Enriching Transportation Survey Datasets Using Big Data and Machine Learning

With an Application for Transferring Attitudinal Variables across Transport Surveys

F. Atiyya Shaw & Patricia L. Mokhtarian
Georgia Institute of Technology
Applied Urban Modelling 2020: Modelling the New Urban World
2 November 2020
Unprecedented changes occurring rapidly... across multiple dimensions

**Technologies**
- Uber, Lyft, car/bike-sharing, etc.
- Long-range electric vehicles
- Autonomous vehicles, drones
- Virtual reality “travel” experiences

**Societal shifts**
- Delayed or deferred marriage and childbearing
- Greater education
- Increasing ethnic diversity
- Shifting values

**Policy instruments**
- Denser, more diverse land uses
- Changing revenue base away from fuel taxes
- Yet-to-be-determined AV policies
Improving behavioral forecasting

Two approaches: functions or data

Behavioral decision-making

\[
\begin{align*}
\text{Mobility Patterns}_{\text{today}} &= f_{\text{today}}(x_{\text{today}}) \\
\text{Mobility Patterns}_{\text{future}} &= f_{\text{today}}(x'_{\text{future}})
\end{align*}
\]

functions:
- machine learning
- latent variable models
- advanced discrete choice models

data:
- passive data streams
- big data: mobile phone location data, etc.
- active data streams
- survey data
Obtaining good survey data is getting harder...

...and all evidence indicates this will continue

Respondents have gotten
-- busier
-- more “jaded”
-- more distracted

Longer surveys
→ lower response rates,
→ increased survey bias

Finding a balance in content
 breadth
 depth

Large-scale surveys like the National Household Travel Survey (NHTS)
-- socio-economic characteristics
-- observed travel behavior attributes
-- across a nationwide sample

Smaller studies
-- broader set of variables
-- smaller samples
-- limited geographic areas
Large-scale household travel surveys

How can we get rich variables like attitudes into them?

Option 1: ask some attitudinal q’s on the survey itself  
- easier said than done

Respondents have gotten
-- busier
-- more “jaded”
-- more distracted

Longer surveys
→ lower response rates,
→ increased survey bias
Large-scale household travel surveys

How can we get rich variables like attitudes into them?

Option 2: “transfer” the info from other surveys

Smaller studies
-- broader set of variables
-- smaller samples
-- limited geographic areas

National/regional surveys like NHTS
-- socio-economic characteristics
-- observed travel behavior attributes
-- sample drawn across a large area

Taking information from richer studies could be useful for lots of difficult-to-measure variable types (e.g. other psychometric variables)
Purpose:
- To understand the impacts of emerging technologies and trends on travel behavior in Georgia (GA)

Details:
- Conducted Fall 2017
- Invited sample: 15 MPO areas
- Current N ~ 3300

Contents:
- A: Attitudes and personality
- B: Technology usage
- C: Key aspects of lifestyle
- D: How you travel
- E: Evolving transportation services
- F: Desires for future travel
- G: Autonomous vehicles
- H: Sociodemographic traits
National Household Travel Survey (NHTS)

The national large-scale survey – the “recipient”

- **Purpose:**
  - To support travel demand modeling & long-range transportation planning across U.S.

- **Details:**
  - Repeated cross-sectional travel behavior survey
  - Georgia subsample used in this study
  - Wave: April 2016 to May 2017
  - *Original* N ~ 8632 (GA respondents)

- **Contents:**
  - *Household data module*
  - *Long distance module*
  - *Vehicle data module*
  - *Person level module*
  - *Person trips module*
  - *Person health module*
  - *Person drive module*
Expanding transportation survey datasets

Using common variables to transfer attitudes

\[
\text{Attitudes}_{\text{GDOT}} = f_{\text{GDOT}}(CV_{\text{GDOT}}, \text{aug}CV_{\text{GDOT}}) + \varepsilon_{\text{GDOT}} \\
\text{Attitudes}_{\text{Fused}} = f_{\text{GDOT}}(CV_{\text{NHTS}}, \text{aug}CV_{\text{NHTS}})
\]

GDOT Dataset
N ~ 2600

NHTS Dataset
N ~ 4500

Fused Dataset
N ~ 4500

CVs
attitudes

CVs
behaviors

CVs
behaviors
predicted attitudes

Common variables:

NHTS/GDOT
Reconciled
Common
Variables

Targeted
Marketing Data

Transit Data

Land Use Data

Native common variables

Augmented/external common variables
Expanding transportation survey datasets

Components of transfer process

- **Transfer variables**: variables of interest being transferred across datasets
- **Features**: inputs to the training algorithm that are used to model/predict the transfer variables
- **Training algorithms**: used to transfer the variables across datasets
Components of the transfer process

Transfer variables

- **Transfer variables**: variables of interest being transferred across datasets
- **Features**: inputs to the training algorithm that are used to model/predict the transfer variables
- **Training algorithms**: used to transfer the variables across datasets

---

Training algorithms / functions

- Dataset integration
- Pre-processing
- Dimension reduction
- Basis expansion
- Varied algorithms
- Hyperparameter tuning
- Training/test set sizes
- Performance metrics

Transfer / dependent variables

- Variable form/transformation
- Dimension reduction and/or latent variable identification methods
# Transfer variables

## Introducing the attitudinal variables for transfer

<table>
<thead>
<tr>
<th>Name</th>
<th>Example statement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lifestyle</strong></td>
<td></td>
</tr>
<tr>
<td>Tech Savvy</td>
<td>Learning how to use new technologies is often frustrating for me (-)</td>
</tr>
<tr>
<td>Work-oriented</td>
<td>... having fun is more important to me than working hard (-)</td>
</tr>
<tr>
<td>Pro-exercise</td>
<td>I am committed to exercising regularly</td>
</tr>
<tr>
<td>Materialistic</td>
<td>I would/do enjoy having a lot of luxury things</td>
</tr>
<tr>
<td>Family-oriented</td>
<td>Family/friends play a big role in how I schedule my time</td>
</tr>
<tr>
<td><strong>Land use</strong></td>
<td></td>
</tr>
<tr>
<td>Pro-suburban</td>
<td>I prefer to live in a spacious home, even if it’s farther away ...</td>
</tr>
<tr>
<td>Urbanite</td>
<td>I like ... having stores, ... mixed among the homes in my n’hood</td>
</tr>
<tr>
<td><strong>Travel</strong></td>
<td></td>
</tr>
<tr>
<td>Non Car Mode</td>
<td>I like the idea of walking [bicycling, PT] as a means of travel for me</td>
</tr>
<tr>
<td>Commute Benefit</td>
<td>My commute is a useful transition between home and work</td>
</tr>
<tr>
<td>Travel Liking</td>
<td>I generally enjoy the act of traveling itself</td>
</tr>
<tr>
<td>Car-owning</td>
<td>I definitely want to own a car</td>
</tr>
<tr>
<td><strong>Personality</strong></td>
<td></td>
</tr>
<tr>
<td>Polychronic</td>
<td>I prefer to do one thing at a time (-)</td>
</tr>
<tr>
<td>Wait Tolerant</td>
<td>Having to wait is an annoying waste of time (-)</td>
</tr>
<tr>
<td>Environmental</td>
<td>Cost or convenience takes priority over environmental impacts ... (-)</td>
</tr>
<tr>
<td>Sociable</td>
<td>I consider myself to be a sociable person</td>
</tr>
</tbody>
</table>
Components of the transfer process

Features

- Transfer variables: variables of interest being transferred across datasets
- Features: inputs to the training algorithm that are used to model/predict the transfer variables
- Training algorithms: used to transfer the variables across datasets

Diagram:

- Features / independent variables:
  - Dataset integration
  - Pre-processing
  - Dimension reduction
  - Basis expansion

- Training algorithms / functions:
  - Varied algorithms
  - Hyperparameter tuning
  - Training/test set sizes
  - Performance metrics

- Transfer / dependent variables:
  - Variable form/transformation
  - Dimension reduction and/or latent variable identification methods
Features

Two types of features available for use

Survey Data Streams
- GDOT Emerging Technologies Survey
  - N ~ 3000
- National Household Travel Survey
  - Georgia Subsample
  - N ~ 8000

External Data Streams
- Targeted Marketing Data
  - N ~ 6000
- All Transit Data
- Land Use Data
- Census
- ACS

Native common variables
- active data sources
  - Exist initially in both datasets
  - Tend to be SED variables
  - Often must be adjusted and recoded across sources

Augmented common variables
- passive/active data sources
  - Obtained from external active or passive datasets
  - Must be appended to both the donor and recipient datasets at either:
    - Household level
    - Individual level
    - Geographic level
Components of the transfer process

**Algorithms**

- **Transfer variables**: variables of interest being transferred across datasets
- **Features**: inputs to the training algorithm that are used to model/predict the transfer variables
- **Training algorithms**: used to transfer the variables across datasets

---

- Dataset integration
- Pre-processing
- Dimension reduction
- Basis expansion
- Varied algorithms
- Hyperparameter tuning
- Training/test set sizes
- Performance metrics
- Variable form/transformation
- Dimension reduction and/or latent variable identification methods
Algorithms

Possible variations within algorithm selection process

• **Algorithm selection**
  • Linear regression, random forest, support vector machine, elastic net regression, lasso regression, ridge regression, random forest, extreme gradient boosting

• **Algorithm tuning**
  • Training/test sample split: 80/20
  • Hyperparameter tuning using k-fold cross validation
  • Final metrics on test/hold-out sample

• **Algorithm performance**
  • Possible metrics: R-squared, correlations (between observed and predicted), mean squared error, misclassification error, etc.
How well are we transferring variables?

Internal validation

Comparison of Feature Subsets

Correlations between observed and predicted attitudes

- Work-oriented
- Tech-savvy
- Pro-exercise
- Materialistic
- Non-car alternatives
- Pro-car owning
- Commute benefit
- Travel-liking
- Pro-suburban
- Modern urbanite
- Pro-environmental
- Polychronic
- Sociable
- Waiting tolerant

- Overall: TM, SED, Land Use
- All TM variables
- SED variables
- EPA SLD + ALL Transit + 50% ACS

26
Are the transferred variables any good?

**External validation:**
How much are they helping in travel behavior models?

**GDOT Survey - Atlanta Region:**
Linear regression model lift due to predicted attitudinal variables

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Dependent variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>SED variables</td>
<td>Lift with predicted attitudinal variables</td>
</tr>
</tbody>
</table>

**NHTS - Atlanta Region:**
Linear regression model lift due to predicted attitudinal variables

<table>
<thead>
<tr>
<th>Explanatory variables</th>
<th>Dependent variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>SED variables</td>
<td>Lift with predicted attitudinal variables</td>
</tr>
</tbody>
</table>
What will we be modeling for external validation?

Atlanta region: Ridehailing frequency

**GDOT Survey**

4. Please indicate how often you typically use each of the following transportation services.

<table>
<thead>
<tr>
<th></th>
<th>Never used / No longer use</th>
<th>Less than once a month</th>
<th>1-3 times a month</th>
<th>1-2 times a week</th>
<th>3 or more times a week</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Carsharing</td>
<td>□ 1</td>
<td>□ 2</td>
<td>□ 3</td>
<td>□ 4</td>
<td>□ 5</td>
</tr>
<tr>
<td>b. On-demand ride service</td>
<td>□ 1</td>
<td>□ 2</td>
<td>□ 3</td>
<td>□ 4</td>
<td>□ 5</td>
</tr>
<tr>
<td>c. Shared on-demand ride service</td>
<td>□ 1</td>
<td>□ 2</td>
<td>□ 3</td>
<td>□ 4</td>
<td>□ 5</td>
</tr>
<tr>
<td>d. Traditional taxi service</td>
<td>□ 1</td>
<td>□ 2</td>
<td>□ 3</td>
<td>□ 4</td>
<td>□ 5</td>
</tr>
</tbody>
</table>

**NHTS**

**RIDESHARE**

Range: 0 - 99
Programmer Note: Asked if subject is at least 16 years of age

In the past 30 days, how many times [SHAVE YOU] purchased a ride with a smartphone rideshare app (e.g. Uber, Lyft, Sidecar)?

<table>
<thead>
<tr>
<th>WEB ATEXT</th>
<th>CATI ATEXT</th>
<th>AVALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENTER NUMBER</td>
<td>ENTER NUMBER</td>
<td></td>
</tr>
<tr>
<td>I don’t know</td>
<td>DON’T KNOW</td>
<td>-8</td>
</tr>
<tr>
<td>I prefer not to answer</td>
<td>REFUSED</td>
<td>-7</td>
</tr>
</tbody>
</table>
How much are the transferred variables helping?

External validation: linear regression, $R^2$ values

**GDOT Survey**

**NHTS**

SED variables | 0.228
How much are the transferred variables helping?

Ext. validation: prediction accuracies for choice models

**GDOT Survey**

<table>
<thead>
<tr>
<th>Model versions</th>
<th>Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCCM: SED + Predicted Atts</td>
<td>86.26%</td>
</tr>
<tr>
<td>LCCM: SED + Observed Atts</td>
<td>99.77%</td>
</tr>
<tr>
<td>Binary Logit: SED + Observed Atts</td>
<td>72.86%</td>
</tr>
<tr>
<td>Binary Logit: SED + Predicted Atts</td>
<td>73.56%</td>
</tr>
<tr>
<td>Binary Logit: SED</td>
<td>69.40%</td>
</tr>
</tbody>
</table>

**NHTS**

<table>
<thead>
<tr>
<th>Model versions</th>
<th>Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCCM: SED + Predicted Atts</td>
<td>97.92%</td>
</tr>
<tr>
<td>Binary Logit: SED + Predicted Atts</td>
<td>86.29%</td>
</tr>
<tr>
<td>Binary Logit: SED</td>
<td>85.62%</td>
</tr>
</tbody>
</table>

**LCCM:** Latent Class Choice Model
Latent class choice model of ridehailing adoption

A preliminary look at some insights!

GDOT model w/ observed attitudes: three latent classes

- Travel liking class
- Family-oriented & car-owning
- Urbanite & tech-savvy

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Travel-liking</th>
<th>Family-oriented</th>
<th>Urbanite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>--</td>
<td>(+)*</td>
<td>(+)**</td>
</tr>
<tr>
<td>Age</td>
<td>(-)***</td>
<td>(-)*</td>
<td>(-)**</td>
</tr>
<tr>
<td>Household income</td>
<td>(+)***</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Household size</td>
<td>--</td>
<td>(-)*</td>
<td>(-)**</td>
</tr>
</tbody>
</table>

Benefits:
- Improved predictive accuracy
- Nuanced interpretation
- Slight increase in model fit

Will the predicted attitudes yield similar benefits for GDOT and NHTS?
# Study overview

## The process in a nutshell...

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1.</strong> DATA ACQUISITION</td>
<td></td>
</tr>
</tbody>
</table>
  - Acquire *source* domain data
  - Acquire *target* domain data
  - Acquire *targeted marketing data* (TMD) for source and target domain |
| **2.** ESTABLISH COMMON VARIABLE SPACE | 
  - Determine *common variables* original to source and target domains
  - Determine *TMD augmented* common variables for source and target domains
  - Determine *land use augmented* common variables for source and target domains |
| **3.** DATA PROCESSING | 
  - Prepare the following datasets for analysis:
    - Source domain data
    - Target domain data
    - TMD data
    - Land-use data |
| **4.** MACHINE LEARNING ALGORITHMS | 
  - Determine ML algorithms for testing
  - Train, optimize, and cross validate ML algorithms
  - Evaluate ML algorithms relative to each other and benchmark measures |
| **5.** EXTERNAL VALIDATION | 
  - Select dependent variable(s) that are possible candidates for external validation
  - Evaluate model improvement due to imputed attitudinal variables |
Study overview

Takeaways

• We can **impute attitudes** reasonably well!
  • ...and we’re getting better

• **Attitudes improve:**
  • Model fit
  • Model interpretability
  • Predictive accuracy (small increases)

• And **interpretations improve:**
  • Latent class choice models of travel behavior in demand forecasting models!
    • Policy implications of more nuanced segments
Study overview

Looking to the future

• In the **short-term**: let’s use what we have!
  • Use trained algorithm to predict attitudes into Atlanta Household Travel Survey
  • How does it affect the final outcomes?
    • Is it moving the outcome in the same direction as the adjustment factors?

• In the **medium-term**: let’s get better at predicting attitudes
  • Continued refinement of machine learning models
  • Additional common variables purchased and integrated into attitudinal prediction functions

• In the **long-term**: let’s talk attitudinal variables!
  • Attitudinal marker statements that could be included on future household travel surveys
    • Can help us obtain improved predictions of attitudes
    • Can be used directly to improve our travel demand models
    • Will not be highly correlated with other explanatory variables
Thank you!

Atiyya Shaw
atiyya@gatech.edu
Patricia Mokhtarian
patmokh@gatech.edu