The Journal of Transport and Land Use jtlu.org
Vol. 17 No. 1 [2024] pp. 369–399

JTLU

Sydney's residential relocation landscape: Machine learning and feature selection methods unpack the whys and whens

Maryam Bostanara (corresponding author)
rCITI, UNSW Sydney
m.bostanara@unsw.edu.au

Amarin Siripanich
rCITI, UNSW Sydney
a.siripanich@unsw.edu.au

Milad Ghasri
UNSW Canberra
m.ghasri@unsw.edu.au

Taha Hossein Rashidi
rCITI, UNSW Sydney
rashidi@unsw.edu.au

Abstract: This study investigates household residential relocation timing, an aspect vital for transport and urban planning. Analyzing a high-dimensional dataset from 1,024 relocations in Sydney, Australia, the research contrasts ten machine learning survival techniques with three classical survival models. Results indicate that when classical models are paired with tree-based automated feature selectors, they align closely with machine learning outcomes. Notably, the GBM, XGBoost, and Random Forest models emerge as standout performers. The study provides a comprehensive comparison between automatic and manual feature selection, shedding light on variables influencing households' duration of stay. While stacked ensemble modeling, which leverages predictions from various models, is used to enhance accuracy, the improvements are marginal, underscoring inherent modeling challenges, particularly the recurring issue of misclassifying specific pairs of households in the concordance index measure. A thorough feature analysis highlights homeownership as the foremost predictor, underscoring the importance of recent life events and accessibility features in relocation decisions. The research emphasizes the importance of considering the accessibility of both current and future homes in relocation models, with 20% feature significance in model outcomes. Building on these foundational insights, the study paves the way for a deeper understanding of individual decision-making processes in sustainable urban planning.

Keywords: Residential relocation, machine learning, survival analysis, residential self-selection, accessibility

Article history:
Received: October 17, 2023
Accepted: January 15, 2024
Available online: May 17, 2024

1 Introduction

The decision of residential mobility is one of the most significant choices households make multiple times throughout their lives. This decision typically involves several sub-decisions, including the decision to leave the current home, choosing the new home's suburb and characteristics, determining the relocation timing, and more (Rashidi &

Copyright 2024 Maryam Bostanara, Amarin Siripanich, Milad Ghasri & Taha Hossein Rashidi
https://doi.org/10.5198/jtlu.2024.2440
ISSN: 1938-7849 | Licensed under the Creative Commons Attribution – Noncommercial License 4.0

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

Ghasri, 2017). While these decisions occur at the household level, collectively they shape the housing market, neighborhood dynamics, city environment, and broader policies (Lerman, 1975). Consequently, residential relocation has captivated researchers across various fields such as transport planning (Aditjandra et al., 2016), geography (Buckle, 2017), and economics (Sánchez & Andrews, 2011) for decades.

From a transport planning viewpoint, understanding how households make residential decisions and how these relate to their transport attitudes and trip generation attributes is crucial. However, modeling residential mobility behavior in a single study presents challenges due to its complexity. Among all sub-topics of residential mobility, this research zeroes in on the topic of residential relocation duration. Understanding this duration is essential for grasping residential mobility, where the timing of households' relocations is modeled in light of various factors, such as socio-demographic attributes, financial status, homeownership, and life-course events (Thomas et al., 2016; Tran et al., 2016), among others. This study seeks to elucidate the relationship between transport-related attitudes and residential mobility behavior, incorporating three key topics: accessibility, daily trip travel-time, and residential self-selection.

Accessibility is an influential feature in home selection (Schirmer et al., 2014). Three main accessibility measures have been employed in this study: 1) Households' Daily Trip Accessibility: This primarily considers the household's daily trips, not necessarily the accessibility of the home or neighborhood. Previous research has highlighted the strong correlation between households' residential relocation dynamics and their daily trip-making behavior (De Vos & Ettema, 2020). Hence, average travel times to work and school by households, as the two main reasons for daily trips (Cervero, 2003), have been estimated and used in the models. 2) Home 30-Minute Accessibility: Centered on home location, this metric counts the number of jobs within a 30-minute range, a recognized accessibility indicator (Levinson, 2019; Srour et al., 2002). This measure not only provides an estimate of the available job opportunities to an individual but also serves as an approximation of the number of jobs accessible for their service. 3) Suburb Land-Use: This reflects the relationship between residence location choice and suburb land-use patterns (Tran et al., 2016).

All these accessibility measures pertain to the current home's situation. This research hypothesizes that both the current and the anticipated next home's accessibility metrics play a significant role in relocation timing. A key contribution of this research is its exploration of households' tendencies to increase or decrease their home accessibility. Supported by a dataset that includes two locations per household, this study predicts the accessibility of the intended home, factoring in both current accessibility and socio-demographic attributes. Yet, this model may be susceptible to endogeneity bias due to residential self-selection.

Residential self-selection endogeneity bias has been extensively studied in the field of transport (Zhang, 2014). The bias is particularly prominent in studies examining the relationships between travel behavior and built environment features (Frank et al., 2006). While neighborhood characteristics clearly correlate with an individual's travel behavior, they don't necessarily cause these behaviors (Hedman & van Ham, 2012). An essential factor to consider is an individual's deliberate choice of a neighborhood. Failing to account for this intentional selection can lead to endogeneity bias in studies, especially when observed and unobserved explanatory variables are correlated (Mokhtarian & Cao, 2008). This self-selection bias is a well-recognized form of endogeneity bias in residential mobility research.

One acknowledged methodology for mitigating such bias is direct questioning, which evaluates the extent to which respondents' preferences and attitudes influence their choice of suburb. This study addresses the self-selection bias using this approach.

Specifically, we've included the households' reported preferences regarding proximity to amenities such as shops, public transport, and workplaces. Consequently, the accessibility of the next home for every household is predicted based on both their stated residential preferences and their current home's accessibility, factored alongside socio-demographic attributes.

Methodologically, this research employs both classical survival analysis tools and ten machine learning survival approaches, given their proficiency in handling censored data. Machine learning, a subset of artificial intelligence, has become a powerful toolset for data analysis in recent decades. Over the years, machine learning has garnered significant attention from both theoretical and applied research communities, elevating its importance and application range. One of the most notable advancements in this domain has been the evolution of algorithms tailored to time-to-event or survival data (Gordon & Olshen, 1985; Sarkar et al., 2021). These advancements present a unique opportunity to enhance classical survival models and maximize the potential of heterogeneous and high-dimensional residential relocation data. Moreover, these models offer the ability to identify non-linear relationships (Kern et al., 2019), automate some challenging modeling steps such as feature selection (Liu et al., 2015), and shed light on the significance of different features. Although machine learning has seen widespread use in transport (Ding et al., 2018; Pineda-Jaramillo & Arbeláez-Arenas, 2022; Xue & Yao, 2022), its application to residential relocation duration remains uncharted. This research endeavors to benchmark machine learning approaches against classical survival models. These results are further juxtaposed with three classical parametric and semi-parametric survival models. In an added layer of analysis, the classical survival models are synergized with the feature selection models to gauge their combined potential.

Additionally, this study undertakes a comprehensive comparison of manual and automatic feature selection methods using ANOVA and bootstrap analysis. We also delve into ensemble models with the aim of enhancing predictive capability. Furthermore, an in-depth analysis of the concordant and discordant pairs within the concordance index of multiple models is conducted to elucidate the nuanced differences between these pairs.

2 Literature review

2.1 Residential relocation

The history of residential mobility research is vast and comprehensive. Rossi (1955) early work highlighted the significance of perceiving households as decision-making entities and understanding their motivations (Dieleman, 2001). This insight paved the way for the emphasis on disaggregated residence mobility modeling, contrasting with the traditional aggregated housing models. Residential relocation has been explored from several practical perspectives. These critical aspects are as follows: firstly, the interplay between residential relocation and travel behavior investigates the influence of moving residences on travel patterns (Lin et al., 2018; Scheiner & Holz-Rau, 2013). Secondly, the relationship between residential relocation and the dynamics of household life-course events (Clark, 2013; Prillwitz et al., 2007) delves into events like job transitions, educational milestones, vehicle transactions, family additions, or household shifts, assessing their potential impact on residential moves. Thirdly, research on relocation timing (Rashidi et al., 2011; Thomas et al., 2016) concentrates on the duration of household relocations and their relation to home and household attributes. Fourthly, the connection between residential relocation and accessibility studies the influence of essential service proximity on relocation choices (Srour et al., 2002; Zhou et al., 2021).

Lastly, residential self-selection (Cao et al., 2009; Mokhtarian & Cao, 2008) addresses often-neglected aspects of home choice and its environmental implications.

2.2 Residential relocation duration

Understanding the duration of residential relocations is pivotal in grasping the dynamics of individual residence mobility. The literature on this topic is replete with studies exploring a myriad of explanatory variables and methodological approaches. Concerning explanatory variables, several have been identified as significant in determining the length of residence durations. Among these, socio-demographic attributes, financial status, homeownership, home payment, and life-course events such as job transitions and educational milestones stand out as the most influential (Bostanara et al., 2023; Bostanara et al., 2021; Rashidi & Ghasri, 2017; Thomas et al., 2016; Tran et al., 2016). From a methodological standpoint, a variety of parametric, semi-parametric, and non-parametric classical survival analysis tools are prevalent, given their efficacy in managing censored data. This study primarily centers on modeling residence duration.

2.3 Residential relocation, accessibility, and self-selection bias

When deciding to relocate, individuals typically weigh various aspects of a potential home. Foremost among these considerations are transport-related attributes and accessibility (Kim et al., 2005) — topics well-documented in existing literature. Evidence suggests that the dynamics of household residential relocations and daily trip-making behaviors are intricately linked (De Vos & Ettema, 2020). Proximity or travel time to workplaces (Sprumont & Viti, 2018) and educational institutions are vital factors in home location decisions, with preferences leaning towards minimized travel (Habib & Miller, 2009; Zhou & Kockelman, 2008). Some studies emphasize automobile travel times exclusively (Zolfaghari et al., 2012), while others underscore the importance of both car and public transport durations (Kim et al., 2005). Although accessibility stands out as a pivotal factor in home selection, its measurement is perceived as intricate (Miller, 2018). Accordingly, scholars have proposed a variety of accessibility metrics, ranging from classic methodologies like the Lowry model (Lowry, 1964) to contemporary measures such as the total number of accessible jobs within specified time or distance thresholds (Wachs & Kumagai, 1973). Srour et al. (2002) conducted a comparative analysis and determined that the latter measure provides the most robust and understandable indicator of accessibility.

This research delves into the correlation between the accessibility of individuals' current homes and that of their previous residences. The dependency of a present home's attributes on those of a former dwelling is an established subject in scholarly discussions. Notably, prior research reveals a pronounced correlation between past and present residential locations (Axhausen et al., 2004; Habib & Miller, 2009). Additionally, numerous studies indicate a tendency for individuals to relocate within proximity to their existing residence (de Palma et al., 2007; Zondag & Pieters, 2005).

Self-selection bias is a recognized form of endogeneity bias in the field of relocation. It occurs when an individual's neighborhood selection is overlooked while modeling travel behavior based on neighborhood characteristics (Hedman & van Ham, 2012). Mokhtarian and Cao (2008) provide a comprehensive overview of seven methodologies to address this bias, which include direct questioning, statistical control, joint discrete choice models, and instrumental variables models. The direct questioning method is particularly potent, as it garners insights directly from respondents regarding their environmental inclinations (Colabianchi, 2009). Earlier research, including work by

Handy and Clifton (2001), has employed this method, affirming that transport-related attitudes play a pivotal role in neighborhood selection for prospective homes.

2.4 Machine learning in transport research

Numerous machine learning approaches have been employed in transport-related research areas, including travel mode choice (Li et al., 2020; Pineda-Jaramillo & Arbeláez-Arenas, 2022), accessibility satisfaction (Cheng et al., 2020), ride-sourcing (Aghaabbasi et al., 2020), parking management (Parmar et al., 2021), land-use (Xu et al., 2019), and transportation safety (Cai et al., 2019). However, there is a scarcity of research in the domain of residential relocation that harnesses machine learning techniques. Notably, Xue and Yao (2022) leveraged a random forest model to discern and quantify the primary drivers of residential relocation. Yi and Kim (2018) examined residential relocation distances using a decision-tree algorithm and juxtaposed their findings with least squares regression results. Scheuer et al. (2021) probed residential location choices employing a random forest model. To the author's knowledge, there hasn't been a prior study that has integrated machine learning algorithms to explore the duration of residential relocations, particularly considering accessibility factors influencing such decisions.

2.5 Contribution aspects

The main contributions of this research are threefold. Firstly, this study is the first in the field to utilize machine learning algorithms to model residential relocation duration. Machine learning algorithms are employed to model the households' duration of stay in their home for high dimensional residential data collected from Sydney, Australia, metropolitan area residents. Secondly, this study evaluates and compares the performance of multiple machine learning and feature selection algorithms against the classical methods typically used in modeling residential relocation duration. Additionally, it offers valuable insights into the variables that influence household residential relocations. The research delves into ensemble models and provides an in-depth exploration of the c-index measure, shedding light on how different household features shape this metric. Thirdly, the future home accessibility model results allow the evaluation of urban planning policies that would be interesting for policymakers.

3 Data

The primary dataset used in this research was extracted from a retrospective survey on residential relocation conducted in Sydney, Australia. In 2019, 512 residents from the metropolitan area participated in this survey, representing a subset of the population of 4,823,991 (based on the 2016 Australian Bureau of Statistics (Australian Bureau of Statistics 2016). The survey, as reported by (Ghasri et al., 2022), aimed to gather information about the participants' current and past residential relocations within Sydney, with a primary focus on understanding the dynamics of household residential relocations and their correlation with life-course events.

The survey incorporated an adaptive observation time window for each household, encompassing both historical and current residence location information. The dataset includes cases of observed move-out dates (failures) as well as cases with non-observed move-out dates (censored data), all of which are utilized in this study. In addition to relocation history, participants were also queried about various socio-demographic factors, such as age, vehicle ownership (including transaction times), occupational

records (contract start dates and income), and education history (commencement and graduation dates).

Table 1. Mean and standard deviation of some of the features

	Feature	Feature description	Mean	Standard Deviation
1	duration	Residing duration	8.61	9.28
2	event	Relocation observed?	0.50	0.50
3	isOwner	Is owner?	0.51	0.50
4	under18	Number of under 18 years old members	1.16	1.31
5	hhIncomeInStart	Household income at the time of relocation	1.31	1.35
6	jobFromHomeSum	Any working from home member?	0.15	0.36
7	jobIsProfessionalSum	Number of members with professional jobs	0.22	0.42
8	commute_drive	Is driving the main commute mode?	0.44	0.50
9	driveLicense	Does own a driver's license?	0.87	0.33
10	driving_main	Is driving the main trip mode?	0.43	0.50
11	public_main	Is public transport the main trip mode?	0.14	0.34
12	last2YearStartJob	A member started a job in last two years?	0.25	0.43
13	last2YearLeftJob	A member left a job in last two years?	0.17	0.37
14	last2YearStartEdu	A member started an education in last two years?	0.06	0.25
15	last2YearLeftEdu	A member completed his/her studies in last two years?	0.06	0.24
16	last2YearStartVeh	A member bought a vehicle in last two years?	0.09	0.29
17	last2YearLeftVeh	A member sold a vehicle in last two years?	0.03	0.16
18	last2YearNewChild	A child was added to the household in last two years?	0.07	0.25
19	tt_work_car_mean	Travel time to work by car in minute	18.55	39.90
20	tt_edu_car_mean	Travel time to school by car in minute	0.70	6.84
21	tt_work_pt_mean	Travel time to work by public transport in minute	138.56	126.69
22	tt_edu_pt_mean	Travel time to school by public transport in minute	112.65	63.54
23	commercial_per	Suburb commercial land-use%	0.06	0.10
24	education_per	Suburb educational land-use%	0.03	0.04
25	parkland_per	Suburb parkland land-use%	0.17	0.15
26	residential_per	Suburb residential land-use%	0.62	0.23
27	transport_per	Suburb transport land-use%	0.02	0.05
28	water_per	Suburb water land-use%	0.002	0.016
29	industrial_per	Suburb industrial land-use%	0.05	0.11
30	hospital_medical_per	Suburb medical land-use%	0.01	0.03
31	acc1	Accessibility in 30 minutes of home by driving	10.65	5.73
32	accessibility_pt	Accessibility in 30 mins of home by public transport	0.38	0.86
33	acc2_pred	Predicted accessibility in 30 mins of next home by driving	10.24	4.47

Furthermore, the survey captured data related to household trip generation activities, intra-household decision-making behavior, and attitudes toward residential location selection. Due to the dataset's richness, it contains numerous variables (features) that cannot all be included in this report's main body. However, a selection of key variables, along with their mean and standard deviation values, is provided in **Table 1**, rows 1-18. For a comprehensive list of variables and further details on the data, please refer to Bostanara et al. (2023). Additionally, Ghasri et al. (2022) provides information on data

representativeness within the population. **Figure 1** visually depicts the spatial distribution of sampled home and job locations for households within the Sydney metropolitan area.

In relation to land-use features, we extracted information about the suburbs' land use from the 2016 Australian Bureau of Statistics dataset (Australian Bureau of Statistics 2016). This dataset provides the percentage/area of land in each suburb allocated for specific purposes, including commercial, educational, residential, parkland, transport, water, industrial, and medical sections. To estimate the land-use status of residential neighborhoods, we used the suburb and postcode of each household's home location. **Table 1**, rows 23-30, displays the average percentage of land use for each purpose within the surveyed areas.

For daily trip accessibility, we calculated travel times to work and school for each household member, assuming a departure time of 8:00 AM on weekdays. These travel times were estimated based on home, work, and school postcodes, considering both vehicle and multi-modal public transport options using the r5r package (Pereira et al., 2021). We then averaged the travel times for household members and incorporated this variable as a predictor in our models. **Table 1** rows 19-22, presents statistics related to the mean travel time variables.

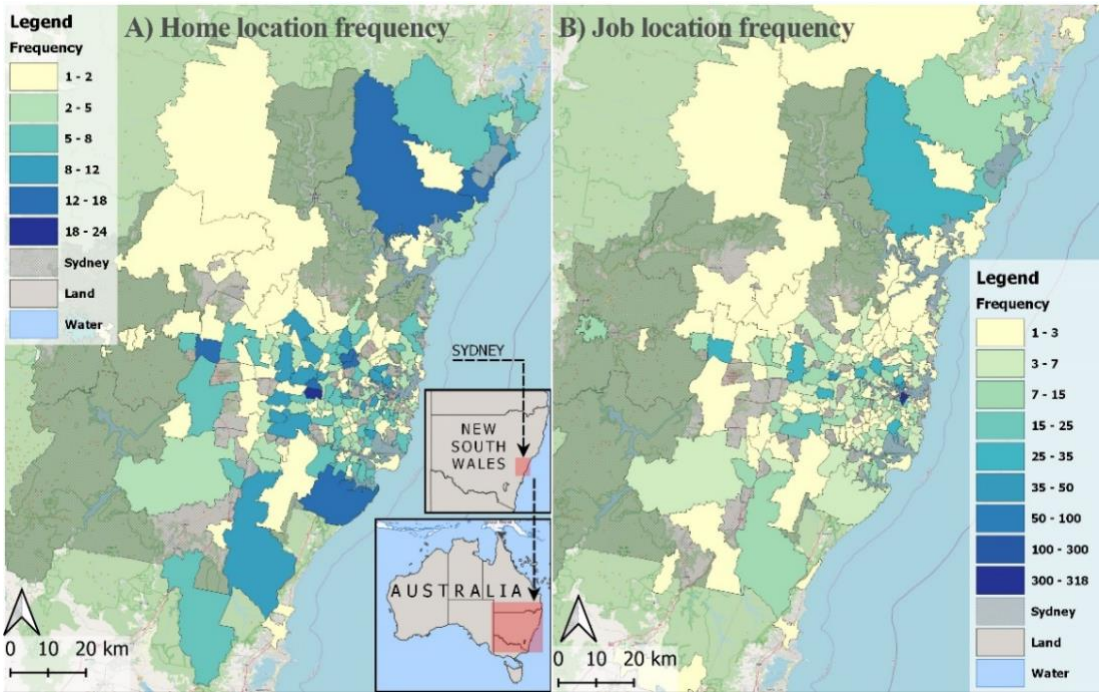

Figure 1. The spatial distribution of A) home locations and B) job locations in Sydney metropolitan area-the numbers in the legend are the number of respondents

One of the accessibility measures utilized in this study is the count of accessible jobs within a 30-minute travel time by car and public transport. These accessibility measures were computed for each suburb in Sydney using the r5r package. **Figure 2** depicts a heatmap of accessibility measures for both car and public transport, illustrating that accessibility is highest in and around the CBD area and gradually diminishes with distance from the CBD. Clearly, there are significant differences between car and public transport accessibility. **Table 1** rows 31-33, provides statistics related to the mean accessibility measure variables.

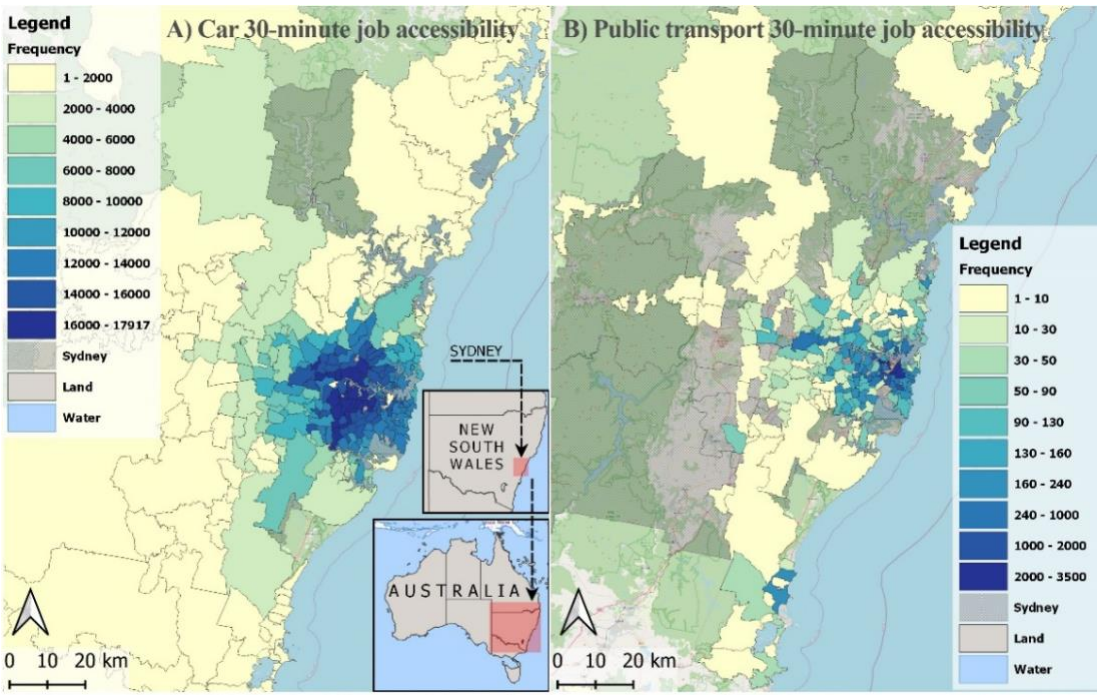

Figure 2. Heatmap of suburbs' accessibility to jobs within 30-minutes using A) car and B) public transport in Sydney metropolitan area—the numbers in the legend are the number of accessible jobs within 30 minutes

Table 2. Frequency of importance rank of each home attributes to households

	RESIDENTIAL ATTITUDE	1 Very Unimportant	2 Unimportant	3 Neutral	4 Important	5 Very Important
rB1	Close to family and friends	150	108	260	274	232
rB2	Close to leisure activities	128	160	304	286	146
rB3	Close to public transportation	106	72	206	304	336
rB4	Close to shops, groceries	56	72	206	370	320
rB5	Close to work	176	100	252	290	206
rB6	Frequent contact with neighbors	300	194	248	164	118
rB7	Good contact with neighbors	192	152	280	258	142
rB8	Neatness and tidiness	80	80	258	352	254
rB9	Appearance of buildings/architecture	74	92	278	364	216
rB10	Presence of bike paths	412	150	192	142	128
rB11	Presence of green areas	136	96	272	292	228
rB12	presence of footpaths	136	148	288	280	172
rB13	Quietness	60	76	258	342	288
rB14	Social Safety, low crime	62	76	192	338	356
rB15	Sufficient parking	104	102	216	314	288
rB16	Traffic safety	102	116	276	318	212

The survey also included questions about the households' attitudes toward residential location selection. Respondents ranked the importance of 16 home attributes on a scale

from "very unimportant" to "very important" when choosing a new home. **Table 2** displays the frequency of respondents' importance rankings for these attributes. Notably, attributes such as "presence of bike paths" and "frequent contact with neighbors" were rated as the least important factors. Conversely, "social safety, low crime," "proximity to public transport," and "proximity to shops and groceries" emerged as the top three most essential home attributes.

4 Methodology

4.1 Survival analysis

Hazard-based modeling, also known as survival analysis, is a well-established field in statistics that focuses on studying the duration between entering a specific state and a subsequent event. In the context of residential relocation, a household enters the "resident" state when they move into a new home location and exits this state when they relocate to a different place. The duration of their stay in a particular location is the primary variable of interest. However, it's important to note that not all households in the dataset will experience the event of moving out. This is often due to the nature of residential stay durations, which tend to be long, while data collection periods are typically shorter. As a result, the residential relocation dataset can contain both "failed" records (where move-out is observed) and "censored" records (where move-out is not observed). Consequently, residential relocation data is often subject to censorship, making survival analysis an ideal tool for analyzing such datasets. Survival analysis deals with a positive dependent variable and can effectively handle censored data.

In all survival analysis approaches, three main functions are considered: the failure function (usually denoted as $f(t)$, representing the probability of failure over time), the survival function (usually denoted as $S(t)$, representing the probability of surviving over time), and the hazard function (usually denoted as $h(t)$, representing the probability of failure at a specific time, given survival up to that point) (Jenkins, 2005).

Classical survival models have a long history and can generally be categorized into three main types: parametric, semi-parametric, and non-parametric models. In this study, we primarily utilize the semi-parametric Cox Proportional Hazard (Cox-PH) model, a representative of classical semi-parametric survival models. Additionally, we employ the Weibull- and Lognormal-based Accelerated Failure Time (AFT) models, which represent classical parametric survival models. Please refer to the appendix for further information.

4.2 Machine learning

Machine learning is a subset of Artificial Intelligence (AI) in which machines are designed to learn and model non-linear, complex relationships between one or more independent variables and a dependent variable using a set of training data. The learned model is then applied to make predictions on a new set of data, known as the test set. In recent years, the field of machine learning has made significant progress, establishing itself as a prominent area of interest in both theoretical and practical research communities. One notable expansion within this field is the development of various learning algorithms tailored for survival data. These advancements present a valuable opportunity for enhancing survival models and leveraging the diverse and high-dimensional residential relocation data.

This study employs ten machine learning algorithms (referred to as "learners") and six feature selection algorithms (referred to as "feature selectors"), all designed to work with censored data. Brief introductions to these algorithms are provided below. Due to the

intricacies and unique features of each algorithm, we are unable to provide a comprehensive introduction within this article. However, interested readers can find thorough reviews of machine learning algorithms for survival analysis in previous studies (Bender et al., 2021; Sonabend et al., 2021; Wang et al., 2019), which offer additional information.

4.2.1 Learners

a. Cross-validation regularized cox proportional-hazards models (cv.glmnet)
   This learner is a penalized maximum likelihood version of the classical Cox-PH model. It is a cross-validated regularized Cox-PH model (Hastie & Qian, 2014). Three main regularized options are available for this learner.
b. Ridge regularization in which coefficients of the correlated features tend to shrink towards each other.
c. Lasso regularization in which only one of the correlated features tends to stay in the model.
d. Elastic net regularization which is a combination of the aforementioned regularization methods.
e. Gradient boosting methods
   Gradient boosting learners is a family of powerful models that are based on an ensemble of weak models (Hothorn et al., 2006) and generate a strong model which often outperforms other models. Some of the learners from this family are listed below.
f. Survival gradient boosting model (GBM).
g. Boosted generalized linear survival model (GLMBoost).
h. Extreme gradient boosting survival model (Tree-based)-also listed in d.
i. Extreme gradient boosting survival model (Linear based) (Chen et al., 2015).
j. Survival tree model
   The decision tree is one of the most straightforward learners in machine learning. These learners are usually improved via ensemble and boosting methods (Gordon & Olshen, 1985).
k. rpart (Therneau et al., 2015).
l. Extreme gradient boosting survival model (Tree-based)-also listed in b.
m. Random forest models
   Decision random forest models are simply ensemble tree learners, which are very powerful (Breiman, 2001).
n. Survival random forest SRC (RFSRC) (Kogalur, 2022).
o. Ranger survival model (Wright & Ziegler, 2015).

4.2.2 Feature selectors

One of the complex stages in modeling is the selection of model features. This process becomes even more challenging when dealing with high-dimensional datasets, as noted by (Spooner et al., 2020). Machine learning algorithms offer valuable assistance in feature selection by identifying a subset of data features that optimize model performance. Feature selection brings several advantages, including enhanced interpretability and computational efficiency, both of which are crucial (Chandrashekar & Sahin, 2014).

Within machine learning, filter methods represent a primary category of feature selectors. These methods employ specific algorithms to assess and rank all available features. Subsequently, a subset of features is chosen based on a predefined threshold, which can be established for various criteria, such as the number of features, a percentage

of features (as utilized in this study), or feature selection performance. Following feature selection, the model is executed using the chosen features. In this study, we have employed the following filter feature selectors.

1. The univariate model score which is based on the univariate statistical tests (UNI) (Jović et al., 2015).
2. Minimum redundancy, maximum relevance (MRMR) is based on selecting the most correlated features with the response variable and least correlated with each other (Radovic et al., 2017).
3. Random Forest feature importance (IMP) (Kogalur, 2022).
4. Random Forest minimal depth (DEPTH) (Kogalur, 2022).
5. Random Forest feature hunting (HUNT) (Kogalur, 2022).
6. Random Forest ranger impurity (RANGER) (Wright & Ziegler, 2015).

In addition, we evaluate the performance of all learners without employing any external feature selector to establish a baseline for comparison.

4.2.3 Evaluation metric

In this study, we employ Harrell's estimator of the Concordance Index (C-Index) to assess the performance of our survival machine learning models. The C-Index excels in quantifying the correlation between hazard predictions and observed event times, effectively consolidating the three crucial aspects of a survival model—hazard, event, and time—into a single metric, as described by (Longato et al., 2020).

To calculate the C-Index, each relevant pair of cases in the dataset is examined. These pairs consist of cases where at least one is not censored, and the duration of the failed case is shorter than that of the censored case. The C-Index assesses whether the predicted order of survival times aligns with the observed order. Pairs that align are termed "concordant," while those that do not are termed "discordant." The index then reports the percentage of concordant pairs relative to the total number of relevant pairs. A higher C-Index value signifies a superior model, with a C-Index of 1 indicating a perfect model where all predictions are correctly ordered.

Furthermore, this metric is utilized to select the best-tuned parameter sets during the inner-resampling iterations, and the average C-Index value across all outer-resampling iterations serves as the model's comprehensive performance measure.

4.2.4 Future accessibility model

Residential relocation hinges on two decisions: leaving the current home and choosing a future one. Both these locations significantly influence the relocation process. Our research emphasizes the role of accessibility in residential relocation timing, raising questions like 1) which location's accessibility holds greater significance in the relocation decision, the current home or the future home? 2) how can we estimate the accessibility of the future home? and 3) is there a correlation between the accessibility of the future home and that of the current home? Our hypothesis suggests that accessibility of both present and future homes greatly affects relocation timing. While current home accessibility is directly measurable, future accessibility is elusive. Our approach estimates it using data from the current home and household behavior predictions. To address the third question, we test whether current home accessibility can serve as a determining factor for future home accessibility.

In this study, we model the increase or decrease in accessibility from the current home to the future home as a function of their respective previous home accessibility and certain socio-demographic attributes. Such a model inherently contains self-selection bias (endogeneity bias) because families may choose to move to a more accessible location

based on their preferences and choices, rather than socio-demographics and future home features (see **Figure 3**).

Self-selection bias is a recognized form of endogeneity bias in the field of residential mobility. It occurs when the influence of a household's home selection is not considered when modeling their subsequent behavior in the new living environment. In other words, individuals may intentionally change their behavior, such as seeking greater accessibility, because it aligns with their preferences, rather than it being solely a result of their new circumstances and environment.

One of the widely accepted methodologies to address this bias is direct questioning, which, in this study, involves directly asking households about the attributes they prioritize in a new home. Some of these attributes reveal the households' inclination toward residing in a more or less accessible location. Therefore, we have included households' attitudes toward residential attributes (introduced in the Data section) in our study.

The dataset used in this study contains information about two locations per household, enabling us to calculate pre (A_{Home1}) and post (A_{Home2}) relocation home accessibility. Accessibility is defined as the number of accessible jobs within a 30-minute car ride. To comprehend households' intentions to increase or decrease their accessibility, we model accessibility change, which is calculated as the accessibility of the second home minus that of the first ($A_{Home2} - A_{Home1}$). This change is modeled as a linear function of the previous home's accessibility (A_{Home1}), households' reported attitudes toward residential attributes (AT_i, where $i = 1, 2, \dots, 16$), and some other explanatory variables (X_i). The final model takes the form of $(A_{Home2} - A_{Home1}) \sim f(A_{Home1}, AT_i, X_j)$. Subsequently, the model coefficients are used to estimate the future home accessibility for all locations in the dataset.

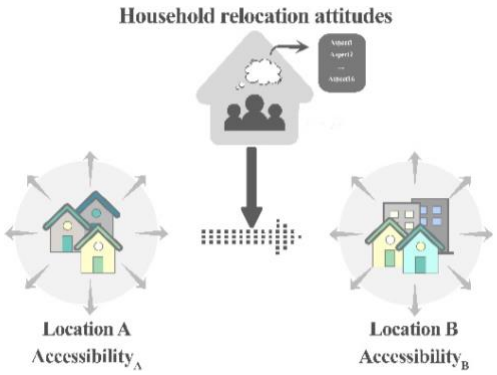

Figure 3. Addressing self-selection bias by the direct questioning method

4.2.5 The machine learning pipeline

This study's machine learning pipeline comprises the following primary steps, as illustrated in **Figure 4** and elaborated below:

A. Data Preparation Steps:
   - A1. Extract the life-course dataset from the residential relocation survey.
   - A2. Create an inter-city travel time matrix.
   - A3. Compile a table of suburbs' land-use structures.
   - A4. Generate 30-minute accessibility tables for car and public transport travel between all suburbs in the study area.

A5. Estimate households' future home accessibility based on their current accessibility, socio-demographic attributes, and home selection attitudes using the future accessibility model.

A6. Combine the cleaned and prepared life-course dataset with the travel time, land-use, and accessibility datasets. The final dataset comprises 163 features (explanatory variables).

B. Cross-Validation Resampling Setup: Preparing a five-fold cross-validation resampling instance to execute all models five times. This involves randomly dividing the dataset into training and test sets five times iteratively (outer-resampling).

C. Model and Feature Selector Selection:

C1. Listing all ten learners, three classical survival models, and the six feature selectors.

C2. Creating a benchmark matrix of 13 models by pairing each learner with each feature selector (13x6 models in total).

D. Model Execution and Evaluation: Execute all 78 models across all five resampling sets and report the aggregate C-Index against the benchmark.

D1. Preparing five-fold cross-validation resampling instances within the outer-resampling training sets to fine-tune all learners five times. This entails randomly splitting each training dataset into training and test sets five times iteratively (inner-resampling).

D2. Performing parameter tuning for the hyper parameters of the learner and feature selector using inner-resampling sets and a predefined grid resolution based on the C-Index measure.

D3. Apply the chosen feature selectors.

D4. Execute the learning process and report the sub-C-Index.

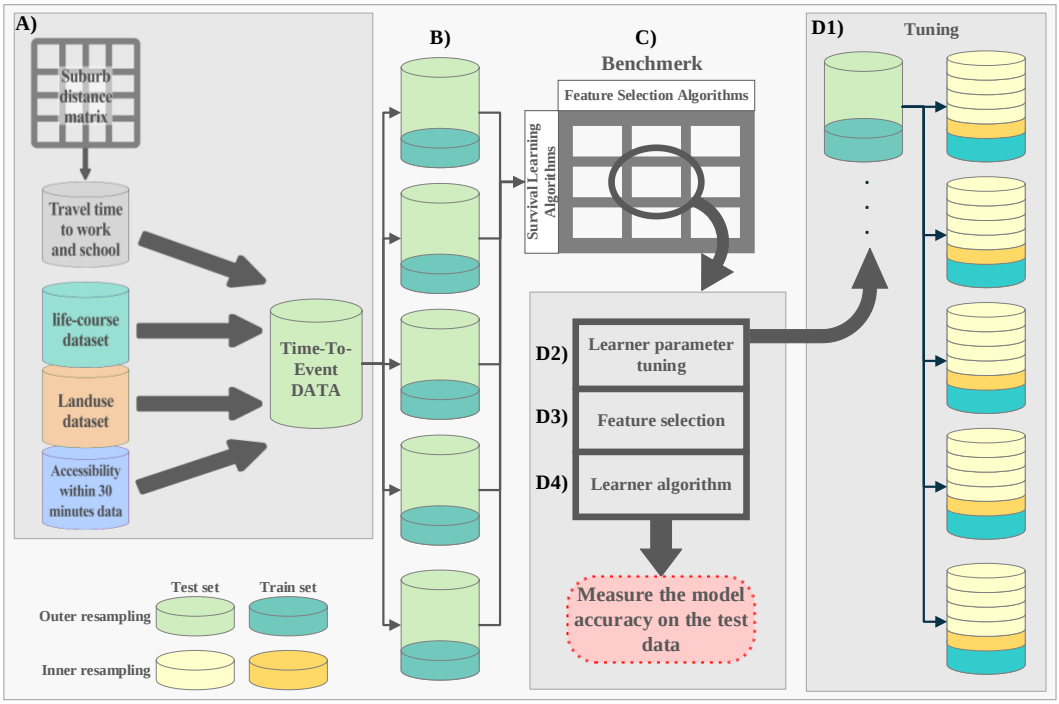

Figure 4. The machine learning pipeline

4.3 Implementation details

R is the primary programming language employed for all tasks in this research, including data preparation, manipulation, analysis, and machine learning modeling. To maintain a reproducible workflow, the "targets" package, as detailed by (Landau, 2021), is utilized. Additionally, the calculation of travel times and accessibility measures is accomplished using the "r5r" package, as introduced by Pereira et al. (2021). For machine learning modeling, the "MLR" package, outlined by Bischl et al. (2016), is employed. The entire code workflow utilized in this study is openly accessible via GitHub[1]. However, it's important to note that the residential relocation data cannot be publicly disclosed due to confidentiality agreements.

5 Results

5.1 Future accessibility model

The outcomes of the linear model analyzing the 30-minute accessibility change from one home to another are summarized in **Table 3**. These findings unveil overarching trends and patterns in households' relocation behavior, which hold valuable implications for policy evaluation and strategy formulation. Notably, all variables exhibit significance, with a minimum threshold of 10%.

Foremost, the accessibility of the initial home emerges as the most pivotal factor contributing to accessibility changes. Higher household income and an increased number of children correlate with relocations to areas with lower accessibility. Additionally, individuals who commute via car tend to relocate to less accessible locations, suggesting a preference for quieter environments and an ability to accommodate longer daily travel times. Furthermore, holding a professional job significantly contributes to increased accessibility, while households with longer travel times to work tend to opt for areas with higher accessibility.

In terms of home selection attitudes, it becomes evident that individuals who prioritize proximity to public transport tend to experience a notable increase in accessibility. Conversely, those with a medium to high preference for proximity to shops and groceries tend to gravitate away from highly accessible areas. This observation aligns with the even distribution of shops, especially grocery stores, across Sydney, making it unnecessary for individuals to relocate to highly accessible regions, such as the CBD, solely for grocery convenience. Similarly, households that assign medium to high importance to proximity to their workplaces tend to migrate toward more accessible suburbs. Conversely, those inclined towards green spaces and peaceful environments tend to move to less accessible locations. Lastly, individuals who consider the presence of footpaths crucial in their choice of residence are inclined to move to more accessible suburbs.

These findings collectively indicate that individuals who can accommodate longer car travel times tend to settle in less accessible and potentially quieter areas, a seemingly prudent household decision. However, when aggregated, such choices contribute to extended car travel times, impacting the city's environment. On the contrary, those whose primary mode of transportation relies on public transport and individuals who value well-developed footpaths tend to relocate to more accessible and potentially busier regions. Public transport accessibility typically aligns with more densely populated areas,

[1] https://github.com/UNSW-rCITI/Survival-ML

necessitating such moves for individuals reliant on public transportation, regardless of their preference for busier areas. Additionally, well-maintained footpaths, a crucial feature for active modes of transportation, are typically found in central districts, further attracting individuals to accessible regions.

Table 3. Future accessibility model

Feature	Importance rank	Estimate	Pr(>|t|)	
(Intercept)	-	32.78	0.00	***
acc1	-	0.68	0.00	***
hhIncomeInStart	-	-2.77	0.08	.
jobIsProfessionalSum	-	8.02	0.09	.
under18	-	-2.31	0.10	.
commute_drive	-	-7.68	0.04	*
tt_work_car_mean	-	0.10	0.03	*
rB3-Close to public transportation	Important	14.64	0.01	**
rB3-Close to public transportation	Very important	18.44	0.00	**
rB4-Close to shops, groceries	Neutral	-20.95	0.00	**
rB4-Close to shops, groceries	Important	-26.65	0.00	***
rB4-Close to shops, groceries	Very important	-23.00	0.01	**
rB5-Close to work	Neutral	11.24	0.03	*
rB5-Close to work	Important	12.01	0.02	*
rB5-Close to work	Very important	16.06	0.01	**
rB11-Presence of green areas	Very important	-14.40	0.01	**
rB12-Presence of footpaths	Neutral	11.63	0.02	*
rB12-Presence of footpaths	Important	15.10	0.01	**
rB12-Presence of footpaths	Very important	13.48	0.06	.
rB13-Quietness	Very important	-12.69	0.01	**

Significance level: '***' ~ 0.001, '**' ~ 0.01, '*' ~ 0.05, '.' ~ 0.1

Furthermore, a significant portion of the population (73%) with a medium to high preference for proximity to their workplaces tends to migrate to areas with greater job opportunities. This poses concerns, particularly when job opportunities are not evenly distributed within the city. Such patterns are neither environmentally sustainable nor conducive to overall well-being (Shen, 2001). This underscores the importance of implementing policies such as: 1) Ensuring public transport equity within the city to allow individuals to remain in their preferred locations without being compelled to move to more accessible areas. This should be accompanied by efforts to enhance the quality of public transport, encouraging car users to transition to public transportation. 2) Promoting the concept of "30-minute cities" to alleviate congestion in neighborhoods and distribute job opportunities more evenly throughout the city. 3) Initiating incentives and programs aimed at reducing private car use. 4) Enhancing footpaths and green spaces across the city to support active modes of transportation and enhance overall citizen well-being. These policy initiatives are essential for fostering a more sustainable urban environment and promoting the well-being of residents.

5.2 Residential relocation model results

The benchmark results and the number of features selected in each model are comprehensively presented in **Table 4** and **Table 5**, respectively. In **Table 4**, the cell values depict the average C-Index performance measures across five-fold resampling sets, obtained from running each pair of learners (rows) and feature selectors (columns). **Table 5** outlines the mean count of features selected and utilized within each model for every pair. In these tables, the first column showcases the learners' performance without the involvement of any external feature selectors. It's worth noting that some learners come equipped with their internal feature selectors, while others do not. This distinction elucidates the varying number of selected features in the initial column of **Table 5.**

Table 4. Average learner and feature selector C-Index performance measure in a five-fold resampling

		None	UNI	MRMR	IMP	DEPTH	HUNT	RANGER
Classical	Cox-PH	0.491	0.757	0.704	0.765	0.765	0.762	0.767
	AFT-Log-normal	0.502	0.759	0.710	0.763	0.766	0.764	0.763
	AFT-Weibull	0.516	0.751	0.707	0.763	0.765	0.752	0.766
Regularized Cox-PH	Ridge	0.732	0.746	0.711	0.753	0.752	0.751	0.752
	Elastic Net	0.752	0.748	0.726	0.757	0.756	0.757	0.753
	Lasso	0.746	0.744	0.716	0.752	0.757	0.750	0.744
Gradient boosting	GBM	0.773	0.772	0.723	0.771	0.768	0.772	0.775
	Glmboost	0.739	0.739	0.724	0.736	0.739	0.738	0.737
	Xgboost lm	0.720	0.725	0.687	0.721	0.727	0.730	0.724
Tree-based	Xgboost tree	0.760	0.766	0.773	0.756	0.760	0.760	0.759
	Rpart	0.696	0.707	0.700	0.716	0.715	0.716	0.713
Random Forest	Random Forest SRC	0.759	0.758	0.771	0.759	0.754	0.753	0.759
	Ranger	0.740	0.750	0.747	0.757	0.750	0.749	0.753

0.4 0.45 0.5 0.55 0.6 0.65 0.7 0.75 0.8

Table 5. Average number of features selected in each model across a five-fold resampling

		None	UNI	MRMR	IMP	DEPTH	HUNT	RANGER
Classical	Cox-PH	0	47.6	56.6	37.6	40.4	26.2	46.4
	AFT-Log-normal	0	43.2	62.2	31.8	33.2	37.6	37.4
	AFT-Weibull	0	47.8	58.0	33.4	31.8	42.0	38.8
Regularized Cox-PH	Ridge	163	56.4	58.0	37.6	42.0	33.2	32.0
	Elastic Net	72	49.0	63.6	41.8	46.2	46.2	49.4
	Lasso	42.6	53.6	62.2	43.2	46.0	43.2	50.6
Gradient boosting	GBM	0	49.0	60.8	43.6	34.8	36.2	53.6
	Glmboost	14.6	57.8	62.2	21.8	27.4	30.4	36.0
	Xgboost lm	163	52.0	52.4	37.6	42.0	40.4	40.6
Tree-based	Xgboost tree	31	78.6	147.0	84.6	71.8	74.8	68.6
	Rpart	39.4	22.8	133.6	45.6	42.2	39.2	25.8
Random Forest	Random Forest SRC	163	52.4	147.0	84.6	55.4	65.0	72.0
	Ranger	163	49.0	147.0	33.0	22.8	26.2	26.2

0 20 40 60 80 100 120 140 160

The first three rows denote the three classical survival models included as a benchmark. When operating without any feature selectors, these classical models exhibit the lowest accuracy based on the C-Index measures, hovering around the 50% mark. This suggests that they can predict only half of the test data accurately, owing to the absence of intra-feature selection methods (i.e., no features are selected).

Nearly all other models surpass the classical models when executed without external feature selectors. Among the feature selectors, the MRMR algorithm emerges as the least potent, albeit generally selecting a larger number of features. Excluding this feature selector, the GBM learner slightly outperforms all other learners. Notably, there appears to be minimal disparity in the performance of the various feature selectors.

An intriguing observation is that tree-based models, such as XGBoost and Random Forest SRC models, outperform all other tree-based models and maintain strong performance irrespective of their feature selector (even with MRMR). Tree-based models are renowned for their capacity to identify non-linear relationships within data. Consequently, the remarkable performance of tree-based models underscores the significance of considering such non-linear relationships.

A particularly noteworthy revelation in the results is the impressive performance of classical models when paired with tree-based feature selectors, as elucidated in the first three rows of **Table 4**. This underscores the efficacy of tree-based feature selectors, particularly when coupled with classical methods. This result is significant, demonstrating that we can closely replicate machine learning results by pairing feature selection algorithms with classical models, thereby retaining the advantages of classical models such as interpretability.

5.3 Feature importance

In this research, the dataset consisted of 163 features. Each feature had the potential to be selected 451 times, given that it could be chosen five times for each combination of learner (13 learners in total) and feature selector (6 external feature selectors plus one intra-learner feature selector), excluding the models with zero features (4 models). **Table 6** illustrates the importance of a selection of these features, while **Figure 5** presents a word cloud map highlighting the most predictive ones.

The feature importance analysis, derived from benchmarking several machine learning and feature selection algorithms, yields intriguing insights across multiple dimensions. Homeownership emerges as crucial, underscored by the "Is owner?" feature, which achieves an importance score of 0.98. This dominance underscores the pivotal role homeownership plays in predictions across the tested algorithms. Delving into life-course events, transformative household changes like "Childbirth during the last two years" and "Purchasing a new vehicle during the last two years" are especially prominent, both achieving scores above 0.90.

Accessibility features also stand out: "Average of travel time to work by public transport" clocks in at 0.87, whereas features denoting "30-minute accessibility" by car or public transport to one's current and expected next home show varied importance, with public transport accessibility consistently scoring higher. In the household realm, the count and age dynamics of household members play a discernible role, with both "Number of under 18" and "Number of over 18" features and their squared values holding significant importance. The student category reveals that the model has a heightened focus on "Number of secondary students in the household," signaling the significance of this demographic.

From an attitudinal perspective, while the importance scores aren't as high as in other categories, nuances like "Being close to public transport" and proximity to shops or

groceries still find representation, showcasing a consistent trend of value placed on convenience and accessibility. Moreover, in the sphere of intra-household decision-making, features suggest a balance, with shared decisions on various daily and significant aspects holding close importance values, underlining the potential influence of collective household decisions on outcomes.

Lastly, in the financial domain, dynamic changes in "Household income" and especially the squared value of its annual increase/decrease achieve prominence, capturing the model's sensitivity to shifts in household financial dynamics. In sum, the benchmark analysis underscores an intertwined interplay of homeownership, major life-course events, demographic details, accessibility, and collective household decisions as driving forces across the tested models.

Table 6. Feature importance of some selected features

	Feature	Importance
Household	Number of under 18	0.50
	(Number of under 18)^2	0.39
	Number of over 18	0.73
	(Number of over 18)^2	0.71
	Max age in household	0.40
	(Max age in household)^2	0.30
	Number of female members	0.33
Home	Is owner?	0.98
	Is house?	0.81
	Rooms less than two	0.74
	Household rent payment	0.83
	(Household rent payment)^2	0.82
	Household mortgage payment	0.62
	(Household mortgage payment)^2	0.63
Life-course	Child birth during last year	0.89
	Child birth during last two years	0.93
	Starting a new job during last year	0.88
	Starting a new job during last two years	0.87
	Leaving a job during last year	0.28
	Leaving a job during last two years	0.33
	Purchasing a new vehicle during last year	0.53
	Purchasing a new vehicle during last two years	0.95
	Selling a vehicle during last two years	0.45
	Comencing a new study during last two years	0.87
	Comencing a new study during last two years	0.97
	Graduating a study during last year	0.53
	Graduating a study during last two years	0.97
Students	Number of primary students in household	0.84
	Number of secondary students in household	0.90
	Number of tertiary students in household	0.50
	(Number of primary students in household)^2	0.81
	(Number of secondary students in household)^2	0.84
	(Number of tertiary students in household)^2	0.42
	Number of students in household	0.48
	(Number of students in household)^2	0.46

	Feature	Importance
Accessibility	Average of travel time to work by car	0.60
	Average of travel time to work by public transport	0.87
	Average of travel time to school by car	0.25
	Average of travel time to school by public transport	0.86
	Current home 30-minute accessibility by car	0.20
	Current home 30-minute accessibility by public transport	0.53
	Expected next home 30-minute accessibility by car	0.16
Job and Financial	Household income	0.55
	Household income increase/decrease over a year	0.85
	(Household income increase/decrease over a year)^2	0.88
	Number of jobs increase/decrease over a year	0.83
	Structure of household is couple without children	0.39
	Number of members with job from home	0.39
	Number of members with professional job	0.36
	Number of members with managerial job	0.26
	Number of members with administrative job	0.32
	(Number of vehicles in household)^2	0.27
Attitudes	Being close to public transport is very important	0.16
	Being close to public transport is important	0.14
	Being close to shops, groceries is very important	0.15
	Being close to shops, groceries is important	0.15
	Home selection attitude-situation being important	0.20
Land-use	%Residential land-use in the home suburb	0.27
	%Education land-use in the home suburb	0.50
	%Industrial land-use in the home suburb	0.36
	%Commercial land-use in the home suburb	0.19
	%Hospital/medical land-use in the home suburb	0.22
	%Parkland land-use in the home suburb	0.24
	%Transport land-use in the home suburb	0.39
Intra-household	Component 1 in PCA of intra-household decision-making	0.26
	Decision on day-to-day spending is shared	0.25
	Decision on large household purchases is shared	0.28
	Decision on hours spent on paid work is shared	0.25
	Decision on savings and investmensts is shared	0.27
	Decision on social life and leisure activities is shared	0.24

5.4 Comparison between machine-learning feature selector and manual feature selection

In this section, we utilized two different approaches for feature selection: manual selection and automatic machine learning-based selection, applying them to a log-normal AFT model. The manual method leverages an updated set of features used in a prior study (Bostanara et al., 2021), a selection that the authors reported required considerable time and effort to optimize. The comparison is presented in **Table 7**. Interestingly, both models share many common features, yet they differ in some that are not necessarily statistically significant. Additionally, the signs of the parameter estimates are consistent between the two models, indicating agreement on whether specific features accelerate or decelerate the rate of relocation. A positive coefficient implies that as the predictor increases, there's a deceleration in the time to the event (i.e., longer survival time), while a negative coefficient indicates an acceleration in the time to the event (i.e., shorter survival time). Both models identify statistically significant predictors that elucidate the reasons and timings for residential relocations. For example, in the manually selected model, variables like "Is house?," "Number of secondary students," and "Purchasing a

new vehicle during last year" are significantly influential but are either absent or insignificant in the automatic model. Conversely, the automatic model includes unique variables such as "Average of travel time to work by public transport" and "Starting a new job during the last two years," along with polynomial terms.

Figure 5. A word cloud map representing the most predictive features (the larger a variable name is, the more time it is picked by a feature selector)

In terms of results interpretation, in summary, homeowners are less likely to relocate, while higher rents accelerate this decision. Recent life changes, such as starting a new job or study, increase the speed of relocation. Longer car commutes to work decrease relocation chances. Moreover, a higher proportion of educational land-use in one's suburb notably increases the likelihood of moving, highlighting the impact of neighborhood characteristics on relocation.

Intriguingly, the automatic model exhibits a slightly lower Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), suggesting a superior model fit. This underscores the efficacy of automatic feature selectors, which, while considering similar features, are able to identify subtler, more potent variables (e.g., polynomial terms) that enhance model fit. To assess the statistical significance of the differences between the two models, we conducted further tests, as detailed below.

5.4.1 ANOVA model comparison

An Analysis of Variance (ANOVA) test was conducted to statistically compare the performance of the two models. The residual degrees of freedom were 993 for the manual feature selection model and 997 for the automatic feature selection model. The test yielded a deviance of -118 and an extremely significant p-value (1.47E-24), indicating that the models are statistically different. This suggests that the variables selected by the automatic feature selection method do, in fact, contribute to the explanation of residential relocation in Sydney in a way that is significantly different from the manual selection model.

Table 7. Manual feature selection versus automatic feature selection models' comparison

	Manual Feature Selection			Automatic Feature Selection		
Variable	**Coefficient**	**p-value**		**Coefficient**	**p-value**	
Intercept	2.191	1.8E-13	***	1.541	2.5E-16	***
Max age in household	0.015	6.9E-01				
Number of under 18	-0.045	2.8E-01				
Number of over 18				0.189	6.9E-02	*
(Number of over 18)^2	-0.055	1.2E-07	***	-0.081	9.7E-05	***
Is owner?	0.786	8.1E-11	***	0.531	1.3E-05	***
Is house?	0.203	7.7E-02	*	0.129	2.2E-01	
Rooms less than two	-0.224	5.1E-02	*	-0.178	9.5E-02	*
Household rent payment	-0.109	3.2E-02	**	-0.215	4.1E-02	**
(Household rent payment)^2				0.023	3.3E-01	
Household income				-0.035	4.0E-01	
Household income increase/decrease over a year	-0.015	7.6E-01		-0.293	3.2E-03	***
(Household income increase/decrease over a year)^2				0.112	3.2E-06	***
Intra-household decision-making being shared	-0.032	5.8E-01				
Number of primary students in household	0.221	2.4E-03	***	0.067	6.6E-01	
Number of secondary students in household	0.140	8.2E-02	**	0.170	3.0E-01	
Number of tertiary students in household	0.263	1.1E-03	***			
(Number of primary students in household)^2				0.005	9.1E-01	
(Number of secondary students in household)^2				-0.008	8.9E-01	
(Number of tertiary students in household)^2				0.081	2.4E-02	**
Number of administrative job workers	0.101	4.4E-01				
Number of working from home workers	0.109	4.2E-01				
Number of jobs changed over a year				0.171	9.1E-02	*
Use of public transport for study trips	-0.144	2.4E-01				
Use of private car for commute trips	-0.002	9.9E-01				
Home selection attitude-situation being important	-0.494	6.6E-02	*			
Childbirth during last year	-0.497	2.1E-02	**	-0.091	7.9E-01	
Childbirth during last two years				-0.439	1.3E-01	.
Starting a new job during last year	-0.660	8.4E-07	***	-0.257	1.4E-01	.
Starting a new job during last two years				-0.562	1.7E-04	***
Commencing a new study during last year	-0.883	5.6E-05	***	-0.091	7.5E-01	
Commencing a new study during last two years				-0.765	8.6E-04	***
Graduating a study during last two years				-0.623	2.7E-04	***
Purchasing a new vehicle during last year	-0.552	2.4E-03	***			
Purchasing a new vehicle during last two years				-0.643	1.3E-06	***
Average of travel time to work by car	-0.004	5.0E-04	***	-0.007	6.8E-08	***
Average of travel time to work by public transport				0.001	5.6E-11	***
Average of travel time to school by car	0.001	8.0E-06	***	0.001	8.0E-10	***
%Commercial land-use in the home suburb	-0.429	3.5E-01		-0.311	4.7E-01	
%Education land-use in the home suburb	-2.102	3.4E-02	**	-2.307	1.2E-02	**
Log(scale)	0.200	4.5E-10	***	0.1203	1.6E-04	
Loglikelihood		-1775			-1716	
AIC		3557			3440	
BIC		3577			3459	
ANOVA						
Residual Degrees of Freedom		993			997	
-2 Log-Likelihood (-2*LL)		3432			3549	
Deviance			-118			
p-value (Pr(>Chi))			1.47E-24			
Bootstrap Analysis: Comparative Evaluation of C-Index Values using Welch Two Sample t-test						
Mean C-index		0.7555			0.7909	
t-statistic (t)			85.668			
Degrees of Freedom (df)			1953.2			
p-value			< 2.2e-16			
95% Confidence Interval			0.0347 - 0.0363			
alternative hypothesis: true difference in means is not equal to 0						

5.4.2 Bootstrap analysis for C-index comparison

A bootstrap analysis using the Welch Two Sample t-test compared the concordance index (C-index) values of the two models. The mean C-index was 0.7555 for the manual feature selection model and 0.7909 for the automatic feature selection model. The t-statistic was calculated to be 85.668 with degrees of freedom at 1953.2, and the p-value was significantly less than 2.2e-16. This result strongly indicates that the model built with automatic feature selection offers superior predictive accuracy in understanding the factors that influence residential relocations in Sydney.

5.4.3 Characteristics of concordant and discordant pairs

In this section, we utilized the results from the machine learning feature selection model to extract both concordant and discordant pairs. We then conducted a descriptive analysis to discern potential differences in the characteristics of these pairs, aiming to identify patterns among those not accurately estimated. **Table 8** presents features with the most significant differences between discordant and concordant pairs. For the majority of variables, the mean differences between these pairs range from -0.03 to 0.03. Notably, discordant pairs feature 11% more households that experienced home relocation (event = 1), indicating the model's reduced proficiency in estimating failure cases (relocated pairs). The model appears better equipped to predict the relocation timing of homeowners as opposed to renters. Additionally, the model indicates that households with shared intra-household decision-making variables are more likely to be estimated concordantly.

Table 8. Comparison between concordant and discordant pairs in automatic feature selection model

				Discordant pairs		Concordant pairs		Difference	
Variable	**Min**	**Max**	**Range**	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**	**Absolute**	**Standardized↑**
Event	0	1	1	0.82	0.38	0.71	0.45	0.11	0.11
Rooms less than two	0	1	1	0.41	0.49	0.38	0.49	0.03	0.03
Is house?	0	1	1	0.59	0.49	0.61	0.49	-0.02	-0.02
Being close to shops, groceries is very important	0	1	1	0.30	0.46	0.32	0.47	-0.02	-0.02
Average of travel time to work by public transport	0	1200	1200	339	353	368	356	-29	-0.02
Decision on day-to-day spending is shared	0	1	1	0.23	0.42	0.25	0.43	-0.03	-0.03
Decision on savings and investments is shared	0	1	1	0.29	0.45	0.31	0.46	-0.03	-0.03
Decision on social life and leisure activities is shared	0	1	1	0.32	0.47	0.34	0.47	-0.03	-0.03
First component in PCA (Principal component) of intra-household decision-making	0	2	2	0.61	0.81	0.68	0.83	-0.06	-0.03
Average of travel time to school by public transport	0	1200	1200	780	482	818	455	-38	-0.03
Household structure: couple with/without children	0	1	1	0.55	0.50	0.59	0.49	-0.04	-0.04
Decision on large household purchases is shared	0	1	1	0.31	0.46	0.35	0.48	-0.04	-0.04
Is owner?	0	1	1	0.43	0.50	0.48	0.50	-0.05	-0.05

5.4.4 Pairs' concordance improvement through automatic feature selection

In this section, we utilized results from both the manual and machine learning feature selection models to extract the concordant and discordant pairs from each model. We listed the pairs that were discordant in the manual feature selection model but concordant in the automatic feature selection model. Additionally, pairs that were concordant in both models were also enumerated. A subsequent descriptive analysis aimed to identify potential differences in the characteristics of these pairs, emphasizing trends that the automatic feature selection could estimate accurately. **Table 9** displays features with the most significant differences between the pairs. Notably, the automatic feature selection model proved more effective in accurately predicting the relocation timing of individuals who had relocated (event = 1). Moreover, the automatic model seems better suited for predicting the relocation timing of those who had experienced a life-course event in the past two years.

Table 9. Comparison between improved concordant pairs with automatic feature selection model

				Discordant pairs improved to concordant pairs		Concordant pairs in both models		Difference	
Variable	**Min**	**Max**	**Range**	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**	**Absolute**	**Standardized↑**
Event	0	1	1	0.76	0.43	0.70	0.46	0.06	0.06
Purchasing a new vehicle during last two years	0	1	1	0.16	0.37	0.12	0.33	0.04	0.04
Average of travel time to work by public transport	0	1200	1200	406	370	363	354	42	0.04
Graduating a study during last two years	0	1	1	0.11	0.32	0.09	0.28	0.03	0.03
Commencing a new study during last year	0	1	1	0.03	0.18	0.06	0.23	-0.02	-0.02
Being close to public transport is very important	0	1	1	0.31	0.46	0.34	0.47	-0.03	-0.03
Being close to shops, groceries is very important	0	1	1	0.29	0.45	0.33	0.47	-0.04	-0.04
Starting a new job during last year	0	1	1	0.12	0.32	0.16	0.37	-0.05	-0.05

5.4.5 Ensemble model of classical models

Building upon previous research, classical models equipped with automated feature selection mechanisms showcased performance levels comparable to machine learning approaches, whilst preserving the intrinsic benefits of the traditional paradigms. To advance accuracy metrics, ensemble methods were engaged, specifically stacking ensemble models, a practice that incorporates predictions from multiple models (base models) to train a higher-level algorithm (meta-model). The semi-parametric model, Cox-PH, was incorporated alongside an array of parametric models, including Weibull, Exponential, Gaussian, Logistic, Log-normal, and Log-logistic, serving as the foundation or base models. These predictions then informed the explanatory variables of a subsequent meta-model. Despite rigorous experimentation with linear regression and finely tuned XG-boost algorithms as potential meta-models, no significant improvement in outcome metrics was observed. As evidenced in **Table 10**, the c-index measures of individual models oscillated around 0.77, with the ensemble model barely nudging to 0.78.

Digging deeper into these outcomes, we shifted our focus to the C-index metric. An analytical exploration was embarked upon to identify patterns in the distribution of concordant and discordant pairs derived from model predictions. Our primary objective was to discern whether different models encountered challenges with similar pairs of observations. Consequently, we scrutinized the overlap of discordant pairs across various models. As delineated in **Table 10**, a significant 64% of discordant pairs were consistent across all models, and an additional 21% were shared amongst three to six models. This consistency suggests that regardless of the model employed, there's a recurring pattern in which pairs are correctly classified.

Table 10. Measures of the base models in the ensemble model

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Average
	Weibull	Exponential	Gaussian	Logistic	Log-normal	Log-logistic	COX-PH	-
Count of concordant/discordant pairs								
Number of relevant	280167	280167	280167	280167	280167	280167	280167	280167
Number of concordant pairs	216148	216216	214568	214962	217612	217615	216364	216212
Number of discordant pairs	64019	63951	65599	65205	62555	62552	63803	63955
C-Index	0.77	0.77	0.77	0.77	0.78	0.78	0.77	0.77
Difference between the duration of the pairs								
Discordant pairs								
Average duration	7.33	7.36	7.34	7.40	7.67	7.60	7.36	
Average duration difference	5.74	5.81	5.98	6.02	6.02	5.96	5.76	
Concordant pairs								
Average duration	8.70	8.69	8.71	8.69	8.60	8.62	8.69	
Average duration difference	10.34	10.31	10.30	10.28	10.23	10.24	10.33	
All pairs								
Average duration	8.61							
Average duration difference	9.27							
Shared discordant pairs between models								
Count	4409	7207	3635	3343	5850	3695	50452	
Percent	0.06	0.09	0.05	0.04	0.07	0.05	0.64	

A meticulous analysis of pair durations offered intriguing revelations. Discordant pairs exhibited an average duration difference ranging from 5.74 to 6.02 across models. In contrast, concordant pairs presented a higher average duration difference, spanning from 10.23 to 10.34. When all pairs were taken into account, this average was 9.27. This trend is also mirrored in the average durations of both discordant and concordant pairs. **Figure 6**'s histogram elucidates this pattern: the duration difference distribution for discordant pairs reveals a pronounced spike around zero, tapering to a slender right tail. In contrast, the distribution for concordant pairs is more evenly dispersed, marked by a significant right tail and a modest peak around zero. This divergence highlights a pervasive challenge: models tend to misclassify pairs that not only are proximate in duration but also have relatively shorter durations.

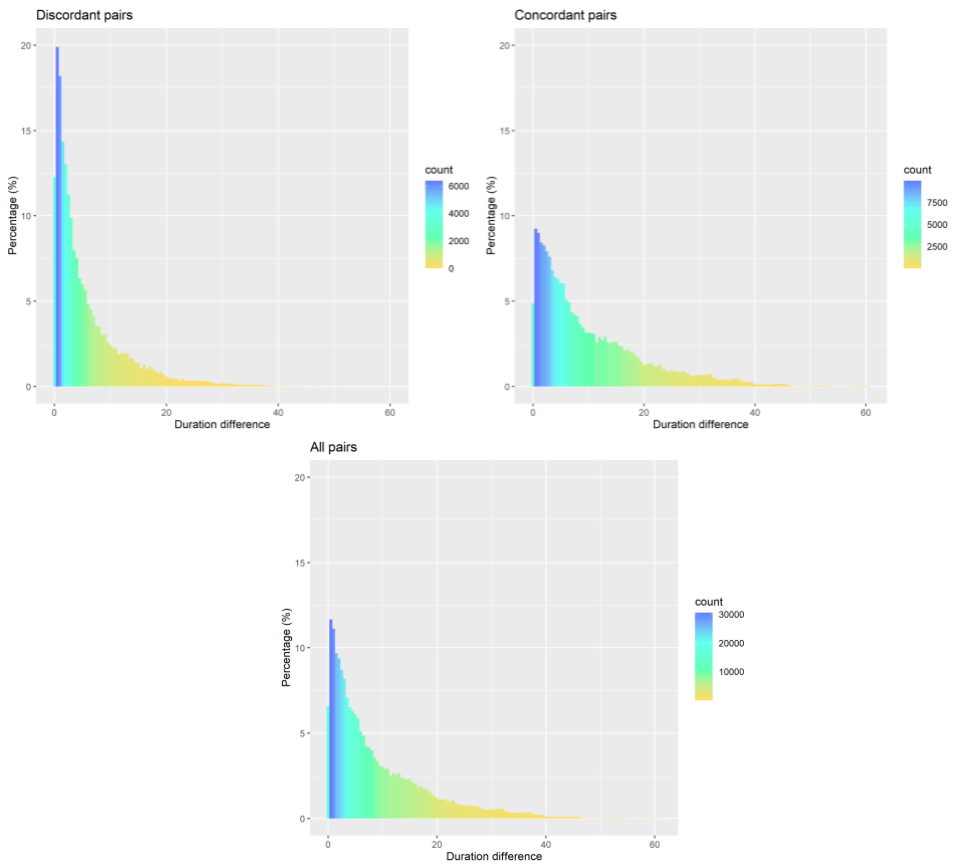

Figure 6. Histogram of duration difference in the pairs considered for C-Index measure

6 Conclusions

Understanding household residential relocation timing is crucial for transport and urban planners. This study analyzed the timing of 1,024 household relocations in Sydney, Australia, using a diverse and high-dimensional dataset. While existing literature explores various methodologies and influential variables for residential relocation duration, recent advancements in machine learning provide new tools tailored for survival data. Such advancements present a prime opportunity to refine classical survival models, particularly when dealing with complex, high-dimensional relocation data. In our research, we benchmarked ten machine learning survival techniques with six feature selection methods against three classic survival models. Findings indicated that while classical models underperformed without feature selections, they closely mirrored machine learning outcomes when integrated with tree-based automated feature selectors. Notably, the GBM, tree-based XGBoost, and Random Forest SRC models emerged as the most potent, with tree-based models exhibiting stability and minimal sensitivity to feature selection algorithms.

In the study, we conducted a comprehensive analysis of 163 features, with each feature having multiple selection opportunities across various learner and feature selector combinations. Key insights from the feature importance analyses reveal home-related features as dominant predictors, with homeownership, indicated by the “Is owner?” feature, being the most predictive one, boasting an importance score of 0.98. Life events, such as “Childbirth” and “Vehicle purchase” within the last two years, along with

accessibility features—particularly those related to public transport, the number of students, and financial situations—emerged as the most crucial features among all.

Further, two feature selection techniques — manual and machine learning-based automatic approaches — were analyzed. Inspired by Bostanara et al. (2021), the manual model generally aligned with the automatic model in many features, though they diverged on some influential variables. Statistical tests, such as ANOVA and bootstrap analysis, emphasized the superior performance of the automatic model, highlighting its increased predictive accuracy. Delving deeper, descriptive analyses on concordant and discordant pairs revealed the automatic model's enhanced capability to predict certain residential relocations, particularly those linked to recent life-course events. Notably, the automatic model excelled in estimating the relocation timing of homeowners compared to renters, and for those households exhibiting shared intra-household decision-making. The automatic approach also displayed a heightened accuracy for individuals having experienced a major life event within the past two years.

In this study, classical models integrated with automated feature selection techniques exhibited comparable performance to advanced machine learning methods, maintaining inherent advantages of conventional approaches. Stacking ensemble models, incorporating predictions from diverse models like Cox-PH and others, aimed to refine accuracy. However, even with rigorous tests using methods like XG-boost, enhancements in outcome metrics were minimal. In-depth analysis revealed consistent challenges across models in categorizing certain observation pairs, especially those with closer and shorter durations, underscoring a common modeling limitation.

Accessibility, a key factor in home selection, was analyzed in this study. We considered three main accessibility metrics: average household travel time to work and school, local land-use structure, and the number of jobs within a 30-minute radius. We hypothesized that considering the accessibility of both current and potential future homes would be crucial for a residential relocation model. A linear model was developed to gauge households' inclination to modify their accessibility post-relocation. The self-selection bias was mitigated by directly asking households about their preferences for home attributes, though such data is rarely available, limiting the model's predictive utility. However, future research could employ ordered logit models to estimate household attitudes towards various home attributes. This study has policy implications, emphasizing the promotion of a 30-minute accessible city and the enhancement of urban greenspaces, footpaths, and public transport equity. In terms of accessibility's impact on residential relocation duration, the accessibility of present and prospective homes was vital in approximately 20% of the models. Additionally, land-use and commute time via public transport were significant in 30% - 55% and 85% of models, respectively.

Our study offers foundational insights into residential relocation dynamics, paving the way for more intricate investigations. By integrating disaggregated prediction methods like agent-based models with synthesized populations, we gain a nuanced understanding of individual decision-making, guiding effective policy and sustainable urban planning. While our research emphasizes relocation timing and accessibility shifts, we haven't detailed household location choices. Recognizing this connection's significance, future research will incorporate discrete choice models to explore location decisions. Merging machine learning with these models also presents a promising avenue for further exploration.

Acknowledgments

This research was supported by an Australian Government Research Training Program (RTP) Scholarship. Map data copyrighted OpenStreetMap contributors and available from https://www.openstreetmap.org.

Author contribution

Maryam Bostanara: Conceptualization, data curation, formal analysis, visualization, methodology, writing—original draft, writing—review and editing. Amarin Siripanich: Data curation, formal analysis, methodology, writing—review and editing. Milad Ghasri: Data curation, writing—review and editing. Taha Hossein Rashidi: Conceptualization, data curation, methodology, writing—review and editing, supervision, funding acquisition, project administration, resources.

Declaration of generative AI in scientific writing

During the preparation of this work the authors used ChatGPT for textual review. This tool was instrumental in conducting thorough grammar checks, enhancing sentence structures, and ensuring overall coherence and cohesion throughout the document. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Appendix

Appendix available as a supplemental file at https://doi.org/10.5198/jtlu.2024.2440.

References

Aditjandra, P. T., Cao, X. Y., & Mulley, C. (2016). Exploring changes in public transport use and walking following residential relocation A British case study. *Journal of Transport and Land Use*, *9*(3), 77-95. https://doi.org/10.5198/jtlu.2015.588

Aghaabbasi, M., Shekari, Z. A., Shah, M. Z., Olakunle, O., Armaghani, D. J., & Moeinaddini, M. (2020). Predicting the use frequency of ride-sourcing by off-campus university students through random forest and Bayesian network techniques. *Transportation Research Part A: Policy and Practice*, *136*, 262-281. https://doi.org/https://doi.org/10.1016/j.tra.2020.04.013

Australian Bureau of Statistics (2016). *'SOCIO-ECONOMIC INDEXES FOR AREAS (SEIFA)'*. Retrieved 27 October 2022 from https://www.abs.gov.au/ausstats/abs@.nsf/mf/2033.0.55.001

Axhausen, K. W., König, A., Scott, D. M., & Jürgens, C. (2004). *Locations, Commitments and Activity Spaces* Human Behaviour and Traffic Networks, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-07809-9_9

Bender, A., Rügamer, D., Scheipl, F., & Bischl, B. (2021). *A General Machine Learning Framework for Survival Analysis* Machine Learning and Knowledge Discovery in Databases, Cham. https://doi.org/10.1007/978-3-030-67664-3_10

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., & Jones, Z. M. (2016). mlr: Machine Learning in R. *The Journal of Machine Learning Research*, *17*(1), 5938-5942.

Bostanara, M., Hossein Rashidi, T., Khan, N. A., Auld, J., Ghasri, M., & Grazian, C. (2023). The co-determination of home and workplace relocation durations using survival copula analysis. *Computers, Environment and Urban Systems*, *99*, 101898. https://doi.org/10.1016/j.compenvurbsys.2022.101898

Bostanara, M., Rashidi, T. H., Auld, J. A., & Ghasri, M. (2021). A comparison between residential relocation timing of Sydney and Chicago residents: A Bayesian survival analysis. *Computers, Environment and Urban Systems*, *89*, 101659. https://doi.org/10.1016/j.compenvurbsys.2021.101659

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Buckle, C. (2017). Residential mobility and moving home. *11*(5), e12314. https://doi.org/https://doi.org/10.1111/gec3.12314

Cai, Q., Abdel-Aty, M., Sun, Y., Lee, J., & Yuan, J. (2019). Applying a deep learning approach for transportation safety planning by using high-resolution transportation and land use data. *Transportation Research Part A: Policy and Practice*, *127*, 71-85. https://doi.org/10.1016/j.tra.2019.07.010

Cao, X., Mokhtarian, P. L., & Handy, S. L. (2009). Examining the Impacts of Residential Self-Selection on Travel Behaviour: A Focus on Empirical Findings. *Transport Reviews*, *29*(3), 359-395. https://doi.org/10.1080/01441640802539195

Cervero, R. (2003). City CarShare: First-Year Travel Demand Impacts. *1839*(1), 159-166. https://doi.org/10.3141/1839-18

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., & Chen, K. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*. https://cran.ms.unimelb.edu.au/web/packages/xgboost/vignettes/xgboost.pdf

Cheng, Z., Wang, W., Lu, J., & Xing, X. (2020). Classifying the traffic state of urban expressways: A machine-learning approach. *Transportation Research Part A: Policy and Practice*, *137*, 411-428. https://doi.org/10.1016/j.tra.2018.10.035

Clark, W. A. V. (2013). Life course events and residential change: unpacking age effects on the probability of moving. *Journal of Population Research*, *30*(4), 319-334. https://doi.org/10.1007/s12546-013-9116-y

Colabianchi, N. (2009). Does the built environment matter for physical activity? *Current Cardiovascular Risk Reports*, *3*(4), 302-307. https://doi.org/10.1007/s12170-009-0046-3

de Palma, A., Picard, N., & Waddell, P. (2007). Discrete choice models with capacity constraints: An empirical analysis of the housing market of the greater Paris region. *Journal of Urban Economics*, *62*(2), 204-230. https://doi.org/10.1016/j.jue.2007.02.007

De Vos, J., & Ettema, D. (2020). Travel and residential change: An introduction. *Travel Behaviour and Society*, *19*, 33-35. https://doi.org/10.1016/j.tbs.2019.11.003

Dieleman, F. M. (2001). Modelling residential mobility; a review of recent trends in research. *Journal of Housing and the Built Environment*, *16*(3-4), 249-265. https://doi.org/10.1023/A:1012515709292

Ding, C., Chen, P., & Jiao, J. (2018). Non-linear effects of the built environment on automobile-involved pedestrian crash frequency: A machine learning approach. *Accident Analysis & Prevention*, *112*, 116-126. https://doi.org/10.1016/j.aap.2017.12.026

Frank, L. D., Sallis, J. F., Conway, T. L., Chapman, J. E., Saelens, B. E., & Bachman, W. (2006). Many Pathways from Land Use to Health: Associations between Neighborhood Walkability and Active Transportation, Body Mass Index, and Air Quality. *Journal of the American Planning Association*, *72*(1), 75-87. https://doi.org/10.1080/01944360608976725

Ghasri, M., Rashidi, T., & Auld, J. (2022). Determinants of residential mobility: an adaptive retrospective survey method. *Transportation Letters*, 1-13. https://doi.org/10.1080/19427867.2022.2038347

Gordon, L., & Olshen, R. A. (1985). Tree-structured survival analysis. *Cancer treatment reports*, *69*(10), 1065-1069. http://europepmc.org/abstract/MED/4042086

Habib, M. A., & Miller, E. J. (2009). Reference-Dependent Residential Location Choice Model within a Relocation Context. *Transportation Research Record*, *2133*(1), 92-99. https://doi.org/10.3141/2133-10

Handy, S. L., & Clifton, K. J. (2001). Local shopping as a strategy for reducing automobile travel. *Transportation*, *28*(4), 317-346. https://doi.org/10.1023/A:1011850618753

Hastie, T., & Qian, J. (2014). Glmnet vignette. *9*(2016), 1-30.

Hedman, L., & van Ham, M. (2012). Understanding Neighbourhood Effects: Selection Bias and Residential Mobility. In M. van Ham, D. Manley, N. Bailey, L. Simpson, & D. Maclennan (Eds.), *Neighbourhood Effects Research: New Perspectives* (pp. 79-99). Springer Netherlands. https://doi.org/10.1007/978-94-007-2309-2_4

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., & Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics*, *7*(3), 355-373. https://doi.org/10.1093/biostatistics/kxj011

Jenkins, S. P. (2005). Survival analysis. *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK*, *42*, 54-56.

Jović, A., Brkić, K., & Bogunović, N. (2015, 25-29 May 2015). A review of feature selection methods with applications. 2015 38th International Convention on

Information and Communication Technology, Electronics and Microelectronics (MIPRO),

Kern, C., Klausch, T., & Kreuter, F. (2019). Tree-based Machine Learning Methods for Survey Research. *Surv Res Methods*, *13*(1), 73-93.

Kim, J. H., Pagliara, F., & Preston, J. (2005). The Intention to Move and Residential Location Choice Behaviour. *Urban Studies*, *42*(9), 1621-1636. https://doi.org/10.1080/00420980500185611

Kogalur, H. I. a. U. B. (2022). Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC). *manual*. https://cran.r-project.org/package=randomForestSRC

Landau, W. M. (2021). The targets R package: a dynamic make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software*, *6*(57), 2959. https://doi.org/10.21105/joss.02959

Lerman, S. R. (1975). *A disaggregate behavioral model of urban mobility decisions* Massachusetts Institute of Technology]. http://hdl.handle.net/1721.1/27388

Levinson, D. M. (2019). *The 30-minute city: designing for access*. Network Design Lab. https://hdl.handle.net/2123/21630

Li, L., Zhu, J., Zhang, H., Tan, H., Du, B., & Ran, B. (2020). Coupled application of generative adversarial networks and conventional neural networks for travel mode detection using GPS data. *Transportation Research Part A: Policy and Practice*, *136*, 282-292. https://doi.org/10.1016/j.tra.2020.04.005

Lin, T., Wang, D., & Zhou, M. (2018). Residential relocation and changes in travel behavior: what is the role of social context change? *Transportation Research Part A: Policy and Practice*, *111*, 360-374. https://doi.org/10.1016/j.tra.2018.03.015

Liu, D.-R., Li, H.-L., & Wang, D. (2015). Feature selection and feature learning for high-dimensional batch reinforcement learning: A survey. *International Journal of Automation and Computing*, *12*(3), 229-242. https://doi.org/10.1007/s11633-015-0893-y

Longato, E., Vettoretti, M., & Di Camillo, B. (2020). A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models. *Journal of Biomedical Informatics*, *108*, 103496. https://doi.org/10.1016/j.jbi.2020.103496

Lowry, I. S. (1964). *A model of metropolis*. https://apps.dtic.mil/sti/citations/tr/AD0603670

Miller, E. J. (2018). Accessibility: measurement and application in transportation planning. *Transport Reviews*, *38*(5), 551-555. https://doi.org/10.1080/01441647.2018.1492778

Mokhtarian, P. L., & Cao, X. (2008). Examining the impacts of residential self-selection on travel behavior: A focus on methodologies. *Transportation Research Part B: Methodological*, *42*(3), 204-228. https://doi.org/10.1016/j.trb.2007.07.006

Parmar, J., Das, P., & Dave, S. M. (2021). A machine learning approach for modelling parking duration in urban land-use. *Physica A: Statistical Mechanics and its Applications*, *572*, 125873. https://doi.org/10.1016/j.physa.2021.125873

Pereira, R. H., Saraiva, M., Herszenhut, D., Braga, C. K. V., & Conway, M. W. (2021). r5r: rapid realistic routing on multimodal transport networks with r 5 in r. *TRANSPORT FINDINGS* 21262.

Pineda-Jaramillo, J., & Arbeláez-Arenas, Ó. (2022). Assessing the Performance of Gradient-Boosting Models for Predicting the Travel Mode Choice Using Household Survey Data. *148*(2), 04022007. https://doi.org/doi:10.1061/(ASCE)UP.1943-5444.0000830

Prillwitz, J., Harms, S., & Lanzendorf, M. (2007). Interactions between Residential Relocations, Life Course Events, and Daily Commute Distances. *Transportation Research Record*, *2021*(1), 64-69. https://doi.org/10.3141/2021-08

Radovic, M., Ghalwash, M., Filipovic, N., & Obradovic, Z. (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics*, *18*(1), 9. https://doi.org/10.1186/s12859-016-1423-9

Rashidi, T. H., & Ghasri, M. (2017). A competing survival analysis for housing relocation behaviour and risk aversion in a resilient housing market. *Environment and Planning B: Urban Analytics and City Science*, *46*(1), 122-142. https://doi.org/10.1177/2399808317703381

Rashidi, T. H., Mohammadian, A., & Koppelman, F. S. (2011). Modeling interdependencies between vehicle transaction, residential relocation and job change. *Transportation*, *38*(6), 909. https://doi.org/10.1007/s11116-011-9359-4

Rossi, P. H. (1955). *Why families move: A study in the social psychology of urban residential mobility*. Free Press. https://doi.org/10.1177/1440783396032001

Sánchez, A. C., & Andrews, D. (2011). Residential Mobility and Public Policy in OECD Countries. *OECD Journal: Economic Studies*, *2011*(1), 1-22. https://doi.org/10.1787/eco_studies-2011-5kg0vswqt240

Sarkar, J. P., Saha, I., Sarkar, A., & Maulik, U. (2021). Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers. *Computers in Biology and Medicine*, *131*, 104244. https://doi.org/10.1016/j.compbiomed.2021.104244

Scheiner, J., & Holz-Rau, C. (2013). A comprehensive study of life course, cohort, and period effects on changes in travel mode use. *Transportation Research Part A: Policy and Practice*, *47*, 167-181. https://doi.org/10.1016/j.tra.2012.10.019

Scheuer, S., Haase, D., Haase, A., Wolff, M., & Wellmann, T. (2021). A glimpse into the future of exposure and vulnerabilities in cities? Modelling of residential location choice of urban population with random forest. *Nat. Hazards Earth Syst. Sci.*, *21*(1), 203-217. https://doi.org/10.5194/nhess-21-203-2021

Schirmer, P. M., van Eggermond, M. A. B., & Axhausen, K. W. (2014). The role of location in residential location choice models: a review of literature. *Journal of Transport and Land Use*, *7*(2), 3-21. https://doi.org/10.5198/jtlu.v7i2.740

Shen, Q. (2001). A Spatial Analysis of Job Openings and Access in a U.S. Metropolitan Area. *Journal of the American Planning Association*, *67*(1), 53-68. https://doi.org/10.1080/01944360108976355

Sonabend, R., Király, F. J., Bender, A., Bischl, B., & Lang, M. (2021). mlr3proba: an R package for machine learning in survival analysis. *Bioinformatics*, *37*(17), 2789-2791. https://doi.org/10.1093/bioinformatics/btab039

Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N. A., Trollor, J., & Brodaty, H. (2020). A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports*, *10*(1), 20410. https://doi.org/10.1038/s41598-020-77220-w

Sprumont, F., & Viti, F. (2018). The effect of workplace relocation on individuals' activity travel behavior. *Journal of Transport and Land Use*, *11*(1). https://doi.org/10.5198/jtlu.2018.1123

Srour, I. M., Kockelman, K. M., & Dunn, T. P. (2002). Accessibility Indices: Connection to Residential Land Prices and Location Choices. *Transportation Research Record*, *1805*(1), 25-34. https://doi.org/10.3141/1805-04

Therneau, T., Atkinson, B., Ripley, B., & Ripley, M. B. (2015). *Package 'rpart'*. https://cran.r-project.org/web/packages/rpart/rpart.pdf

Thomas, M. J., Stillwell, J. C., & Gould, M. I. (2016). Modelling the duration of residence and plans for future residential relocation: A multilevel analysis. *Transactions of the Institute of British Geographers*, *41*(3), 297-312. https://doi.org/10.1111/tran.12123

Tran, M. T., Zhang, J., Chikaraishi, M., & Fujiwara, A. (2016). A joint analysis of residential location, work location and commuting mode choices in Hanoi, Vietnam. *Journal of Transport Geography*, *54*, 181-193. https://doi.org/10.1016/j.jtrangeo.2016.06.003

Wachs, M., & Kumagai, T. G. (1973). Physical accessibility as a social indicator. *Socio-economic Planning Sciences*, *7*(5), 437-456. https://doi.org/10.1016/0038-0121(73)90041-4

Wang, P., Li, Y., & Reddy, C. K. (2019). Machine Learning for Survival Analysis: A Survey. *51*(6 %J ACM Comput. Surv.), Article 110. https://doi.org/10.1145/3214306

Wright, M. N., & Ziegler, A. (2015). *ranger: A fast implementation of random forests for high dimensional data in C++ and R*. https://doi.org/10.48550/arXiv.1508.04409

Xu, T., Gao, J., & Li, Y. (2019). Machine learning-assisted evaluation of land use policies and plans in a rapidly urbanizing district in Chongqing, China. *Land Use Policy*, *87*, 104030. https://doi.org/10.1016/j.landusepol.2019.104030

Xue, F., & Yao, E. (2022). Adopting a random forest approach to model household residential relocation behavior. *Cities*, *125*, 103625. https://doi.org/10.1016/j.cities.2022.103625

Yi, C., & Kim, K. (2018). A Machine Learning Approach to the Residential Relocation Distance of Households in the Seoul Metropolitan Region. *10*(9), 2996. https://www.mdpi.com/2071-1050/10/9/2996

Zhang, J. (2014). Revisiting residential self-selection issues: A life-oriented approach. *Journal of Transport and Land Use*, *7*(3), 29-45. https://doi.org/10.5198/jtlu.v7i3.460

Zhou, B., & Kockelman, K. M. (2008). Microsimulation of Residential Land Development and Household Location Choices: Bidding for Land in Austin, Texas. *Transportation Research Record*, *2077*(1), 106-112. https://doi.org/10.3141/2077-14

Zhou, M., Le, D.-T., Nguyen-Phuoc, D. Q., Zegras, P. C., & Ferreira, J. (2021). Simulating impacts of Automated Mobility-on-Demand on accessibility and residential relocation. *Cities*, *118*, 103345. https://doi.org/10.1016/j.cities.2021.103345

Zolfaghari, A., Sivakumar, A., & Polak, J. W. (2012). Choice set pruning in residential location choice modelling: a comparison of sampling and choice set generation approaches in greater London. *Transportation Planning and Technology*, *35*(1), 87-106. https://doi.org/10.1080/03081060.2012.635420

Zondag, B., & Pieters, M. (2005). Influence of Accessibility on Residential Location Choice. *Transportation Research Record*, *1902*(1), 63-70. https://doi.org/10.1177/0361198105190200108