Journal Title
XX(X):1–10

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

ActiveCA: Time Use Data from the General Social Survey of Canada to Study Active Travel

Anon1, Anon2, Anon3

Abstract

This paper describes the open data product `ActiveCA`, with Canadian time use data about active travel. `ActiveCA` is an `R` data package that contains analysis-ready data from cycles General Social Survey cicles regarding active travel in Canada. The package provides data set detailing active trip origins, destinations, and duration, covering a wide variety of locations, such as home, work or school, libraries, museums, restaurants, bars, sports centers, health clinics, place of worship, and others. The package also details the respondent's region characteristics, specifying wheter they live in a metropolitan area and their province of residency. `ActiveCA` also provides pre-estimated impedance functions for active travel for Canadian Metropolitan Area and Census Agglomerations, which can be used in accessibility analysis to calculate the cost of travel between different locations. The package will continue to expand with contributions from the authors and the broader community through requests in the future. `ActiveCA` is freely accessible for exploration and download from the associated Github repository, where the documentation and code involved in creating and manipulating data are detailed.

Keywords

Active; mobility; walking; cycling; travel time; impedance; transportation; R

Introduction

This paper presents the open data product `ActiveCA`. Open data products (ODPs) are the outcome of a transparent process that transforms raw data (open or not) into analysis-ready data, in which all stages of development follow open principles (Arribas-Bel et al., 2021). ODPs differ from general open data due to their high utility, added value and open availability.

`ActiveCA` is an `R` data package that provides analysis-ready data from the General Social Survey of Canada Cycles 2 (1986), 7 (1992), 12 (1998), 19 (2005), 24 (2010), and 29 (2015) (Canada, 2024) regarding active travel. These data can be used to study walking and cycling in Canada for various purposes over time, and to calculate impedance functions for accessibility analysis. The package provides pre-estimated impedance functions for cycling and walking trips for Census Agglomerations and Census Metropolitan Areas.

To create this `R` data package, we collected, cleaned and processed the Time Use collections from the GSS surveys to make them ready for analysis. An `R` data package contains code, data and documentation in a standardized format that can be installed by `R` users via a centralized software repository, such as CRAN (Comprehensive R Archive Network) and GitHub. Although the GSS surveys are publicly sourced managed by Statistic Canada (Canada, 2024), preparing them for analysis can be time-consuming, tedious and perhaps not even possible for those who try, due to a lack of documentation or prior knowledge.

The aim of this paper is to walk readers through the data sets and invite others to experiment in its uses and applications. `ActiveCA` is freely available on GitHub for all to install and freely use in the spirit of open and reproducible research. In addition of enabling to obtain impedance functions, `ActiveCA` can be adopted in various applications that even go beyond the range of possibilities we have imagined. Not only the data, but also all the code documenting the processing methodology is available for consultation and evaluation in its repository. This package contributes to reducing the barrier to using the information contained in GSS surveys to provide data-driven decisions in transportation analysis.

General Social Survey (GSS) collection

Statistics Canada (2024) conducts GSS surveys to obtain data on social trends to track changes in Canadians' living conditions and well-being over time. This survey is used to understand how citizens spend and manage their time and what factors contribute to their happiness and stress. Created in 1985, the survey is part of a series of independent, annual, and cross-sectional surveys.

In addition to the main topic, each GSS cycle includes new content that addresses emerging and policy-relevant issues. Every five to seven years, the Time Use Surveys (Canada, 2022) collect data on respondents' participation and time spent on a wide range of everyday activities using a 24-hour retrospective diary, with information on the location of these activities (e.g. at home, at work, etc.) and, for non-personal activities, the people who were present with the respondent at the time of the activity. In addition, time-use surveys also cover topics related to leisure time, work-life balance, health, commuting, culture and sports, and many others.

The most recent time use survey was carried out in 2022 (Wray, 2024). However, the 2022 dataset has not been fully published and, because of this, our analysis focused on the surveys from 1986 to 2015 (1986, 1992, 1998, 2005, 2010 and 2015). Time Use surveys are composed of two data sets, the main one and the episode file, explained in the following subsections.

The Main File

The main file compiles an large array of aggregated data, summarizing the answers to the questionnaire and derived variables that summarize the respondents' time use across different activities, locations, and social interactions. This file documents the time and duration that respondents allocate to each activity and location. The Main File provides a overview of daily routines and social dynamics, not focusing on individual activity episodes. Additionally, this file categorizes activities into bigger groups and subcategories, facilitating the data's analytical utility with additional metrics such as total transit time, duration spent with household members, and counts of activities and episodes.

The Episode File

The episode file records detailed data for each activity episode reported by respondents. The entry includes the start and end times, duration, location, and accompanying social context, informing when and where activities occurred and with whom. The file distinguishes itself by focusing on individual episodes rather than respondents, with the data structured around the numerous activity instances that compose a day of the respondent. Although respondent-specific characteristics are not included within the episode file, it is possible to link the main file and the episode file by using an identifiable variable present in both data sets.

Methodology

For each selected cycle of the GSS surveys, we reviewed the episode files to identify cases with activities listed as walking or cycling, selecting the activities immediately before and after the mobility episode. After that, we labeled the code variables with their appropriate descriptions, identifying each origin and destination, transportation mode, as well the province and urban classification of the respondent's residency.

Additionally, for Census Agglomerations and Census Metropolitan Areas we pre-estimated impedance functions for cycling and walking trips. Impedance functions reveals important information about the travel behavior of the population, by describing the relationship between the population at an origin and their likelihood or ability to travel to specific destinations to access opportunities (Soukhov and Paez, 2024).

Impedance functions are commonly used in accessibility analysis to calculate the cost of travel between different locations (Hansen, 1959; Páez et al., 2012; Palacios and El-geneidy, 2022). However, these functions need to be calibrated in order to accurately represent travel behavior. One effective way of calibrating an impedance function is by using the trip length distribution (TLD) obtained from origin-destination data (Soukhov and Paez, 2024). The TLD represents the probability that a certain proportion of trips will be made at a specific cost, such as travel time. In this data set, low travel times are associated with a higher proportion of trips, while high travel times are associated with a lower proportion of trips.

For each combination of year, destination, and transportation mode (walking or cycling), we fitted the most suitable impedance function based on empirical data from the GSS surveys. We used the `fitdistrplus` package (Delignette-Muller and Dutang, 2015) to estimate the functions, selecting the distribution with the lowest Akaike information criterion (AIC) among exponential, gamma, log-normal, normal, and uniform types.

Results

Descriptive statistics

Considering GSS Cycles analyzed, we identified 21748 episodes that recorded active travel episodes, with trip duration ranging from 0 to 900 minutes, to twelve different destinations. `ActiveCA` includes all these episodes ready for analysis. Table 1 presents descriptive statistics on walking and cycling trips between 1986 and 2015, including metrics such as the count of recorded trips (count), and measures of trip duration in minutes: maximum (max), mean, median, and minimum (min). The 1986 survey did not include bicycle trips.

Table 1. Descriptive statistics for episodes with active transport records

		Year					
Mode	Statistic	1986	1992	1998	2005	2010	2015
Walking	Count	4347	1500	1670	5533	4379	3251
	Maximum	660	300	255	515	480	900
	Mean	21	19	11	12	12	17
	Median	10	10	5	10	8	10
	Minimum	1	1	1	0	0	5
	Standard deviation	31	25	17	16	17	27
Cycling	Count		135	119	333	236	245
	Maximum		240	90	180	153	120
	Mean		31	21	19	21	24
	Median		20	15	15	15	15
	Minimum		5	2	1	1	5
	Standard deviation		36	18	18	23	20

Table 1 shows that the median values for walking trips range between 5 and 10 minutes, while cycling trips have a consistent median of 15 minutes since 1998. The table also highlights very high maximum values, particularly for walking trips, with recorded episodes exceeding 4 hours in all cases.

Table 2 and 3 provide descriptive statistics for the two modes of transportation, split by destination categories, from 1986 to 1998 and from 2005 to 2015, respectively. In Table 2, one can observed that in 1986 and 1992, walking trips destined for `home` had the highest medians. However, by 1998, the highest medians shifted to trips to `work or school`, a transition that also occurred for cycling trips between 1992 and 1998. Table 3 indicates that the median duration for trips to `home` and `work or school` remained at 10 minutes.

Table 2. Comparison of travel statistics by mode and destination: 1986, 1992, 1998

		1986				1992				1998			
Destination	**Mode***	**Min***	**Med***	**Max***	**(%)***	**Min**	**Med**	**Max**	**(%)**	**Min**	**Med**	**Max**	**(%)**
Cycling	Home					5	20	240	55.6	2	15.0	90	52.9
	Other's home					5	10	145	18.5	2	10.0	80	17.6
	Work or school					5	15	45	25.9	5	20.0	75	29.4
Walking	Home	1	15	330	46.4	1	10	300	59.5	1	5.0	255	51.6
	Other's home	1	10	660	42.3	1	5	135	21.3	1	5.0	120	28.1
	Work or school	1	10	450	11.3	2	10	60	19.2	1	6.5	75	20.4

Note:
* The symbols used in this table represent the following: 'Min' denotes the minimum time to reach the destination; 'Max' denotes the maximum time to reach the destination; '(%)' indicates a percentage of the total time to reach the destination; 'Med' refers to the median time to reach the destination

Table 3. Comparison of travel statistics by mode and destination: 2005, 2010, 2015

		2005				2010				2015			
Destination	**Mode**	**Min***	**Med***	**Max***	**(%)***	**Min**	**Med**	**Max**	**(%)**	**Min**	**Med**	**Max**	**(%)**
Cycling	Cultural venues	10	12.5	15	0.6	10	25	30	1.3	15	15.0	15	0.8
	Grocery store	2	10.0	30	10.2	5	10	75	8.9	5	15.0	80	6.5
	Health clinic									10	15.0	90	2.0
	Home	1	15.0	180	48.9	1	15	135	50.4	5	20.0	120	46.9
	Neighbourhood									10	30.0	45	1.2
	Other's home	1	15.0	35	9.0	5	10	45	9.3	5	15.0	40	5.3
	Outdoors	5	15.0	45	6.0	3	10	115	3.8	15	20.0	30	1.2
	Place of worship	20	20.0	20	0.3					15	15.0	15	0.4
	Restaurant	5	20.0	35	3.0	10	15	153	2.1	10	17.5	60	4.1
	Sport area									10	15.0	15	2.9
	Work or school	1	15.0	90	21.9	1	15	100	24.2	5	15.0	120	28.6
Walking	Business									5	10.0	30	0.2
	Cultural venues	5	12.5	40	0.6	2	10	40	0.7	5	10.0	40	1.5
	Grocery store	1	10.0	90	12.5	1	8	105	13.2	5	10.0	130	11.8
	Health clinic									5	10.0	130	1.0
	Home	0	10.0	515	44.4	0	10	270	43.6	5	10.0	900	45.3
	Neighbourhood									5	10.0	60	2.1
	Other's home	1	5.0	300	11.7	0	5	140	11.3	5	10.0	120	7.3
	Outdoors	1	5.0	295	3.6	0	10	480	5.2	5	10.0	135	2.8
	Place of worship	1	10.0	30	0.8	1	8	60	0.9	5	15.0	45	1.1
	Restaurant	0	5.0	85	9.3	1	5	153	10.0	5	10.0	120	8.4
	Sport area									5	10.0	45	3.3
	Work or school	0	10.0	175	17.1	0	10	150	15.0	5	10.0	190	15.1

Note:
* The symbols used in this table represent the following: 'Min' denotes the minimum time to reach the destination; 'Max' denotes the maximum time to reach the destination; '(%)' indicates a percentage of the total time to reach the destination; 'Med' refers to the median time to reach the destination

`ActiveCA` also enables visual analysis of active travel in Canada using traditional exploratory data analysis techniques. Figures 1 and 2 show walking and cycling trips

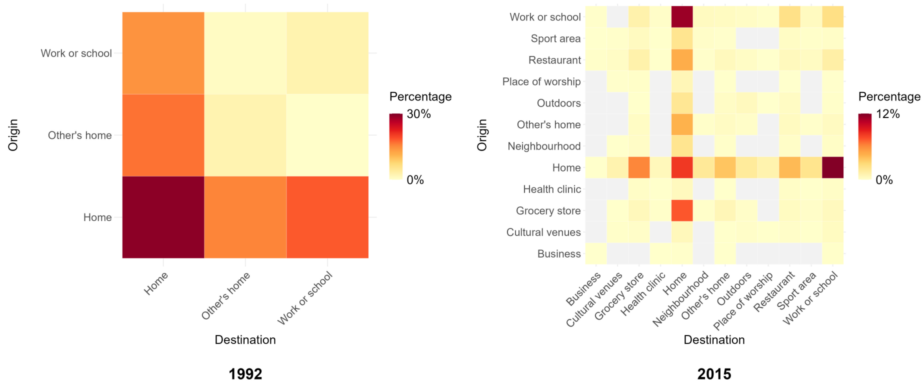

Figure 1. Percentage of walking trips categorized by origin and destination

from 1992 and 2015 through heat maps. These maps use color gradients to represent the percentage of trips between various origins and destinations, with darker colors indicating higher percentages and lighter colors representing less frequent routes. To avoid overwhelming the reader, we omitted the heat maps for the other years analyzed.

In 1992, walking trips with `home` as both the origin and destination made up the majority, accounting for about 30% of all walking trips. These trips often involved leisure activities, like short walks or dog walking. Following this, trips from `home` to `work or school` comprised 18% of walking trips. Overall, `home` emerged as a crucial hub, either as an origin or destination, with only 5% of trips not involving `home.` By 2015, `home` remained a significant node, but new locations distributed the proportion of trips to areas not considered in 1992. In 2015, the highest proportion of trips were from `home` to `work or school` (12%) and vice versa (11%). `home` to `home` accounted for 8% of trips, and `grocery stores` became a notable destination for those leaving `home` (6%), surpassing trips to `other's home` (4%).

For cycling trips, Figure 2, shows that in 1992, when this mode of transportation was first included as an activity, the majority of trips were from `home` to `work or school`, accounting for about 25% of cases. This pattern remained in 2015, with these trips representing 30% of the cases. However, a notable change occurred in `home` to `home` trips, which decreased significantly from 19% in 1992 to 5% in 2015.

Pre-estimated impedance functions for active travel in Canada

`ActiveCA` includes a total of 64 pre-estimated impedance functions for cycling and walking trips for Census Agglomerations and Census Metropolitan Areas. Table 5 shows the best impedance functions for walking trips to `Work or school` across survey years, where all distributions, except for a gamma function in 1998, are log-normal.

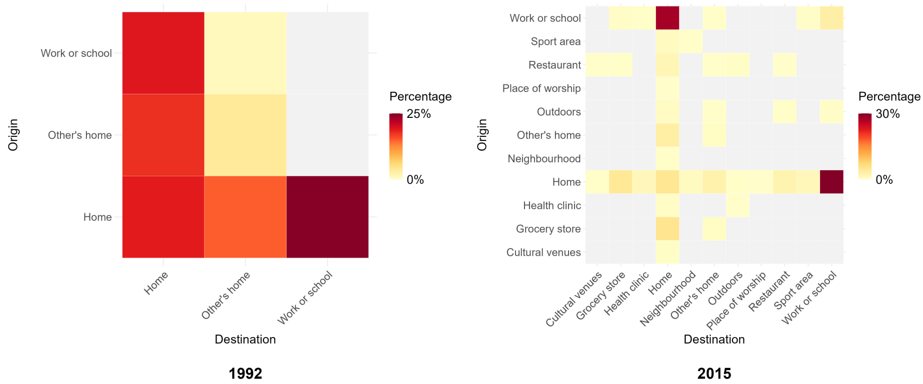

Figure 2. Percentage of cycling trips categorized by origin and destination

Table 4. Impedance functions and AIC for 'Walking' trips considering 'Work or school' as destination.

Table 5. Impedance functions and AIC for 'Cycling' trips considering 'Work or school' as destination.

Mode	Year	Impedance function	Parameter 1*	Parameter 2*	AIC	Count
Cycling	1992	Gamma	3.00	0.17	433582	19
	1998	Gamma	3.37	0.10	481536	19
	2005	Lognormal	2.93	0.70	888655	64
	2010	Lognormal	2.65	0.77	1292760	53
	2015	Lognormal	3.03	0.41	1162876	63
Walking	1992	Lognormal	2.38	0.70	2319400	113
	1998	Gamma	1.23	0.09	2318752	109
	2005	Lognormal	2.13	0.79	8182691	724
	2010	Lognormal	2.21	0.78	7917431	494
	2015	Lognormal	2.55	0.64	6612061	407

Note:
*For lognormal distributions, 'Parameter 1' and 'Parameter 2' refer to the mean and standard deviation of the distribution on the logarithmic scaler, respectively. For gamma distribution, 'Parameter 1' and 'Parameter 2' refer to the rate and shape of the distribution, respectively. 'AIC' means Akaike information criterion.

Figure 3 illustrates these fitted functions (red line) alongside histograms of empirical travel times (grey bars). As for the others pre-estimated functions, these examples allow to calculate gravity-based accessibility measures for active travel across various destinations and temporal scales in Canadian urban areas.

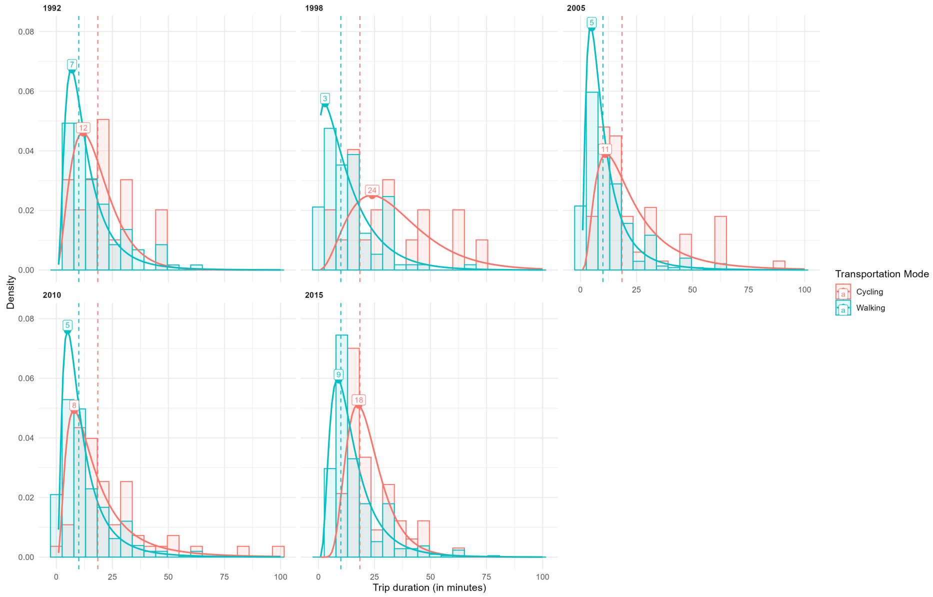

Figure 3. Empirical data and impedance functions fitted for ‘work or school‘ as destination.

Concluding remarks

This article presents `ActiveCA`, an open data product that provides analysis-ready data from Cycles 2 (1986), 7 (1992), 12 (1998), 19 (2005), 24 (2010), and 29 (2015) of the GSS surveys on active travel in Canada. In the form of an `R` data package, `ActiveCA` was developed after collecting, cleaning, and processing the survey data, providing information on origins, destinations, and duration of active travel, as well other information. Additionally, the package includes a series of pre-estimated impedance functions for walking and cycling trips, considering various destinations and time periods, for Canadian Metropolitan and Census Agglomeration areas.

The value of `ActiveCA` lies in its transparency, accessibility, and open infrastructure, which facilitates the addition of complementary data sets in the future. R users can seamlessly explore GSS walking and cycling episodes along with calibrated impedance functions, with the option to suggest enhancements to the package as needed. This article adopts the structure proposed by Anastasia and Páez (2023), whose work provided essential guidance for the creation of this package. Similarly, we aim to contribute to the academic community by promoting transparent research practices that encourage replication and innovation in related fields. We believe that `ActiveCA` will serve as a basis for further research on GSS and for the integration of additional data by the authors or the wider open source community.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Social Sciences and Humanities Research Council of Canada (*More description about the funding source after the review process*).

ORCID

Author 1
Author 2
Author 3

Data availability statement

The {ActiceCA} R data package can be found and installed on Github (*link*).

References

Arribas-Bel D, Green M, Rowe F and Singleton A (2021) Open data products-a framework for creating valuable analysis ready data. *Journal of Geographical Systems* 23(4): 497–514. DOI:10.1007/s10109-021-00363-5. URL https://doi.org/10.1007/s10109-021-00363-5.

Canada S (2022) Time use survey. Technical report. URL https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4503. Last Modified: 2024-06-04.

Canada S (2024) Statistics canada: Canada's national statistical agency. Technical report. URL https://www.statcan.gc.ca/en/start.

Delignette-Muller ML and Dutang C (2015) fitdistrplus: An r package for fitting distributions. *Journal of Statistical Software* 64(4): 1–34. DOI:10.18637/jss.v064.i04. URL https://www.jstatsoft.org/index.php/jss/article/view/v064i04.

Hansen WG (1959) How accessibility shapes land use. *Journal of the American Institute of Planners* 25(2): 73–76. DOI:10.1080/01944365908978307. URL https://doi.org/10.1080/01944365908978307. Publisher: Routledge _eprint: https://doi.org/10.1080/01944365908978307.

Palacios MS and El-geneidy A (2022) Cumulative versus gravity-based accessibility measures: Which one to use? *Findings* DOI:10.32866/001c.32444. URL https://findingspress.org/article/32444-cumulative-versus-gravity-based-accessibility-measures-which-one Publisher: Findings Press.

Páez A, Scott DM and Morency C (2012) Measuring accessibility: positive and normative implementations of various accessibility indicators. *Journal of Transport Geography* 25: 141–153. DOI:10.1016/j.jtrangeo.2012.03.016. URL https://www.sciencedirect.com/science/article/pii/S0966692312000798.

Soukhov A and Paez A (2024) Accessibility analysis for planning applications. Technical report. URL https://github.com/soukhova/MJ-Accessibility-Blogs.

Soukhov A and Páez A (2023) Tts2016r: A data set to study population and employment patterns from the 2016 transportation tomorrow survey in the greater golden horseshoe area, ontario, canada. *Environment and Planning B: Urban Analytics and City Science* 50(2): 556–563. DOI:10.1177/23998083221146781. URL https://doi.org/10.1177/23998083221146781. Publisher: SAGE Publications Ltd STM.

Wray D (2024) Telework, time use, and well-being: Evidence from the 2022 time use survey. Technical report. URL https://www150.statcan.gc.ca/n1/daily-quotidien/240605/dq240605a-eng.htm.