نردبان بازتولیدپذیری با برآورد دوبارهٔ طراحی‌آگاه برای پژوهش‌های سفر مبتنی بر پیمایش: مطالعهٔ موردی بیست‌ودو مطالعه از پیمایش ملی سفر خانوار آمریکا

سید صابر ناصرعلوی
گروه مهندسی عمران، دانشگاه شهید باهنر کرمان، کرمان ۷۶۱۶۹۱۴۱۱۱، ایران
saber_naseralavi@uk.ac.ir

اکرم مظاهری
گروه مهندسی عمران و محیط زیست، دانشگاه تربیت مدرس، تهران ۱۴۱۱۹۴۴۹۶۱، ایران

چکیده

برنامه‌ریزان حمل‌ونقل از ضرایب رگرسیون و امتیازهای پیش‌بینی پیمایش‌های ملی سفر برای ارزیابی اثر سواری‌خواهی اینترنتی، حمل‌ونقل عمومی یا خودروهای برقی بر الگوی سفر بهره می‌گیرند. این عددها به انتخاب‌هایی بستگی دارند که مقاله‌ها کمتر کامل شرح می‌دهند: کدام رکوردها نگه داشته شد، آیا وزن‌های پیمایش به کار رفت، عدم‌قطعیت چگونه محاسبه شد و مدل پیش‌بینی چگونه آزموده شد. ما نردبانی چهارپله‌ای برای بازتولیدپذیری پیشنهاد می‌کنیم (بازسازی نمونه، بازسازی جدول‌های توصیفی، برازش مدل منتشرشده، و برازش آن با طرح نمونه‌گیری خود پیمایش) و آن را بر ۲۲ مطالعهٔ منتشرشده از پیمایش ملی سفر خانوار آمریکا (NHTS) به کار می‌بریم؛ فایل‌های عمومی اجازهٔ برازش دوبارهٔ مدل ۱۳ مطالعه را داد. نمونهٔ ۹ مطالعه خوب بازسازی شد و ۸۱ درصد از ۱٬۴۵۴ مقدار توصیفی چاپ‌شده بازتولید شد. هر جا برازش وزنی و بدون وزن عددهای متفاوتی بازتولید می‌کنند، برآوردگر اعلام‌شدهٔ مقاله مقدارهای چاپی‌اش را بازتولید کرد: در یک مطالعه نسبت‌های بخت وزنی ۵۷ درصد از مقدارهای چاپی را بازتولید کرد و نسبت‌های بدون وزن ۳ درصد. با وزن‌های تکرار همراه پیمایش، خطای معیار به‌طور میانه ۲٫۲ برابر خطای معیار برازش بدون وزن بود. این هزینه تقریباً یکسره از نابرابری وزن‌ها می‌آمد (۲٫۲ برابر)؛ همخوشگی اشخاص در خانوار (۱٫۰۶ برابر) و فرمول وزن تکرار (۰٫۹۹ برابر) به‌طور میانگین کم افزودند و خطای معیار تحلیل بدون وزن تنها ۱٫۱۰ برابر کمتر از واقع بود. زیر استنباط اعلام‌شدهٔ هر مطالعه، ۲۴ درصد از ضرایب معنادار (بازهٔ دقیق ۱۸ تا ۳۱ درصد) زیر طرح معنادار نبودند و ۸ درصد ضرایب میان برازش بدون وزن و وزنی علامتشان عوض شد. برای مدل‌های طبقه‌بندی، بیش‌نمونه‌گیری از رستهٔ کمیاب پیش از تقسیم داده، سطح زیر منحنی مشخصهٔ عملکرد گیرنده (ROC)، امتیازی که شانس را پنجاه درصد و کمال را صد درصد می‌داند، را برای یک جنگل تصادفی از ۶۳ درصد به ۹۹ درصد رساند؛ مقدار چاپی با هیچ‌یک از دو خط لوله بازتولید نشد. امتیاز رگرسیونیِ منتشرشده (سهم واریانس توضیح‌داده‌شده) برابر ۷۱ درصد بازتولید نشد (۲۰ درصد روی خانوارهای کنار گذاشته‌شده). این ابزار به‌صورت دو ماژول از بستهٔ متن‌باز پایتون pynhts منتشر شده است.

واژه‌های کلیدی: پیمایش ملی سفر خانوار؛ بازتولیدپذیری؛ وزن‌های پیمایش؛ واریانس با وزن‌های تکرار؛ نشت داده؛ رگرسیون دوجمله‌ای منفی

مقدمه

پیمایش ملی سفر خانوار آمریکا (NHTS) مهم‌ترین منبع عمومی دربارهٔ چگونگی سفر آمریکایی‌هاست. موج‌های ۲۰۰۱، ۲۰۰۹، ۲۰۱۷ و ۲۰۲۲ آن هر کدام میان حدود هشت هزار (۲۰۲۲) تا صد و پنجاه هزار خانوار را دربرمی‌گیرند و برای هر عضو، سفرهای یک روز تعیین‌شده، خودروها و فهرست بلندی از ویژگی‌های خانوار و فرد ثبت شده است. پژوهشگران از

این فایل‌ها برای برآورد این‌که چه کسی سوار می‌شود، چه کسی هزینه می‌دهد و چه کسی آلایندگی ایجاد می‌کند استفاده می‌کنند و برنامه‌ریزان ضرایب رگرسیون، نسبت‌های بخت و امتیازهای پیش‌بینی حاصل را شاهدی دربارهٔ سواری‌خواهی اینترنتی، دوچرخهٔ اشتراکی، حمل‌ونقل عمومی، خودروهای برقی و خودروکیلومتر پیموده می‌خوانند (Alberini و همکاران, 2021, Conway و همکاران, 2018, Wu و MacKenzie, 2021). جمله‌ای مانند «هر بار استفادهٔ بیشتر از سواری‌خواهی در ماه با سفرهای روزانهٔ بیشتر همراه است» جمله‌ای کوتاه است که بر زنجیره‌ای بلند از تصمیم‌ها تکیه دارد: کدام پاسخ‌گویان وارد تحلیل شدند، آیا وزن‌های پیمایش اعمال شد، خطای معیار چگونه ساخته شد و، برای مدل پیش‌بینی، چگونه روی داده‌ای که ندیده بود آزموده شد.

این تصمیم‌ها از آن رو اهمیت دارند که NHTS نمونهٔ تصادفی ساده نیست. خانوارها در لایه‌ها انتخاب می‌شوند، اشخاص درون خانوار خوشه‌ای‌اند، نرخ پاسخ در گروه‌ها فرق دارد، و به همین دلیل فایل‌ها وزن فرد و خانوار و، برای موج‌های ۲۰۰۱، ۲۰۰۹ و ۲۰۱۷، میان نود و هشت تا صد وزن تکرار برای هر رکورد دارند تا تحلیل‌گر بتواند برآوردی از واریانس که طرح نمونه‌گیری را رعایت کند به دست آورد (Rust و Rao, 1996, Wolter, 2007). این‌که رگرسیون وزن‌دار شود یا نه پرسشی دیرینه در آمار پیمایشی است: وزن‌ها وقتی انتخاب به پیامد وابسته باشد می‌توانند اریبی را بردارند، اما وقتی وابسته نباشد کارایی را کاهش می‌دهند (Bollen و همکاران, 2016, Solon ,2007, Gelman و همکاران, 2015, Winship و Radbill, 1994). این‌که هنگام محاسبهٔ خطای معیار از طرح نمونه‌گیری استفاده شود کمتر محل بحث است: نادیده‌گرفتن خوشه‌بندی و لایه‌بندی عدم‌قطعیت را کمتر از واقع نشان می‌دهد (Lumley, 2004). نمونه‌هایی از پیمایش‌های سلامت نشان می‌دهند ارتباطی که در تحلیل بدون وزن معنادار است ممکن است با رعایت وزن و طرح معنادار نباشد (Birrell و همکاران, 2019, Gard و همکاران, 2023). رگرسیون روی داده‌های NHTS امری عادی است: گردایهٔ رسمی کاربردهای NHTS از FHWA/ ORNL از ۲۰۰۶ تا ۲۰۲۵ شمار ۴٬۳۰۱ اثر متمایز را فهرست می‌کند و چکیدهٔ ۱٬۰۳۹ اثر از آن‌ها تحلیل رگرسیونی یا یادگیری ماشین را توصیف می‌کند (۳۴۲ مورد از این‌ها در چکیده نام NHTS را می‌برند؛ قواعد واژه‌ای، مواد تکمیلی). این‌که انتخاب‌های بالا تا چه اندازه نتیجهٔ چنین تحلیل‌هایی را عوض می‌کند هنوز اندازه‌گیری نشده است.

نگرانی دوم به بازتولیدپذیری در معنای وسیع‌تر برمی‌گردد. بازآفرینی نتیجه‌های محاسباتی از روی شرح منتشرشده اغلب دشوار است (Hardwicke و همکاران, 2021, Peng, 2011, Stodden و همکاران, 2018)، و وقتی تحلیل‌گران مستقل داده و پرسش یکسانی دریافت می‌کنند، انتخاب‌های موجه اما متفاوت می‌تواند به نتیجه‌های متفاوت و گاه در جهت‌های مخالف بینجامد (Schweinsberg و همکاران, 2021, Silberzahn و همکاران, 2018, Simonsohn و همکاران, 2020, Steegen و همکاران, 2016). مدل‌های پیش‌بینی لایه‌ای دیگر می‌افزایند. یادگیری ماشین با این ادعا که مدل‌های انعطاف‌پذیر از مدل‌های کلاسیک بهترند وارد پژوهش رفتار سفر شده است (Hagenauer و Helbich, 2017, Martín-Baos و همکاران, 2023, Wang و همکاران, 2024)، در حالی که ادبیات گسترده‌تر نشان می‌دهد نشت میان داده‌های آموزش و آزمون در بسیاری از رشته‌ها امتیازهای خوش‌بینانه می‌سازد (Kapoor و همکاران, 2024, Kapoor و Narayanan, 2023, Rosenblatt و همکاران, 2024). دو نشت اینجا ویژه اهمیت دارد. ساختن نمونه‌های مصنوعی از رستهٔ کمیاب پیش از تقسیم داده، نسخه‌هایی از ردیف‌های آزمون را وارد آموزش می‌کند و امتیازها را به‌شدت بالا می‌برد (Blagus و Lusa, 2013, Chawla و همکاران, 2002, Demircioğlu, 2024). تقسیم تصادفی ردیف‌ها وقتی ردیف‌ها خانوار، مکان یا فرد مشترک دارند، شبه‌نسخه‌ها را در دو سوی تقسیم قرار می‌دهد و طرح‌های اعتبارسنجی که این ساختار را رعایت می‌کنند امتیازهای پایین‌تر و صادقانه‌تر می‌دهند (Ploton و همکاران, 2020, Roberts و همکاران, 2017, Vabalas و همکاران, 2019). ردیف‌های NHTS دقیقاً از این نوع‌اند: اشخاص یک خانوار تقریباً همه چیز مهم را با هم دارند.

کارهای دربارهٔ خود NHTS بیشتر کیفیت داده را بررسی کرده‌اند، برای نمونه اریبی انتخاب میان پاسخ‌گویانی که پذیرفتند در پیمایش‌های پیگیری شرکت کنند (Wang و همکاران, 2022)، یا از آن برای پاسخ به پرسش‌های محتوایی بهره برده‌اند (Conway و همکاران, 2018). خط لوله‌های باز برای تقاضای سفر مصنوعی از پیمایش‌های ملی دیگر هم وجود دارد (Hörl و Balac, 2021). در جست‌وجوهای OpenAlex و Crossref، یک آزمون تکراری‌بودن موضوع، یک مرور ScienceDirect LeapSpace و یک ممیزی واژه‌ای گردایهٔ رسمی (همگی در مواد تکمیلی شرح داده شده‌اند) مطالعه‌ای

نیافتیم که مجموعه‌ای از تحلیل‌های منتشرشدهٔ NHTS را بردارد، نمونه‌ها و جدول‌هایشان را بازسازی کند، مدل‌هایشان را دوباره برازش دهد و بپرسد آیا نتیجه‌هایشان از طرح نمونه‌گیری خود پیمایش و از اعتبارسنجی بدون نشت جان به‌در می‌برد. پیشینهٔ پژوهش در ادامه شواهد این گزاره را می‌دهد. این شکافی دوگانه است: اندازهٔ اثر نادیده‌گرفتن طرح در رگرسیون‌های حمل‌ونقل ناشناخته است و میزان وابستگی امتیازهای پیش‌بینیِ منتشرشده به خط لولهٔ اعتبارسنجی نیز ناشناخته است.

این مقاله شکاف را با یک روش و یک مطالعهٔ موردی پر می‌کند. روش، نردبانی چهارپله‌ای برای بازتولیدپذیری است که هر پله‌اش با عددهای چاپی درجه‌بندی می‌شود، همراه با برآورد دوبارهٔ طراحی‌آگاه که سه شیوهٔ برازش یک مدل را مقایسه می‌کند و امتیازدهی دوبارهٔ طبقه‌بندها در چهار خط لوله. مطالعهٔ موردی آن را بر ۲۲ مطالعهٔ منتشرشده از NHTS، که به‌صورت مجموعه‌ای در دسترس گردآوری شده، به کار می‌برد و فایل‌های عمومی اجازهٔ برازش دوبارهٔ ۱۳ مطالعه را داد. چهار پرسش داریم: چه مقدار از هر مطالعه را می‌توان بازسازی کرد؛ آیا عددهای چاپی با برآوردگرِ اعلام‌شدهٔ مقاله بازتولید می‌شوند؛ نتیجه‌ها زیر طرح نمونه‌گیری چقدر تغییر می‌کنند؛ و امتیاز پیش‌بینی تا چه اندازه به خط لوله وابسته است. همین ابزار به‌صورت دو ماژول از بستهٔ متن‌باز pynhts منتشر شده تا بررسی به‌جای چند هفته در چند خط کد انجام شود.

پیشینهٔ پژوهش

طرح پیمایش، وزن‌ها و واریانس در پژوهش‌های NHTS

NHTS یکی از خانوادهٔ پیمایش‌های ملی سفر است که طرحشان از کشوری به کشور دیگر فرق می‌کند و میان موج‌ها هم تغییر می‌کند (McGuckin, 2021, Porschen و Kuhnimhof, 2026, Svaboe و همکاران, 2024). در مجموعهٔ NHTS موج ۲۰۰۱ خانوارها را از چارچوب تلفنی و موج ۲۰۱۷ از چارچوب نشانی‌مبنا برگزید؛ فایل‌های عمومی موج‌های ۲۰۰۱، ۲۰۰۹ و ۲۰۱۷ وزن تکرار دارند، در حالی که فایل‌های موج ۲۰۲۲ لایه و خوشه برای خطی‌سازی می‌دهند. اشخاص، خودروها و سفرها جدول‌هایی تودرتو در خانوارند، پس هر تحلیل فرد یا سفر در واقع تحلیل رکوردهای خوشه‌ای با وزن‌های نابرابر است.

تصویرهای توصیفی ملی از سفر شهری، روستایی، سالمندان و صندلی عقب به‌طور معمول وزن‌های نهایی را اعمال می‌کنند (Collia و همکاران, 2003, Krumm, 2012, Pucher و Renne, 2005, Trowbridge و Kent, 2009, Wang و Renne, 2023). وزن‌ها بازنمایی را بهتر می‌کنند اما پوشش ازدست‌رفته را جبران نمی‌کنند: پژوهش‌هایی دربارهٔ این‌که پیمایش‌های سفر خانوار تا چه اندازه سفرکنندگان آسیب‌پذیر، افراد دارای معلولیت، پاسخ‌گویانی که هیچ سفری گزارش نمی‌کنند و گروه‌هایی را که پیمایش‌های پیگیری از دست می‌دهند ثبت می‌کنند نشان می‌دهد کسان و سفرهایی که وزن‌ها باید جایگزینشان شوند دشوارترین موارد برای مشاهده‌اند (Christensen و همکاران, 2025, Esekhaigbe و همکاران, 2026, Shaw و همکاران, 2021, Wang و همکاران, 2022, Zhang و Maruyama, 2026). بنابراین وزن‌دادن برای گزارهٔ ملی لازم است اما برای درستی آن کافی نیست.

وقتی تحلیل از توصیف به رگرسیون یا مدل انتخاب می‌رسد، رفتار با طرح نمونه‌گیری کمتر دیده می‌شود. مدل‌های انتخاب وزنیِ سواری‌خواهی (Zou و Cirillo, 2021) استثنا هستند؛ رگرسیون‌های دوچرخه‌سواری، سفر مهاجران، سفر سالمندان و خرید اینترنتی روی NHTS (Barajas, 2021, Kim و Ulfarsson, 2021, Porter و همکاران, 2020, Tribby و Tharp, 2019, Tumula و همکاران, 2026, Yang, 2025) رایج‌اند، و از هشت مطالعهٔ رگرسیونیِ فهرست خودمان که توانستیم دوباره برآورد کنیم تنها یکی برای واریانس از وزن‌های تکرار استفاده کرده بود (جدول 5). برای دیدن این‌که طرح تا چه اندازه در متن چاپی دیده می‌شود، چکیدهٔ ورودی‌های گردایهٔ رسمی کاربردهای NHTS از FHWA/ORNL را برچسب‌گذاری کردیم؛ دوازده ویرایش چاپی و برون‌داد جست‌وجوی تعاملی آن از ۲۰۰۶ تا ۲۰۲۵ ۴٬۶۱۹ ورودی دارد که ۴٬۳۰۱ مورد متمایزند. از ۱٬۰۳۹ چکیده‌ای که تحلیل رگرسیونی یا یادگیری ماشین را توصیف می‌کنند، ۳۱ مورد (۳٫۰ درصد) وزن پیمایش و ۴ مورد (۰٫۴ درصد) برآوردگر واریانس مبتنی بر طرح را نام می‌برند؛ یعنی وزن ۷٫۸ برابر بیشتر از برآوردگر واریانس نام برده می‌شود و ۹۷ درصد هیچ‌یک را نام نمی‌برند. در میان ۳۴۲ چکیده‌ای که نام NHTS را می‌برند، شمارها ۱۷ و ۴ است. خواندن همهٔ موردهای مثبت نشان داد ۲۹ مورد از ۳۱ برچسب وزن و ۲ مورد از ۴ برچسب واریانس به خود پیمایش مربوط است. چکیده راهنمای خوبی برای آنچه تحلیل کرده است نیست؛ این شمارها نشان می‌دهند طرح چقدر کم نام برده می‌شود، نه

چقدر کم به کار رفته است، و دلیل بازسازی تحلیل‌ها همین است. کار طراحی‌آگاه بیشتر در برآورد ناحیهٔ کوچک است که طرح موضوع مقاله است (Vaish و همکاران, 2010) و در شمار اندکی مدل انتخاب وزنی.

زمینهٔ آماری قدیمی است. واریانس یک برآورد زیر طرح پیچیده با خطی‌سازی تیلور (Woodruff, 1971) یا با تکرار (Rust و Rao, 1996, Wolter, 2007) به دست می‌آید، کارایی گزینه‌ها و هزینهٔ نادیده‌گرفتن طرح دهه‌هاست مطالعه می‌شود (Cohen و همکاران, 1986, Johnson و Elliott, 1998, Korn و Graubard, 1999)، و نرم‌افزار آن در R (Lumley, 2004) و بسته‌های تجاری هست که تفاوت‌ها و سوءاستفاده‌هایشان نیز مستند شده است (Bergtold و همکاران, 2017, Brogan, 2005 ,de Sousa و da Silva, 2000, Richardson و همکاران, 2019). تا آنجا که می‌دانیم پشتهٔ علمی پایتون پیاده‌سازی آمادهٔ واریانس وزن تکرار برای مدل‌های شمارشی ندارد و به همین دلیل موتور به‌کاررفته در این مقاله در pynhts منتشر شده است.

بازتولیدپذیری، انعطاف تحلیلی و نشت اعتبارسنجی

فراتر از پیمایش، علوم محاسباتی گسترده‌تر ثبت کرده‌اند که چه اندازه نتیجه‌های منتشرشده از روی شرح منتشرشده بازآفرینی‌پذیر نیست (Hardwicke و همکاران, 2021, Peng, 2011, Stodden و همکاران, 2018) و نتیجهٔ تحلیل‌گران مستقل از یک داده چگونه به انتخاب‌های تحلیلیِ موجه بستگی دارد (Schweinsberg و همکاران, 2021, Silberzahn و همکاران, 2018, Simonsohn و همکاران, 2020, Steegen و همکاران, 2016). پژوهش سفر هم در معرض همین پراکندگی است: مقایسه میان پیمایش‌های ملی یا میان شیوه‌های پیمایش بر تصمیم‌های پردازشی تکیه دارد که کم گزارش می‌شوند (Nishikawa و Tawfik, 2022, Svaboe و همکاران, 2024, Verreault و Morency, 2016).

مدل‌های پیش‌بینی لایه‌ای خاص خود را می‌افزایند. محک‌های یادگیری ماشین در برابر مدل‌های انتخاب گسسته، و مدل‌های سفر یادگیری‌ماشینی برازش‌شده بر NHTS، گزارش می‌کنند مدل‌های انعطاف‌پذیر می‌توانند با مدل‌های کلاسیک برابری کنند یا از آن‌ها بهتر باشند (Hagenauer و Helbich, 2017, Martín-Baos و همکاران, 2023, Qawasmeh و همکاران, 2024, Wang و همکاران, 2024)، در حالی که ممیزی‌های رشته‌های دیگر بسیاری از امتیازهای خوش‌بینانه را به نشت میان داده‌های آموزش و آزمون برمی‌گردانند (Kapoor و همکاران, 2024, Kapoor و Narayanan, 2023, Rosenblatt و همکاران, 2024). دو نشت به تفصیل شناخته‌شده است: بیش‌نمونه‌گیری رستهٔ کمیاب پیش از تقسیم (Blagus و Lusa, 2013, Chawla و همکاران, 2002, Demircioğlu, 2024) و تقسیم تصادفی ردیف‌هایی که گروه مشترک دارند (Ploton و همکاران, 2020, Roberts و همکاران, 2017, Vabalas و همکاران, 2019). هر دو در داده‌های سفر خانوار محتمل‌اند، چون پیامد کمیابی مانند مالکیت خودروی برقی مادهٔ همیشگی بیش‌نمونه‌گیری است و اعضای خانوار ردیف‌هایی با ورودی‌های تقریباً یکسان‌اند. در گردایه ۲۱۵ چکیده مطالعه‌های یادگیری ماشین را توصیف می‌کنند که ۱۰۵ مورد از آن‌ها در ویرایش‌های ۲۰۲۳ تا ۲۰۲۵ آمده است؛ ۹ مورد از بیش‌نمونه‌گیری، SMOTE یا ناترازی رده‌ها یاد می‌کنند و ۱۱ مورد دقت یا سطح زیر منحنی دست‌کم نود درصد را بیان می‌کنند.

شکاف و سهم مقاله

بنابراین برای NHTS سه پرسش باز است. اثر نادیده‌گرفتن وزن و طرح بر رگرسیون‌های منتشرشده چه اندازه است؟ کدام لایهٔ طرح آن را می‌سازد: وزن‌های نابرابر، خوشه‌بندی در خانوار یا فرمول تکرار؟ و امتیازهای پیش‌بینیِ منتشرشده تا چه اندازه به خط لولهٔ اعتبارسنجی وابسته‌اند؟ مطالعه‌ای نیافتیم که با بازسازی مجموعه‌ای از تحلیل‌های منتشرشدهٔ NHTS به آن‌ها پاسخ دهد. سهم این مقاله روشی درجه‌بندی‌شده برای این کار، تجزیهٔ تفاوت خطای معیار به لایه‌هایش، امتیازدهی دوبارهٔ بدون نشتِ طبقه‌بندها و ابزارهای هر سه است.

مواد و روش‌ها

مجموعهٔ مطالعه‌های موردی و داده

مجموعه از ۲۲ مطالعهٔ منتشرشده میان سال‌های ۲۰۱۱ تا ۲۰۲۵ تشکیل شده که NHTS را تحلیل می‌کنند و عدد کافی برای مقایسه چاپ کرده‌اند (جدول 1). این مطالعه‌ها (Alberini و همکاران, 2021, Barua و همکاران, 2021, Deka و Fei, 2019, Fu و همکاران, 2020, Jia, 2019, Jia و همکاران, 2020, Jiao و همکاران, 2020, Krizek و McGuckin, 2019, Lidbe و همکاران, 2021, Naseralavi و همکاران, 2025, Quinn و همکاران, 2016, Sabouri و همکاران, 2020, Sadeghvaziri و Tawfik, 2020, Shirgaokar و Nobler, 2021, Su و همکاران, 2021, Wu و MacKenzie, 2021, Xue و همکاران, 2021, Yang و همکاران, 2011; Zhang و Zhou; Zhang, 2018a, 2018b, 2018c و Wang, 2014) هنگام ساختن بستهٔ pynhts گردآوری شدند، بنابراین مجموعه‌ای در دسترس است و نمونهٔ سامانمند از ادبیات نیست؛ دربارهٔ سهم همهٔ مقاله‌های NHTS که چنین رفتاری دارند نتیجه‌ای نمی‌گیریم. با جست‌وجوی عنوان در گردایهٔ رسمی کاربردهای NHTS، ۱۹ مورد از ۲۲ مطالعه در آن فهرست شده‌اند، در ۶ فصل از فصل‌های برنامه‌ریزی آن (جدول 2)؛ این فصل‌ها ۵۱ درصد مطالعه‌های گردایه را دربر دارند و ۳ مطالعه‌ای که فهرست نشده‌اند نشان می‌دهد خود گردایه راهنمایی ناقص است. یکی از مطالعه‌ها را ما دو نفر با همکارانمان نوشته‌ایم و نویسندهٔ اول توسعه‌دهندهٔ بسته نیز هست؛ آن مطالعه در فهرست آمده اما برای پرهیز از درجه‌بندی کار خودمان دوباره برآورد نشد و مدل‌های آن برای مقاله‌ای جدا نگه داشته شده است. فایل‌های عمومی NHTS برای موج‌های ۲۰۰۱، ۲۰۰۹ و ۲۰۱۷ از پایگاه پیمایش و از راه pynhts دریافت شد که هر فایل را با کدنامه‌اش راستی‌آزمایی می‌کند. هر ۲۲ مطالعه از موج ۲۰۰۱، ۲۰۰۹ یا ۲۰۱۷ استفاده می‌کنند.

جدول 1: بیست‌ودو مطالعهٔ موردی و پله‌ای که هر یک به آن رسید. L۱ شمار عددهای توصیفی چاپی درجه‌بندی‌شده و در پرانتز سهم عددهای EXACT یا CLOSE را می‌دهد.

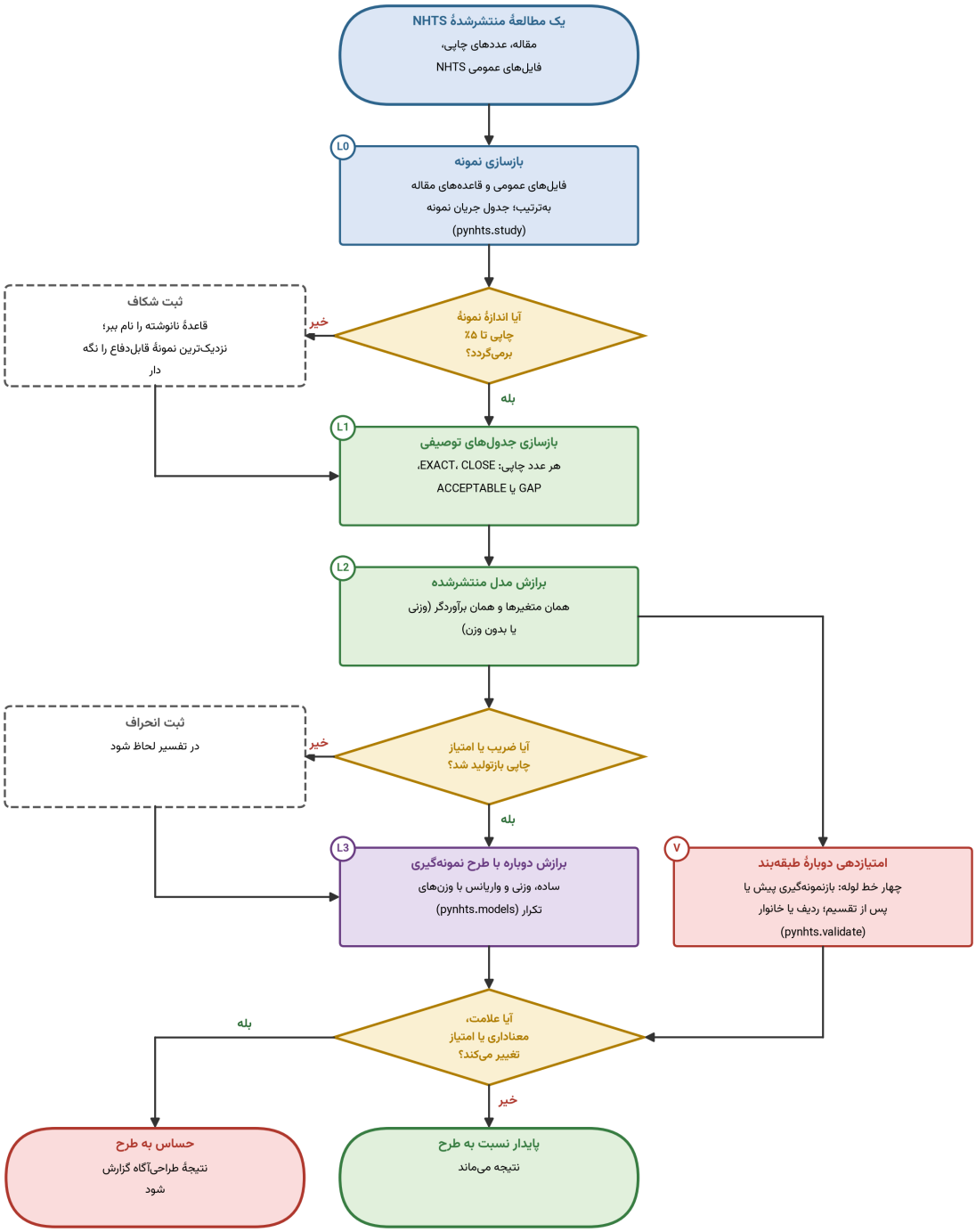
شناسه	مطالعه	موج	مدل	نمونه (L0)	عددهای چاپی منطبق (L1)	مدل (L2)
P01	Yang ۲۰۱۱	۲۰۰۱	logit	کامل	۱۵۴ (۹۵٪)	بازتولید شد
P02	Zhou ۲۰۱۴	۲۰۰۹	SEM	جزئی	۲۳ (۴۸٪)	بازتولید نشد
P03	Quinn ۲۰۱۶	۲۰۰۹	logit	جزئی	۲۱ (۴۸٪)	جزئی
P05	Deka ۲۰۱۹	۲۰۱۷	ZINB	جزئی	۸۶ (۷۰٪)	اجرا نشد
P06	Jia ۲۰۱۹	۲۰۱۷	ML classifiers	کامل	۴۲ (۸۶٪)	بازتولید نشد
P07	Krizek ۲۰۱۹	۲۰۱۷	descriptive	جزئی	۱۸ (۲۲٪)	اجرا نشد
P09	Fu ۲۰۲۰	۲۰۱۷	simulation	ناکافی	۱۳ (۸٪)	اجرا نشد
P10	Lidbe ۲۰۲۱	۲۰۰۱/۲۰۱۷	descriptive	کامل	۳۳۶ (۹۰٪)	اجرا نشد
P11	Alberini ۲۰۲۱	۲۰۱۷	OLS	کامل	۵۰ (۹۴٪)	جزئی
P12	Xue ۲۰۲۱	۲۰۱۷	SEM	ناکافی	۲۰ (۳۵٪)	اجرا نشد
P13	Zhang ۲۰۱۸a	۲۰۱۷	ZINB	جزئی	۸۰ (۶۹٪)	بدون هدف معتبر
P14	Zhang ۲۰۱۸b	۲۰۱۷	ZINB	کامل	۹۷ (۹۷٪)	جزئی
P15	Zhang ۲۰۱۸c	۲۰۱۷	ZINB	کامل	۸۵ (۹۶٪)	بدون هدف معتبر
P16	Sabouri ۲۰۲۰	۲۰۱۷	multilevel Poisson	جزئی	۳۰ (۷۰٪)	اجرا نشد
P17	Jia ۲۰۲۰	۲۰۱۷	ML classifiers	کامل	۴۰ (۸۵٪)	بازتولید نشد
P18	Jiao ۲۰۲۰	۲۰۱۷	NB	کامل	۵۵ (۸۷٪)	بازتولید شد
P19	Sadeghvaziri ۲۰۲۰	۲۰۰۹/۲۰۱۷	descriptive	جزئی	۱۱ (۸۲٪)	اجرا نشد
P20	Shirgaokar ۲۰۲۱	۲۰۱۷	NB	جزئی	۶۳ (۴۱٪)	جزئی
P21	Barua ۲۰۲۱	۲۰۰۹/۲۰۱۷	GBM	جزئی	۴۰ (۴۸٪)	بازتولید نشد
P22	Wu ۲۰۲۱	۲۰۱۷	PSM	کامل	۷۱ (۹۷٪)	بازتولید نشد
P23	Su ۲۰۲۱	۲۰۱۷	motif clustering	جزئی	۶۵ (۶۵٪)	اجرا نشد
P24	Naseralavi ۲۰۲۵	۲۰۱۷	ML per state	جزئی	۵۴ (۹۱٪)	اجرا نشد

جدول 2: مطالعه‌های موردی در فصل‌های گردایهٔ رسمی کاربردهای NHTS (فصل نخستین ظهور). ستون دوم همهٔ مطالعه‌های متمایز گردایه در آن فصل را می‌شمارد.

فصل گردایه	مطالعه‌ها در فصل	مطالعه‌های موردی	شمار
پیمایش، ترکیب داده و کاربردهای دیگر	۴۷۶	P24 P09’	۲
رفتار سفر	۴۷۰	P18’ P16’ P12’ P11’ P07’ P22 P21’	۷
گروه‌های ویژهٔ جمعیتی	۳۹۴	P23 P20’ P19’ P10’	۴
مطالعه‌های دوچرخه و پیاده	۳۴۴	P03 P01’	۲
تحلیل روند و بخش‌بندی بازار	۳۳۴	P17 P02’	۲
برنامه‌ریزی حمل‌ونقل همگانی	۱۷۰	P13 P05’	۲
در گردایه فهرست نشده	-	P15 P14’ P06’	۳
همهٔ مطالعه‌های موردی	۴۳۰۱		۲۲

نردبان بازتولیدپذیری

نردبان چهار پله دارد (شکل 1). **L0**: نمونهٔ تحلیل را از فایل‌های عمومی با اعمال قاعده‌های مقاله به همان ترتیب بازسازی و اندازهٔ نمونه را با عدد چاپی مقایسه کن. **L1**: جدول‌های توصیفی را بازسازی کن و هر مقدار چاپی را EXACT (برابر در دقت چاپی)، CLOSE (تا دو درصد)، ACCEPTABLE (تا پنج درصد) یا GAP درجه‌بندی کن. **L2**: مدل منتشرشده را با همان متغیرها و برآوردگرِ اعلام‌شدهٔ مقاله (وزنی یا بدون وزن) برازش بده و هر ضریب یا امتیاز چاپی را به همان شیوه درجه‌بندی کن، با رواداری دست‌کم یک رقم آخر چاپی. **L3**: مدل را زیر طرح نمونه‌گیری دوباره برازش بده. هیچ قاعده‌ای برای رسیدن به عدد چاپی تغییر نکرد: تفاوت به‌صورت شکاف ثبت می‌شود و طبیعی‌ترین خوانش یک قاعدهٔ نانوشته همراه با شاهدی که آن را پشتیبانی می‌کرد مستند می‌شود. پله‌های L0 و L1 پیش از اجرای مدل‌های این مقاله کامل شده بودند و درجه‌بندی‌شان از پروندهٔ ممیزی مخزن بسته بازتولید می‌شود. طرح تحلیل پس از اجرای سه مطالعه (یانگ، کوین، جیائو) هنگام توسعهٔ موتور تثبیت شد؛ این سه حدود نیمی از ضرایب و هر سه نمونهٔ اصلی را دارند و انحراف‌ها از طرح، از جمله بازنویسی فرضیهٔ چهارم، در آن ثبت و در مواد تکمیلی خلاصه شده است.



شکل 1: روش مقاله. پله‌های L0 تا L3 بر هر مطالعهٔ منتشرشده اعمال می‌شود؛ جعبهٔ V که طبقه‌بند را دوباره امتیاز می‌دهد برای مطالعه‌های یادگیری ماشین است. لوزی‌ها تصمیم‌اند؛ جعبه‌های خط‌چین آنچه را به مرحلهٔ بعد منتقل می‌شود ثبت می‌کنند.

چهار شیوه برای برازش یک مدل

هر رگرسیون به چهار شیوه برازش شد. برازش بدون وزن از بیشینهٔ درست‌نمایی با خطای معیار مبتنی بر مدل استفاده می‌کند، که همان است که بیشتر مطالعه‌ها گزارش می‌کنند. برازش خوشه‌ای همان برآورد بدون وزن را نگه می‌دارد و واریانس را با ساندویچی می‌سازد که در آن مشاهده‌های یک خانوار یک خوشه‌اند. برازش وزنی از وزن فرد، خانوار یا فرد-روز مشاهده (مقیاس‌شده به اندازهٔ نمونه برای برآورد) با همان ساندویچ خوشه‌ایِ خانوار بهره می‌برد؛ همین برازش با هر مشاهده به‌عنوان خوشهٔ خودش، هزینهٔ نابرابری وزن‌ها به‌تنهایی را می‌سنجد. برازش طراحی‌آگاه همان برآورد وزنی را با طرح واریانس موج به کار می‌برد. این برازش‌ها به پرسش‌های متفاوتی پاسخ می‌دهند: برآورد بدون وزن ارتباط را میان پاسخ‌گویان توصیف می‌کند و برآورد وزنی ارتباط را در جمعیتی که وزن‌ها نماینده‌اش‌اند؛ پس مقایسهٔ خطاهای معیارشان تغییر برآوردشونده را با تغییر دقت درهم می‌آمیزد و به همین دلیل خطای معیار را به‌صورت زنجیره‌ای از نسبت‌ها (وزن، خوشه، فرمول تکرار) گزارش می‌کنیم و نه فقط حاصل‌ضربشان. برای موج‌هایی که وزن تکرار دارند این روش جک‌نایف است: مدل با هر یک از R وزن تکرار w_r دوباره برازش می‌شود و واریانس ضریب $\hat{\beta}$ برابر است با

$$\hat{V}(\hat{\beta}) = f \sum_{r=1}^{R} \left(\hat{\beta}_r - \hat{\beta}\right)^2,$$

با $R = ۹۸$ و $f = 6/7$ برای ۲۰۱۷، $R = ۱۰۰$ و $f = 99/100$ برای ۲۰۰۹ و $R = ۹۹$ و $f = 98/99$ برای ۲۰۰۱، مانند راهنمای کاربر. رگرسیون‌های لجستیک، خطی و پواسون از آغاز گرم دقیقاً دوباره برازش می‌شوند. رگرسیون‌های دوجمله‌ای منفی و دوجمله‌ای منفی صفرافزوده که برازش دوبارهٔ آن‌ها دقیقه‌ها طول می‌کشد، از گام‌های نیوتن از بهینهٔ نمونهٔ کامل بهره می‌گیرند، گامی که با برازش‌های دوبارهٔ دقیق سنجیدیم. درست‌نمایی مدل‌های شمارشی و گرادیان آن‌ها از خود ماست و با تفاضل متناهی و با کتابخانهٔ statsmodels سنجیده شد (Lambert, 1992, Zeileis و همکاران, 2008). برای سنجش‌های مستقل، برازش‌های لجستیک، پواسون و خطی روی بزرگسالان فایل فرد ۲۰۱۷ با بستهٔ survey در R (طرح وزن تکرار با همان ضریب) مقایسه شد: ضریب‌ها تا یک در صد هزار و خطای معیارها تا ۰٫۲ درصد برابر بودند؛ واریانس سری تیلور موج ۲۰۲۲ نیز به همین شیوه با همان بسته سنجیده شد. روی مدل دوجمله‌ای منفی آخر هفتهٔ یک مطالعه، ۹۸ برازش دوبارهٔ دقیق خطای معیارهایی تا ۰٫۲ درصدِ مقدارهای نیوتن داد. رژیمی که عددهای چاپی یک مطالعه در آن بازتولید شد برآوردگر اعلام‌شدهٔ آن مطالعه گزارش می‌شود.

مدل‌های مطالعه‌های موردی

هشت مطالعهٔ رگرسیونی منتشرشده دوباره برازش شد: رگرسیون لجستیک پیاده‌روی و دوچرخه‌سواری (دو مطالعه)؛ کمترین مربعات معمولی با اثر ثابت برای سازندهٔ خودرو، ایالت و چگالی (یک)؛ رگرسیون دوجمله‌ای منفی سفرهای روزانه (دو)؛ و رگرسیون دوجمله‌ای منفی صفرافزودهٔ استفادهٔ ماهانه از تحرک اشتراکی (سه). یک مدل معادلات ساختاری با دو گرایش پنهان، همتاسازی امتیاز گرایش گروه‌های کاربر سواری‌خواهی، رگرسیون تقویت گرادیان خرید اینترنتی خانوار و پنج طبقه‌بند پذیرش خودروی سوخت جایگزین و برقی مجموعه را کامل می‌کنند. هر جا جدول ضرایب چاپی وجود دارد هر ضریب چاپی هدف L2 است؛ هر جا فقط یک نتیجهٔ کلیدی چاپ شده (اثر حمل‌ونقل عمومی در دو مطالعهٔ ZINB، کشش در مطالعهٔ کمترین مربعات)، همان نتیجه هدف است و هر جا از درصدی در چکیده به دست آمده در مواد تکمیلی گفته شده است.

مدل‌های یادگیری ماشین، طرح‌های اعتبارسنجی و تنظیم

پنج طبقه‌بند دوباره روی نمونهٔ ۳۰٬۵۱۱ مالک خودروی نو در ۲۷٬۴۱۳ خانوار اجرا شد، با پیامد خودروی سوخت جایگزین (۱٬۲۳۶ AFV، مالک) یا خودروی برقی (۳۸۹ EV، مالک). قاعدهٔ نمونه (نوترین خودروی هر راننده اصلی، حداکثر دو سال مدل، متغیرهای کلیدی کامل) از اندازهٔ نمونهٔ چاپی بازیابی شد و اندکی کوچک‌تر از نمونهٔ چاپی است. مدل‌ها در جدول 3 توضیح داده شده‌اند. رگرسیون لجستیک جمع وزنی ورودی‌ها را به احتمال تبدیل می‌کند؛ بیز ساده فرض می‌کند هر ورودی در هر رده زنگوله‌ای و مستقل از دیگر ورودی‌هاست؛ درخت تصمیم رشته‌ای از پرسش‌های بله/خیر دربارهٔ ورودی‌ها می‌پرسد؛ جنگل تصادفی رأی درختان بسیاری را که روی بخش‌های تصادفی داده رشد کرده‌اند میانگین می‌گیرد؛ و ماشین بردار پشتیبان

با هستهٔ شعاعی رده‌ها را با مرزی هموار که از نزدیک‌ترین نقطه‌های آموزشی ساخته می‌شود از هم جدا می‌کند. خانوادهٔ ششم، تقویت گرادیان، که درخت‌های کوچک را یکی پس از دیگری می‌افزاید تا هر یک خطای مجموع پیشین را اصلاح کند، برای رگرسیون خرید اینترنتی در سطح خانوار همراه با رگرسیون خطی، رگرسیون بردار پشتیبان و جنگل تصادفی به کار رفت. این چهار مدل رگرسیونی تنظیم نشدند: با تنظیمات ثابت و مستند اجرا شدند، همان‌ها که مطالعه چاپ کرده بود (هزینهٔ ده و پهنای هستهٔ یک‌صدم برای رگرسیون بردار پشتیبان، کمینهٔ بیست‌وپنج ردیف در برگ برای جنگل)، زیرا پرسش این بود که آیا با تنظیمات چاپی می‌توان به امتیازهای چاپی نزدیک شد.

جدول 3: پنج طبقه‌بند و فضای جست‌وجوی تنظیمات آن‌ها (برای همهٔ مدل‌ها همان جست‌وجوی تصادفی ده‌آزمایشی).

مدل	ایده به زبان ساده	ابرپارامترهای جست‌وجوشده (ده آزمایش تصادفی برای هر مدل)
LR	رگرسیون لجستیک: جمع وزنی ورودی‌ها که به احتمال تبدیل می‌شود	قدرت جریمه C (۰٫۰۰۱ تا ۱۰۰، مقیاس لگاریتمی)؛ نوع جریمه (L1 یا L2)
NB	بیز ساده با توزیع گاوسی: هر ورودی در هر رده زنگوله‌ای و مستقل از دیگران فرض می‌شود	کف واریانس (۱e-۱۱ تا ۱e-۳، مقیاس لگاریتمی)
DT	درخت تصمیم: رشته‌ای از پرسش‌های بله/خیر دربارهٔ ورودی‌ها که به برگی با یک رده می‌رسد	بیشینهٔ عمق (۳، ۵، ۸، ۱۲، ۲۰)؛ کمینهٔ اندازهٔ برگ (۱، ۵، ۲۰، ۵۰)؛ معیار شکست (جینی یا آنتروپی)
RF	جنگل تصادفی: درختان بسیار روی بخش‌های تصادفی داده که رأی‌شان میانگین‌گیری می‌شود	شمار درختان (۵۰، ۱۰۰، ۲۰۰)؛ بیشینهٔ عمق (۵، ۱۰، ۲۰، ۴۰)؛ کمینهٔ اندازهٔ برگ (۱، ۲، ۵، ۲۰)؛ سهم ورودی‌های آزموده در هر شکست (۰٫۳، ۰٫۵، ۱)
SVM	ماشین بردار پشتیبان با هستهٔ شعاعی: رده‌ها را با مرزی هموار که از نزدیک‌ترین نمونه‌های آموزشی ساخته می‌شود جدا می‌کند	هزینهٔ C (۰٫۱ تا ۱۰۰، مقیاس لگاریتمی)؛ پهنای هسته گاما (۰٫۰۰۱ تا ۱، مقیاس لگاریتمی)؛ آموزش روی حداکثر ۱۲٬۰۰۰ ردیف

هر مدل یکسان تنظیم شد. فضای جست‌وجوی هر مدل در جدول 3 آمده است. روش جست‌وجو، جست‌وجوی تصادفی با بودجهٔ یکسان (ده آزمایش) برای هر مدل است؛ هر آزمایش با اعتبارسنجی سه‌بخشی درون ردیف‌های آموزش امتیاز می‌گیرد، در خط لوله‌هایی که خانوارها را تقسیم می‌کنند گروه‌بندی‌شده بر پایهٔ خانوار و در خط لوله‌های تقسیم ردیف بر پایهٔ ردیف، و بیش‌نمونه‌گیری تنها درون بخش‌های آموزش درونی در خط لوله‌هایی که پس از تقسیم بیش‌نمونه‌گیری می‌کنند؛ ملاک سطح زیر منحنی ROC است. پیکربندی برگزیده روی همهٔ ردیف‌های آموزش دوباره برازش می‌شود و یک بار روی ردیف‌های کنار گذاشته‌شده امتیاز می‌گیرد (اعتبارسنجی تودرتو). آزمایش با دو بذر اصلی اجرا شد. چهار خط لوله فقط در جای بیش‌نمونه‌گیری و آنچه تقسیم می‌شود فرق دارند. در بیش‌نمونه‌گیری و سپس تقسیم ردیف‌ها رستهٔ کمیاب روی کل فایل با درون‌یابی میان همسایه‌های نزدیک بیش‌نمونه‌گیری می‌شود (Chawla و همکاران, 2002) و سپس ردیف‌ها تصادفی تقسیم می‌شوند؛ در تقسیم ردیف‌ها و سپس بیش‌نمونه‌گیری تقسیم اول می‌آید و تنها ردیف‌های آموزش بیش‌نمونه‌گیری می‌شوند؛ در تقسیم خانوارها و سپس بیش‌نمونه‌گیری هر خانوار کاملاً در یک سو قرار دارد؛ و در تقسیم خانوارها بدون بیش‌نمونه‌گیری ردیف‌های آموزش همان‌گونه که هستند به کار می‌روند و امتیاز آزمون با وزن پیمایش وزن‌دهی می‌شود. در هر تقسیم از طرح‌های خانوار حدود ۹٬۱۸۹ ردیف از حدود ۲۷٬۴۱۳ خانوار کنار گذاشته می‌شود (جدول 4) و هیستوگرام فراوانی رستهٔ کمیاب در همهٔ ردیف‌ها، ردیف‌های آموزش و ردیف‌های کنار گذاشته‌شده در شکل 8 مواد تکمیلی آمده است. چون خانوارهای کنار گذاشته‌شده زیرمجموعه‌ای تصادفی از همهٔ خانوارها هستند، این طرح‌ها تعمیم به خانوارهای تازه در همان نمونه را می‌آزمایند؛ هیچ گروه کنار گذاشته‌شده‌ای بیرون از دامنهٔ گروه‌های آموزش نیست، پس هیچ‌کدام به معنای پوش محدب برون‌یابی نیست. برای مطالعهٔ رگرسیونی طرح دوم کل ایالت‌ها را کنار می‌گذارد که آزمونی سخت‌تر برای انتقال است.

جدول 4: ردیف‌های کنار گذاشته‌شده و نگه‌داشته‌شده در تقسیم خانوار و سهم هر رستهٔ خودرو در آن‌ها (فراوانی رستهٔ کمیاب پیش و پس از کنار گذاشتن).

گروه	ردیف	سهم AFV	سهم EV
همهٔ مالکان خودروی نو	۳۰٬۵۱۱	۴٫۱٪	۱٫۳٪
خانوارهای آموزش (میانگین سه تقسیم)	۲۱٬۳۲۲	۴٫۱٪	۱٫۳٪
خانوارهای کنار گذاشته‌شده (میانگین سه تقسیم)	۹٬۱۸۹	۳٫۹٪	۱٫۳٪

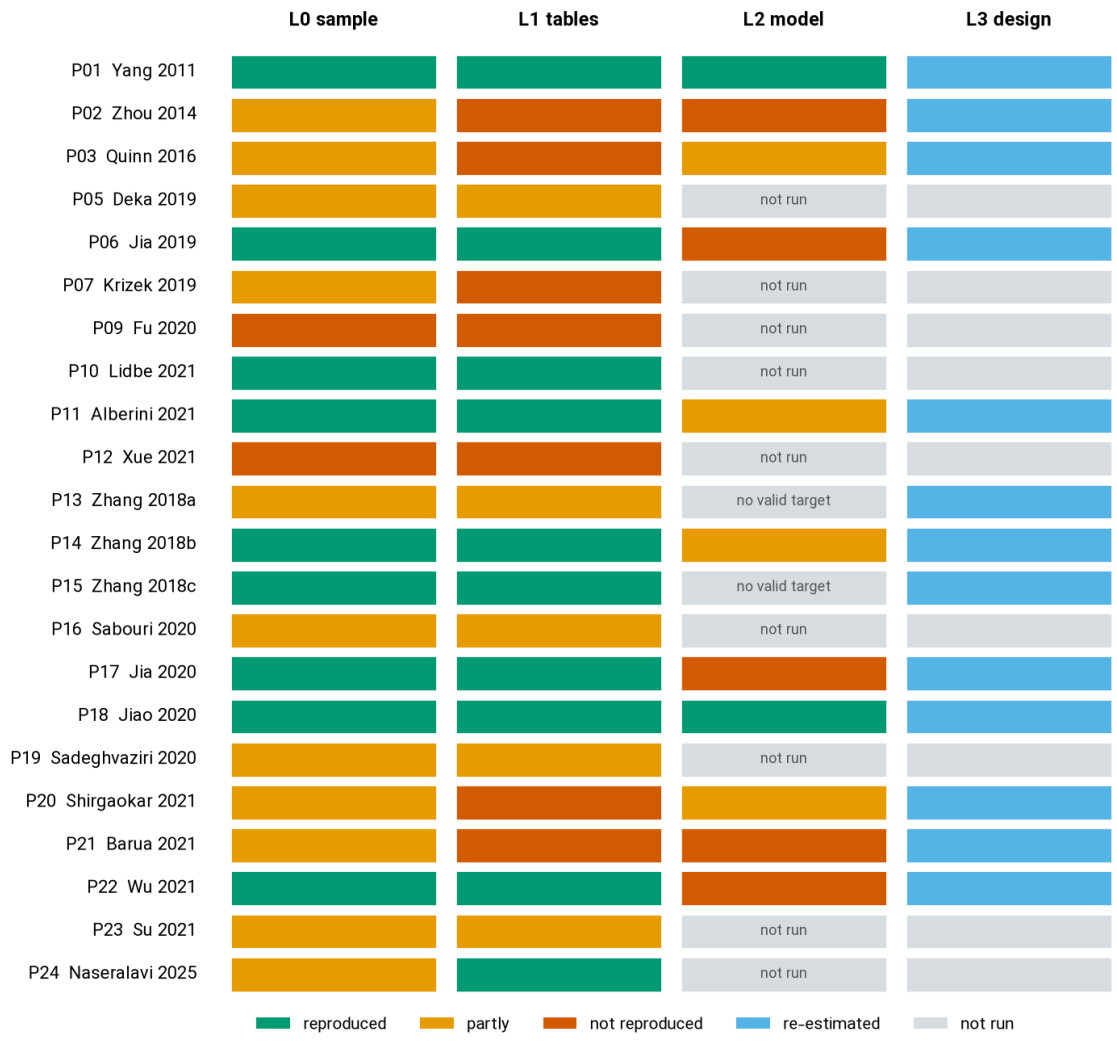
نرم‌افزار و بازتولیدپذیری

همهٔ عددها، جدول‌ها و شکل‌ها با کد از پرونده‌های نتیجه ساخته می‌شوند؛ متن هیچ نتیجهٔ دست‌تایپ‌شده‌ای ندارد. کد تحلیل، موتور رژیم‌های رگرسیون و خط لوله‌های اعتبارسنجی در مخزن است و با یک فرمان اجرا می‌شود، و ابزارهای رگرسیون و اعتبارسنجی بخشی از pynhts هستند (ماژول‌های `pynhts.models` و `pynhts.validate`) با آزمون واحد، آزمون دادهٔ واقعی در برابر R و مستندات. طرح تحلیل، با فرضیه‌ها و آستانه‌ها، پیش از اجرای مدل‌های نیمهٔ دوم مطالعه‌ها نوشته شد و تاریخ دارد.

نتایج

آنچه بازسازی شد

نردبان در شکل 2 دیده می‌شود. از ۲۲ مطالعه، ممیزی فایل‌های عمومی را برای بازسازی داده‌های ۹ مطالعه کافی، برای ۱۱ مطالعه تا حدی کافی و برای ۲ مطالعه ناکافی دانست که در آن‌ها جامعه یا قاعدهٔ نمونه‌گیری ننوشته شده است. در مجموع، ۸۷۹ مقدار از ۱٬۴۵۴ مقدار توصیفی چاپی دقیقاً در دقت چاپی بازتولید شد و ۸۱ درصد دقیقاً یا تا دو درصد. شکاف‌ها بیشتر از خود مقاله‌ها برخاستند: قاعده‌های نانوشتهٔ حذف رکورد، حساب‌هایی که در جدول جمع نمی‌شود و متغیرهایی که در فایل عمومی نیستند. ۱۳ مطالعه را می‌شد تا پلهٔ مدل برد؛ دلیل دیگر مطالعه‌ها در جدول 1 آمده است.



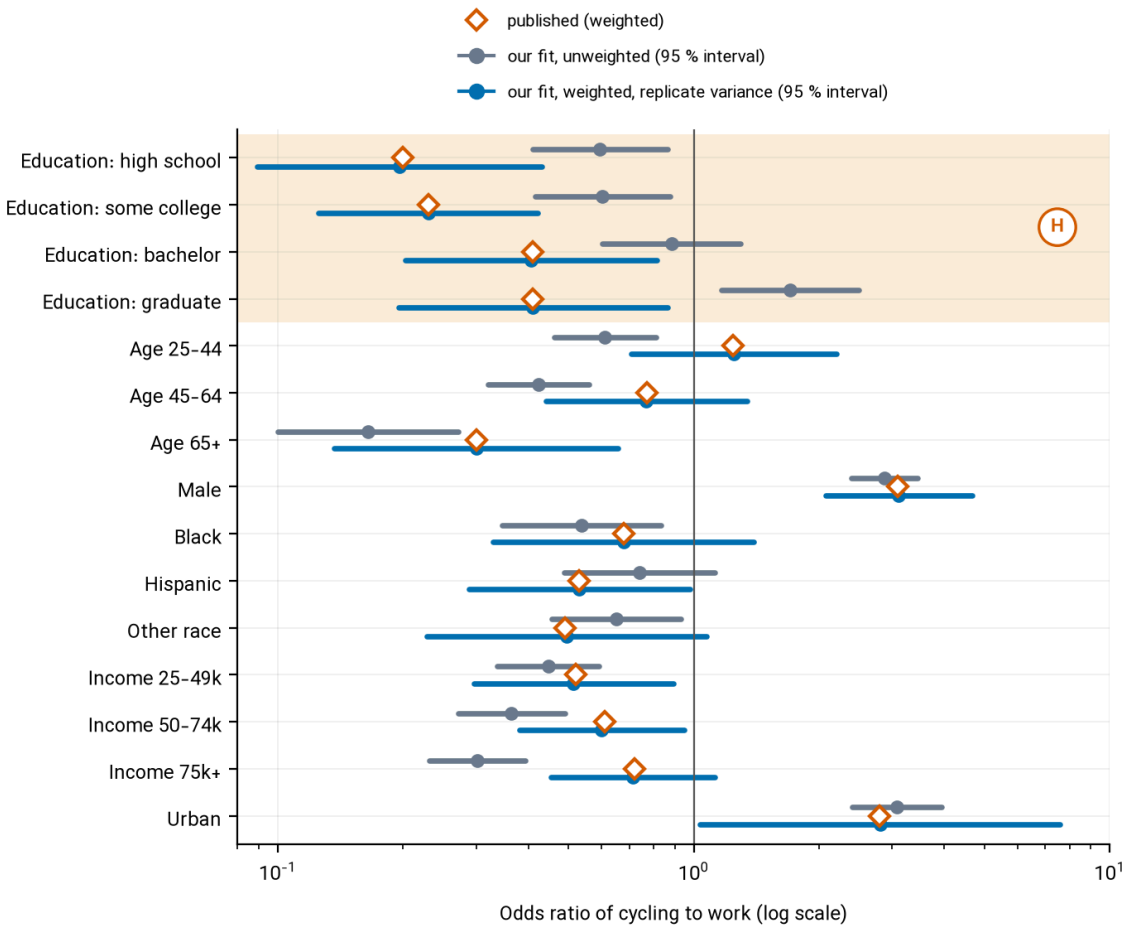
شکل 2: نردبان برای ۲۲ مطالعه. هر سطر یک مطالعه و هر ستون یک پله است؛ رنگ‌ها درجه را نشان می‌دهند. ستون L3 نشان می‌دهد کدام مدل‌ها دوباره برآورد شدند و درجه نیست.

برآوردگر اعلام‌شده عددهای چاپی را بازتولید می‌کند

در مطالعه‌هایی که برازش بدون وزن و وزنی در شمار مقدارهای چاپی بازتولیدشده آشکارا فرق دارند، برآوردگر اعلام‌شده بیشتر بازتولید کرد. مطالعهٔ پیاده‌روی و دوچرخه‌سواری کوین و همکاران می‌گوید وزن‌های جمعیتی و جک‌نایف به کار رفته است: نسبت‌های بخت وزنی ۳۴ مقدار از ۶۰ مقدار چاپی را بازتولید کرد و نسبت‌های بدون وزن ۲ مقدار؛ برای دو مدل رفت‌وآمد، ۲۹ مقدار از ۳۰ مقدار تا دقت چاپی یا دو درصد برابر شد، اما برای دو مدل سفر کمتر برابر شد که آن را به نمونه نسبت می‌دهیم (نمونهٔ چاپی را نتوانستیم بسازیم). در مطالعهٔ فعالیت بدنی بر پایهٔ فصل، نسبت‌های بخت وزنی ۲۱ مقدار از ۲۱ مقدار چاپی را بازتولید کرد و نسبت‌های بدون وزن ۶ مقدار. در جهت دیگر، مطالعهٔ دوجمله‌ای منفی تحرک اشتراکی و سفرها، که برازش بدون وزن در R گزارش می‌کند، با برازش بدون وزن بازتولید شد (۳۰ ضریب از ۳۲ ضریب تا دو درصد، ۲۰ ضریب در دقت چاپی) و مطالعهٔ صفرافزودهٔ مالکیت خودرو و سواری‌خواهی برای ۱۵ ضریب از ۳۲ ضریب تا دو درصد و ۲۵ ضریب تا پنج درصد. در مطالعه‌های دیگر هیچ‌یک از دو برازش بسیار بازتولید نکرد: کشش مایل سالانه نسبت به هزینهٔ سوخت در هر مایل برای یکی از ۴ مقدار چاپی برابر شد (−۰٫۵۱۵ در برابر −۰٫۵۱۰). فرضیهٔ نخست بنابراین

برای چهار مطالعه‌ای که دو برآوردگر در آن‌ها از هم تشخیص‌پذیرند پشتیبانی می‌شود و برای بقیه آزموده نشده است؛ آن را به همهٔ مطالعه‌ها تعمیم نمی‌دهیم.

برآوردگر تنها در مطابقت نقش ندارد، در محتوا هم مؤثر است. برجسته‌ترین نمونه اثر تحصیلات بر دوچرخه‌سواری به محل کار است. نسبت بخت چاپی برای مدرک تحصیلات تکمیلی در برابر کمترین گروه تحصیلی ۰٫۴۱ است، برازش وزنی ۰٫۴۱ می‌دهد و برازش بدون وزن ۱٫۷۱: علامت ارتباط با تحصیلات برعکس می‌شود (شکل 3، حرف مرجع H). همین در اثر منطقهٔ میانه‌غربی بر سفر فعال در ۲۰۰۱ دیده می‌شود، با ۰٫۶۷ چاپی، ۰٫۶۷ وزنی و ۰٫۸۳ بدون وزن. برازش بدون وزن ارتباط را میان پاسخگویان برآورد می‌کند و این نتایج نشان می‌دهد پاسخگویان به شیوه‌هایی که برای این ارتباط مهم است با جمعیت فرق دارند؛ آگاهی‌بخش بودن وزن‌ها را به‌صورت رسمی نیازمودیم.



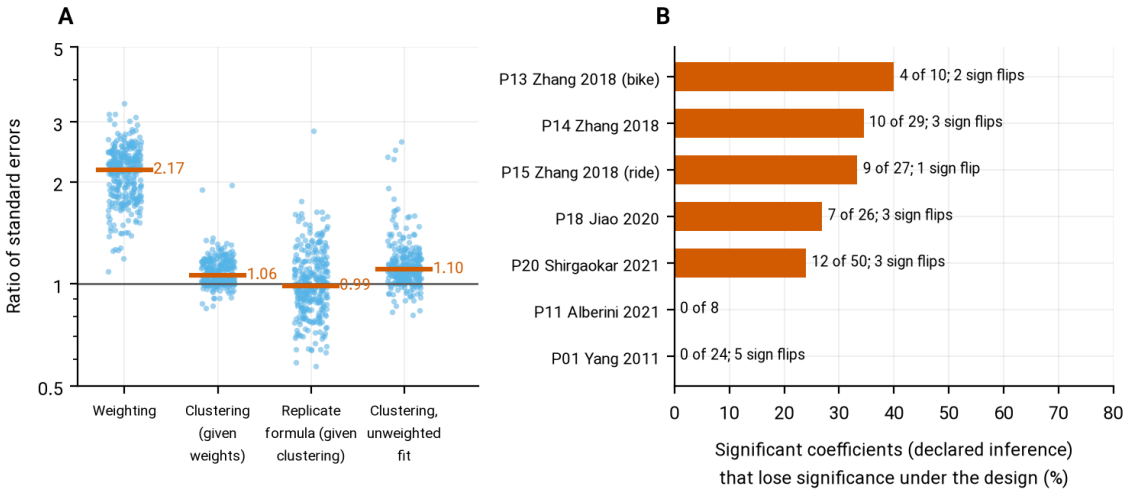
شکل 3: نسبت‌های بخت دوچرخه‌سواری به محل کار در مطالعهٔ کوین و همکاران: مقدار چاپی، برازش بدون وزن ما و برازش وزنی ما با واریانس وزن تکرار. سطرهای سایه‌دار (حرف مرجع H) اثرهای تحصیلات‌اند که در متن بحث شده است.

خطای معیار و معناداری

در ۳۱۰ ضریبِ دوباره برآوردشدهٔ هشت مطالعهٔ رگرسیونی، خطای معیار طراحی‌آگاه به‌طور میانه ۲٫۱۹ برابر خطای معیار بدون وزن بود (دامنهٔ میان‌چارکی ۱٫۸۶ تا ۲٫۶۶). شکل 4 (پنل A) نشان می‌دهد این از کجا می‌آید. وزن‌دهی به‌تنهایی، با هر مشاهده به‌عنوان خوشهٔ خودش، خطای معیار را به‌طور میانه ۲٫۱۷ برابر کرد؛ خوشه‌کردن اشخاص در خانوار ۱٫۰۶ برابر دیگر روی آن افزود و فرمول وزن تکرار ۰٫۹۹ برابر دیگر (صدک پنجم تا نود و پنجم ۰٫۷۰ تا ۱٫۴۴). برازش بدون وزنی

که خطای معیارش بر پایهٔ خانوار خوشه‌ای شد تنها با ضریب میانهٔ ۱٫۱۰ از خطای معیار مبتنی بر مدلش فرق داشت. بنابراین تحلیل‌های بدون وزن دربارهٔ دقت چیزی که برآورد می‌کنند، یعنی ارتباط میان پاسخگویان، اشتباه نبودند؛ خطای معیارهای بزرگ بهای برآورد ارتباط جمعیتی با وزن‌های نابرابر است.

زیر استنباط اعلام‌شدهٔ هر مطالعه (بدون وزن برای شش مطالعه، وزنی با خطای خوشه‌ایِ خانوار برای یکی؛ مطالعه‌ای که خود از جک‌نایف استفاده کرد کنار گذاشته شد)، ۴۲ ضریب از ۱۷۴ ضریبی که در سطح پنج درصد معنادار بودند در برازش طراحی‌آگاه معنادار نبودند، یعنی ۲۴ درصد با بازهٔ دقیق نودوپنج‌درصدی ۱۸ تا ۳۱ درصد؛ برای همهٔ مطالعه‌ها در برابر برازش بدون وزن این اعداد ۷۸ از ۲۴۴ است. سهم به مطالعه بستگی دارد (شکل 4، پنل B؛ جدول 5). برآورد نقطه‌ای ۲۶ ضریب (۸٫۴ درصد) میان برازش بدون وزن و وزنی در علامت فرق دارد، اما تنها ۸ مورد از آن‌ها در برازش بدون وزن معنادار بود و ۱ مورد در هر دو، پس بیشتر تغییرهای علامت به ضرایب نزدیک به صفر مربوط است. فرضیهٔ دوم (میانهٔ نسبت بیش از یک‌ونیم) و سوم (بیش از یک‌دهم ضرایب معنادار معناداری خود را از دست می‌دهند، اینجا با کران پایین بازه بالاتر از یک‌دهم) پشتیبانی می‌شوند.

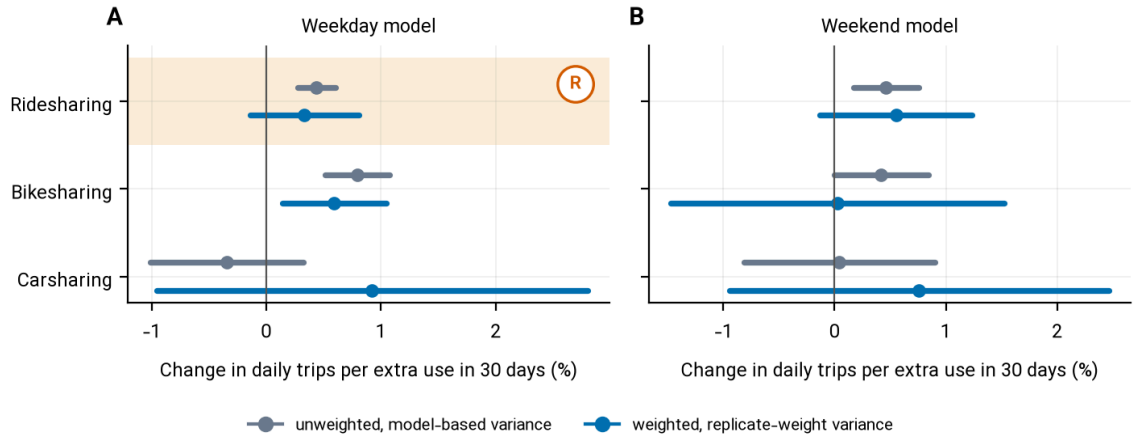


شکل 4: منشأ تفاوت خطای معیار و آنچه عوض می‌کند. پنل A: نسبت خطاهای معیار روی همهٔ ضرایب دوباره برآوردشده، به تفکیک منبع (نارنجی: میانه). پنل B: سهم ضریب‌های معنادار زیر استنباط اعلام‌شدهٔ هر مطالعه که در برازش طراحی‌آگاه معناداری خود را از دست می‌دهند، به تفکیک مطالعه.

جدول 5: برآورد دوبارهٔ هشت مطالعهٔ رگرسیونی. سه ستون نسبت، میانهٔ نسبت خطاهای معیار است: وزنی مبتنی بر مدل بر بدون وزن؛ وزنی خوشه‌ای خانوار بر وزنی مبتنی بر مدل؛ وزن تکرار بر وزنی خوشه‌ای خانوار. معناداری ازدست‌رفته: ضرایب معنادار زیر استنباط اعلام‌شدهٔ مطالعه که در برازش طراحی‌آگاه نیستند (مطالعه‌ای که خود از جک‌نایف استفاده کرد موردی ندارد). تغییر علامت: علامت برآورد بدون وزن و طراحی‌آگاه فرق دارد.

مطالعه	ضرایب	مقدار چاپی بازتولیدشده: بدون وزن / وزنی / چاپی	وزن	خوشه	فرمول تکرار	معناداری ازدست‌رفته (استنباط اعلام‌شده)	تغییر علامت
P01 Yang ۲۰۱۱	۵۰	۲۱ / ۲۱ / ۶	۱٫۷۸	۱٫۱۲	۰٫۹۴	۲۴ / ۰	۵
P03 Quinn ۲۰۱۶	۷۲	۶۰ / ۳۴ / ۲	۲٫۳۳	۱٫۰۱	۰٫۹۹	-	۹
P11 Alberini ۲۰۲۱	۸	۴ / ۰ / ۱	۲٫۲۸	۱٫۰۱	۰٫۷۵	۸ / ۰	۰
P13 Zhang ۲۰۱۸ (bike)	۳۰	-	۱٫۶۷	۱٫۰۲	۱٫۰۹	۱۰ / ۴	۲
P14 Zhang ۲۰۱۸	۳۲	۳۲ / ۴ / ۱۵	۲٫۱۹	۱٫۱۲	۱٫۰۵	۲۹ / ۱۰	۳
P15 Zhang ۲۰۱۸ (ride)	۳۰	-	۲٫۱۸	۱٫۰۹	۱٫۰۶	۲۷ / ۹	۱
P18 Jiao ۲۰۲۰	۳۲	۳۲ / ۱۲ / ۳۰	۲٫۱۲	۱٫۰۴	۱٫۱۵	۲۶ / ۷	۳
P20 Shirgaokar ۲۰۲۱	۵۶	۱۴ / ۶ / ۸	۲٫۳۴	۱٫۱۶	۰٫۹۴	۵۰ / ۱۲	۳

نمونهٔ ملموس، مطالعهٔ دوجمله‌ای منفی تحرک اشتراکی و سفرهاست که نتیجه‌اش این است که استفاده از سواری‌خواهی پیش‌بین معنادار سفرهای روزانه است. برازش بدون وزن ضریب چاپی را بازتولید می‌کند (۰٫۴۴ درصد سفر بیشتر برای هر بار استفادهٔ اضافی در ماه گذشته در روزهای هفته) و p کمتر از ۰٫۰۰۱ می‌دهد. برآورد وزنی کوچک‌تر است (۰٫۳۳ درصد، تغییر −۲۴ درصد)؛ p خوشه‌ایِ خانوار آن برابر ۰٫۰۴۵ است و خطای معیار وزن تکرار ۰٫۰۰۲۴ (در برابر ۰٫۰۰۰۸ بدون وزن) p برابر ۰٫۱۷ می‌دهد؛ در آخر هفته p وزن تکرار برابر ۰٫۱۱ است (شکل 5، حرف مرجع R). پس معناداری، که نتیجه بر آن استوار است، هم به وزن‌دهی و هم به برآوردگر واریانس بستگی دارد. نمونهٔ دوم اثر راننده‌بودن بر تعداد سواری‌ها در مدل صفرافزوده است: برآورد از −۰٫۴۳ به −۰٫۱۹ می‌رسد و p از کمتر از ۰٫۰۰۱ (بدون وزن) به برابر ۰٫۰۱۵ (وزنی، خوشه‌ای) و برابر ۰٫۱۴ (وزن تکرار).



شکل 5: جمله‌های تحرک اشتراکی در مدل‌های سفر روزهای هفته (پنل A) و آخر هفته (پنل B) در مطالعهٔ جیائو و همکاران، با بازه‌های نودوپنج‌درصدی از برازش بدون وزن و طراحی‌آگاه. حرف مرجع R سطر سواری‌خواهی را که در متن بحث شده نشان می‌دهد.

کدام لایهٔ طرح اهمیت دارد

تجزیهٔ شکل 4 (پنل A) می‌گوید وزن‌ها تفاوت را می‌سازند و، به‌طور میانگین، خوشه‌بندی خانوار و فرمول تکرار کم می‌افزایند. این اجازهٔ نادیده‌گرفتن فرمول تکرار نیست. به تفکیک مطالعه، میانهٔ نسبت خطای معیار وزن تکرار به خطای معیار خوشه‌ایِ خانوار از ۰٫۷۵ تا ۱٫۱۵ بود و برای ضریب‌های جداگانه از ۰٫۷۰ تا ۱٫۴۴ (صدک پنجم تا نود و پنجم)؛ از ۱۸۲ ضریب معنادار زیر برازش وزنی خوشه‌ایِ خانوار، ۱۶ ضریب (۹ درصد) زیر واریانس تکرار معناداری خود را از دست داد، از جمله ضریب سواری‌خواهی بالا. هر جا وزن تکرار هست باید به کار رود؛ برازش وزنی خوشه‌ایِ خانوار برای موج ۲۰۲۲ که طرح با لایه و خوشه بیان می‌شود جایگزین معقولی است، اما تقریبی است و درست در بعضی موردهای مهم شکست می‌خورد.

مدل‌های بدون جدول ضرایب چاپی

مدل معادلات ساختاری خرید اینترنتی و سفرهای خرید، که ضرایب استانداردشده چاپ می‌کند، با همان متغیرها به جز دو متغیر که از منبع دیگرند روی ۱۰۶٬۱۳۶ نفر برازش شد، در برابر ۸۵٬۶۶۳ نفر چاپی. دو اثر متقابل آن در برازش وزنی ما −۰٫۰۸ و ۰٫۱۷ بود در برابر ۰٫۱۲ و ۰٫۲۱ چاپی، با خطای معیار طراحی حدود ۰٫۰۲. یکی از اثرهای متقابل علامت مخالف دارد و دیگری اندازهٔ دیگری، و قرارداد علامت متغیرهای پنهان نمی‌تواند هر دو را اصلاح کند؛ مدل را بازتولیدنشده درجه‌بندی می‌کنیم. برای مطالعهٔ همتاسازی امتیاز گرایش سواری‌خواهی و خودروکیلومتر، همتاسازی حریصانهٔ نزدیک‌ترین همسایه در همان لایه‌ها ۲٬۶۶۴ سه‌تایی همتاشده از اشخاص دارای دسترسی به خودروی خانوار داد، در برابر حدود پنج هزار چاپی. میانگین مایل روزانهٔ همتاشده با میانگین‌های چاپی فرق داشت (برای کاربران پرتکرار ۲۹٫۸ در برابر ۲۴٫۸ مایل) و فقط یکی از سه تضاد خودرو به مقدار چاپی نزدیک بود (از غیرکاربر به کاربر گاه‌به‌گاه −۰٫۰۶۷ در برابر −۰٫۰۶۵)؛ مطالعه را بازتولیدنشده درجه‌بندی می‌کنیم. مقایسهٔ طراحی‌آگاه نمونه‌های همتاشده سرراست نیست، چون وزن‌دهی جداگانهٔ دو بازو جفت‌شدگی را می‌شکند و واریانس تکرار نمونهٔ همتاشده را ثابت می‌گیرد؛ فقط گزارش می‌کنیم که وقتی بازوها وزن‌دار شوند هیچ‌یک از سه تضاد خودرو آمارهٔ t بالاتر از دو ندارد (۰٫۹۵، ۰٫۰۰ و ۱٫۶۲)، بی‌آن‌که با آمارهٔ زوجی مقاله (۲٫۴۵ برای تضاد نخست) مقایسه کنیم.

مدل‌های یادگیری ماشین

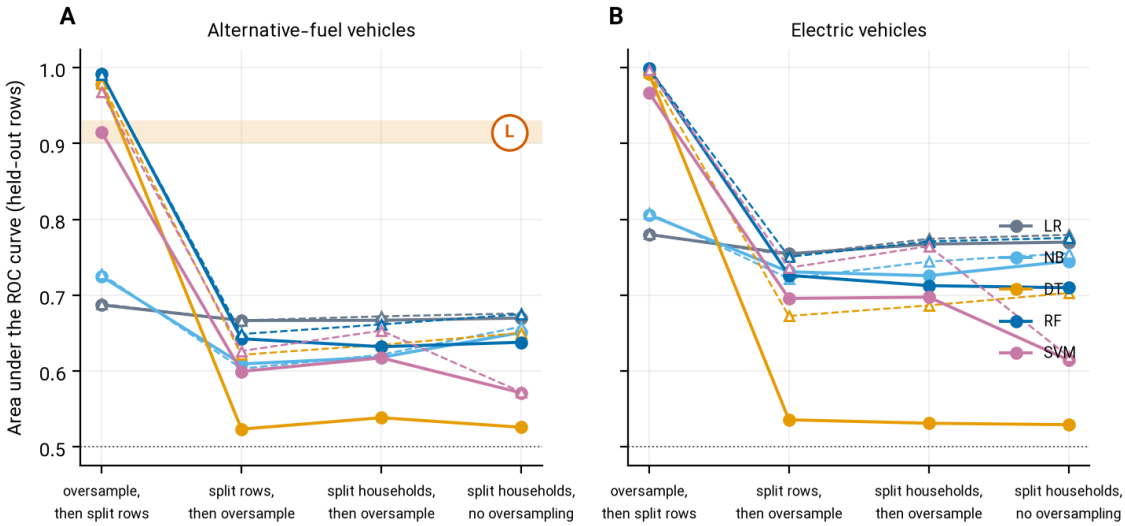
طبقه‌بندهای پذیرش خودروی سوخت جایگزین و برقی بیشترین وابستگی را به خط لوله نشان دادند (شکل 6). وقتی رستهٔ کمیاب پیش از تقسیم ردیف‌ها بیش‌نمونه‌گیری شد، سطح زیر منحنی جنگل تصادفی ۰٫۹۹ (خودروی سوخت جایگزین) و ۱٫۰۰

(خودروی برقی) و درخت تصمیم ۰/۹۸ و ۰/۹۹ بود. وقتی تقسیم اول آمد، همان مدل‌ها برای جنگل تصادفی ۰/۶۴ و ۰/۷۳ و برای درخت ۰/۵۲ و ۰/۵۴ گرفتند؛ تقسیم بر پایهٔ خانوار این را کم عوض کرد (۰/۶۳ و ۰/۷۱)، که انتظار می‌رفت چون تنها ۲۰ درصد ردیف‌ها خانوار را با ردیفی دیگر شریک‌اند و نشتی که اینجا مهم است بیش‌نمونه‌گیری است. با وزن پیمایش، جنگل تصادفی ۰/۶۵ و ۰/۶۹ گرفت. بادکردگی جنگل تصادفی ۰/۳۶ (سوخت جایگزین) و ۰/۲۹ (برقی) بود؛ درخت تصمیم که در خط لولهٔ نشت‌دار ۰/۹۸ گرفت، هنگام ارزیابی صادقانه به‌ندرت بهتر از شانس است (۰/۵۴).

مقدارهای چاپی برای خودروی سوخت جایگزین، ۰/۹۲ برای جنگل تصادفی و ۰/۹۱ برای درخت تصمیم، میان امتیازهای نشت‌دار و صادقانه می‌افتند و با هیچ‌کدام بازتولید نمی‌شوند (شکل 6، حرف مرجع L)؛ نمونهٔ ما نیز از نمونهٔ چاپی کوچک‌تر است. دقت‌های چاپی، بالای نود درصد برای هر پنج مدل، با آزمون صادقانهٔ یک رستهٔ کمیاب سازگارند، که پیش‌بینی رستهٔ پرتعداد ۰/۹۶ می‌دهد، و با خط لولهٔ نشت‌دار نه، که دقت آن ۰/۶۴ برای رگرسیون لجستیک، ۰/۶۶ برای بیز ساده و ۰/۸۳ برای ماشین بردار پشتیبان بود. بنابراین ادعا نمی‌کنیم خط لولهٔ منتشرشده نشت داشته است؛ نشان می‌دهیم بیش‌نمونه‌گیری پیش از تقسیم امتیاز سه مدل از پنج مدل را بیش از پانزده واحد درصد باد می‌کند (جنگل تصادفی ۰/۳۶، درخت تصمیم ۰/۴۴ و ماشین بردار پشتیبان ۰/۳۰) و دو مدل دیگر، رگرسیون لجستیک و بیز ساده، را ۰/۰۲ و ۰/۱۱. فرضیهٔ چهارم با واژه‌بندی نخستین (امتیاز چاپی فقط در خط لولهٔ نشت‌دار بازتولید می‌شود) پشتیبانی نمی‌شود؛ بادکردگیِ آن پشتیبانی می‌شود. تنظیم یکسان همهٔ مدل‌ها شکاف را از میان نبرد. جنگل تصادفی تنظیم‌شده وقتی بیش‌نمونه‌گیری مقدم بود ۰/۹۹ و با تقسیم بر پایهٔ خانوار ۰/۶۶ گرفت (بادکردگی ۰/۳۳؛ برای خودروی برقی ۰/۲۳)، و درخت تصمیم تنظیم‌شده از ۰/۵۴ به ۰/۶۳ رسید اما بسیار پایین‌تر از امتیاز نشت‌دارش (۰/۹۸) ماند. بزرگ‌ترین انحراف معیار میانگین امتیاز روی بذرها ۰/۰۳۳ برای تنظیمات پیش‌فرض و ۰/۰۳۹ (۰/۰۷۵ برای امتیاز وزنی) برای مدل‌های تنظیم‌شده بود.

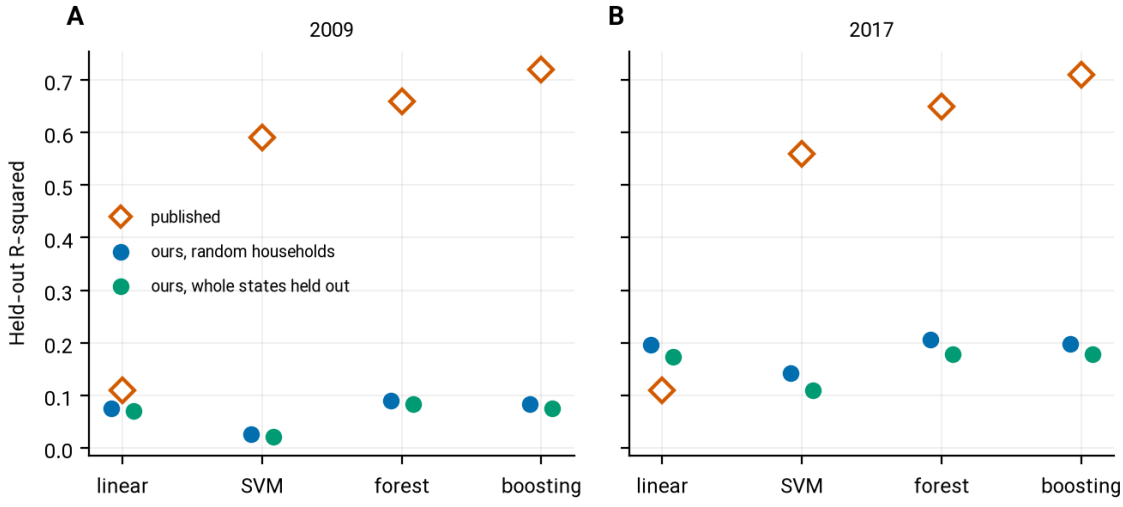
جدول 6: سطح زیر منحنی ROC روی ردیف‌های کنار گذاشته‌شده (میانگین سه بذر برای تنظیمات پیش‌فرض و دو بذر برای مدل‌های تنظیم‌شده). نشت‌دار: بازنمونه‌گیری و سپس تقسیم ردیف‌ها. ردیف‌ها: تقسیم ردیف‌ها و سپس بازنمونه‌گیری. خانوارها: تقسیم خانوارها و سپس بازنمونه‌گیری. وزنی: تقسیم خانوارها بدون بازنمونه‌گیری و امتیاز وزنی.

پیامد	مدل	نشت‌دار	ردیف‌ها	خانوارها	وزنی	تنظیم‌شده، نشت‌دار	تنظیم‌شده، خانوارها
سوخت جایگزین	LR	۰/۶۹	۰/۶۷	۰/۶۷	۰/۶۶	۰/۶۹	۰/۶۷
سوخت جایگزین	NB	۰/۷۲	۰/۶۱	۰/۶۲	۰/۶۳	۰/۷۳	۰/۶۲
سوخت جایگزین	DT	۰/۹۸	۰/۵۲	۰/۵۴	۰/۵۲	۰/۹۸	۰/۶۳
سوخت جایگزین	RF	۰/۹۹	۰/۶۴	۰/۶۳	۰/۶۵	۰/۹۹	۰/۶۶
سوخت جایگزین	SVM	۰/۹۱	۰/۶۰	۰/۶۲	۰/۵۶	۰/۹۷	۰/۶۵
برقی	LR	۰/۷۸	۰/۷۵	۰/۷۷	۰/۷۷	۰/۷۸	۰/۷۷
برقی	NB	۰/۸۱	۰/۷۳	۰/۷۳	۰/۷۳	۰/۸۱	۰/۷۴
برقی	DT	۰/۹۹	۰/۵۴	۰/۵۳	۰/۵۳	۰/۹۹	۰/۶۹
برقی	RF	۱/۰۰	۰/۷۳	۰/۷۱	۰/۶۹	۱/۰۰	۰/۷۷
برقی	SVM	۰/۹۷	۰/۷۰	۰/۷۰	۰/۶۳	۱/۰۰	۰/۷۶



شکل 6: سطح زیر منحنی ROC پنج طبقه‌بند در چهار خط لوله، برای خودروهای سوخت جایگزین (پنل A) و خودروهای برقی (پنل B). دایره‌های توپر: تنظیمات پیش‌فرض آزمایش نخست؛ مثلث‌های توخالی: همهٔ مدل‌ها یکسان تنظیم‌شده. حرف مرجع L امتیاز چاپی جنگل تصادفی مطالعه را نشان می‌دهد.

مطالعهٔ رگرسیونی خرید اینترنتی خانوار ضریب تعیین کنار گذاشته‌شدهٔ ۰/۷۱ را برای تقویت گرادیان در ۲۰۱۷ و ۰/۷۲ را در ۲۰۰۹ چاپ کرده بود. مدل تقویت گرادیان ما روی متغیرهای عمومی متناظر با فهرست چاپی به ۰/۲۰ و ۰/۰۸ روی خانوارهای تصادفی کنار گذاشته‌شده و ۰/۱۸ هنگام کنار گذاشتن کل ایالت‌ها رسید (شکل 7). طرح اعتبارسنجی شکاف را توضیح نمی‌دهد: حتی روی خانوارهای آموزش مدل در ۲۰۱۷ تنها ۰/۲۹ را توضیح داد و رگرسیون خطی ما ۰/۲۰ گرفت در برابر ۰/۱۱ چاپی، پس هدف یا متغیرها فرق دارند. جدول متغیرهای چاپی درصدهای بالاتر از صد و متغیری دوبار فهرست‌شده دارد، بنابراین نمی‌توانیم بگوییم کدام متغیرها وارد مدل چاپی شدند و دربارهٔ درستی آن یا دربارهٔ نشت ادعایی نمی‌کنیم.



شکل 7: ضریب تعیین کنار گذاشته‌شدهٔ چهار مدل رگرسیونی برای خرید اینترنتی خانوار در ۲۰۰۹ (پنل A) و ۲۰۱۷ (پنل B): مقدارهای چاپی و مقدارهای ما برای کنار گذاشتن تصادفی و کنار گذاشتن کل ایالت‌ها.

بحث

نردبان چه می‌گوید. برای یک رگرسیون معمولی NHTS، عددهای چاپی از فایل‌های عمومی بازیافتنی‌اند، به شرط آن‌که خواننده بداند نویسندگان وزن داده‌اند یا نه، که مقاله‌ها را گاه باید دوبار خواند تا بدهند؛ و هر جا دو برآوردگر از هم تشخیص‌پذیرند، برآوردگر اعلام‌شده عددهای چاپی را بازتولید می‌کند. پلهٔ بعد جایی است که یافته‌ها جابه‌جا می‌شوند. خطای معیار طراحی‌آگاه به‌طور میانه ۲٫۲ برابر خطای معیار بدون وزن است و زیر استنباط اعلام‌شدهٔ خود مطالعه‌ها ۲۴ درصد ضرایب معنادار زیر طرح معنادار نیستند. این‌ها خطا در مقاله‌های اصلی نیستند، که بیشترشان آنچه کرده‌اند گفته‌اند؛ نشان می‌دهند گزینش میان تحلیل بدون وزن و تحلیل وزنیِ طراحی‌آگاه پیامدهایی دارد که مقاله‌ها کم دربارهٔ آن بحث می‌کنند.

خطای معیار چه معنایی دارد. تحلیل‌های بدون وزن ارتباطی را میان پاسخ‌گویان برآورد می‌کنند و خطای معیارهایشان برای آن برآوردشونده تقریباً درست است (به‌طور میانه فقط ۱٫۱۰ برابر کمتر از واقع). خطای معیارهای طراحی‌آگاه بزرگ‌اند چون برآورد ارتباط جمعیتی با وزن‌های نابرابر دقت کمتری دارد. برنامه‌ریزی که می‌خواهد تقاضا را پیش‌بینی کند ارتباط جمعیتی را لازم دارد و بنابراین برازش وزنی با عدم‌قطعیت صادقانه‌اش را؛ پژوهشگری که به پاسخ‌گویان علاقه دارد می‌تواند موجه بدون وزن بماند، اما باید بگوید. در نمونهٔ سواری‌خواهی ضریب زیر هر دو برازش کوچک است، معناداری‌اش، که سرخط بر آن استوار است، به برآوردگر بستگی دارد و برآورد وزنی یک‌چهارم کوچک‌تر از بدون وزن است. برای مطالعهٔ همتاسازی خودروکیلومتر، نمونه‌های همتاشدهٔ ما تضادهای چاپی را بازتولید نمی‌کنند، پس دربارهٔ کنار گذاشتن خودرو چیزی نمی‌گوییم.

چرا خوشه‌بندی اینجا کم اهمیت است. اشخاص یک خانوار مکان، درآمد و خودرو را مشترک دارند، اما پیامدهای مطالعه‌شده (شمار سفر، استفاده از وسیله) پس از ورود متغیرها در مدل درون خانوار اندکی همبسته‌اند و بسیاری ردیف‌ها عضو دیگری در نمونه ندارند. اثر وزن‌ها بر خطای معیار بسیار بزرگ‌تر است چون وزن‌های NHTS میان اشخاص به‌شدت فرق می‌کنند. وزن‌های تکرار لایه‌بندی و انتخاب نابرابر را می‌افزایند و حکم ضریب‌های جداگانه را عوض می‌کنند، پس هر جا هست باید به کار رود.

امتیازهای پیش‌بینی. بیش‌نمونه‌گیری پیش از تقسیم امتیاز درخت تصمیم، جنگل تصادفی و ماشین بردار پشتیبان را بیش از پانزده واحد درصد باد می‌کند، نشتی که در رشته‌های دیگر ثبت شده است (Demircioğlu, 2024, Kapoor و Narayanan, 2023). امتیازهای چاپی با هیچ‌یک از دو خط لوله بازتولید نمی‌شوند و دقت‌های چاپی علیه خط لولهٔ نشت‌دارند، بنابراین نتیجه هشدار است نه حکم دربارهٔ مطالعه‌های اصلی. خواننده می‌تواند هر خط لوله را با `pynhts.validate.audit_classifier` بسنجد که چند ثانیه می‌برد. وقتی داده پیش از بیش‌نمونه‌گیری بر پایهٔ خانوار تقسیم شود، جنگل تصادفی پذیرش سوخت جایگزین سطح زیر منحنی ۰٫۶۳ دارد، نزدیک رگرسیون لجستیک (۰٫۶۷)، که متوسط است و پیش از به‌کارگیری مدل برای هدف‌گذاری سیاست پذیرش دانستنی. شکاف امتیاز رگرسیونی مطالعهٔ خرید اینترنتی از نوع دیگری است: طرح اعتبارسنجی آن را توضیح نمی‌دهد.

پیامد برای بسته. این مطالعه به ابزارهایی نیاز داشت که در بسته نبود: رگرسیونی که برآورد وزنی را با طرح واریانس موج برمی‌گرداند، بررسی نشت در خط لوله‌های پیش‌بینی، و راهی برای مقایسهٔ عدد چاپی با عدد بازسازی‌شده. دو ابزار نخست در pynhts با نام‌های `pynhts.models` (اکنون با برازش‌های بدون وزن، خوشه‌ای، وزنی و طراحی‌آگاه و هشدار هنگام اختلاف حکم‌ها) و `pynhts.validate` منتشر شده‌اند، برای طرح‌های وزن تکرار و تیلور در برابر R اعتبارسنجی شده‌اند و قاعده‌های راهنمای عامل بسته می‌گوید از آن‌ها استفاده شود. سه توسعهٔ دیگر از شواهد برمی‌آید و برنامه‌ریزی شده است: رابط فرمول و خانواده‌های ترتیبی و چندحالته برای مقاله‌های انتخاب گسسته که مجموعهٔ کنونی ندارد؛ مدل‌های چندسطحی در سطح خانوار با ادغام جزئی برای مقایسه‌های منطقه‌ای؛ و کتابخانه‌ای از دستورهای بازتولید، یکی برای هر مطالعهٔ منتشرشده، تا شکاف‌های ثبت‌شده به آزمون‌های رگرسیونی تبدیل شوند که هر کاربر بتواند اجرا کند. بسته گردایهٔ کاربردها را نیز با نام `pynhts.compendium` دربر دارد: ۴٬۳۰۱ مطالعهٔ متمایز با ویرایش‌ها، فصل‌ها و برچسب‌های واژه‌ایِ به‌کاررفته در این مقاله، جست‌وجو و خروجی BibTeX، ممیزی چگونگی تجزیهٔ گردایه، و یک مشاور که برای موضوعی مانند پذیرش خودروی برقی یا دوچرخهٔ اشتراکی فهرست می‌کند مطالعه‌های پیشین چه کرده‌اند و کدام یک از دام‌های اندازه‌گیری‌شده در

این مقاله به آن موضوع مربوط است. ۱۹ مطالعهٔ ممیزی‌شده‌ای که گردایه فهرست کرده به پلهٔ خود در نردبان پیوند خورده است تا هر ورودی دیگر نیز به همین شیوه بتواند دستور بازتولید بگیرد.

محدودیت‌ها. مجموعهٔ مطالعه‌ها مجموعه‌ای در دسترس از ۲۲ مطالعه است و ۱۳ مورد از آن‌ها دوباره برازش شد؛ سهم نتیجه‌های شکننده در همهٔ مقاله‌های NHTS بنابراین برآورد نشده است. سرشماری تلفیقی مطالعه‌هایی با اندازه‌های متفاوت را می‌آمیزد؛ نسخهٔ استنباط اعلام‌شده مقایسه را به آنچه هر مطالعه کرد محدود می‌کند. دو مطالعهٔ صفرافزوده را نتوانستیم درجه‌بندی کنیم چون نتیجهٔ کلیدی چاپی‌شان به ضریب نگاشته نمی‌شود؛ فقط اثر طرح را سهم می‌دهند. برازش‌های ما بعضی مقدارهای چاپی را فقط تقریبی بازتولید می‌کنند چون نمونه‌های چاپی دقیقاً بازسازی نمی‌شوند. مطالعهٔ همتاسازی از همتاسازی حریصانه استفاده کرد، نه همتاسازی بهینهٔ اصلی. آزمایش‌های یادگیری ماشین از پنج طبقه‌بند با ده آزمایش تنظیم برای هر کدام استفاده می‌کنند، جست‌وجویی متوسط که مقایسه را منصفانه می‌کند اما جامع نه، و ردیف‌های فرد را با وزن خانوار امتیاز می‌دهند. آگاهی‌بخش بودن وزن‌ها دربارهٔ پیامدها را نیازمودیم، که مبنای رسمی گزینش میان برآوردشوندهٔ وزنی و بدون وزن می‌بود (Bollen و همکاران, 2016). گزاره‌ها دربارهٔ چگونگی گزارش وزن و واریانس در ادبیات گسترده‌تر بر فهرست خودمان و برچسب‌های واژه‌ایِ چکیدهٔ گردایه تکیه دارند، نه بر مرور متن کامل: روی موردهای مثبتی که خواندیم دقت برچسب‌ها برای وزن ۹۴ درصد و برای برآوردگر واریانس ۵۰ درصد است (بازیابی سنجیده نشد)، چکیده‌ای که وزن را نام نمی‌برد ممکن است آن را به کار برده باشد، و گردایه فهرستی رسمی است نه همهٔ ادبیات. موج ۲۰۲۲ که وزن تکرار ندارد از خطی‌سازی استفاده می‌کند؛ پیاده و در برابر R سنجیده شده اما روی مطالعه‌های موردی به کار نرفته است. سرانجام، نتیجه‌ها را حساس به طرح می‌نامیم؛ نمی‌گوییم نادرست‌اند.

توصیه‌ها. برای نویسندگان: متغیر وزن را نام ببرید (`WTPERFIN` برای افراد، `WTHHFIN` برای خانوارها، `WTTRDFIN` برای سفرها)، بگویید وزن به کار رفته یا نه و چرا، و ضرایب برای کدام جمعیت‌اند؛ خطای معیار را از طرح پیمایش گزارش کنید (وزن‌های تکرار، یا لایه و خوشه برای موج ۲۰۲۲) یا، هر جا هیچ‌یک به کار نمی‌رود، با خطای معیار خوشه‌ایِ خانوار و این اعلام که تقریبی است؛ زیرگروه را به‌صورت یک حوزه از نمونهٔ کامل تحلیل کنید، نه با کنار گذاشتن ردیف‌های دیگر؛ و جدول جریان نمونه را با شمار رکوردهایی که هر قاعده حذف می‌کند بدهید. برای مدل‌های پیش‌بینی: پیش از هر کار بر پایهٔ خانوار تقسیم کنید، تنها ردیف‌های آموزش را بیش‌نمونه‌گیری کنید، امتیاز وزنی گزارش کنید و هر مدل را یکسان تنظیم کنید. برای داوران: ستارهٔ معناداری از رگرسیون بدون وزن NHTS گزاره‌ای دربارهٔ پاسخ‌گویان است و سطح زیر منحنی از تقسیم تصادفی ردیف‌های بیش‌نمونه‌گیری‌شده باید پرسشی برانگیزد.

نتیجه‌گیری

نردبان چهارپلهٔ بازتولیدپذیری با برآورد دوبارهٔ طراحی‌آگاه نگرانی کلی دربارهٔ پژوهش حمل‌ونقل مبتنی بر پیمایش را به عدد تبدیل می‌کند. بر ۲۲ مطالعهٔ منتشرشدهٔ NHTS، عددهای چاپی با برآوردگر اعلام‌شده، هر جا برآوردگرها از هم تشخیص‌پذیر بودند، بازیافتنی بودند؛ خطای معیار طراحی‌آگاه حدود ۲٫۲ برابر خطای معیار بدون وزن بود، تقریباً یکسره به‌سبب نابرابری وزن‌ها؛ ۲۴ درصد ضرایبی که زیر استنباط خود مطالعه‌ها معنادار بودند زیر طرح معنادار نبودند؛ و بیش‌نمونه‌گیری پیش از تقسیم امتیاز طبقه‌بندهای پذیرش خودرو را تا ۰٫۴۴ سطح زیر منحنی باد کرد، در حالی که امتیازهای چاپی با هیچ‌یک از دو خط لوله بازتولید نشدند. این ابزار به‌صورت `pynhts.models` و `pynhts.validate` منتشر شده تا دیگر پژوهشگران همین بررسی را روی تحلیل‌های خود انجام دهند.

اعلان‌ها

اعلان استفاده از هوش مصنوعی مولد و فناوری‌های یاری‌رسان در تهیهٔ دست‌نوشته. در تهیهٔ این اثر نویسندگان از Claude (Anthropic) به‌عنوان دستیار برای برنامه‌ریزی مطالعه، نوشتن و اجرای کد تحلیل، غربالگری منابع، تجزیه و برچسب‌گذاری گردایهٔ کاربردها، و پیش‌نویس و ویرایش متن، و از ScienceDirect LeapSpace (Elsevier) برای جست‌وجوی ادبیات و پیش‌نویس بازنگری مقدمه و پیشینه استفاده کردند. پس از استفاده از این ابزارها، نویسندگان مواد تولیدشده (کد، نتایج، منابع و متن) را بازبینی، راستی‌آزمایی و ویرایش کردند و مسئولیت کامل محتوای اثر را می‌پذیرند. عددهای متن را کد از

پرونده‌های نتیجه وارد می‌کند و هر منبع در Crossref بازیابی و از نظر پس‌گرفته‌شدن بررسی شده است. [[AUTHOR-TODO: فهرست کامل ابزارهای هوش مصنوعی تأیید شود؛ پیش‌نویس بازنگری‌ای که برای این نسخه رسید نام Gemini را هم آورده است.]]

بیانیهٔ مشارکت نویسندگان (CRediT). [[AUTHOR-TODO: نقش‌های CRediT را خود نویسندگان بنویسند؛ تاکنون چیزی اعلام نشده است.]]

تعارض منافع. نویسندهٔ اول توسعه‌دهندهٔ بستهٔ pynhts است که ماژول‌هایش اینجا شرح شده و هر دو نویسنده در نوشتن یکی از مقاله‌های موردی همکاری داشته‌اند؛ آن مقاله در فهرست آمد و دوباره برآورد نشد. [[AUTHOR-TODO: جملهٔ استاندارد نبود تعارض منافع دیگر تأیید شود.]]

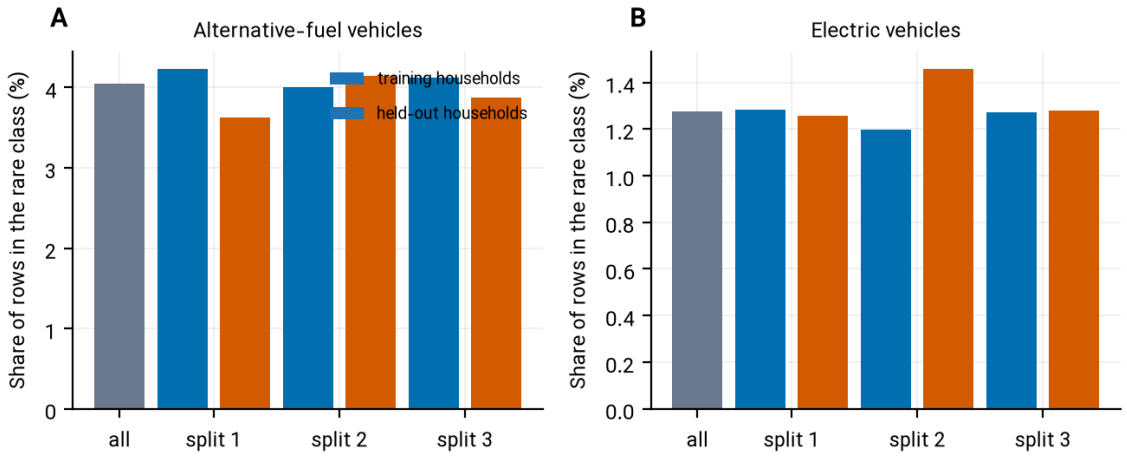
منابع مالی. این پژوهش از هیچ نهاد تأمین مالی در بخش دولتی، تجاری یا غیرانتفاعی کمک مالی ویژه‌ای دریافت نکرده است. [[AUTHOR-TODO: تأیید شود.]]

دسترسی به داده و کد. فایل‌های عمومی NHTS از پایگاه پیمایش در دسترس‌اند. ویرایش‌های گردایهٔ کاربردها نیز از پایگاه پیمایش در دسترس‌اند؛ داده‌های کتاب‌شناختی و برچسب‌های گردایه (بدون متن چکیده) همراه بسته منتشر می‌شوند. کد، جدول‌های نتیجه و بسته در مخزن باز pynhts (https://github.com/sabernaseralavi-60/pynhts) است و با شناسهٔ پایدار در Zenodo بایگانی شده است.

مواد تکمیلی

انحراف‌ها از طرح تحلیل (گزارش کامل در `analysis_plan.md`): سه مطالعه پیش از تثبیت طرح اجرا شده بودند؛ رواداری درجه‌بندی کفی برابر دو واحد رقم آخر چاپی داشت؛ فرضیهٔ سوم در برابر استنباط اعلام‌شدهٔ هر مطالعه ارزیابی شد (پیش‌نویس نخست در برابر برازش بدون وزن تلفیق می‌کرد)؛ فرضیهٔ چهارم پس از آن‌که معلوم شد امتیاز چاپی با هیچ خط لوله بازتولید نمی‌شود بازنویسی شد؛ اجرای تنظیم‌شده دو بذر داشت؛ تجزیهٔ نسبت خطای معیار، برازش خوشه‌ایِ بدون وزن و سنجش نیوتن در برابر برازش دوبارهٔ دقیق پس از یک بازبینی بیرونی افزوده شد؛ هدف‌های چاپی دو مطالعهٔ صفرافزوده کنار گذاشته شد چون به ضریب نگاشته نمی‌شوند.

جست‌وجوی ادبیات. پیشینهٔ پژوهش بر سه جست‌وجو در اکتبر ۲۰۲۶ و یک ممیزی فهرستی رسمی تکیه دارد: جست‌وجوی واژه‌ای و ارجاعی در OpenAlex و Crossref دربارهٔ وزن‌های پیمایش، واریانس تکرار، بازتولیدپذیری و نشت (گزارش پرسش‌ها در مخزن، `lit/search_log.md`)؛ آزمون تکراری‌بودن پرسش مقاله در برابر دوازده مقالهٔ همسایه که مقاله‌ای را نیافت که رگرسیون‌ها یا طبقه‌بندهای منتشرشدهٔ NHTS را زیر طرح پیمایش دوباره برآورد کرده باشد؛ و مرور ScienceDirect LeapSpace دربارهٔ چگونگی رفتار مطالعه‌های NHTS با طرح، وزن و واریانس. منبع چهارم ممیزی فهرستی رسمی است، یعنی گردایهٔ کاربردهای NHTS از FHWA/ORNL: دوازده ویرایش PDF آن (۲۰۰۶ تا ۲۰۲۳) صفحه به صفحه و با لنگر جدول مطالب هر ویرایش تجزیه شد و با برون‌داد جست‌وجوی تعاملی ۲۰۲۳ تا ۲۰۲۵ ادغام گردید؛ ورودی‌ها با عنوان حذف تکرار شدند، با قواعد واژه‌ای برای موج‌های NHTS، خانواده‌های مدل، وزن، برآوردگرهای واریانس و واژه‌های اعتبارسنجی برچسب خوردند، و شمارهای تجزیه‌شده با شمارش‌های مستقل هر ویرایش مقایسه شد (جدول 8). کد، قواعد، جمع‌های آزمون و خوانش موردهای مثبت در مخزن است (`compendium/`، `tools/compendium/`). جست‌وجوهای OpenAlex و LeapSpace برای یافتن ادبیات بودند؛ برای شمردن این‌که وزن یا برآوردگر واریانس چند بار نام برده می‌شود به کار نرفتند. هر منبع ذکرشده در Crossref با فراداده‌اش بازیابی و در OpenAlex از نظر پس‌گرفته‌شدن بررسی شده است (`lit/verified.csv`، `lit/verify_refs.py`)؛ منبع‌های بازیابی‌نشده و نامزدهایی که موضوعشان از جمله‌ای که برایش پیشنهاد شده بود پشتیبانی نمی‌کرد کنار گذاشته شدند. این مرور سامانمند رسمی نیست.



شکل 8: هیستوگرام فراوانی رستهٔ کمیاب (سهم خودروهای سوخت جایگزین، پنل A، و برقی، پنل B) در همهٔ ردیف‌ها، در خانوارهای آموزش و در خانوارهای کنار گذاشته‌شدهٔ هر یک از سه تقسیم.

جدول 7: ابرپارامترهای برگزیدهٔ مدل‌های تنظیم‌شده در خط لولهٔ تقسیم خانوار (بذر اصلی نخست). C: هزینه یا قدرت جریمه؛ vs: کف واریانس؛ crit، leaf، depth: عمق درخت، کمینهٔ اندازهٔ برگ و معیار شکست؛ mf، n: شمار درختان و سهم ورودی در هر شکست؛ gamma: پهنای هسته.

پیامد	مدل	مقدارهای برگزیده	AUC اعتبارسنجی درونی
سوخت جایگزین	LR	penalty_l1 = False C = ۰٫۰۲۵۶’	۰٫۶۷
سوخت جایگزین	NB	vs = ۱٫۷۹e-۰۹	۰٫۶۳
سوخت جایگزین	DT	crit = entropy leaf = ۵۰’ depth = ۳’	۰٫۶۳
سوخت جایگزین	RF	mf = ۰٫۳ leaf = ۱’ depth = ۱۰’ n = ۲۰۰’	۰٫۶۵
سوخت جایگزین	SVM	gamma = ۰٫۰۰۴۰۷ C = ۲٫۵۹’	۰٫۶۲
برقی	LR	penalty_l1 = True C = ۰٫۰۱۷۵’	۰٫۷۵
برقی	NB	vs = ۱٫۷۹e-۰۹	۰٫۷۲
برقی	DT	crit = entropy leaf = ۵۰’ depth = ۳’	۰٫۶۸
برقی	RF	mf = ۰٫۳ leaf = ۱’ depth = ۱۰’ n = ۲۰۰’	۰٫۷۴
برقی	SVM	gamma = ۰٫۰۰۴۰۷ C = ۲٫۵۹’	۰٫۷۲

جدول 8: ممیزی استخراج گردایه: ورودی‌های استخراج‌شده از هر ویرایش PDF (و ردیف‌های برون‌داد جست‌وجوی ۲۰۲۳ تا ۲۰۲۵) در برابر شمارش‌های مستقل. خط تیره یعنی ویرایش چنین شمارشی ندارد.

ویرایش	ورودی‌های استخراج‌شده	لنگرهای چکیده	برچسب عنوان / ۲	شمار اعلام‌شده در ویرایش	بیشترین اختلاف
۲۰۰۶–۲۰۱۱	۱۰۶	۱۰۶	-	-	۰
۲۰۱۲	۲۱۰	۲۰۸	-	۲۱۰	۰
۲۰۱۳	۲۸۳	۲۸۳	-	۲۸۲	۱
۲۰۱۴	۳۲۲	۳۲۱	-	۳۲۲	۱
۲۰۱۵	۳۷۷	۳۷۷	-	۳۷۹	۲
۲۰۱۶	۳۷۶	۳۷۶	۳۷۶	۳۷۵	۱
۲۰۱۸	۱۷۲	۱۷۲	۱۷۲	-	۰
۲۰۱۹	۱۹۲	۱۹۲	۱۹۲	-	۰
۲۰۲۰	۴۱۲	۴۱۲	۴۱۲	-	۱
۲۰۲۱	۴۵۴	۴۵۴	۴۵۴	-	۱
۲۰۲۲	۴۲۰	۴۲۰	۴۲۰	-	۱
pdf-۲۰۲۳	۲۱۴	۲۱۴	۲۱۴	-	۰
۲۰۲۳	۳۸۱	-	-	-	۰
۲۰۲۴	۳۶۳	-	-	-	۰
۲۰۲۵	۳۳۷	-	-	-	۰

کتاب‌نامه

Alberini, A., Teja Burra, L., Cirillo, C., Shen, C., 2021. Counting vehicle miles traveled: What can we learn from the NHTS?. Transportation Research Part D: Transport and Environment 98, 102984.. https://doi.org/10.1016/j.trd.2021.102984

Barajas, J.M., 2021. The effects of driver licensing laws on immigrant travel. Transport Policy 105, 22–34.. https://doi.org/10.1016/j.tranpol.2021.02.010

Barua, L., Zou, B., Zhou, Y., Liu, Y., 2021. Modeling household online shopping demand in the U.S.: a machine learning approach and comparative investigation between 2009 and 2017. Transportation 50, 437–476.. https://doi.org/10.1007/s11116-021-10250-z

Bergtold, J.S., Pokharel, K.P., Featherstone, A.M., Mo, L., 2017. On the examination of the reliability of statistical software for estimating regression models with discrete dependent variables. Computational Statistics 33, 757–786.. https://doi.org/10.1007/s00180-017-0776-5

Birrell, C.L., Steel, D.G., Batterham, M.J., Arya, A., 2019. How to use replicate weights in health survey analysis using the National Nutrition and Physical Activity Survey as an example. Public Health Nutrition 22, 3315–3326.. https://doi.org/10.1017/s1368980019001927

Blagus, R., Lusa, L., 2013. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 14, 106.. https://doi.org/10.1186/1471-2105-14-106

Bollen, K.A., Biemer, P.P., Karr, A.F., Tueller, S., Berzofsky, M.E., 2016. Are Survey Weights Needed? A Review of Diagnostic Tests in Regression Analysis. Annual Review of Statistics and Its Application 3, 375–392.. https://doi.org/10.1146/annurev-statistics-011516-012958

Brogan, D., 2005. Software for Sample Survey Data, Misuse of Standard Packages. Encyclopedia of Biostatistics.. https://doi.org/10.1002/0470011815.b2a16072

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16, 321–357.. https://doi.org/10.1613/jair.953

Christensen, K.M., Chamberlain, B., Park, K., Abrishami, M., Sheen, J.C., Larsen, T., 2025. Surveying people with disabilities: Insights on methods and challenges. Transportation Research Interdisciplinary Perspectives 34, 101702.. https://doi.org/10.1016/j.trip.2025.101702

Cohen, S.B., Burt, V.L., Jones, G.K., 1986. Efficiencies in Variance Estimation for Complex Survey Data. The American Statistician 40, 157–164.. https://doi.org/10.1080/00031305.1986.10475383

Collia, D.V., Sharp, J., Giesbrecht, L., 2003. The 2001 National Household Travel Survey: A look into the travel patterns of older Americans. Journal of Safety Research 34, 461–470.. https://doi.org/10.1016/j.jsr.2003.10.001

Conway, M.W., Salon, D., King, D.A., 2018. Trends in Taxi Use and the Advent of Ridehailing, 1995–2017: Evidence from the US National Household Travel Survey. Urban Science 2, 79.. https://doi.org/10.3390/urbansci2030079

de Sousa, M.H., da Silva, N.N., 2000. Comparação de softwares para análise de dados de levantamentos complexos. Revista de Saúde Pública 34, 646–653.. https://doi.org/10.1590/s0034-89102000000600013

Deka, D., Fei, D., 2019. A comparison of the personal and neighborhood characteristics associated with ridesourcing, transit use, and driving with NHTS data. Journal of Transport Geography 76, 24–33.. https://doi.org/10.1016/j.jtrangeo.2019.03.001

Demircioğlu, A., 2024. Applying oversampling before cross-validation will lead to high bias in radiomics. Scientific Reports 14, 11563.. https://doi.org/10.1038/s41598-024-62585-z

Esekhaigbe, E., Bills, T.S., Broaddus, A., 2026. Group representation of vulnerable travelers in household travel surveys: an assessment of 43 U.S. surveys. Transportation.. https://doi.org/10.1007/s11116-026-10795-x

Fu, Z., Liu, W., Zhu, J., Yuan, Y., 2020. Trip Chain Simulation of Electric Vehicles based on NHTS Data. 2020 12th IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC) 1–6.. https://doi.org/10.1109/appeec48164.2020.9220726

Gard, A.M., Hyde, L.W., Heeringa, S.G., West, B.T., Mitchell, C., 2023. Why weight? Analytic approaches for large-scale population neuroscience data. Developmental Cognitive Neuroscience 59, 101196.. https://doi.org/10.1016/j.dcn.2023.101196

Gelman, A., 2007. Struggles with Survey Weighting and Regression Modeling. Statistical Science 22, 153–164.. https://doi.org/10.1214/088342306000000691

Hagenauer, J., Helbich, M., 2017. A comparative study of machine learning classifiers for modeling travel mode choice. Expert Systems with Applications 78, 273–282.. https://doi.org/10.1016/j.eswa.2017.01.057

Hardwicke, T.E., Bohn, M., MacDonald, K., Hembacher, E., Nuijten, M.B., Peloquin, B.N., deMayo, B.E., Long, B., Yoon, E.J., Frank, M.C., 2021. Analytic reproducibility in articles receiving open data badges at the journal Psychological Science : an observational study. Royal Society Open Science 8, 201494.. https://doi.org/10.1098/rsos.201494

Hörl, S., Balac, M., 2021. Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data. Transportation Research Part C: Emerging Technologies 130, 103291.. https://doi.org/10.1016/j.trc.2021.103291

Jia, J., 2019. Analysis of Alternative Fuel Vehicle (AFV) Adoption Utilizing Different Machine Learning Methods: A Case Study of 2017 NHTS. IEEE Access 7, 112726–112735.. https://doi.org/10.1109/ACCESS.2019.2934780

Jia, J., Shi, B., Che, F., Zhang, H., 2020. Predicting the Regional Adoption of Electric Vehicle (EV) With Comprehensive Models. IEEE Access 8, 147275–147285.. https://doi.org/10.1109/ACCESS.2020.3014851

Jiao, J., Bischak, C., Hyden, S., 2020. The impact of shared mobility on trip generation behavior in the US: Findings from the 2017 National Household Travel Survey. Travel Behaviour and Society 19, 1–7.. https://doi.org/10.1016/j.tbs.2019.11.001

Johnson, D.R., Elliott, L.A., 1998. Sampling Design Effects: Do They Affect the Analyses of Data from the National Survey of Families and Households?. Journal of Marriage and the Family 60, 993.. https://doi.org/10.2307/353640

Kapoor, S., Cantrell, E.M., Peng, K., Pham, T.H., Bail, C.A., Gundersen, O.E., Hofman, J.M., Hullman, J., Lones, M.A., Malik, M.M., Nanayakkara, P., Poldrack, R.A., Raji, I.D., Roberts, M., Salganik, M.J., Serra-Garcia, M., Stewart, B.M., Vandewiele, G., Narayanan, A., 2024. REFORMS: Consensus-based Recommendations for Machine-learning-based Science. Science Advances 10, eadk3452.. https://doi.org/10.1126/sciadv.adk3452

Kapoor, S., Narayanan, A., 2023. Leakage and the reproducibility crisis in machine-learning-based science. Patterns 4, 100804.. https://doi.org/10.1016/j.patter.2023.100804

Kim, S., Ulfarsson, G.F., 2021. Staying home or going places: Mobility factors of older minority women's daily trip making in the United States. Journal of Transport & Health 21, 101031.. https://doi.org/10.1016/j.jth.2021.101031

Korn, E.L., Graubard, B.I., 1999. Analysis of Health Surveys. Wiley.. https://doi.org/10.1002/9781118032619

Krizek, K.J., McGuckin, N., 2019. Shedding NHTS Light on the Use of 'Little Vehicles' in Urban Areas. Transport Findings.. https://doi.org/10.32866/10777

Krumm, J., 2012. How People Use Their Vehicles: Statistics from the 2009 National Household Travel Survey. SAE Technical Paper Series 1, 2012–01–0489.. https://doi.org/10.4271/2012-01-0489

Lambert, D., 1992. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. Technometrics 34, 1.. https://doi.org/10.2307/1269547

Lidbe, A., Adanu, E.K., Penmetsa, P., Jones, S., 2021. Changes in the travel patterns of older Americans with medical conditions: A comparison of 2001 and 2017 NHTS data. Transportation Research Interdisciplinary Perspectives 11, 100463.. https://doi.org/10.1016/j.trip.2021.100463

Lumley, T., 2004. Analysis of Complex Survey Samples. Journal of Statistical Software 9.. https://doi.org/10.18637/jss.v009.i08

Martín-Baos, J.Á., López-Gómez, J.A., Rodriguez-Benitez, L., Hillel, T., García-Ródenas, R., 2023. A prediction and behavioural analysis of machine learning methods for modelling travel mode choice. Transportation Research Part C: Emerging Technologies 156, 104318.. https://doi.org/10.1016/j.trc.2023.104318

McGuckin, N., 2021. The National Household Travel Survey Data Series (NPTS/NHTS). International Encyclopedia of Transportation 493–495.. https://doi.org/10.1016/B978-0-08-102671-7.10372-0

Naseralavi, S., Soltanirad, M., Ranjbar, E., Lucero, M., Gorzin, F., Hakiminejad, Y., Azimi, S., Baghersad, M., Mazaheri, A., 2025. Machine Learning Modeling of Household Trip Generation by State Using NHTS Data. Urban Science 9, 353.. https://doi.org/10.3390/urbansci9090353

Nishikawa, A., Tawfik, A.M., 2022. An Exploratory Comparison of International Travel Behavior Differences: Travel Activity, Travel Time, and Travel Distance in Great Britain, New Zealand, and the United States. International Conference on Transportation and Development 2022 122–132.. https://doi.org/10.1061/9780784484340.011

Peng, R.D., 2011. Reproducible Research in Computational Science. Science 334, 1226–1227.. https://doi.org/10.1126/science.1213847

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., Pélissier, R., 2020. Spatial

validation reveals poor predictive performance of large-scale ecological mapping models. Nature Communications 11, 4540.. https://doi.org/10.1038/s41467-020-18321-y

Porschen, M., Kuhnimhof, T., 2026. The State of National Travel Surveys: Design Characteristics, Impact of the Pandemic and the Move to New Technologies. Transportation Research Procedia 97, 99–109.. https://doi.org/10.1016/j.trpro.2026.04.147

Porter, A.K., Kontou, E., McDonald, N.C., Evenson, K.R., 2020. Perceived barriers to commuter and exercise bicycling in U.S. adults: The 2017 National Household Travel Survey. Journal of Transport & Health 16, 100820.. https://doi.org/10.1016/j.jth.2020.100820

Pucher, J., Renne, J.L., 2005. Rural mobility and mode choice: Evidence from the 2001 National Household Travel Survey. Transportation 32, 165–186.. https://doi.org/10.1007/s11116-004-5508-3

Qawasmeh, B., Qawasmeh, S., Al Tawil, A., Qawasmeh, D., 2024. Estimation of Trip-based Generation Models and Calibration of Mode Choice Models for the American Travel Behavior. The Open Transportation Journal 18, e26671212348473.. https://doi.org/10.2174/0126671212348473241010122638

Quinn, T.D., Jakicic, J.M., Fertman, C.I., Barone Gibbs, B., 2016. Demographic factors, workplace factors and active transportation use in the USA: a secondary analysis of 2009 NHTS data. Journal of Epidemiology and Community Health 71, 480–486.. https://doi.org/10.1136/jech-2016-207820

Richardson, S., Lin, T., Li, Y., Niu, X., Xu, M., Stander, V., Tu, X.M., 2019. Guidance for use of weights: an analysis of different types of weights and their implications when using SAS PROCs. General Psychiatry 32, e100038.. https://doi.org/10.1136/gpsych-2018-100038

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929.. https://doi.org/10.1111/ecog.02881

Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S., Scheinost, D., 2024. Data leakage inflates prediction performance in connectome-based machine learning models. Nature Communications 15, 1829.. https://doi.org/10.1038/s41467-024-46150-w

Rust, K., Rao, J., 1996. Variance estimation for complex surveys using replication techniques. Statistical Methods in Medical Research 5, 283–310.. https://doi.org/10.1177/096228029600500305

Sabouri, S., Brewer, S., Ewing, R., 2020. Exploring the relationship between ride-sourcing services and vehicle ownership, using both inferential and machine learning approaches. Landscape and Urban Planning 198, 103797.. https://doi.org/10.1016/j.landurbplan.2020.103797

Sadeghvaziri, E., Tawfik, A., 2020. Using the 2017 National Household Travel Survey Data to Explore the Elderly's Travel Patterns. International Conference on Transportation and Development 2020 86–94.. https://doi.org/10.1061/9780784483169.008

Schweinsberg, M., Feldman, M., Staub, N., Akker, O.R. van den, Aert, R.C. van, Assen, M.A. van, Liu, Y., Althoff, T., Heer, J., Kale, A., Mohamed, Z., Amireh, H., Venkatesh Prasad, V., Bernstein, A., Robinson, E., Snellman, K., Amy Sommer, S., Otner, S.M., Robinson, D., Madan, N., Silberzahn, R., Goldstein, P., Tierney, W., Murase, T., Mandl, B., Viganola, D., Strobl, C., Schaumans, C.B., Kelchtermans, S., Naseeb, C., Mason Garrison, S., Yarkoni, T., Richard Chan, C., Adie, P., Alaburda, P., Albers, C., Alspaugh, S., Alstott, J., Nelson, A.A., Rubia, E. Ariño de la, 2021. Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. Organizational Behavior and Human Decision Processes 165, 228–249.. https://doi.org/10.1016/j.obhdp.2021.02.003

Shaw, F.A., Wang, X., Mokhtarian, P.L., Watkins, K.E., 2021. Supplementing transportation data sources with targeted marketing data: Applications, integration, and internal validation. Transportation Research Part A: Policy and Practice 149, 150–169.. https://doi.org/10.1016/j.tra.2021.04.021

Shirgaokar, M., Nobler, E., 2021. Differences in daily trips between immigrants and US-born individuals: Implications for social integration. Transport Policy 105, 103–114.. https://doi.org/10.1016/j.tranpol.2021.03.008

Silberzahn, R., Uhlmann, E.L., Martin, D.P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M.A., Dalla Rosa, A., Dam, L., Evans, M.H., Flores Cervantes, I., Fong, N., Gamez-Djokic, M., Glenz, A., Gordon-McKeon, S., Heaton, T.J., Hederos, K., Heene, M., Hofelich Mohr, A.J., Högden, F., Hui, K., Johannesson, M., Kalodimos, J., Kaszubowski, E., Kennedy, D.M., Lei, R., Lindsay, T.A., Liverani, S., Madan, C.R., Molden, D., Molleman, E., Morey, R.D., 2018. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. Advances in Methods and Practices in Psychological Science 1, 337–356.. https://doi.org/10.1177/2515245917747646

Simonsohn, U., Simmons, J.P., Nelson, L.D., 2020. Specification curve analysis. Nature Human Behaviour 4, 1208–1214.. https://doi.org/10.1038/s41562-020-0912-z

Solon, G., Haider, S.J., Wooldridge, J.M., 2015. What Are We Weighting For?. Journal of Human Resources 50, 301–316.. https://doi.org/10.3368/jhr.50.2.301

Steegen, S., Tuerlinckx, F., Gelman, A., Vanpaemel, W., 2016. Increasing Transparency Through a Multiverse Analysis. Perspectives on Psychological Science 11, 702–712.. https://doi.org/10.1177/1745691616658637

Stodden, V., Seiler, J., Ma, Z., 2018. An empirical analysis of journal policy effectiveness for computational reproducibility. Proceedings of the National Academy of Sciences 115, 2584–2589.. https://doi.org/10.1073/pnas.1708290115

Su, R., Xiao, J., McBride, E.C., Goulias, K.G., 2021. Understanding senior's daily mobility patterns in California using human mobility motifs. Journal of Transport Geography 94, 103117.. https://doi.org/10.1016/j.jtrangeo.2021.103117

Svaboe, G.B.A., Tørset, T., Lohne, J., 2024. A comparative study of national travel surveys in six European countries. Transportation Planning and Technology 47, 400–418.. https://doi.org/10.1080/03081060.2024.2311081

Tribby, C.P., Tharp, D.S., 2019. Examining urban and rural bicycling in the United States: Early findings from the 2017 National Household Travel Survey. Journal of Transport & Health 13, 143–149.. https://doi.org/10.1016/j.jth.2019.03.015

Trowbridge, M.J., Kent, R., 2009. Rear-Seat Motor Vehicle Travel in the U.S.. American Journal of Preventive Medicine 37, 321–323.. https://doi.org/10.1016/j.amepre.2009.05.021

Tumula, R.R., Burra, L.T., Gkritza, K., 2026. Who continues to shop online? Sociodemographic determinants of post-COVID e-commerce. Transportation Research Interdisciplinary Perspectives 38, 102084.. https://doi.org/10.1016/j.trip.2026.102084

Vabalas, A., Gowen, E., Poliakoff, E., Casson, A.J., 2019. Machine learning algorithm validation with a limited sample size. PLOS ONE 14, e224365.. https://doi.org/10.1371/journal.pone.0224365

Vaish, A.K., Chen, S., Sathe, N.S., Folsom, R.E., Chandhok, P., Guo, K., 2010. Small area estimates of daily person-miles of travel: 2001 National Household Transportation Survey. Transportation 37, 825–848.. https://doi.org/10.1007/s11116-010-9279-8

Verreault, H., Morency, C., 2016. Integration of a phone-based household travel survey and a web-based student travel survey. Transportation 45, 89–103.. https://doi.org/10.1007/s11116-016-9726-2

Wang, S., Mo, B., Zheng, Y., Hess, S., Zhao, J., 2024. Comparing hundreds of machine learning and discrete choice models for travel demand modeling: An empirical benchmark. Transportation Research Part B: Methodological 190, 103061.. https://doi.org/10.1016/j.trb.2024.103061

Wang, X., Renne, J.L., 2023. Socioeconomics of urban travel in the U.S.: Evidence from the 2017 NHTS. Transportation Research Part D: Transport and Environment 116, 103622.. https://doi.org/10.1016/j.trd.2023.103622

Wang, X., Shaw, F.A., Mokhtarian, P.L., Watkins, K.E., 2022. Response willingness in consecutive travel surveys: an investigation based on the National Household Travel Survey using a sample selection model. Transportation 50, 2339–2373.. https://doi.org/10.1007/s11116-022-10312-w

Winship, C., Radbill, L., 1994. Sampling Weights and Regression Analysis. Sociological Methods & Research 23, 230–257.. https://doi.org/10.1177/0049124194023002004

Wolter, K.M., 2007. Introduction to Variance Estimation. Springer.. https://doi.org/10.1007/978-0-387-35099-8

Woodruff, R.S., 1971. A Simple Method for Approximating the Variance of a Complicated Estimate. Journal of the American Statistical Association 66, 411–414.. https://doi.org/10.1080/01621459.1971.10482279

Wu, X., MacKenzie, D., 2021. Assessing the VMT effect of ridesourcing services in the US. Transportation Research Part D: Transport and Environment 94, 102816.. https://doi.org/10.1016/j.trd.2021.102816

Xue, C., Wu, Q., Sun, M., Bai, P., Chen, Y., 2021. The Interaction between E-Shopping and Shopping Trips: An Analysis with 2017 NHTS. Complexity 2021, 8247158.. https://doi.org/10.1155/2021/8247158

Yang, Y., 2025. Public transit use among American older adults: Insights from a nationally representative survey. Journal of Transport & Health 41, 101982.. https://doi.org/10.1016/j.jth.2025.101982

Yang, Y., Diez Roux, A.V., Bingham, C.R., 2011. Variability and seasonality of active transportation in USA: evidence from the 2001 NHTS. International Journal of Behavioral Nutrition and Physical Activity 8, 96.. https://doi.org/10.1186/1479-5868-8-96

Zeileis, A., Kleiber, C., Jackman, S., 2008. Regression Models for Count Data in R. Journal of Statistical Software 27.. https://doi.org/10.18637/jss.v027.i08

Zhang, Y., Zhang, Y., 2018a. Associations between Public Transit Usage and Bikesharing Behaviors in The United States. Sustainability 10, 1868.. https://doi.org/10.3390/su10061868

Zhang, Y., Zhang, Y., 2018b. Examining the Relationship between Household Vehicle Ownership and Ridesharing Behaviors in the United States. Sustainability 10, 2720.. https://doi.org/10.3390/su10082720

Zhang, Y., Zhang, Y., 2018c. Exploring the Relationship between Ridesharing and Public Transit Use in the United States. International Journal of Environmental Research and Public Health 15, 1763.. https://doi.org/10.3390/ijerph15081763

Zhang, Z., Maruyama, T., 2026. Immobility or soft refusal? An empirical analysis of the association between respondents' diligence and reported immobility in household travel surveys. Transportation Research Part A: Policy and Practice 204, 104815.. https://doi.org/10.1016/j.tra.2025.104815

Zhou, Y., Wang, X.(., 2014. Explore the relationship between online shopping and shopping trips: An analysis with the 2009 NHTS data. Transportation Research Part A: Policy and Practice 70, 1–9.. https://doi.org/10.1016/j.tra.2014.09.014

Zou, Z., Cirillo, C., 2021. Does ridesourcing impact driving decisions: A survey weighted regression analysis. Transportation Research Part A: Policy and Practice 146, 1–12.. https://doi.org/10.1016/j.tra.2021.02.006