA	Independence from Irrelevant Alternatives
IID	Independent and Identically Distributed
LIME	Local Interpretable Model-agnostic Explanations
LRP	Layerwise Relevance Propagation
LSTM	Long Short-Term Memory
ML	Machine Learning
MNL	Multinomial Logit
MNP	Multinomial Probit
NL	Nested Logit
NLP	Natural Language Processing
ODiN	Onderzoek Onderweg in Nederland
OViN	Onderzoek Verplaatsingen in Nederland
RATE	Relative Aggregate Transfer Error
REM	Relative Error Measure
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RUM	Random Utility Maximization
SHAP	Shapley Additive Explanations
SVM	Support Vector Machines
TAZ	Traffic Analysis Zone
TDM	Travel Demand Management
TL	Transfer Learning
TS	Transfer Scaling
TUS	Time Use Survey