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

Journal Title
XX(X):1–10
©The Author(s) 0000
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/

SAGE

Anon1, Anon2, Anon3

Abstract

This paper describes {ActiveCA}, and open data product with Canadian time use data. {ActiveCA} is an R data package that contains analysis-ready data related to active travel spanning almost 40 years, extracted from cycles 2 (1986), 7 (1992), 12 (1998), 19 (2005), 24 (2010), and 29 (2015) of Canada's General Social Survey. Active travel is characterized by mode, with walking being part of every cycle and bicycling starting in 1992. The attributes of active trips are the types of locations of origins and destinations, the duration of trips, and episode weights for expanding the trips to population-wide estimates. Based on the year of the survey, a variety of locations are coded. In earlier cycles, these include home, work or school, and other's home, whereas in later cycles these are augmented with locations such as grocery stores, restaurants, outdoor destinations, and others. The geographical resolution includes the province and whether the episode was in an urban or rural setting.

Keywords

Active; mobility; walking; cycling; travel time; time-use; General Social Survey

Introduction

The objective of this paper is to introduce {ActiveCA}, an open data product with time use data from Canada's General Social Surveys. Open data products (ODPs) are the outcome of a process that transforms raw data (open or not) into analysis-ready data, following a transparent process in which all stages of development follow open principles (Arribas-Bel et al., 2021). ODPs, while still open, differ from general open data in their

degree of ease of access, their heightened usability, and potentially the value they add to the raw data.

{ActiveCA} provides analysis-ready data concerning active travel in Canada spanning a period of almost 40 years. The source of these data is Canada's program of General Social Surveys (GSSs). This program is designed to provide cross-sectional data on topics of interest to improve the well-being of Canadians. As part of this program, every five to seven years the survey is done on the topic of time use. Concretely, {ActiveCA} covers Cycles 2 (1986), 7 (1992), 12 (1998), 19 (2005), 24 (2010), and 29 (2015) of the GSS. Time use data in these surveys is coded using a very fine grain, from time spent in chores, leisure, and sleeping, to time spent working or at school. These surveys have proved valuable in investigations of mobility and quality of life (Spinney et al., 2009), the relationship between active travel and transit use (Lachapelle and Pinto, 2016), and travel behavior and time poverty (Kim et al., 2024), to name but a few examples.

For {ActiveCA} and using GSS Public Use Microdata Files (PUMFs), we extracted all data needed to characterize active travel in Canada, namely, episodes where the activity was identified as moving between an origin and a destination, either by walking or cycling. Although Statistics Canada offers Public Use Microdata Files and documentation for the GSS program (see Canada, 2024), accessing these files, and preparing them for analysis is not a straightforward matter, given their size and complexity. The process of extracting information of interest from the source files is time-consuming, tedious, and challenging and/or prone to error due to the expertise required to work with these files. To create {ActiveCA} we collected, cleaned, and processed the cycles of the GSS surveys concerning time use to make them ready for analysis.

{ActiveCA} is distributed as an R package with a number of data objects and their documentation. R packages contain code, data, and documentation in a standardized format that can be installed by R users via a software repository, such as CRAN (Comprehensive R Archive Network) or GitHub, which makes them an adroit medium to distribute analysis-ready data.

Given the level of interest in active travel (e.g., McCurdy et al., 2023), reducing the barriers to using data contained in rich, but difficult to access and use surveys, such as Canada's GSSs, is a worthy endeavour that can only improve data-driven decisions in transportation, urban, and health policy.

The rest of this paper discusses the sources of data, and the process implemented to retrieve and package them. Then, we explain how to use the package and show some selected examples of analysis to whet the imagination of potential users. This ODP provides not only data that are easy to use, but also all the code and documentation that make this a reproducible research project. In summary, {ActiveCA} aims to implement and inspire the best principles of open spatial sciences (Páez, 2021; Brunsdon and Comber, 2021).

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. The series of survey

on time use are used to understand how Canadian residents spend and manage their time, and what factors contribute to their happiness and stress. The GSS program was created in 1985, and is serialized to provide a collection of annual, representative cross-sectional surveys.

The topics of the survey cycle every few years to cover topics that include family, health, social identity, and every five to seven years time use. The first Canadian time use survey done as part of the GSS program was conducted in 1986, and the most recent was completed in 2015. These 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 Public Use Microdata are released by Statistics Canada in two files: a *Main File* and an *Episode File*. The files are linked by keys that identify households, individuals, and episodes (i.e., activities) conducted by individuals. We discuss these files in more detail in the following section.

The Main File

The main file of the time use survey compiles a large array of aggregated data, summarizing the answers to the questionnaire that describe households and individuals, as well as derived variables that summarize the respondents' use of 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, time spent with household members, and counts of activities and episodes.

The table 2 shows the first ten lines and the first 6 variables of the GSS PUMF 2015 main file (Cycle 29). Each line of the table refers to a respondent in a survey and the columns refer respectively to the record identification (PUMFID), the person's weight (WGHT_PER), the month the survey data was collected (SURVMNTH), the respondent's age group (AGEGR10), the respondent's sex (SEX), and the respondent's marital status (MARSTAT). In total, the main file of the 2015 GSS surveys has 17,390 respondents corresponding to 2.9766399×10^7 people and 848 variables. For the discrete variables, Statistic Canada has created codes for the possible values and each value has its own label. For example, for the variable SURVMNTH, the value 1 means January 2016, while 2 means February 2016, 3 is March 2016 and so on. As you can see from the example in Table 2, the variables are not labeled. In addition, the format of the tables (comma separated) does not allow you to specify the type of each variable (whether it is continuous or discrete, for example), which can cause confusion for an analyst with relatively little experience in handling PUMFs.

Table 1. Visualization of the first ten lines and first six columns of the main file of the 2015 GSS.

PUMFID	WGHT_PER	SURVMNTH	AGEGR10	SEX	MARSTAT
10000	616.6740	7	5	1	5
10001	8516.6140	7	5	1	1
10002	371.7520	1	4	2	1
10003	1019.3135	3	6	2	5
10004	1916.0708	9	2	1	6
10005	1952.2015	4	1	1	6
10006	5761.5528	8	1	1	6
10007	466.0426	6	5	2	3
10008	2479.2991	2	2	2	1
10009	1436.1641	8	6	1	3

Note:
Legend: ‘PUMFID’: record identification. ‘WGHT_PER’: person weight. ‘SURVMNTH’: survey month of data collection. ‘AGEGR10’: age group of the respondent. ‘SEX’: sex of the respondent. ‘MARSTAT’: marital status of the respondent.

The Episode File

The episode is a much bigger file that records detailed data for each activity episode reported by respondents. Each episode represents a single activity and its duration, and the sum of all episodes throughout the day adds up to 24 hours. Each entry in this file includes the start and end times of the activity, the duration, location, and accompanying social context, informing when and where activities occurred and with whom. The focus of the Episode File is not on the characteristics of the respondents, but on the characteristics of the activities, and the data are 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 a key present in both the Main and Episode Files.

Similarly to Table 2 that displayed an example of the main file structure, Table ?? shows the first seven episodes for the record identification number 10041 and some variables of the GSS PUMF 2015 episode file (Cycle 29). Each line of the table refers to an episode of the especified record identification (PUMFID = 10041), the episode’s weight (WGHT_EPI), the episode number (EPINO), the activity code (TUI_01), the episode’s duration (DURATION) and the episode’s location (LOCATION). In total, the episode file of the 2015 GSS surveys has 274,108 records corresponding to 4.6183762×10^8 episodes and 527 variables that detail the episodes. As similar to the main file, Statistic Canada has created codes for the possible values of the discrete variables with each value has its own label. Considering the case displayed in Table ??, this person started the diary description while sleeping at home (TUI_01 = 1 and LOCATION = 300) for 210 minutes, then engaged in personal hygiene for 40 minutes (TUI_01 = 2), then spent 15 minutes getting personal care (which can be getting ready for school, supervising homework, reading, playing, reprimanding, educational, or emotional help, sinalized by the TUI_01 = 27). After this, this person had an activity travel episode,

walking for 15 minutes ($TUI_01 = 7$ and $LOCATION = 315$) in which the origin and destination was their own house. Cases like this, where the trip starts and finishes at home refers to recreational and leisure trips. After that, the person spent 3 hours looking for a job ($TUI_01 = 9$), had break for a lunch of 15 minutes duration ($TUI_01 = 6$) and cleaned the house ($TUI_01 = 18$) for two hours. Table ?? shows only six variables from the 527 possible cases. As you can see, since that data set is not labeled, figuring out the codes for every variable can be time consuming and difficult.

Table 2. Visualization of the seven first episodes of the record number ‘10041’.

PUMFID	EPINO	WGHT_EPI	TUL01	DURATION	LOCATION
10041	1	1353.818	1	210	300
10041	2	1353.818	2	40	300
10041	3	1353.818	27	15	300
10041	4	1353.818	7	15	315
10041	5	1353.818	9	180	300
10041	6	1353.818	6	15	300
10041	7	1353.818	18	120	300

Note:

Legend: ‘PUMFID’: record identification. ‘EPINO’: episode number. WGHT_EPI: episode’s weight. TUL01: activity code. DURATION: episode’s duration. LOCATION: episode’s location.

Data extraction

For each selected cycle of the GSS surveys, we reviewed the episode files to identify episodes of movement that involved walking or cycling. This allowed us to select the activities immediately before and after the movement episode. After that, we labeled the code variables with their appropriate descriptions, identifying each origin and destination, mode of travel, as well the province and urban classification of the episode.

How to use {ActiveCA}

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 3 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 3. 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 3 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 4 and 5 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 4, 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 5 indicates that the median duration for trips to home and work or school remained at 10 minutes.

Table 4. Comparison of travel statistics by mode and destination: 1986, 1992, 1998

Destination	Mode*	1986				1992				1998			
		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 5. Comparison of travel statistics by mode and destination: 2005, 2010, 2015

Destination	Mode	2005				2010				2015			
		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 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. For conciseness, 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.

Python integration

CAN WE PROVIDE PYTHON INTEGRATION? See here:

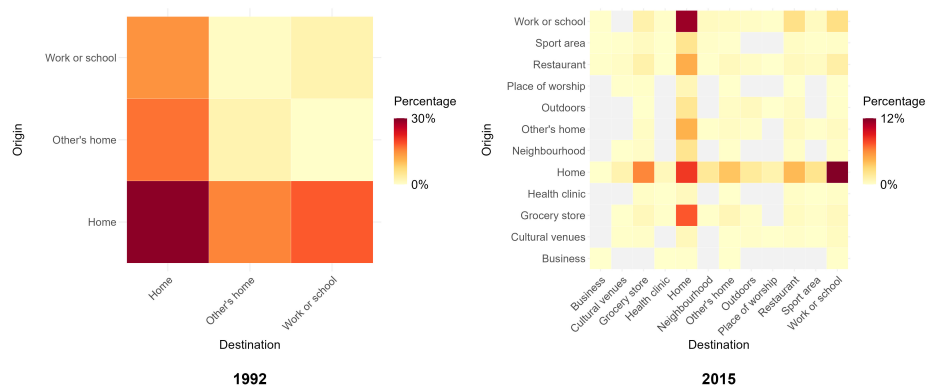


Figure 1. Percentage of walking trips categorized by origin and destination

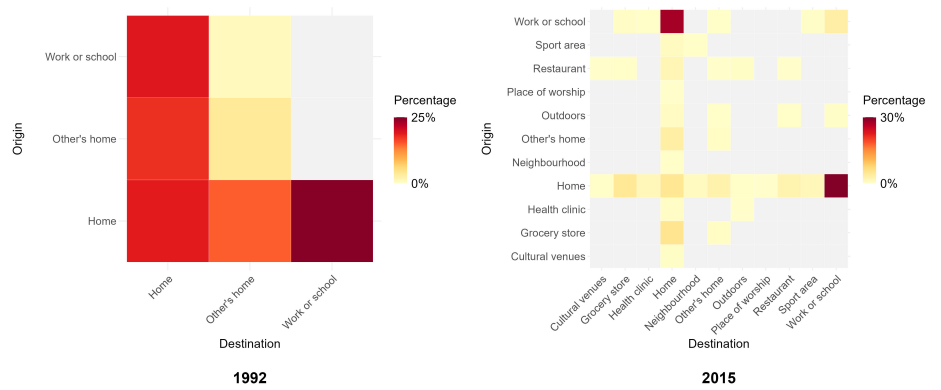


Figure 2. Percentage of cycling trips categorized by origin and destination

<https://github.com/paezha/idealista18/tree/master/Python>

Concluding remarks

This paper 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.

It is important to remark that the present version of {ActiveCA} covers all Canadian time use surveys up to 2015. While the most recent time use survey was carried out in 2022 (Wray, 2024), the Public Use Microdata Files are currently unavailable, and at the moment it is estimated that they will only be published in the later part of 2025. The R package will be updated once these files become available.

The value of {ActiveCA} lies in its transparency, accessibility, and ease of use, which facilitates the addition of complementary data sets in the future. R users can seamlessly explore GSS walking and cycling episodes, 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 {ActiveCA} 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>.

- Brunsdon C and Comber A (2021) Opening practice: supporting reproducibility and critical spatial data science. *Journal of Geographical Systems* 23(4): 477–496. DOI:10.1007/s10109-020-00334-2.
- 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>.
- Kim SO, Palm M, Han S and Klein NJ (2024) Facing a time crunch: Time poverty and travel behaviour in Canada. *Transportation Research Part D: Transport and Environment* 126: 104028. DOI:10.1016/j.trd.2023.104028. URL <https://www.sciencedirect.com/science/article/pii/S136192092300425X>.
- Lachapelle U and Pinto DG (2016) Longer or more frequent walks: Examining the relationship between transit use and active transportation in Canada. *Journal of Transport & Health* 3(2): 173–180. DOI:10.1016/j.jth.2016.02.005. URL <https://www.sciencedirect.com/science/article/pii/S2214140516000153>.
- McCurdy A, Faulkner G, Cameron C, Costas-Bradstreet C and Spence JC (2023) Support for Active Transport Policy Initiatives Among Canadian Adults: The Canadian National Active Transportation Survey. *Active Travel Studies* 3(2). DOI:10.16997/ats.1450. URL <https://activetravelstudies.org/article/id/1450/>. Number: 2 Publisher: University of Westminster Press.
- Páez A (2021) Open spatial sciences: an introduction. *Journal of Geographical Systems* 23(4): 467–476. DOI:10.1007/s10109-021-00364-4. URL <https://doi.org/10.1007/s10109-021-00364-4>.
- 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.
- Spinney JEL, Scott DM and Newbold KB (2009) Transport mobility benefits and quality of life: A time-use perspective of elderly Canadians. *Transport Policy* 16(1): 1–11. DOI: 10.1016/j.tranpol.2009.01.002.
- 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>.