Machine learning applications in land use and travel behavior analysis

Xinyu (Jason) Cao

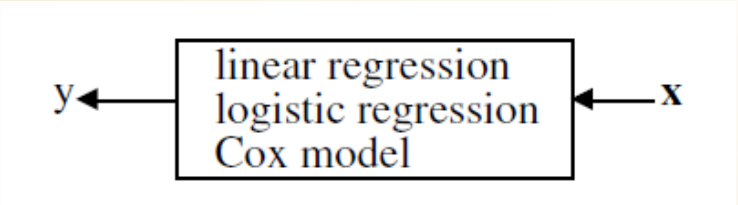
cao@umn.edu

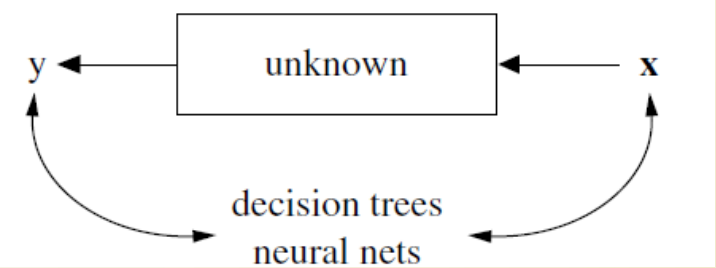
HUMPHREY SCHOOL OF PUBLIC AFFAIRS

UNIVERSITY OF MINNESOTA

Two Cultures of Modeling

- Breiman (2001): Real-world data are generated through an unknown and complex process.





- Data modeling is to estimate model parameters from data, based on the premise that the data are generated through a causal pathway with a known distribution. It may produce "irrelevant theory and questionable scientific conclusions".
- Machine learning explores multiple functions and finds the one that performs best.
- How data fit a model vs. use data to predict a model

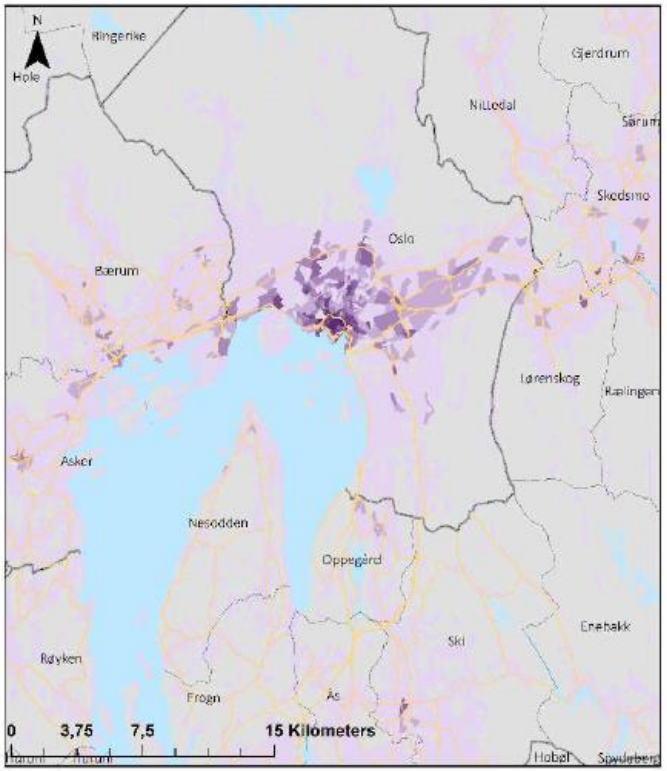
Case studies

- Questionable conclusions from data modeling
- Important, plausible, but infeasible land use interventions
- Interaction effects without priori knowledge

Introduction

- Monocentric to polycentric urban form
  - City center**, regional center, local center
  - How do centers affect activity location choice? (Petter Naess)
    - The best facility vs. proximity
- Does distance to the city center have a linear influence?
- Research questions
  - How important is the effect of sub-centers, relative to the city center, on driving distance?
  - Do they have nonlinear relationships?

Oslo



Legend

Employees per km2

- 0 - 1900
- 1901 - 5900
- 5901 - 12500
- 12501 - 24300
- 24301 - 43100
- 43101 - 80600
- Uninhabitated area

- Municipal border
- County border
- Road

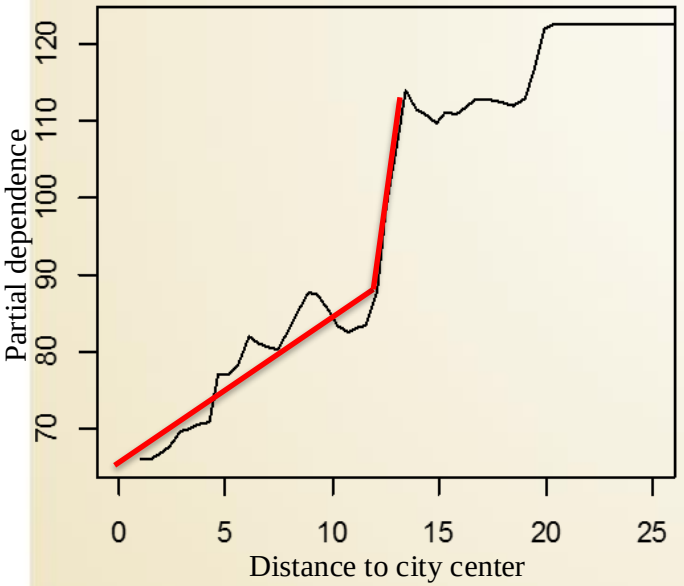
Method

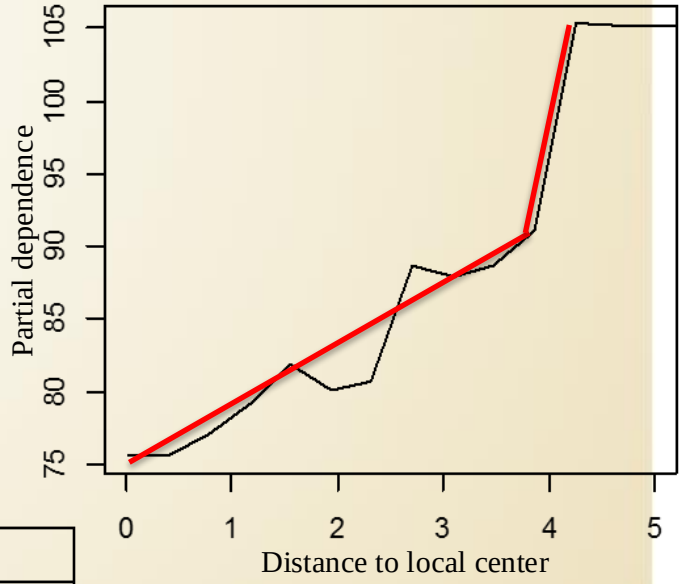
- Self-administered online survey in 2015
  - 1,517 respondents in Oslo
- Dependent variable
  - Weekly driving distance
- Built environment variables
  - Distances to the main (second-order, local) center
  - Population and job densities – density
  - Transit zone
- Gradient boosting decision trees (the “gbm” package in R)

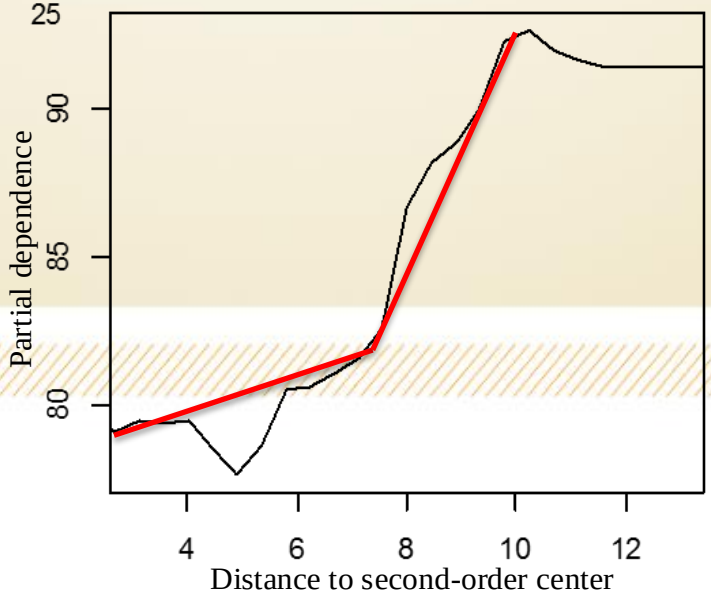
Feature importance

Categories	Variable	Weekly driving distance	
		Rank	Importance (%)
Demographic characteristics	Children	12	0.61
	Teenager	14	0.55
	Education	10	3.17
	Personal income	6	10.23
	Household income	7	7.15
	Age	2	14.37
	Household size	11	2.25
	Female	9	4.43
	Workforce	15	0.53
Built environment attributes	Distance to city center	1	18.03
	Distance to local center	4	10.56
	Distance to second-order center	5	10.25
	Population density	3	12.03
	Employment density	8	5.26
	Transit zone	13	0.58

Partial Dependence Plots







HUMPHREY SCHOOL OF PUBLIC AFFAIRS
UNIVERSITY OF MINNESOTA

Key results

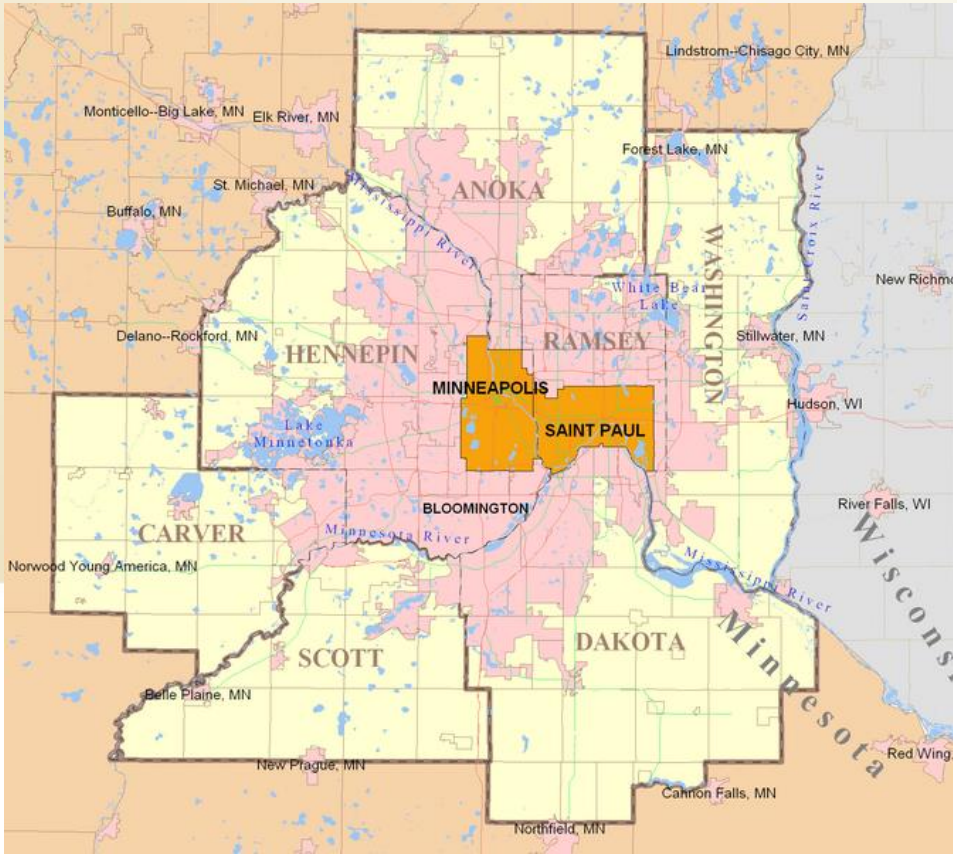
- The importance of centers
  - City center
  - Second-order center + local center
- Why does machine learning produce different results?
  - Nonlinear relationships
  - Multicollinearity among the three distance variables
- Notes
  - An efficient tool to estimate irregular nonlinear relationships as no priori knowledge is required
  - Thresholds → planning guidelines

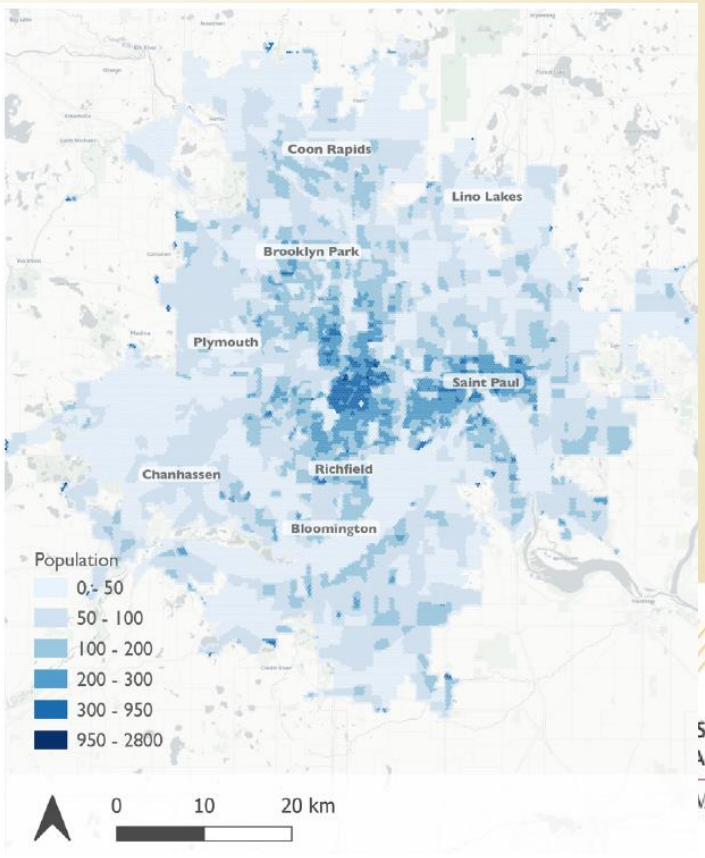
Case studies

- Questionable conclusions from data modeling
- Important, plausible, but infeasible land use interventions
- Interaction effects without priori knowledge

Minneapolis-St. Paul

- Data
  - Regional household travel survey in the Twin Cities (2019)





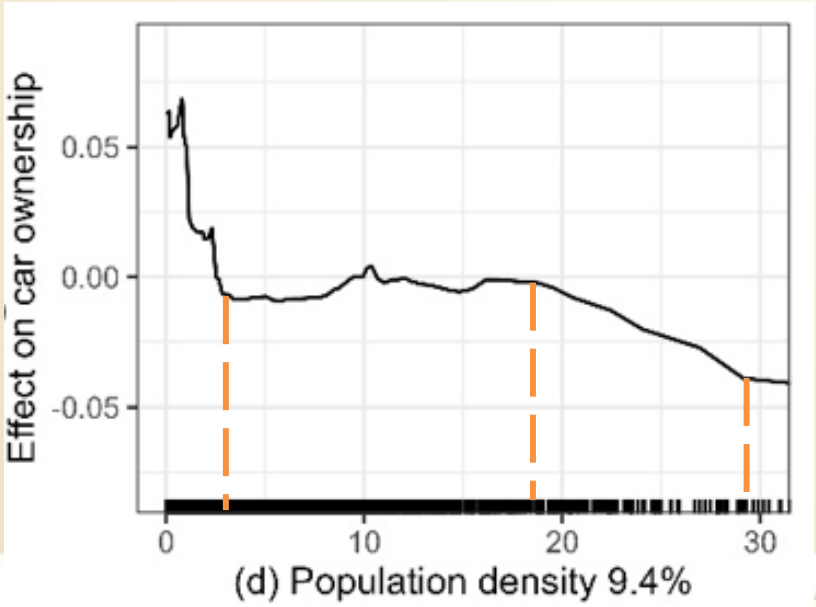
SCHOOL
AFFAIRS
MINNESOTA

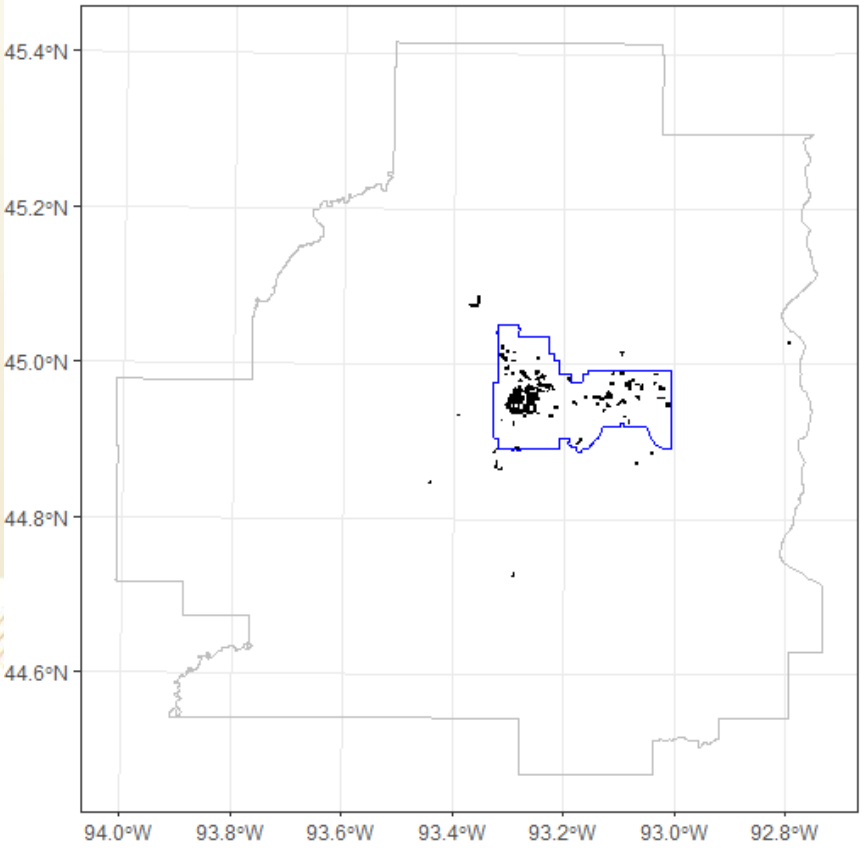
Method

- Variables
  - Dependent variable: Number of vehicles per licensed household member
  - Independent variables
    - Demographics
    - Built environment **D**imensions: **population density**, land use mix, intersection density, transit stop density, distance to Minneapolis, distance to St. Paul, and job accessibility
  - Method
    - Linear regression: significant and negative; implications
    - Gradient boosting decision trees

Key results

- Relative importance: 9.4%; the 2nd most important predictor
- Accumulated local effects (ALE) Plot: negative





Case studies

- Questionable conclusions from data modeling
- Important, plausible, but infeasible land use interventions
- Interaction effects without priori knowledge

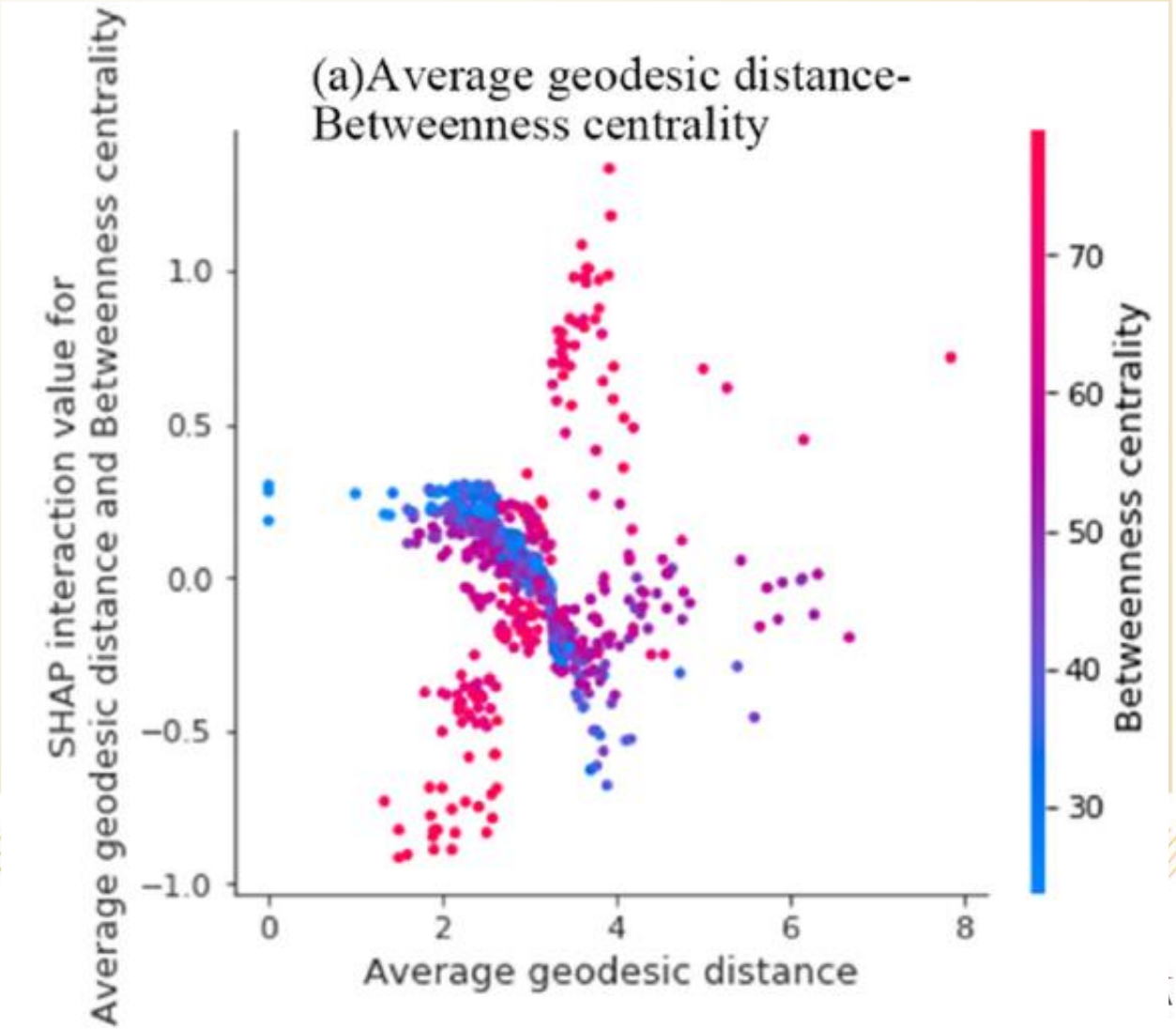
Method

Ji et al. (2022)

- Bicycle travel survey in Xi'an (2016)
- Dependent variable: cycling distance
- Key independent variables
  - Average geodesic distance: a smaller value represents a more direct connection.
  - Betweenness centrality: a smaller value means more accessible and interconnected.
  - Built environment attributes
- Method: XGBoost and SHAP value

Relationships differ by betweenness centrality

- Overall
  - Positive
- Centrality
  - Small
  - Negative
  - Medium
  - V-shaped
  - Large
  - Positive



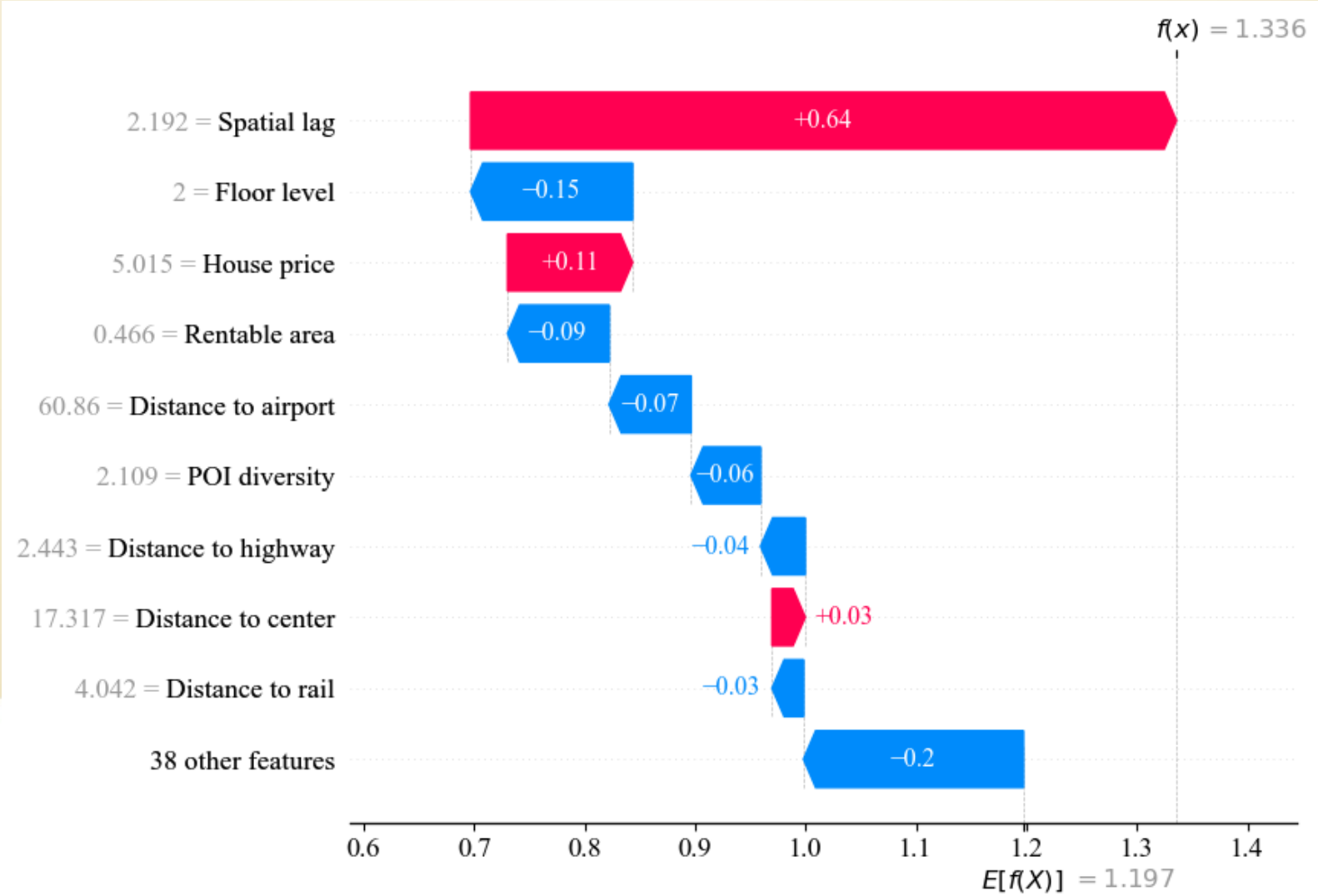
Embrace machine learning

- Challenge the conventional understanding of the relationship between variables
  - Questionable conclusions based on unreal assumptions
- Improve the understanding of the relationship
  - Interactions without a priori
- Inform planning practices
  - Plausible but infeasible

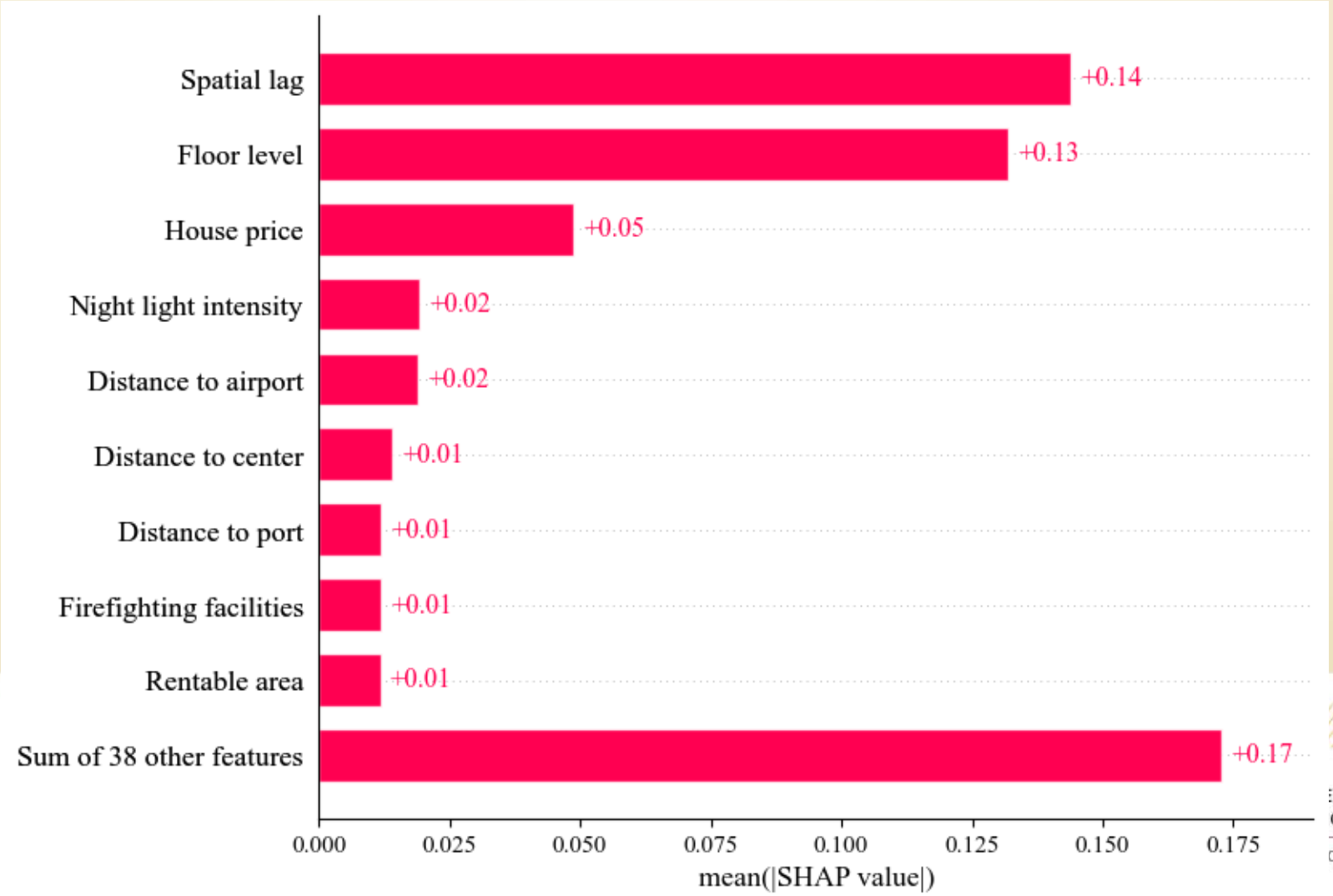
References

- Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)." *Statistical Science* 16 (3):199-231.
- Cao, Jason and Tao Tao. 2024. “Can an identified environmental correlate of car ownership serve as a practical planning tool?” under review.
- Ding, Chuan, Xinyu Cao, and Petter Næss. 2018. "Applying gradient boosting decision trees to examine non-linear effects of the built environment on driving distance in Oslo." *Transportation Research Part A* 110:107-117.
- Ji, Shujuan, Xin Wang, Tao Lyu, Xiaojie Liu, Yuanqing Wang, Eva Heinen, and Zhenwei Sun. 2022. "Understanding cycling distance according to the prediction of the XGBoost and the interpretation of SHAP: A non-linear and interaction effect analysis." *Journal of Transport Geography* 103:103414.

The rental price of a warehouse



Contribution to the predictions of all warehouses



Machine learning applications in land use and travel behavior analysis by Prof. Jason Cao, 10/18/24

Niaz Mahmud Zafri

Part I. Literature (for further reading)

Ding, C., Cao, X. J., & Næss, P. (2018). Applying gradient boosting decision trees to examine non-linear effects of the built environment on driving distance in Oslo. *Transportation Research Part A: Policy and Practice*, *110*, 107-117. https://doi.org/10.1016/j.tra.2018.02.009

Ding, C., Cao, X., & Wang, Y. (2018). Synergistic effects of the built environment and commuting programs on commute mode choice. *Transportation Research Part A: Policy and Practice*, *118*, 104-118. https://doi.org/10.1016/j.tra.2018.08.041

Wu, X., Tao, T., Cao, J., Fan, Y., & Ramaswami, A. (2019). Examining threshold effects of built environment elements on travel-related carbon-dioxide emissions. *Transportation Research Part D: Transport and Environment*, *75*, 1-12. https://doi.org/10.1016/j.trd.2019.08.018

Ji, S., Wang, X., Lyu, T., Liu, X., Wang, Y., Heinen, E., & Sun, Z. (2022). Understanding cycling distance according to the prediction of the XGBoost and the interpretation of SHAP: A non-linear and interaction effect analysis. *Journal of Transport Geography*, *103*, 103414. https://doi.org/10.1016/j.jtrangeo.2022.103414

Tong, Z., An, R., Zhang, Z., Liu, Y., & Luo, M. (2022). Exploring non-linear and spatially non-stationary relationships between commuting burden and built environment correlates. *Journal of Transport Geography*, *104*, 103413. https://doi.org/10.1016/j.jtrangeo.2022.103413

Ashik, F. R., Sreezon, A. I. Z., Rahman, M. H., Zafri, N. M., & Labib, S. M. (2024). Built environment influences commute mode choice in a global south megacity context: Insights from explainable machine learning approach. *Journal of Transport Geography*, *116*, 103828. https://doi.org/10.1016/j.jtrangeo.2024.103828

Part II. Recent News

Mary Scott Nabers. (2024, October 24). T*ransit-Oriented Development Projects are Launching Nationwide.* Construction Citizen. https://constructioncitizen.com/blog/transit-oriented-development-projects-are-launching-nationwide/2410241

Michael J. Autuori. (2024, September 21). *Transit-oriented development, a reality check.* News Times. https://www.newstimes.com/opinion/article/transit-oriented-development-reality-check-19769015.php

Laurel Demkovich. (2024, January 19). *The fight over building denser housing near transit.* Washington State Standard. https://washingtonstatestandard.com/2024/01/19/the-fight-over-building-denser-housing-near-transit/

Andrew Kenney. (2024, April 23). *In a push for more housing density near transit lines, highway dollars have become a political football.* CRP News. https://www.cpr.org/2024/04/23/housing-density-near-transit-oriented-areas-highway-dollars/

Aaron Short. (2024, October 23). *Sprawl order: Presidential candidates are talking about new housing. But where?*. Greater Greater Washington. *https://ggwash.org/view/97369/sprawl-order-presidential-candidates-are-talking-about-new-housing*

Part III. Questions and answers

Jinhua's Questions:

Q1: Jinhua questioned whether past studies that assumed linear relationships may have overlooked threshold effects and non-linear relationships, asking if researchers should revisit prior conclusions in the literature.

A1: Jason agreed, suggesting that traditional linear models might indeed miss non-linear relationships, which often results in findings that show non-significant or counterintuitive relationships between the built environment and travel behavior. He emphasized that revisiting prior findings with more flexible modeling assumptions could yield different results, potentially revealing a more practical and relevant connection between the built environment and travel behavior. Jason advocated for conducting more studies across various contexts using relaxed assumptions to achieve more reliable conclusions.

Q2: Jinhua raised the issue of data standardization in transportation research, asking if the field could benefit from open-source, standardized datasets similar to those in computer science. He added that this would allow for the validation of new methodologies and verification of results by other researchers.

A2: Jason agreed that transportation research would benefit significantly from improved data standardization and open accessibility. He pointed out that most transportation research relies on travel survey data, which typically includes similar variables and follows similar data collection procedures. Additionally, he noted that some researchers encourage the collection of attitudinal data and analysis using standardized formats, which would make results more comparable and suitable for meta-analysis. Such efforts could enhance comparability and reproducibility, supporting more rigorous and impactful research.

Q3: Jinhua questioned the role of hypotheses in this type of research, contrasting two approaches: one where researchers have a clear hypothesis or theory to test, and another where they use data-driven approaches to explore patterns without a prior hypothesis.

A3: Jason recommended integrating both theory-driven and data-driven approaches. For exploring non-linear relationships between the built environment and travel behavior, he indicated that he begins with a hypothesis but finds that traditional statistical tools may not fully address it. In such cases, he leverages machine learning as a tool to verify his hypothesis. He also acknowledged that when theoretical insights are limited, a data-driven approach can reveal patterns that might otherwise be missed, especially during exploratory phases.

Q4: Jinhua expressed concerns about the misuse of machine learning in transportation studies, noting instances where researchers apply machine learning models to data without a clear purpose, often resulting in meaningless findings.

A4: Jason emphasized that machine learning should be grounded in a clear theoretical basis to ensure meaningful results. He advised researchers to critically assess unexpected findings, examining their validity and relevance. Ideally, findings should be validated against theoretical expectations or tested on multiple datasets. This validation process ensures that machine learning insights are genuinely impactful rather than random noise. When machine learning outputs align with established theories or offer theoretically plausible insights, they become more relevant and useful for policy and planning.

Q5: Jinhua inquired about the future directions of research in the field of built environment and travel behavior.

A5: Jason emphasized the need to revisit the existing literature on the relationship between the built environment and travel behavior. He highlighted the importance of conducting studies in diverse contexts, relaxing linear model assumptions to verify past findings and achieve more reliable conclusions. Furthermore, he suggested exploring new data sources, such as mobile phone data and big data, and employing novel tools to deepen understanding in this field.

Audience Questions

Q1: What actionable advice would you give mayors and city planners based on your research?

A1: These studies highlight the range within which the built environment significantly affects travel behavior and where these effects diminish. By recognizing these thresholds, policymakers can make informed decisions to encourage sustainable development within effective ranges.

Q2: Are Americans effectively stuck with the car?

A2: Jason argued that while density alone may not reduce car dependency, directing growth toward existing developed areas, instead of sprawling outward, and enhancing transit service, accessibility, and proximity to activity centers could gradually reduce car use, although it won't be an immediate shift.

Q3: Have walkability or transit metrics been used to predict car ownership?

A3: This study didn't include walkability metrics. While transit metrics were considered, Jason noted that walkability likely has a greater impact in urban areas than in suburban ones, where walking is often limited to recreational use due to fewer nearby destinations.

Q4: Could geography and topography affect the study's results?

A4: Jason acknowledged this possibility, suggesting that research in diverse locations—coastal, mountainous, urban, and suburban—could help validate findings. He advocated for further studies to assess generalizability.

Q5: Is using only non-spatial variables in machine learning sufficient for understanding urban travel?

A5: Jason emphasized the importance of integrating geospatial data in models. While methods like spatial lag models and distance-based weighted matrices have been used, he noted that the field is evolving and that more advanced approaches incorporating geospatial data should improve model insights over time.

Q6: Can large language models help create a transferable urban model?

A6: Jason shared that while early experiments, such as PlanGPT, show potential, LLMs alone may not yield complete insights. He emphasized the importance of human interpretation and the political context in effectively applying model insights to urban planning.

Part IV. Summary of Memos.

Themes from Other Memos

1. *Dynamic and Temporal Models*: Adding time-sensitive elements to models to capture shifts in the urban environment, transit policies, and technology (e.g., autonomous vehicles) could lead to more accurate explanations of trends and outcomes.
2. *Probabilistic Programming*: By embracing probabilistic models, researchers could better address uncertainties in consumer behavior and operational planning, allowing for more adaptable and insightful planning tools.
3. *Explainability for Stakeholder Trust*: Clear, interpretable models are essential for building trust with planners and decision-makers, who need actionable insights rather than complex "black-box" predictions.
4. *Hybrid Approaches*: Combining machine learning with traditional inferential methods could yield more robust, theory-driven insights, enhancing both accuracy and interpretability.
5. *Diverse and Longitudinal Data*: Using varied, long-term data across different regions can improve model reliability and deepen understanding of built environment and travel behavior interactions.

These directions aim to make machine learning models more practical, trustworthy, and valuable in real-world urban planning.

My Reflection

The relationship between the built environment (BE) and travel behavior has long intrigued researchers and practitioners in transportation, urban planning, design, public health, and related fields. Over the past three decades, substantial literature has accumulated on this topic. While many studies suggest that BE factors influence travel behavior, the findings regarding the nature and magnitude of these effects remain mixed. One key reason is the reliance on traditional linear models, which may have underestimated the true impact of BE on travel behavior.

The relationship between BE and travel behavior may not be fully linear, and threshold effects—specific values of BE factors that trigger different travel outcomes—can influence the strength of these relationships. As such, linear models often provide inaccurate estimates and fail to capture the complexity of these effects. In the recent presentation, Dr. Jason Cao discussed these non-

linear and threshold effects, highlighting how features like activity centers influence driving distance and how BE factors affect travel-related emissions.

Methodologically, recent studies have employed machine learning (ML) techniques to better explore non-linearities and threshold effects. Unlike traditional models, which assume linearity and are vulnerable to multicollinearity, self-selection bias, and endogeneity issues, ML models help address these challenges. Explainable ML tools such as partial dependence plots (PDPs), accumulated local effects (ALE), and SHAP values provide deeper insights into the actual nature of BE-travel relationships.

However, while ML models offer flexibility and efficiency in identifying non-linear patterns, they also have limitations compared to traditional models. Specifically, ML methods often lack the capacity to provide clear impact magnitudes (e.g., coefficients) and significance levels. Therefore, integrating ML with traditional inferential models could yield more robust results—leveraging ML's ability to detect patterns without presuppositions alongside the inferential strengths of linear models in estimating effect sizes and statistical significance.

It is also important to note that ML models are highly sensitive to hyperparameter tuning, and the relationships or thresholds they uncover may not generalize well over time. Observed associations could become inconsistent, unstable, or unreliable outside the specific spatiotemporal context in which they were identified. A promising future direction could involve combining ML with traditional modeling approaches by incorporating non-linear effects directly into traditional methods and using data from more diverse places and time points (longitudinal data) to achieve more comprehensive insights.

For built environment data in the USA, one may visit this webpage: https://www.epa.gov/smartgrowth/smart-location-mapping

Part V. Other Information

N/A