Universität Zürich^UZH

Bachelor Thesis at the Department of Informatics

Lifestyle-based Analysis of Travel Behavior

Analyzing the applicability of Claritas PRIZM Premier for classifying travel behavior through sequence and state analysis on daily activity schedules

Professor: Prof. Dr. Lorenz M. Hilty

Supervisor: Jan C. T. Bieser

Author: David Wyss (14-734-453)

Deadline: 02. June 2019

Abstract. Many factors influence the decisions that lead to an individual's daily schedule; the resulting lifestyle greatly influences the individual's environmental impact. This schedule refers to the individual's daily time-use pattern, i.e. the sequence of activities the individual performs on a given day. As travel is an activity with high environmental impacts, categorizing individuals with respect to their travel behavior is one of the key requirements in predicting and affecting individual behavior on a larger scale. We adapt and apply an existing lifestyle classification, Claritas PRIZM Premier, to sequential time use data from the Multinational Time-Use Study (MTUS). Using R with the sequence analysis tool TraMineR, we analyze categorized daily activity schedules to gain insights on whether the variables that went into creating said lifestyle classification affect the typical travel patterns found within the dataset. Additionally, travel activities are categorized into recreational, work-related and maintenance travel, meaning that different impacts of lifestyle could be measured for each travel type. Results show significant correlations between lifestyle-determining factors and work travel in both frequency and distribution, while recreational travel and maintenance travel show little to no dependency on lifestyle categories. While the applied lifestyle categorization significantly influences travel behavior, it is not sufficient for accurately predicting individual travel schedules.

Table of Contents

List of Figures

1 Introduction

Today, travelling on a daily basis has become a necessity for many people. However, this is not restricted to the traditional definition of travel; in this context, travel can be described as any activity that leads to time individuals spend in transportation between different places. Susilo & Dijst (2010) describe the need for travel as follows: "It is well recognised that travel is a derived demand that is based on the needs and desires of individuals and households. In order to take part in activities, individuals often have to travel between places. Whilst there are some routine activities that generate routine travel patterns to routine locations, such as commuting, there are also activities that are not necessarily everyday and do not necessarily occur in the same location, such as (non-daily) shopping and leisure trips.". This claim is also supported by Gangrade, Pendyala & McCullough (2002), who also describe the demand for travel as a derived demand instead of a travel being a demand in itself. Travel behavior carries a significant environmental impact. Nilsson & Küller (2000) state that "Despite the fact that technical improvements […] have reduced the pollution per vehicle, the environmental gains have been lost through the more extensive use of private cars.".

Therefore, analyzing the type of activities that lead to time spent travelling between places might be an important step in preventing emissions caused by avoidable travel. Bieser & Hilty (2018) state that "Analyzing lifestyles form a time-use perspective can provide a more comprehensive understanding about the indirect environmental impact of ICT […] because individual lifestyles (how do people spend their time) are a major determinant of environmental impacts,"

This thesis draws on data gathered in the Multinational Time Use Study (MTUS) conducted by the Department of Sociology at the University of Oxford, as well as the lifestyle classification Claritas PRIZM Premier to draw conclusions on the relationship between various lifestyle factors and typical travel activity sequences.

2 Research Goals, Hypotheses and Approach

2.1 Research Goals

The goal of this thesis is to provide an insight into whether existing lifestyle classifications can be used to adequately group people's travel behavior into distinct categories. Using an adapted version of Claritas PRIZM Premier, we aim to analyze the correlations between the factors influencing typical lifestyles and their respective travel behaviors. Froemelt, Dürrenmatt & Hellweg (2018) state that in order to devise accurately targeted environmental policies, an understanding of the differences between consumption patterns among the population is needed. Thus, an accurate categorization of various lifestyles is a key ingredient in accurately understanding and predicting travel behavior among large populations.

2.2 Hypotheses

The core assumption behind this thesis is that on a larger scale, people's lifestyles can be classified based on a combination of attributes such that other people sharing similar attributes will exhibit similar travel behavior patterns. In order to further define the topic at hand, we establish the following hypotheses:

1) Marketing-based classifications (such as a simplified version of Claritas PRIZM Premier) are sufficiently accurate for classifying travel behavior and typical travel behavior patterns can be found in available MTUS data.
2) There exists a correlation between the factors determining lifestyle classification (such as age, urbanicity and wealth) and a person's travel behavior. That is, the factors that go into classifying a person's lifestyle could also be suitable for differentiating travel behavior among people.

In the following chapters, we analyze and prepare our dataset with the goal of verifying or rejecting these two hypotheses.

2.3 Approach

In order to be able to verify the hypotheses outlined in chapter 2.2, we take the approach of first applying an existing lifestyle classification, Claritas PRIZM Premier, to a specific part of the MTUS dataset. In doing so, we assign a lifestyle category to each individual diarist within the dataset. We then prepare the dataset for sequence analysis with regard to various variables within the dataset that we want to consider. Afterwards, using sequence analysis tools, we identify common daily activity schedules among diarists of different lifestyle categories, which we then use to draw conclusions about the differences between the typical lifestyles found for each lifestyle category. We also quantitatively analyze various types of travel activities and their correlations to different lifestyle-determining factors.

Further details on the exact statistical methodology used towards these goals are found in the following chapters; specifically, chapter 7.1 focuses on the sequence analysis methodology that was used in this thesis.

3 Existing Classifications and Adaptation

3.1 Introduction

Before delving into the data analysis that was undertaken to verify the hypotheses made in chapter 2.2, this chapter shall serve as a brief overview on preexisting classifications of lifestyles. Lin, Lo & Chen (2009) define a lifestyle as follows: "A lifestyle represents individual orientation or preference about his/her daily decisions regarding activity and travel, which are in general related to socio-economic and demographic variables such as household structure, work participation, and housing type, etc." Creating a classification that fully encompasses all the variables that shape a lifestyle is well beyond the scope of this thesis, which is why this chapter is intended to provide an overview of existing lifestyle classifications and what went into creating them. Contrary to Lin et al. (2009), who derived lifestyle clusters from census data as well as activity-travel patterns, our approach relies on modifying an already existing classification for the purpose of analyzing travel behavior. As preexisting classifications are created with the factors described beforehand in mind, they fit into Lin et al. (2009)'s definition of lifestyles and are thus suitable for further analysis. With this goal in mind, we will be making adaptations and simplifications to the chosen classification, PRIZM Premier by Claritas Inc., to make it more suitable for statistical analysis of our dataset.

3.2 Claritas PRIZM Premier

Originally developed by Claritas Inc. for the US market, Claritas PRIZM Premier is a marketing-driven approach in dividing the population into segments with the goal of helping marketers define and reach their targeted customers more easily. In order to achieve this, PRIZM Premier divides the population into 68 distinct segments. This segmentation is based on factors such as urbanization, socioeconomic rank, age, presence of children at home, occupation, income, education and home value. This is done by dividing the population into 14 Social groups, which are based on urbanization and socioeconomic rank, as well as 11 Lifestage groups, which are based on socioeconomic rank, age, and the presence of children at home (Claritas Inc., 2016).

While PRIZM Premier was developed for marketing purposes, the classification is very in-depth regarding many factors that might influence travel behavior. However, for the purpose of this thesis, the division into 68 distinct segments is too fine; further adaptation of the PRIZM Premier classification (as described in chapter 3.4) is needed.

Claritas PRIZM Premier is predated by Claritas PRIZM, which dates back to 1980, while PRIZM Premier represents an adapted, modernized and expanded revision of the original PRIZM classification.

3.3 Other Classifications

While there are other classifications for lifestyle segmentation, PRIZM Premier is very in-depth, considers a wide range of factors that determine lifestyle and should be quite relevant for more recent data. Thus, we focus on adapting PRIZM Premier to suits our analytical purposes rather than focusing on whether the classification as such is directly suitable in this context.

3.4 Adaptation of PRIZM Premier

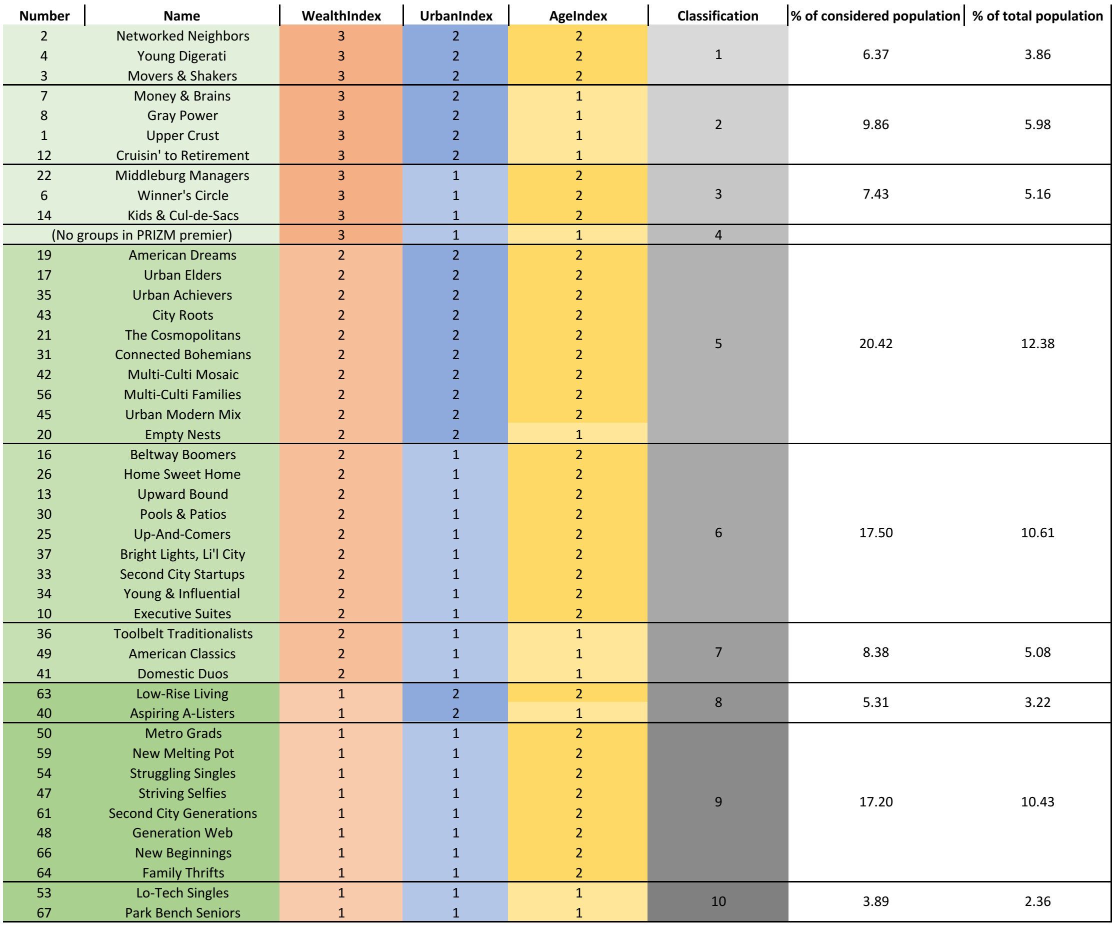
For the purpose of analyzing travel behavior among different population groups, adapting the very fine-grained classification that is PRIZM Premier is needed; this is to avoid having subgroups that only make up a very small percentage of the dataset and would thus be unsuitable for analysis. We base the process of adapting PRIZM Premier to our use case on the following assumptions and prerequisites:

- We assume non-comparability between heavily urbanized European countries and the US-based population that PRIZM Premier was developed for. To alleviate these purported differences, we will not be taking rural population categories into account.
- Based on the original reasoning behind the segmentation in PRIZM Premier, we assume the three primary attributes determining travel behavior to be income, urbanicity and age. These parameters also appear in the MTUS dataset, making them suitable for analysis.
- Sub-groups shall not make up less than 3% of the total considered population; thus, some categories may be grouped together based on more loose parameters if they only account for a small percentage of the population.

The following steps were taken for grouping the originally 68 segments specified by PRIZM Premier into 10 distinct categories for further statistical analysis; any statistical data about PRIZM Premier segments used herein was derived from https://support.geopath.io/hc/en-us/sections/360000917931-Market-Segmentation.

1) From the original set of 68 segments, we remove any groups where urbanicity is specified as "Rural" or "Town", since we (as described beforehand) only consider diarists living in urban or semi-urban areas. This step then leaves us with 44 segments.
2) From there, we group the remaining segments into three categories based on average household income. This categorization is made by dividing the population into the top 25%, the middle 50% and the bottom 25%. We assign a Wealth Index (from one to three, with three representing the highest income category) to each of the 44 segments, depending into which of the income groups they fall.
3) We then assign an Urbanicity Index to each segment. This Index has a value of either 1 or 2, with 1 signifying medium urbanicity and 2 signifying high urbanicity. The categories "Urban" and "Suburban" are assigned an Urbanicity Index of two, whereas "Metro Mix" and "Second City" are assigned an Index of one.
4) Grouping the 44 segments by both Urbanicity Index (*urbanIndex*, either 1 or 2) and Wealth Index (*wealthIndex*, a value from 1 to 3) leaves us with six categories, one for each possible combination. However, as we initially assumed the parameter "Age" to be a contributing factor as well, further sub-categories are needed.
5) We therefore assign an Age Index (*ageIndex*, either 1 or 2) to each of the 44 segments. An Age Index of one denotes the average age of the group being less than 55 years, whereas an Age Index of two corresponds to an average age of more than 55 years. The reasoning behind choosing 55 years as the cutoff for our age index is that by this age, many diarists' children will already be adults; thus, it represents an important step in the lifestages as described by Claritas.
6) Based on the three indices we set beforehand, this would leave us with 12 distinct groups. However, as established beforehand, sub-groups shall not make up less than 3% of the total considered population. As this is not the case for age-divided subgroups with the combinations (*wealthIndex*=2, *urbanIndex*=2) and (*wealthIndex*=1, *urbanIndex*=2), age division is omitted for these two groups. This leaves us with a total of 10 subgroups which will then be used for further classification.

Following these steps, we have grouped the original 68 population segments into 10 distinct categories which still take the three primary parameters wealth, urbanicity and age into account. For the purpose of identifying statistically significant patterns within our dataset, this classification should prove much more meaningful, as it does not contain any groups that account for less than 3% of the population, which we consider to be the minimum within our dataset.



Number	Name	WealthIndex	UrbanIndex	AgeIndex	Classification	% of considered population	% of total population
2	Networked Neighbors	3	2	2	1	6.37	3.86
4	Young Digerati	3	2	2			
3	Movers & Shakers	3	2	2			
7	Money & Brains	3	2	1	2	9.86	5.98
8	Gray Power	3	2	1			
1	Upper Crust	3	2	1			
12	Cruisin' to Retirement	3	2	1			
22	Middleburg Managers	3	1	2	3	7.43	5.16
6	Winner's Circle	3	1	2			
14	Kids & Cul-de-Sacs	3	1	2			
(No groups in PRIZM premier)		3	1	1	4		
19	American Dreams	2	2	2	5	20.42	12.38
17	Urban Elders	2	2	2			
35	Urban Achievers	2	2	2			
43	City Roots	2	2	2			
21	The Cosmopolitans	2	2	2			
31	Connected Bohemians	2	2	2			
42	Multi-Culti Mosaic	2	2	2			
56	Multi-Culti Families	2	2	2			
45	Urban Modern Mix	2	2	2			
20	Empty Nests	2	2	1			
16	Beltway Boomers	2	1	2	6	17.50	10.61
26	Home Sweet Home	2	1	2			
13	Upward Bound	2	1	2			
30	Pools & Patios	2	1	2			
25	Up-And-Comers	2	1	2			
37	Bright Lights, Li'l City	2	1	2			
33	Second City Startups	2	1	2			
34	Young & Influential	2	1	2			
10	Executive Suites	2	1	2			
36	Toolbelt Traditionalists	2	1	1	7	8.38	5.08
49	American Classics	2	1	1			
41	Domestic Duos	2	1	1			
63	Low-Rise Living	1	2	2	8	5.31	3.22
40	Aspiring A-Listers	1	2	1			
50	Metro Grads	1	1	2	9	17.20	10.43
59	New Melting Pot	1	1	2			
54	Struggling Singles	1	1	2			
47	Striving Selfies	1	1	2			
61	Second City Generations	1	1	2			
48	Generation Web	1	1	2			
66	New Beginnings	1	1	2			
64	Family Thrifts	1	1	2			
53	Lo-Tech Singles	1	1	1	10	3.89	2.36
67	Park Bench Seniors	1	1	1			

Figure 1: Classification of PRIZM Premier groups

The figure above shows that PRIZM Premier does not contain any subgroups corresponding to the combination *wealthIndex*=3, *urbanIndex*=1, *ageIndex*=1. As this combination appears in our data, we assign a classification type (number 4) to this combination.

4 Tool Requirements and Selection

4.1 Tool Requirements

In order to fulfill the requirements that stem from the hypotheses made in chapter 2.2, a suitable statistical software with sufficient capabilities regarding sequence analysis is needed to analyze the MTUS dataset. This software must at a minimum offer the following capabilities:

1) Restructuring and reorganizing the dataset.
2) Merging both MTUS datasets in order to obtain additional information about individual diarists.
3) Appending the three indices *wealthIndex*, *urbanIndex* and *ageIndex* conditionally based on different variables as well as determining the lifestyle category for each individual diarist.
4) Categorizing activities into various activity types.
5) Conducting sequence analysis with the goal of finding common patterns both for the entire dataset and individual lifestyle categories.
6) Conducting state analysis on activity sequences within the MTUS dataset.
7) Quantitatively analyzing travel time for each individual category.

In addition to the criteria outlined above, a suitable tool shall offer good ease of use, expandability as well as documentation.

4.2 Tool Selection

While there exists a wide array of software for statistical analysis, we chose to use the free statistical software *R*, as it provides all required functionalities, a large online community as well as expandability via packages. *R* finds widespread usage in many fields, such as academia and industry (Zhao, 2015). Additionally, there exist various extension packages for sequence analysis in *R*; we chose to use the package *TraMineR* developed by the Institute of Demography and Socioeconomics (IDESO) at the University of Geneva. *TraMineR* offers a wide variety of tools for analyzing sequences within datasets, however this will require some adaptation of the MTUS dataset, which is described in more detail in the following chapter. More details on how *TraMineR* can be used for statistical analysis as well as its concrete usage to our topic are found in chapter 7.1.

5 Data Understanding

5.1 Data Description

As described in chapter 1, this thesis is based on data gathered in the Multinational Time Use Study (MTUS) conducted by the Department of Sociology at the University of Oxford. The MTUS provides a variety of datasets about activities undertaken during an individual's daily routine. The Harmonised Aggregate Files (HAF) provide a large variety of variables about the diarist, but lack chronological order for activities undertaken by the diarist. Therefore, we focus on the Harmonised Episode Files (HEF). These files contain entries for distinct activity sequences of individual diarists, as well as some other variables including sex and age. The activities within these diaries are grouped into 69 activity categories, which will be analyzed more in-depth in the following chapters. Since the Harmonised Episode Files do not contain more variables about the diarist (such as wealth, urbanicity etc.), joining both HAF and HEF datasets together is needed to establish a more complete picture of the diarist in order be able to properly analyze the dependencies between travel times and other factors. This is possible due to the fact that a unique ID can be assigned to each diarist based on a combination of variables, meaning distinct persons are identifiable across both HEF and HAF datasets. An in-depth user guide to the MTUS dataset, upon which the following processes were based, is available at: https://www.timeuse.org/MTUS-User-Guide.
In terms of individual countries that are featured in the HEF dataset, there are large differences in the amounts of available data. The following figure outlines the numbers of individual diarists that completed at least one diary available in the HEF dataset for any year after 1990.

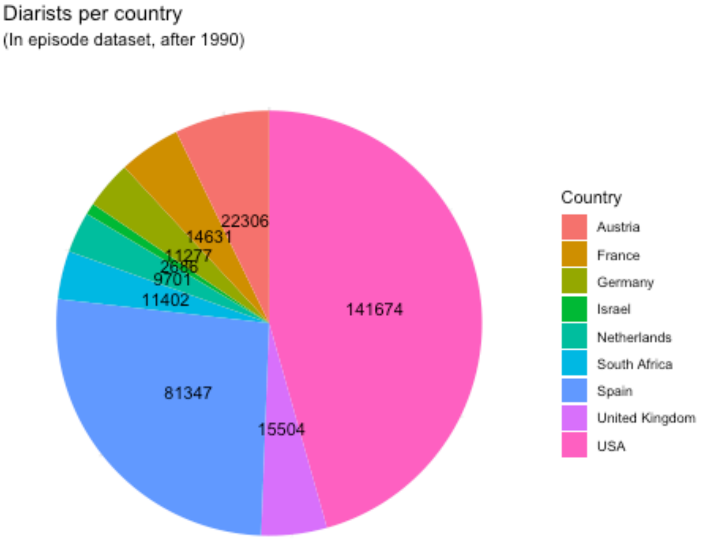


Figure 2: Diarists per country in the HEF dataset

Additionally, we analyzed the number of diarists that completed at least one diary available in the HEF dataset on a per-year-basis. The following figure shows comparably small amounts of available data for the years preceding 1990, while the following years contain high quantities of data within the HEF dataset.

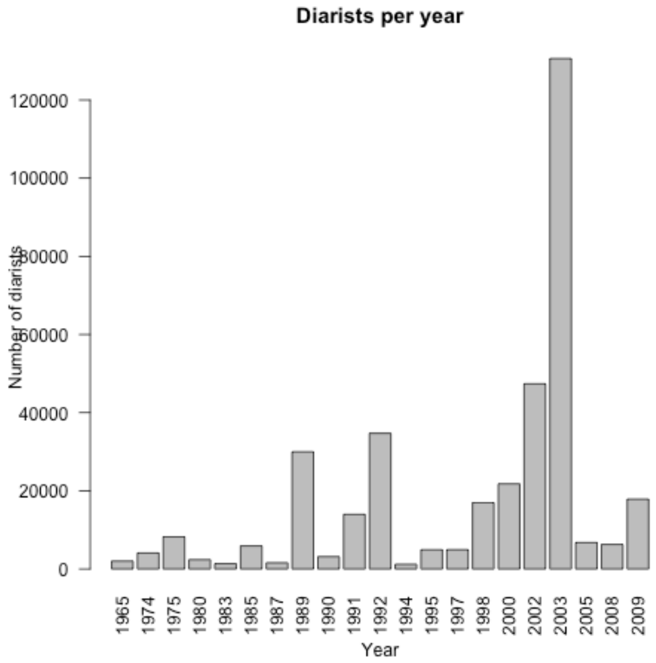


Figure 3: Diarists per year in the HEF dataset

5.2 Data Exploration

The MTUS dataset contains data from 23 different countries, but only for 11 of them, HEF data is available. As described beforehand, our focus lies on west European countries with relatively high degrees of urbanization. Since the HEF dataset does not contain sufficient amounts of data for all countries, we chose to analyze only data from the Netherlands (NL) and the United Kingdom (UK). For both of these countries, sufficient amounts of diaries are available in the HEF dataset. The HEF dataset contains a total of 15'504 diarists (UK) and 6'544 diarists (NL) respectively for a total of 22'048 unique diarists. Note that there may be multiple diaries for each diarist, with each one corresponding to the diarist's activities during a single day.

Behrens & Mistro (2010) state that travel behavior changes significantly over time as a result of policy and system changes. Due to this influence an ever-changing environment has on individual travel behavior, we focus on data gathered from studies that were conducted after 1990. As seen in figure 3, there is also a significant amount of data available for the year 1989. However, as the only study conducted in 1989 that resulted in HEF files took place in Italy, we do not consider this data to be relevant for our research.

5.3 Data Quality

As the data available within the MTUS dataset has already been preprocessed as well as harmonized, we find relatively usable data for our analysis purposes. As we do not consider many of the columns available within the dataset (instead focusing on full diaries), we have no need for many of the variables that exhibit high percentages of missing values. Thus, for our purposes, we considered only a small percentage of the variables, meaning omission of any missing values will not lead to considerable quantities of lost data. Therefore, we later opt to remove any entries presenting missing values for the variables we use in our analysis.

6 Data Preparation

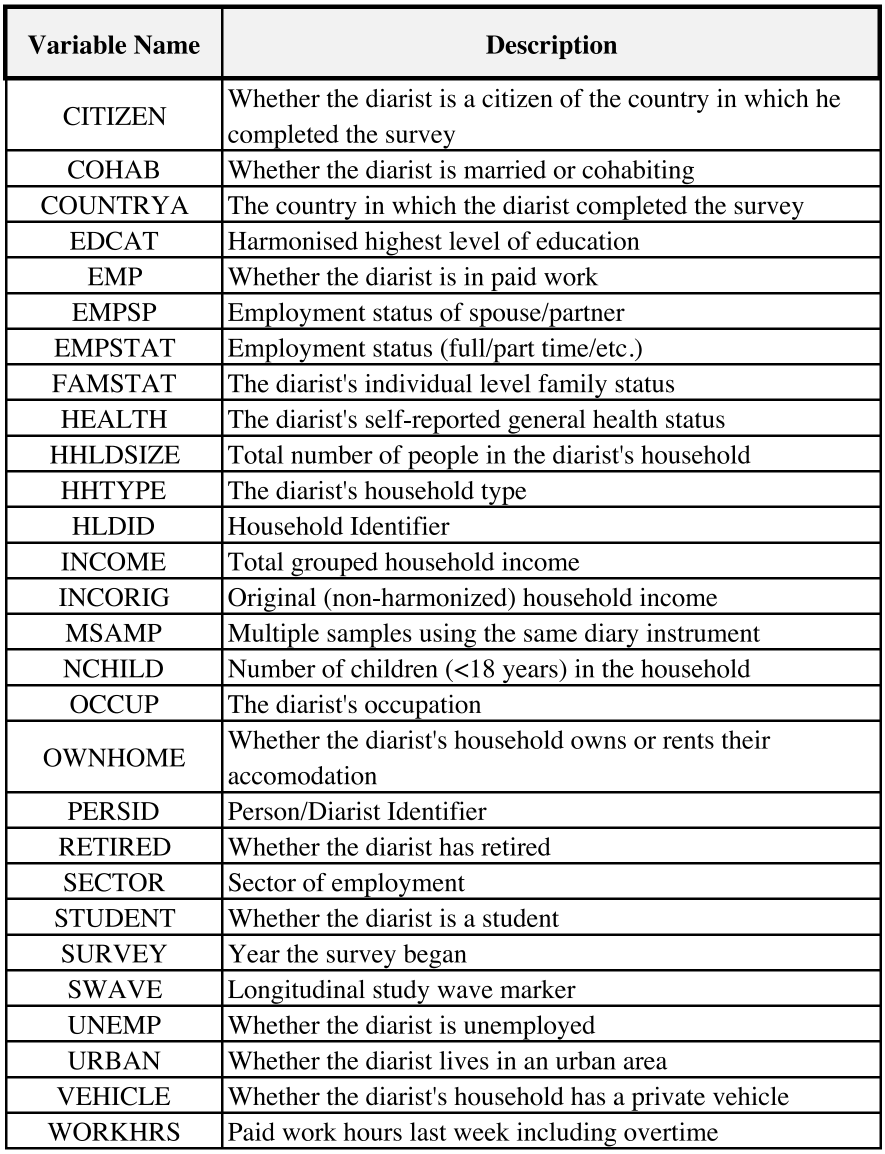
6.1 Data Selection and Cleaning

In the process of Data Selection and Cleaning, we exclude all parts of the data that are deemed unfit for analysis. Thus, in our case, we omit all entries with missing values for the variable "*urban*", as it represents one of the variables needed for lifestyle categorization based on the adapter version of PRIZM Premier, as described in chapter 3.4. Additionally, we remove any rows with missing values for the Variable "*age*", as well as any entries where the diarist's age was below 18 years. Lastly, any rows with missing entries for "*main*", the primary activity indicator, were removed. The fact that we do not rely on most of the variables contained within the HEF dataset allows us to maintain most of the diaries that fit our criteria.

6.2 Data Construction

The process of Data Construction is described by Brown (2014) as a process in which new columns and/or rows are created and aggregated within a dataset. This represents perhaps the biggest part in preparing the MTUS datasets for sequence analysis. As described in chapter 5.1, the HEF dataset does not contain many variables about the diarists themselves; this necessitates the gathering of additional information about individual diarists by merging the HEF and HAF datasets together with respect to the relevant variables. Note that while we won't be using most of these variables for further analysis, they might prove suitable for further research into this topic.

The following table shows the variables that we chose to import from the HAF dataset (in alphabetical order) as well as their descriptions. Note that while we chose to import a relatively large number of variables, not all of them are used for lifestyle classification.



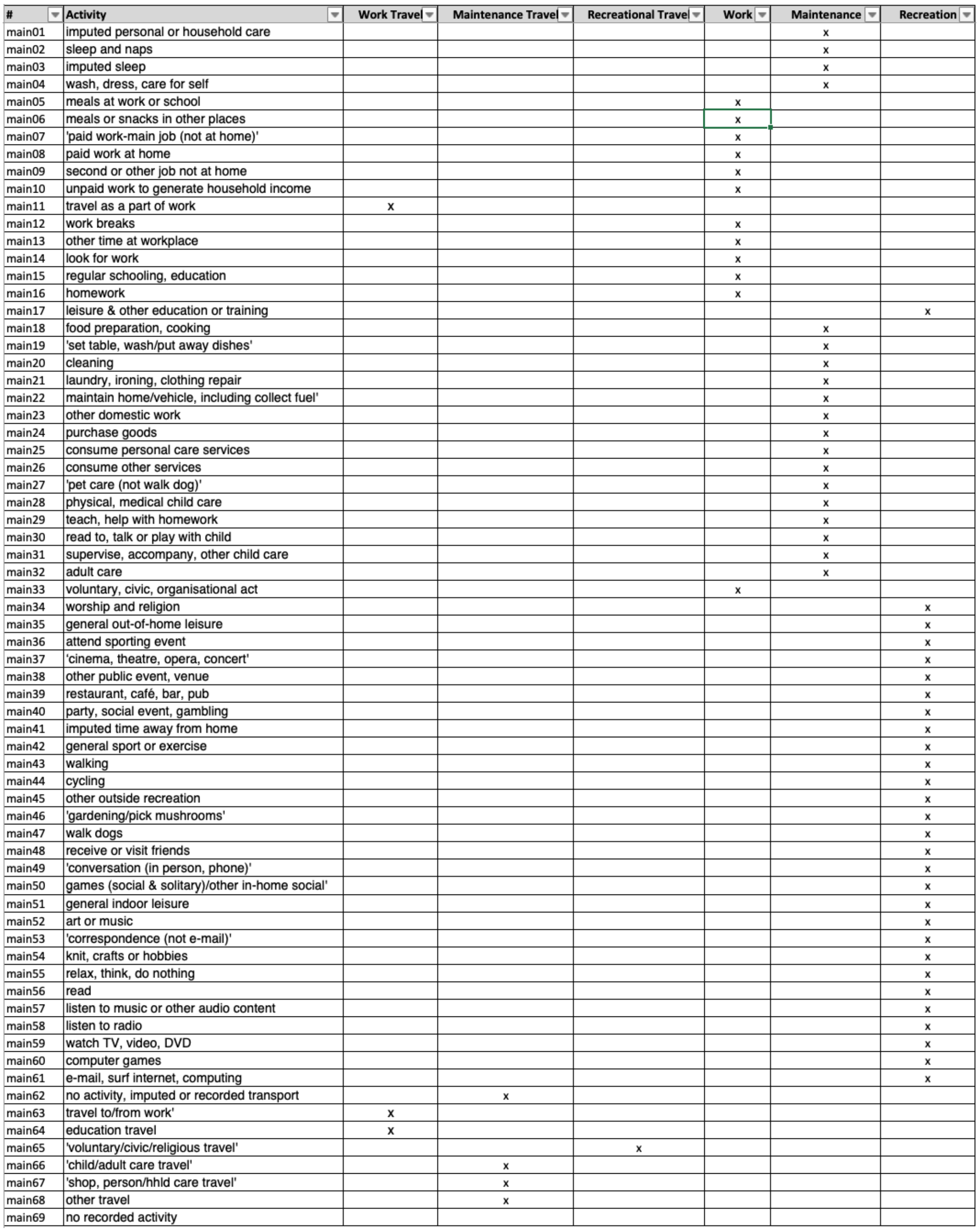
Variable Name	Description
CITIZEN	Whether the diarist is a citizen of the country in which he completed the survey
COHAB	Whether the diarist is married or cohabiting
COUNTRYA	The country in which the diarist completed the survey
EDCAT	Harmonised highest level of education
EMP	Whether the diarist is in paid work
EMPSP	Employment status of spouse/partner
EMPSTAT	Employment status (full/part time/etc.)
FAMSTAT	The diarist's individual level family status
HEALTH	The diarist's self-reported general health status
HHLDSIZE	Total number of people in the diarist's household
HHTYPE	The diarist's household type
HLDID	Household Identifier
INCOME	Total grouped household income
INCORIG	Original (non-harmonized) household income
MSAMP	Multiple samples using the same diary instrument
NCHILD	Number of children (<18 years) in the household
OCCUP	The diarist's occupation
OWNHOME	Whether the diarist's household owns or rents their accomodation
PERSID	Person/Diarist Identifier
RETIRED	Whether the diarist has retired
SECTOR	Sector of employment
STUDENT	Whether the diarist is a student
SURVEY	Year the survey began
SWAVE	Longitudinal study wave marker
UNEMP	Whether the diarist is unemployed
URBAN	Whether the diarist lives in an urban area
VEHICLE	Whether the diarist's household has a private vehicle
WORKHRS	Paid work hours last week including overtime

Figure 4: HAF variable selection

For the next part of data construction, our goal is to uniquely identify each diarist to correctly assign the selected HAF variable values to the corresponding rows in the HEF dataset. In order to fulfill this, we assign a unique ID value (labeled "*uniqueId*") to each diarist, determined by the following variables:
persid, hldid, countrya, survey, swave, msamp
The resulting unique ID has the following format:
1_0_22_1995_0_0
This example represents a diarist with person ID of 1, household ID of 0, in the Netherlands (country code of 22) that participated in a study conducted in the year 1995.
This process is applied to both the HEF and HAF datasets; this unique ID allows us to individually identify any diarist that participated in any MTUS study in both datasets. Additionally, we remove all variables that were used to generate the unique ID (except for "*countrya*", as this variable will still be used) from the HAF and HEF datasets. After generating unique ID values for both the HEF and HAF datasets, we merge them based on the value of "*uniqueId*". In doing so, we have expanded the episode dataset to contain additional information about each diarist. This then allows us to categorize all diarists in the HEF dataset based on the criteria outlined in chapter 3.4. This is done by assigning the three indices for age, urbanicity and wealth to each diarist in the dataset (with the variables called "*ageIndex*", "*urbanIndex*" and "*wealthIndex*"). Based on these three indices, we then determine which of the 10 categories in our adapted version of PRIZM Premier the diarist belongs to. This results in a new column called "*classification*", whose values range from 1 to 10.
Next, we categorize all 69 activity types (as used in the "*main*" column in the HEF dataset) into six distinct categories. This is done in order to facilitate sequence analysis as well as simplify the high number of different activities. As our focus lies on analyzing travel behavior, we introduce a distinction between three different types of travel: Work Travel, Maintenance Travel and Recreational Travel. This division is largely based on the work of Susilo & Dijst (2010), who separate travel activities into Work, Maintenance and Leisure Travel. Additionally, the remaining activities are grouped into Work activities, Maintenance activities and Recreational activities.
We define the six activity categories as follows:

1) Work Travel (*travel_work*)
2) Maintenance Travel (*travel_maint*)
3) Recreational Travel (*travel_recr*)
4) Work (*work*)
5) Maintenance (*maint*)
6) Recreation (*recr*)

Additionally, activity number 69 (undefined / no recorded activity) was assigned its own category, Unknown (*unkn*). This category will also be used to fill any gaps in diaries where no recorded activity information is present, should this case occur. The following table shows how the 69 main activities were assigned to the previously described activity categories.



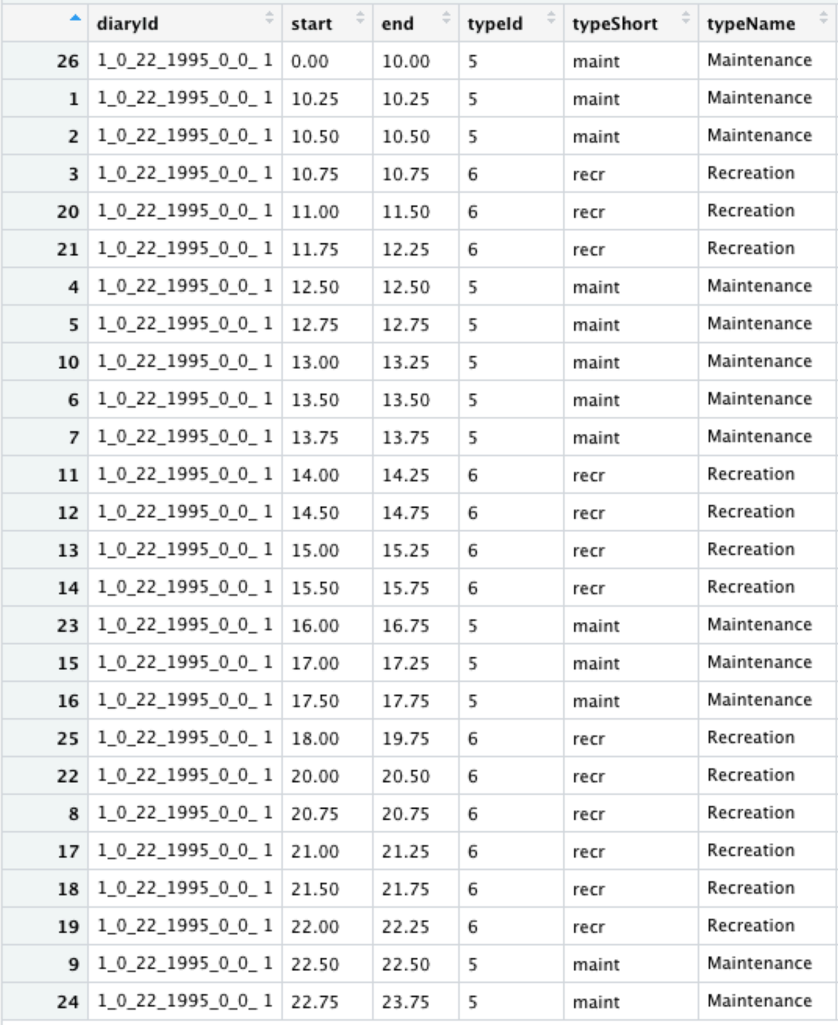
#	Activity	Work Travel	Maintenance Travel	Recreational Travel	Work	Maintenance	Recreation
main01	imputed personal or household care					x	
main02	sleep and naps					x	
main03	imputed sleep					x	
main04	wash, dress, care for self					x	
main05	meals at work or school				x		
main06	meals or snacks in other places				x		
main07	'paid work-main job (not at home)'				x		
main08	paid work at home				x		
main09	second or other job not at home				x		
main10	unpaid work to generate household income				x		
main11	travel as a part of work	x					
main12	work breaks				x		
main13	other time at workplace				x		
main14	look for work				x		
main15	regular schooling, education				x		
main16	homework				x		
main17	leisure & other education or training						x
main18	food preparation, cooking					x	
main19	'set table, wash/put away dishes'					x	
main20	cleaning					x	
main21	laundry, ironing, clothing repair					x	
main22	maintain home/vehicle, including collect fuel'					x	
main23	other domestic work					x	
main24	purchase goods					x	
main25	consume personal care services					x	
main26	consume other services					x	
main27	'pet care (not walk dog)'					x	
main28	physical, medical child care					x	
main29	teach, help with homework					x	
main30	read to, talk or play with child					x	
main31	supervise, accompany, other child care					x	
main32	adult care					x	
main33	voluntary, civic, organisational act				x		
main34	worship and religion						x
main35	general out-of-home leisure						x
main36	attend sporting event						x
main37	'cinema, theatre, opera, concert'						x
main38	other public event, venue						x
main39	restaurant, café, bar, pub						x
main40	party, social event, gambling						x
main41	imputed time away from home						x
main42	general sport or exercise						x
main43	walking						x
main44	cycling						x
main45	other outside recreation						x
main46	'gardening/pick mushrooms'						x
main47	walk dogs						x
main48	receive or visit friends						x
main49	'conversation (in person, phone)'						x
main50	games (social & solitary)/other in-home social'						x
main51	general indoor leisure						x
main52	art or music						x
main53	'correspondence (not e-mail)'						x
main54	knit, crafts or hobbies						x
main55	relax, think, do nothing						x
main56	read						x
main57	listen to music or other audio content						x
main58	listen to radio						x
main59	watch TV, video, DVD						x
main60	computer games						x
main61	e-mail, surf internet, computing						x
main62	no activity, imputed or recorded transport		x				
main63	travel to/from work'	x					
main64	education travel	x					
main65	'voluntary/civic/religious travel'			x			
main66	'child/adult care travel'		x				
main67	'shop, person/hhld care travel'		x				
main68	other travel		x				
main69	no recorded activity						

Figure 5: Activity Categorization

This categorization of the 69 main activities is then applied to the extended HEF dataset, adding the columns "*typeId*" (number, from 1 to 6), "*typeShort*" (short category name) and "*typeName*" (full category name).

In the next part of Data Construction, our goal is to format the extended episode dataset to contain exactly one row per unique diary, which contains a sequence of categorized activities. As it stands, each line in the HEF dataset represents a single activity from a diarist's diary. There is no unique identifier for individual diaries, as the variable "*diary*" represents the individual diarist's diary count and is therefore not unique. Thus, we construct a new column called "*diaryId*" whose values uniquely identify each diary. This is achieved by combining the values of the "*diary*" variable with the diarist's

unique ID, as each diary ID only appears once per diarist. The resulting diary ID has the following format:
1_0_22_1995_0_0_1
This extends the unique ID example described beforehand by appending "*_1*" at the end, signifying the diarist's first diary. Based on this unique diary ID, we then generate a new dataset containing exactly one row per diary, called "*episode_individual_diaries*". Within this newly generated dataset, we create a total of 96 additional columns; each representing a fifteen-minute interval. These columns are named in the format "*time_0*", "*time_0.25*", "*time_0.5*" etc. until "*time_23.75*". Each increment of 0.25 represents an interval of fifteen minutes. All activity end times are rounded down, such that an activity being in one column, for example "time_12.5" would correspond to the activity having a duration of fifteen minutes.
We then prefill these columns with the value "*unkn*" to avoid having null values for time slots where no activity was recorded. All activity start/end times, recorded in the variables "*start*" and "*end*", are then rounded to fifteen minute increments to prevent overlap between activities and to correspond exactly to the columns that were generated beforehand. This prepared dataset for one individual diary might look as follows before being entered into the corresponding columns:



	diaryId	start	end	typeId	typeShort	typeName
26	1_0_22_1995_0_0_ 1	0.00	10.00	5	maint	Maintenance
1	1_0_22_1995_0_0_ 1	10.25	10.25	5	maint	Maintenance
2	1_0_22_1995_0_0_ 1	10.50	10.50	5	maint	Maintenance
3	1_0_22_1995_0_0_ 1	10.75	10.75	6	recr	Recreation
20	1_0_22_1995_0_0_ 1	11.00	11.50	6	recr	Recreation
21	1_0_22_1995_0_0_ 1	11.75	12.25	6	recr	Recreation
4	1_0_22_1995_0_0_ 1	12.50	12.50	5	maint	Maintenance
5	1_0_22_1995_0_0_ 1	12.75	12.75	5	maint	Maintenance
10	1_0_22_1995_0_0_ 1	13.00	13.25	5	maint	Maintenance
6	1_0_22_1995_0_0_ 1	13.50	13.50	5	maint	Maintenance
7	1_0_22_1995_0_0_ 1	13.75	13.75	5	maint	Maintenance
11	1_0_22_1995_0_0_ 1	14.00	14.25	6	recr	Recreation
12	1_0_22_1995_0_0_ 1	14.50	14.75	6	recr	Recreation
13	1_0_22_1995_0_0_ 1	15.00	15.25	6	recr	Recreation
14	1_0_22_1995_0_0_ 1	15.50	15.75	6	recr	Recreation
23	1_0_22_1995_0_0_ 1	16.00	16.75	5	maint	Maintenance
15	1_0_22_1995_0_0_ 1	17.00	17.25	5	maint	Maintenance
16	1_0_22_1995_0_0_ 1	17.50	17.75	5	maint	Maintenance
25	1_0_22_1995_0_0_ 1	18.00	19.75	6	recr	Recreation
22	1_0_22_1995_0_0_ 1	20.00	20.50	6	recr	Recreation
8	1_0_22_1995_0_0_ 1	20.75	20.75	6	recr	Recreation
17	1_0_22_1995_0_0_ 1	21.00	21.25	6	recr	Recreation
18	1_0_22_1995_0_0_ 1	21.50	21.75	6	recr	Recreation
19	1_0_22_1995_0_0_ 1	22.00	22.25	6	recr	Recreation
9	1_0_22_1995_0_0_ 1	22.50	22.50	5	maint	Maintenance
24	1_0_22_1995_0_0_ 1	22.75	23.75	5	maint	Maintenance

Figure 6: Example activity entries for one diary

Next, we iteratively insert the task category short names (as shown above in the column "*typeShort*") into the time columns that were generated beforehand. As this conversion process is relatively computationally expensive, it takes considerable time. Running on a machine with a six core, twelve thread 3.3GHz Intel Xeon processor, the dataset conversion takes around 40 minutes for our dataset with a total of 47’903 diaries. However, this step is necessary to facilitate sequence analysis with *TraMineR*, which is described in the following chapter.

7 Data Analysis

7.1 Methodology

As described in chapter 4.2, we will be using the free statistics software *R* as well as the extension package *TraMineR*. With these tools, we aim to analyze the dataset prepared in chapter 6.3. The goal of this part is to determine whether there are distinguishable differences in typically recognized activity and travel patterns between the previously categorized population groups (referred to as "lifestyle categories"), thus verifying the hypotheses made in chapter 2.2. Additionally, we also compare the amounts of time spent on different kinds of travel by different population groups.
Brzinsky-Fay, Kohler & Luniak (2006) describe a sequence as "an ordered list of elements, where an element can be a certain status (e.g., employment or marital status), a physical object […], or an event […].". In our particular case, the event described herein corresponds to the activity taken in a single episode in our dataset, e.g. "*travel_work*".

7.2 Overview

To provide a general overview of the gathered sequence data, we first analyze the dataset with *TraMineR* without any further filtering based on lifestyle categories. Extracting the ten most common sequences from the dataset results in the following plot:

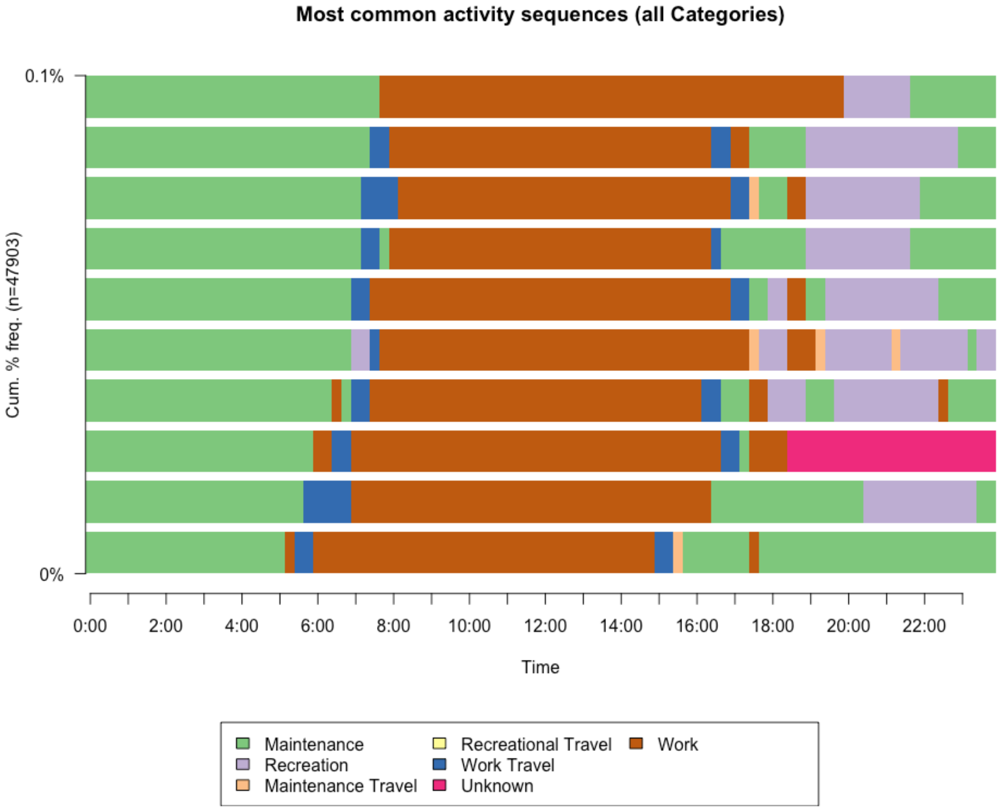


Figure 7: Most common activity sequences (all categories)

As is evident from this plot, the most common sequences within our dataset mostly describe a typical day structure consisting of Maintenance (e.g. sleep or preparation for work), followed by Work Travel, which is then followed by Work for an extended time. After this, in most cases, there is another instance of Work Travel that is succeeded by either Recreation (e.g. time spent on hobbies or with family) or Maintenance (such as shopping, household work etc.). The diary usually concludes with Maintenance activities, which can be explained by the diarist going to sleep.

However, this graph exhibits a problem: As we only analyze whether complete day patterns (down to fifteen minute increments) are found repetitively, the resulting numbers of matching patterns are relatively low; in some cases, a diary might differ from another only by a few episodes, but this would lead to the diaries being non-identical and thus not considered as more common, even though some patterns might be found in both.

One way around this issue is to split the dataset into two equal halves; one corresponding to the times before noon (12 o'clock) and one to the times after noon. From this, we can then derive two separate groups of common sequences; one for each half of the day. As the following two graphs show, we now find much higher numbers of repetitions for each of the most common sequences – in some cases, a sequence appears up to 36 times more often in the split dataset.

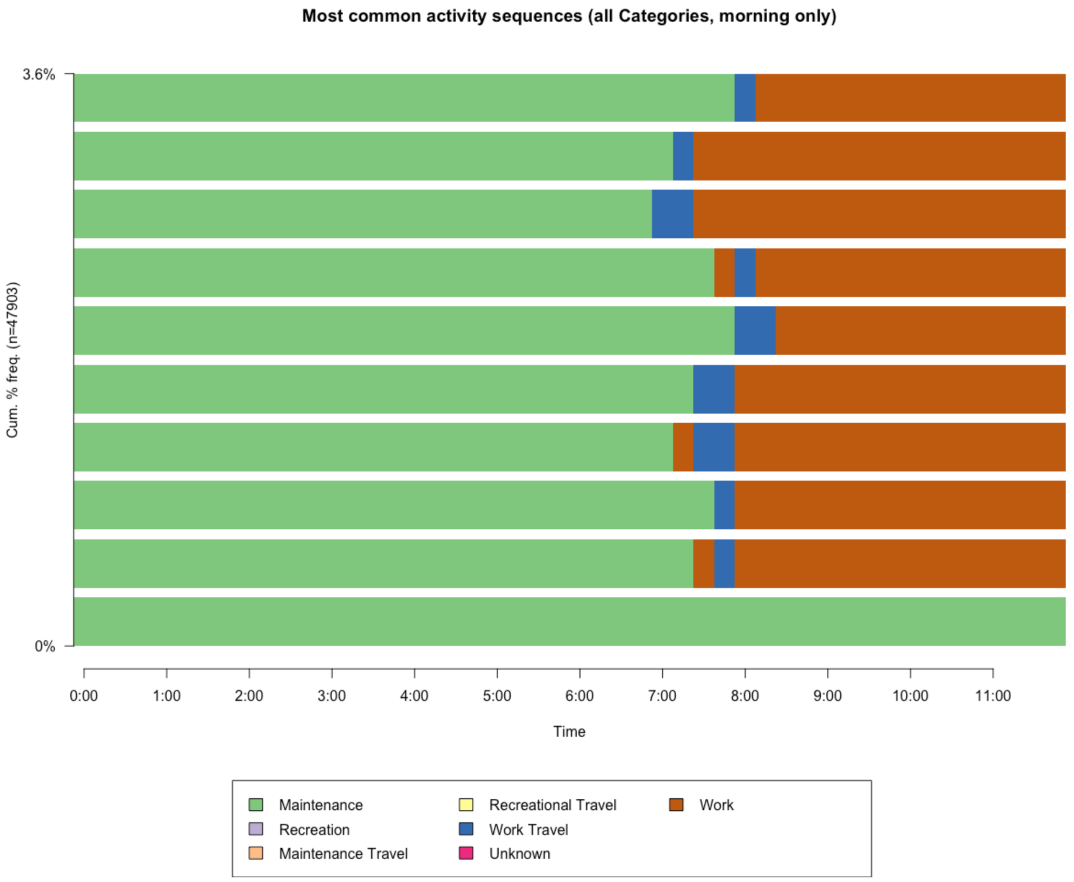


Figure 8: Most common activity sequences (all Categories, morning only)

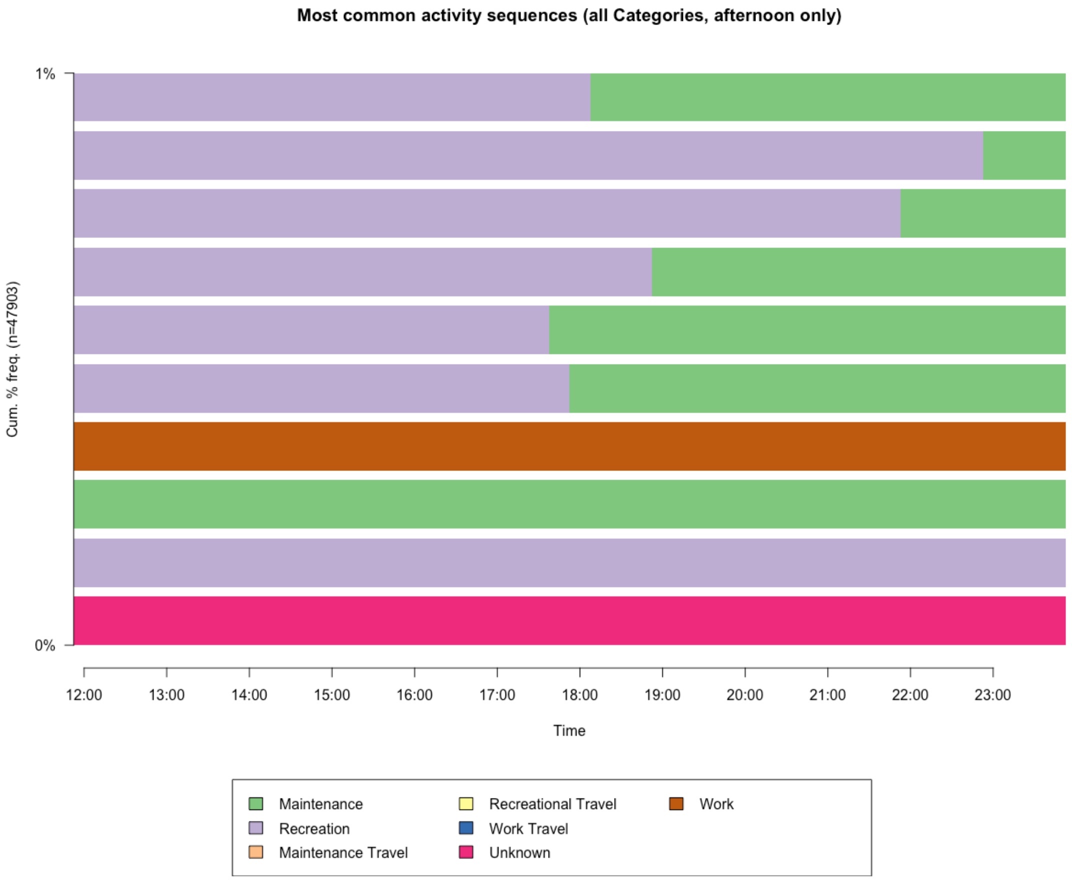


Figure 9: Most common activity sequences (all Categories, afternoon only)

An interesting information that can be derived from these two graphs is that while morning activity sequences are less varied (and thus, the most common sequences are found quite often within our dataset), the variation in afternoon activity sequences is much larger, thus leading to a smaller number of identical sequences within the dataset. Nevertheless, we also find the most common activity sequences for the afternoon (as shown in Figure 9) up to ten times more often than when considering the entire day as a whole.

Another method that can be used to alleviate the issue of only slightly variated patterns not being recognized as similar is using state distribution analysis instead of sequence analysis. This method has the advantage of showing not only the more commonly found patterns, but instead taking into account all diaries with respect to the frequency in which certain activities appear. The disadvantage of this method however is that state transitions for individual diaries (e.g. *Work* → *Work Travel)* are not visible. We consider the current state in a diary to be the activity category to which the recorded activity belongs, e.g. if the diarist is sleeping, the recorded state would be *Maintenance*. Plotting the daily state distribution for all diarists results in the following figure:

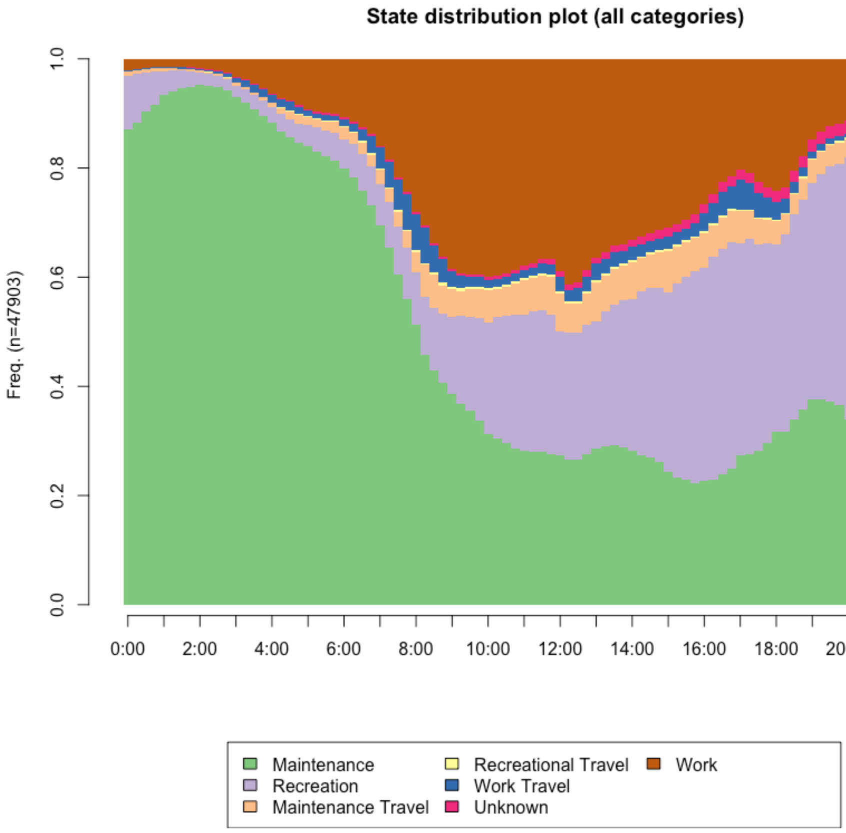


Figure 10: State distribution plot (all categories)

This state distribution plot shows clear trends for any given time during the day. Certain trends can be intuitively explained, for example the sharp decline in *Maintenance* activities at around 6am, followed by a significant increase in *Work Travel* and *Work* stems from the fact that around that time, many diarists transitioned from sleeping to going to work or actually working. We also see a significant rise of recreational activities towards the evening, indicating that diarists have more free time later in the day.

Regarding travel, we see high amounts of work travel around 7am and 5pm, which coincides with typical commuter schedules. However, Frias-Martinez, Soguero, & Frias-Martinez (2012) state that commuter schedules differ between different cultures, which would mean that applying the same analysis technique on diarists from different countries would not necessarily exhibit similar times where work travel is found in high frequencies.

Another information that can be derived from this graph is the fact that we see a rise in *Unknown* activities over the course of the day; this might be due to the diarist not recording his activities correctly after a certain time. However, these diaries were intentionally left within the dataset, since (as shown above) they can still be used for determining common patterns for parts of the day.

In addition to visualizing the states of diarists during the day, we can also compute the transition probabilities for all combinations of activity types. Using *TraMineR*, this is done by computing the percentage of transitions from any given state to any other among all sequences. Thus, every row in the transition probability matrix adds up to a total of 100%. We then visualize the transition probability matrix:

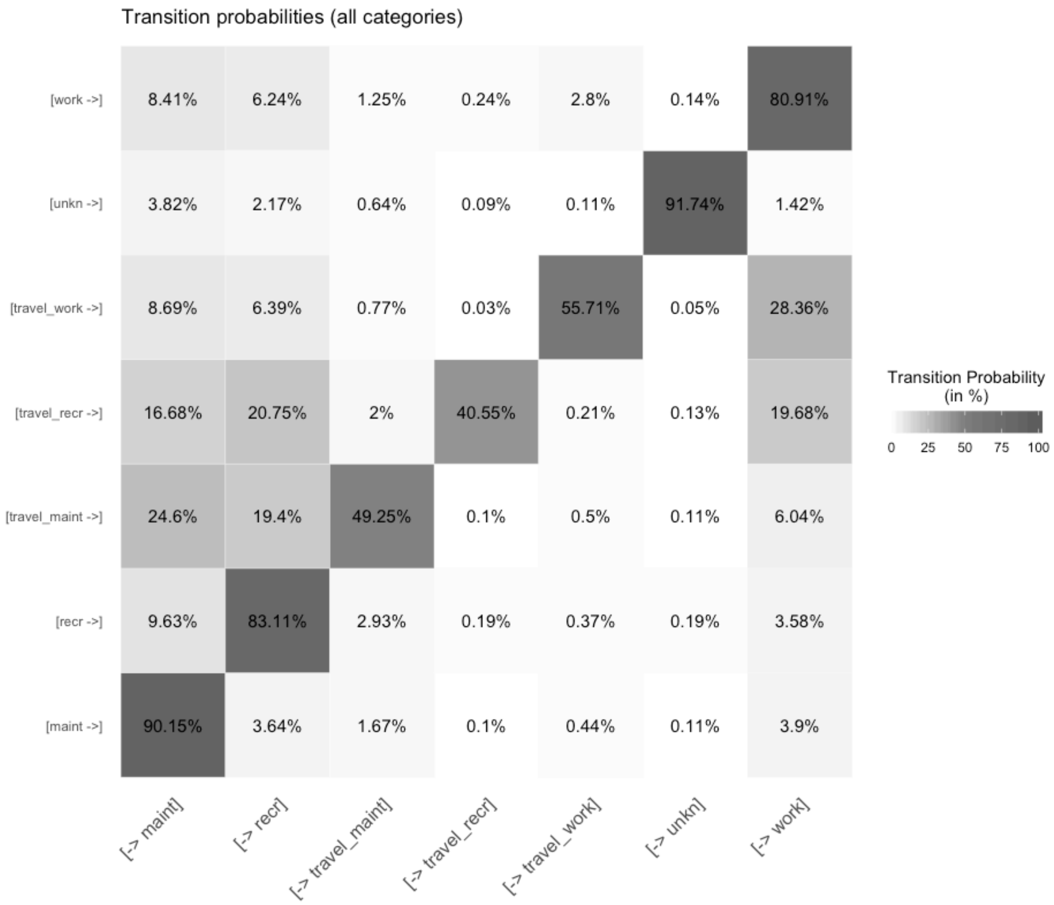


Figure 11: Transition probabilities (all categories)

The main point we can deduct from this graph is that for most activity types, the most likely transition is to another activity of the same type. This intuitively makes sense, as the transition rates are computed for fifteen minute intervals; thus, any activity that takes longer than a single fifteen-minute interval will include a transition to its own state (e.g. 30 minutes of work followed by fifteen minutes of recreation would result in the following sequence: *work* → *work* → *recreation*, which includes a transition from *work* to *work*.) In addition, we also see higher transition probabilities for each of the three travel types to the corresponding activity type:

travel_work → *work* : 28.36%
travel_recr → *recr* : 20.75%
travel_maint → *maint*: 24.6%

It can also be noted that for all three travel types, the probability to stay within the same activity type (e.g. *travel_work* → *travel_work*) is notably lower than for the other activity types; the highest value found here is 55.71% for *travel_work*, while the lowest value for non-travel activities is 80.91% for *work*. This can be considered as an indication that the average duration of travel activities is lower than that of non-travel activities.

7.3 Sequence Analysis by Lifestyle Category

In this chapter, we apply the process of finding the most common daily activity sequences using *TraMineR* to each of the lifestyle categories within our dataset. To do this, we split the previously analyzed dataset into ten individual datasets, one for each lifestyle category. Then, using the same algorithm, we look for the most common full-day as well as half-day activity patterns within the data. Note that all graphs presented in the following chapters are available in larger size in the thesis' appendix.

7.3.1 Lifestyle Category 1: "Older Wealthy Urbanites"

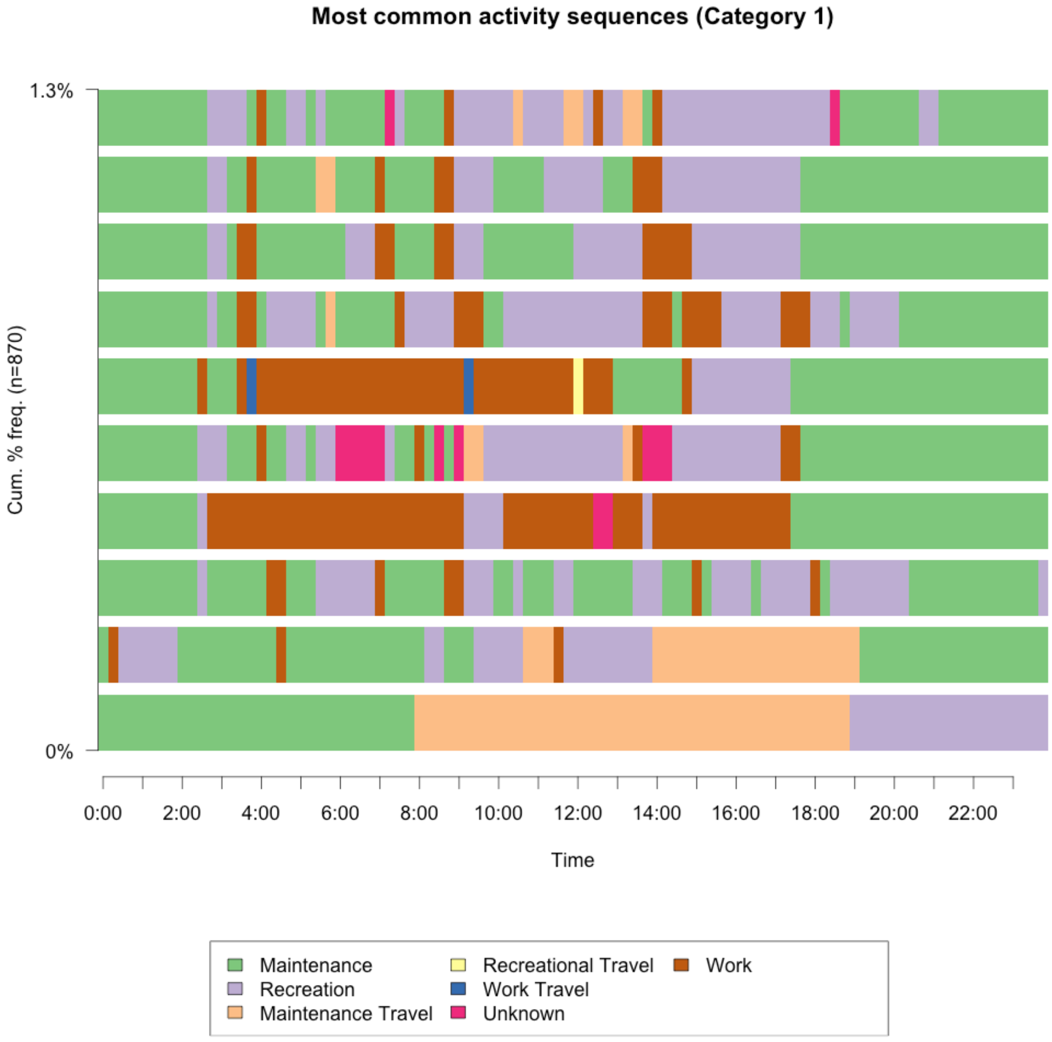


Figure 12: Most common activity sequences (Category 1)

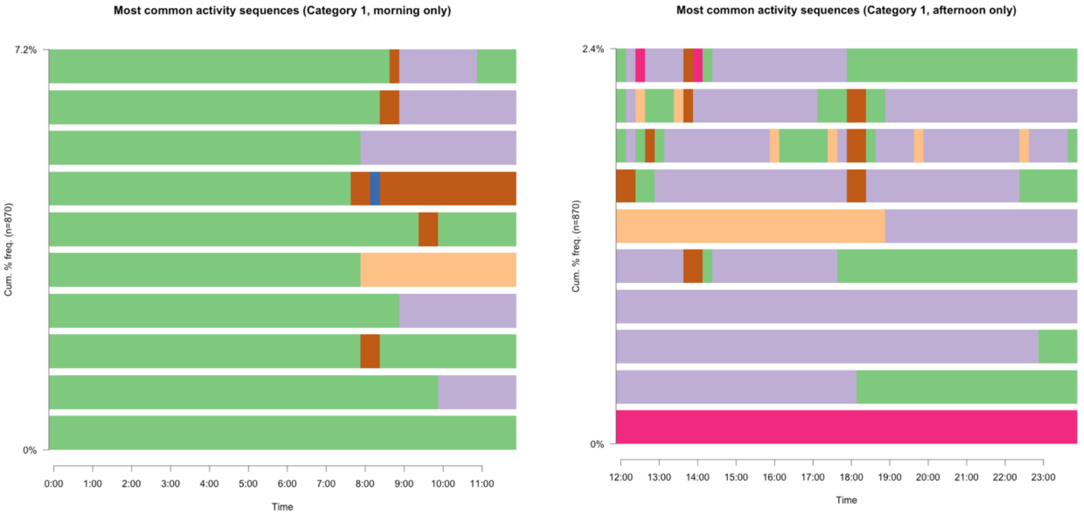


Figure 13: Most common half-day activity sequences (Category 1)

For category 1 (*WealthIndex* = 3, *UrbanIndex* = 2, *AgeIndex* = 1), we analyzed a total of 870 diaries. We observe a great variety in patterns, as well as a high number of state transitions for the most common full-day activity sequences. Notable observations for these graphs include:

1) We observe a relatively low amount of work hours, which are divided into relatively short episodes.
2) The patterns exhibit high amounts of recreational activities.
3) Work travel is found very rarely within the most common activity sequences.
4) We find medium to large amounts of maintenance travel, which are arranged into very irregular patterns.

7.3.2 Lifestyle Category 2: "Younger Wealthy Urbanites"

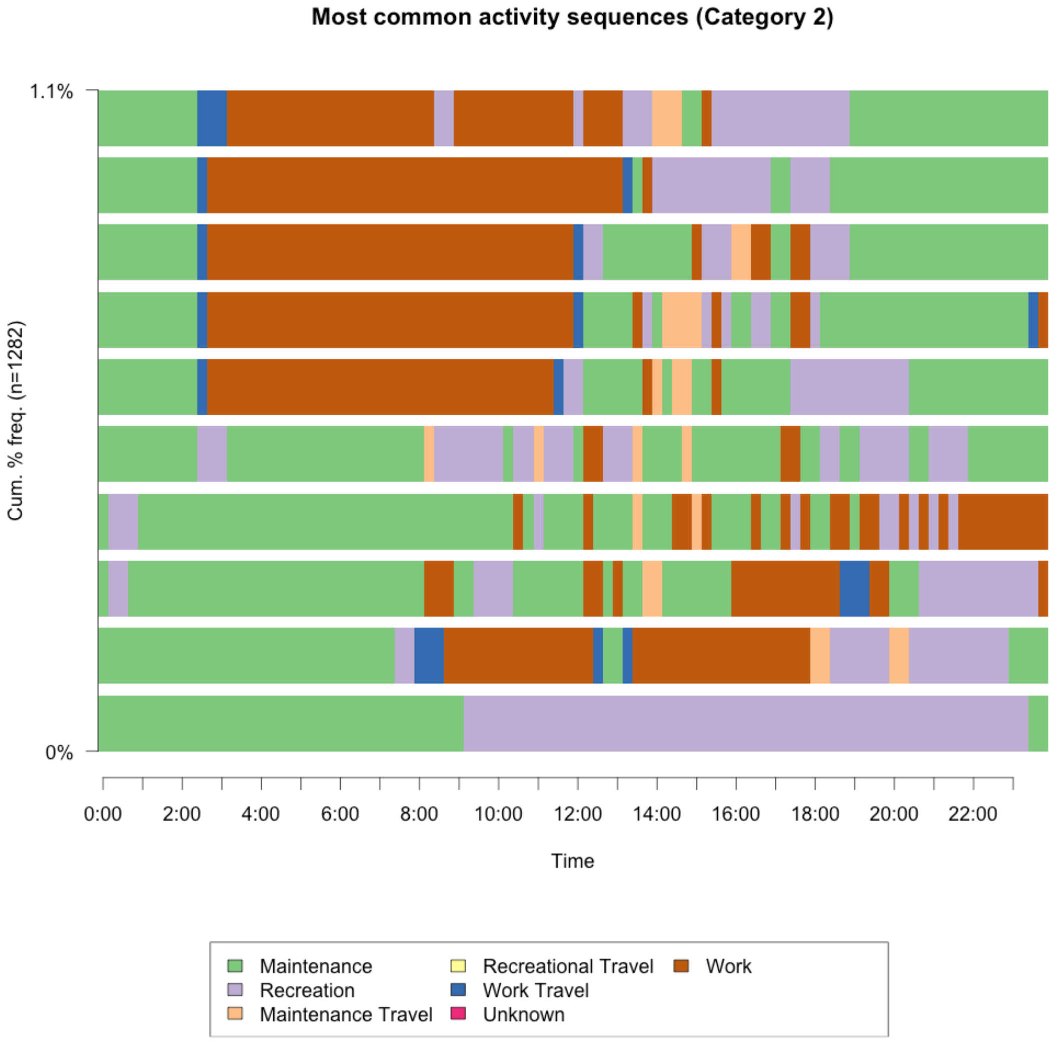


Figure 14: Most common activity sequences (Category 2)

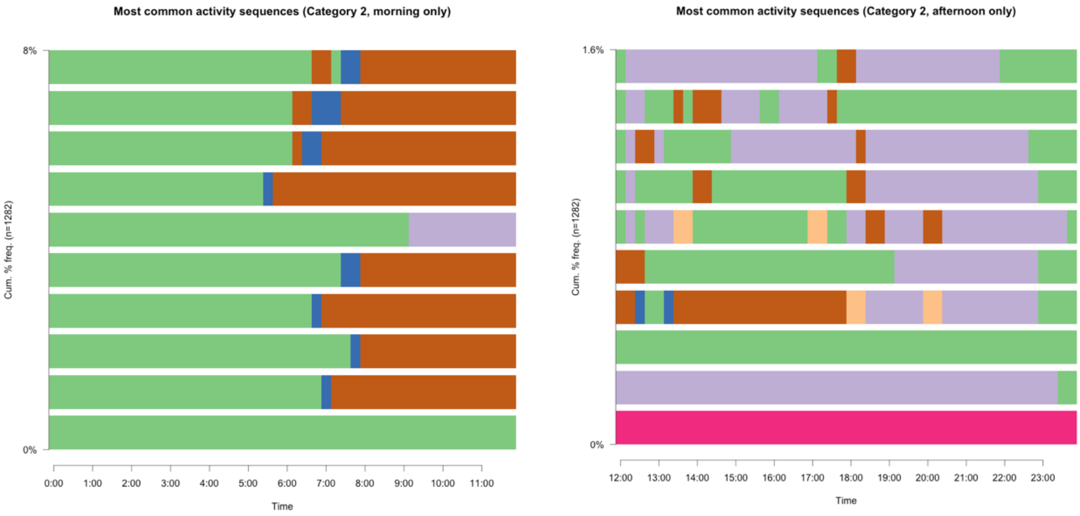


Figure 15: Most common half-day activity sequences (Category 2)

For category 2 (*WealthIndex* = 3, *UrbanIndex* = 2, *AgeIndex* = 1), we analyzed a total of 1282 diaries. We observe a lower variety in patterns than for category 1, coupled with a similarly high number of state transitions for the most common full-day activity sequences (indicating a large number of different activities undertaken during a day). Notable observations for these graphs include:

1) We observe a comparably higher amount of work hours, which are divided into longer episodes that are almost always preceded and succeeded by short episodes of work travel.
2) The start- and end-times of common *work travel* → *work* → *work travel* sequences are very similar for all of the most common patterns; also, compared to the full dataset, we observe relatively early work start times (averaging between 5:50am and 7:30am).
3) Work travel episodes are arranged very regularly and occur at similar times for most diarists.
4) We observe medium amounts of maintenance travel, which are arranged into very irregular patterns.
5) As with the full dataset, we observe a much higher variety in afternoon activity sequences compared to morning activity sequences.

7.3.3 Lifestyle Category 3: "Older Wealthy Suburbanites"

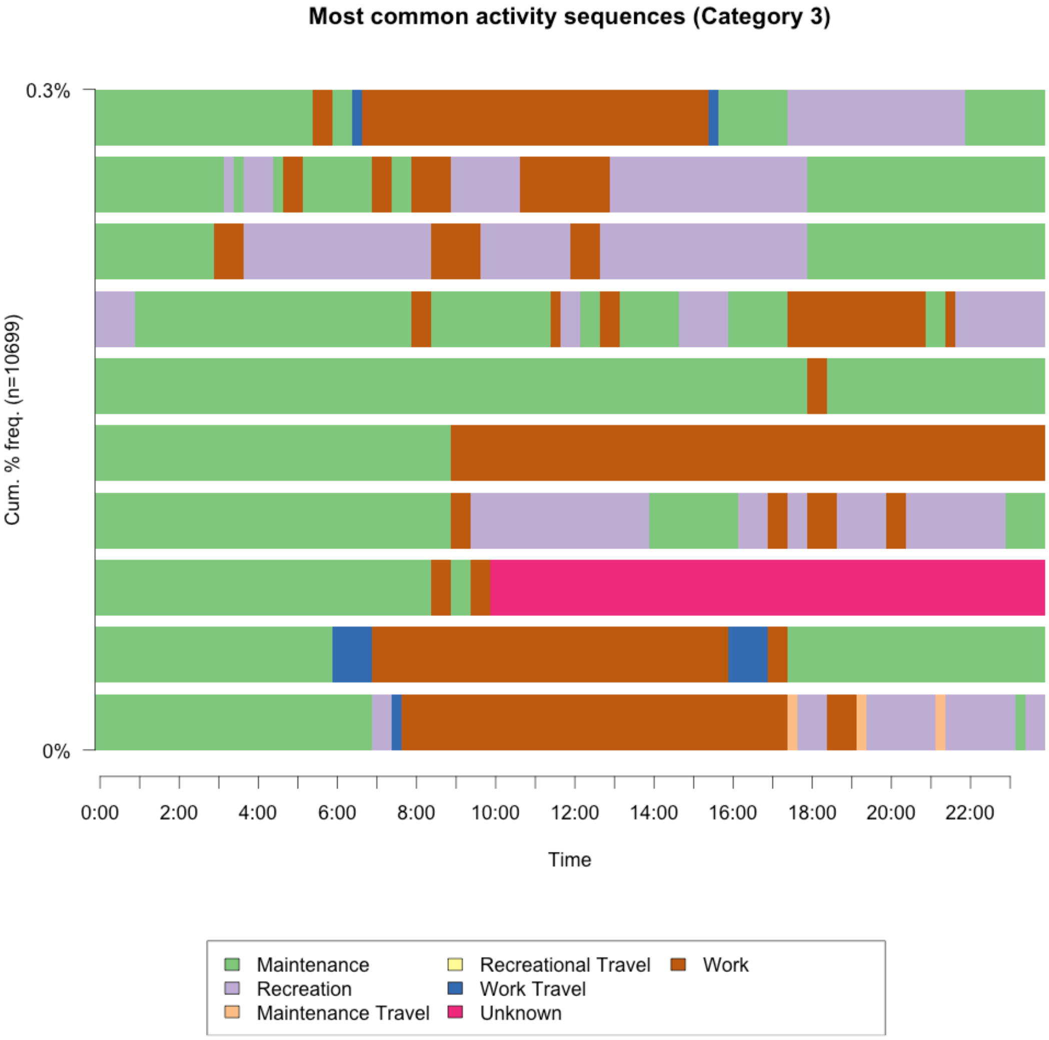


Figure 16: Most common activity sequences (Category 3)

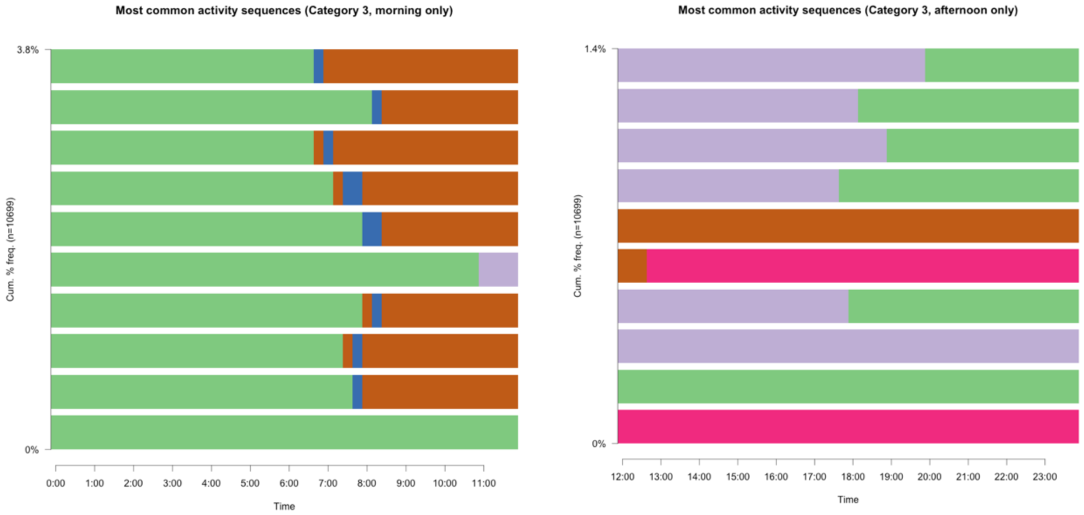


Figure 17: Most common half-day activity sequences (Category 3)

For category 3 (*WealthIndex* = 3, *UrbanIndex* = 1, *AgeIndex* = 2), we analyzed a total of 10’699 diaries. We observe a relatively high variety in patterns and a medium number of state transitions for the most common full-day activity sequences. Notable observations for these graphs include:

1) While identical full-day sequences are relatively rare within this category, a relatively high number of similar morning activity sequences can be found, indicating that for most diarists, afternoon schedules differ more than morning schedules.
2) As with category 2, the start times of common *work travel* → *work* sequences are very similar for most of the observed patterns; we however see slightly later start times than in category 2.
3) Work travel is relatively low in duration and is found very regularly before and after longer work episodes.
4) While we see a large number of longer work episodes, we also observe a significant number of diaries wherein work hours are divided among a larger number of short work episodes.
5) Maintenance travel is very limited in both frequency and duration.

7.3.4 Lifestyle Category 4: "Younger Wealthy Suburbanites"

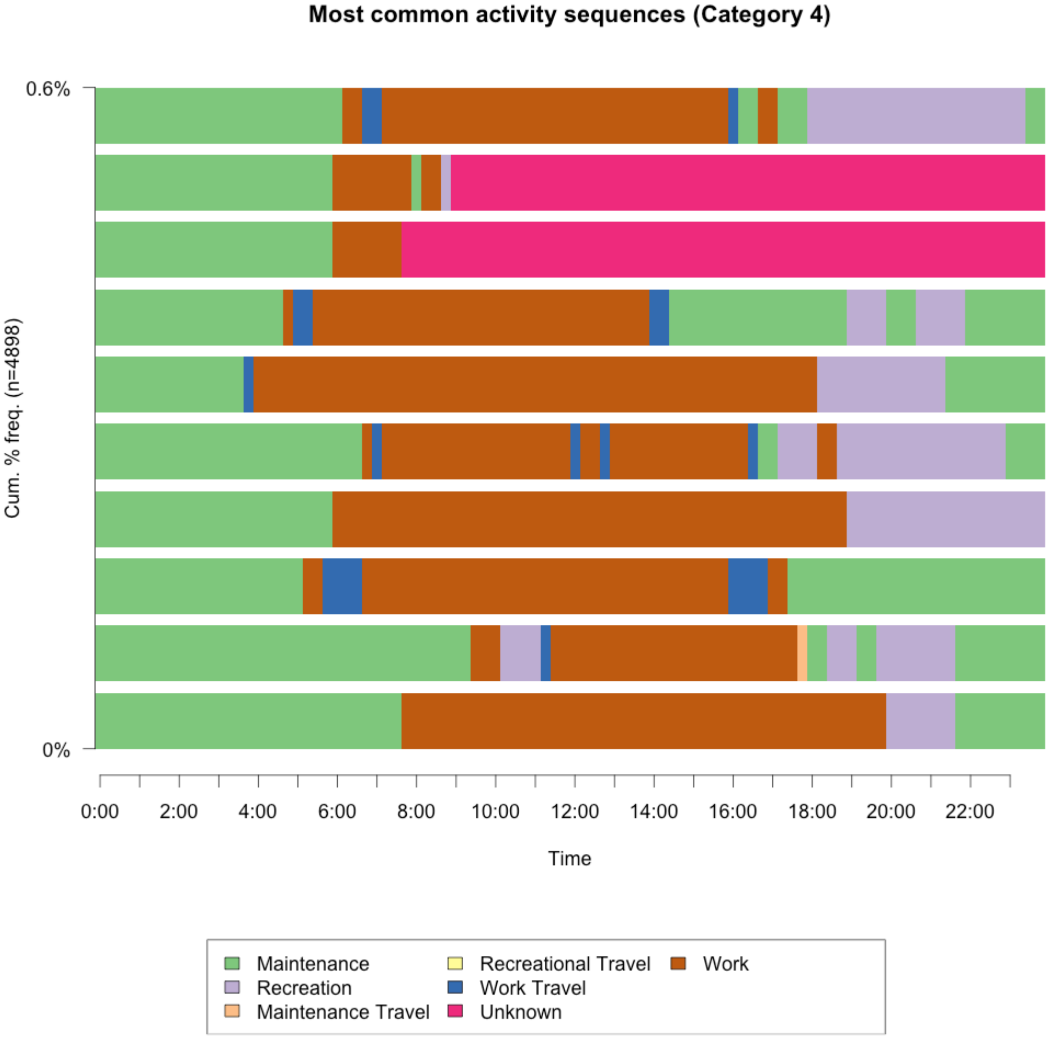


Figure 18: Most common activity sequences (Category 4)

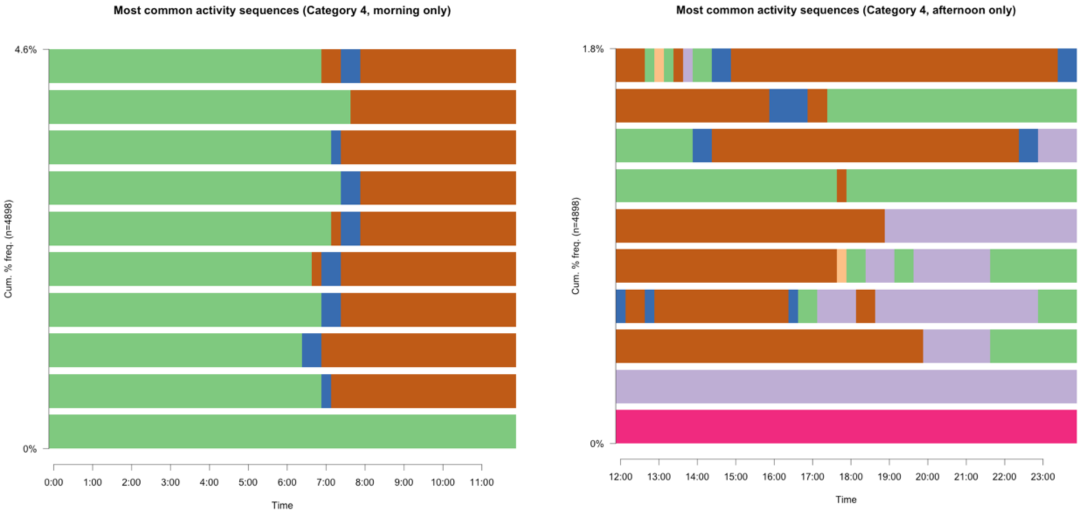


Figure 19: Most common half-day activity sequences (Category 4)

For category 4 (*WealthIndex* = 3, *UrbanIndex* = 1, *AgeIndex* = 1), we analyzed a total of 4898 diaries. We observe relatively regular patterns exhibiting typical work-day structures for this category. Notable observations for these graphs include:

1) We observe regular, long work episodes which are almost always preceded and succeeded by short to medium-length work travel episodes.
2) The patterns exhibit relatively low amounts of recreational activities, found almost exclusively towards the evening.
3) For many diaries, work travel is preceded by short work episodes.
4) Much like in category 2, the start- and end-times of common *work travel* → *work* → *work travel* sequences are very similar for the most common patterns.

7.3.5 Lifestyle Category 5: "Urban Middle-Class"

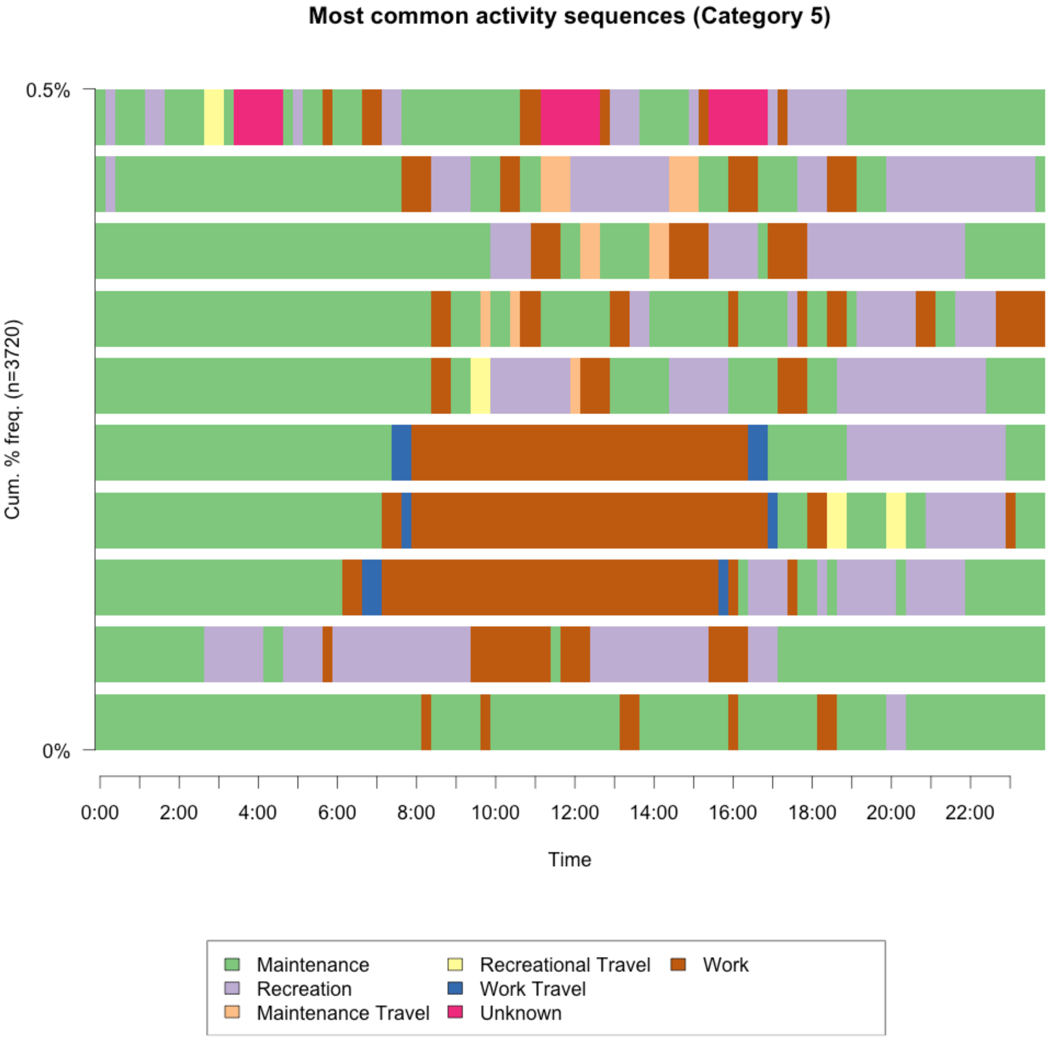


Figure 20: Most common activity sequences (Category 5)

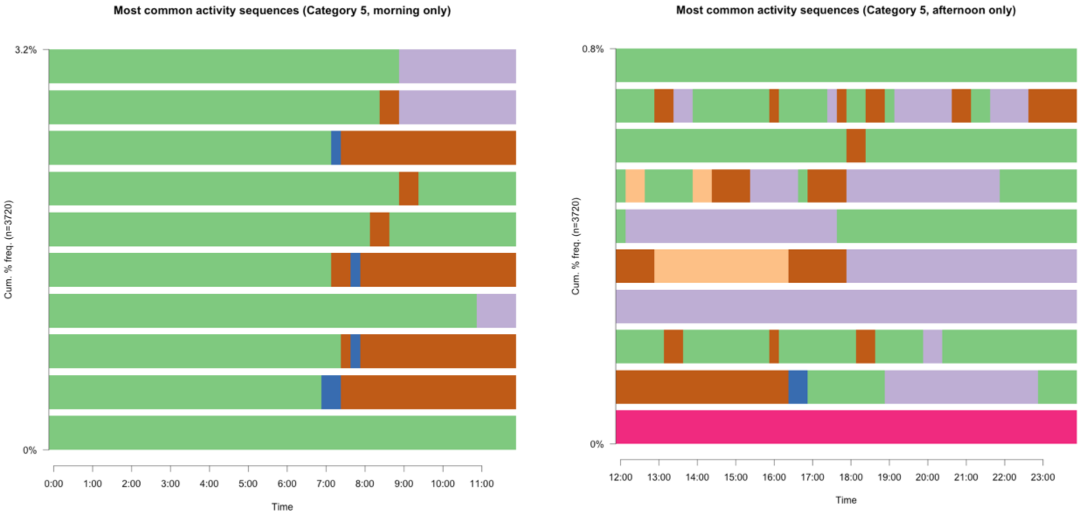


Figure 21: Most common half-day activity sequences (Category 5)

For category 5 (*WealthIndex* = 2, *UrbanIndex* = 2, *AgeIndex* = 1 & 2), we analyzed a total of 3720 diaries. We observe very irregular patterns with high numbers of different activities during the day. Notable observations for these graphs include:

1) We observe mostly irregular, short work episodes which are mostly not preceded by any work travel; for some longer work episodes, low amounts of preceding work travel can be observed.
2) The patterns exhibit relatively high numbers of recreational activities, which are mostly divided into episodes with a duration of less than an hour.
3) We also irregularly find medium duration episodes of recreational travel.
4) An aspect that might influence the relatively high pattern regularity for this category is the fact that both older and younger diarists (Age Indices 1 and 2) are included in these observations.

7.3.6 Lifestyle Category 6: “Older Suburban Middle-Class”

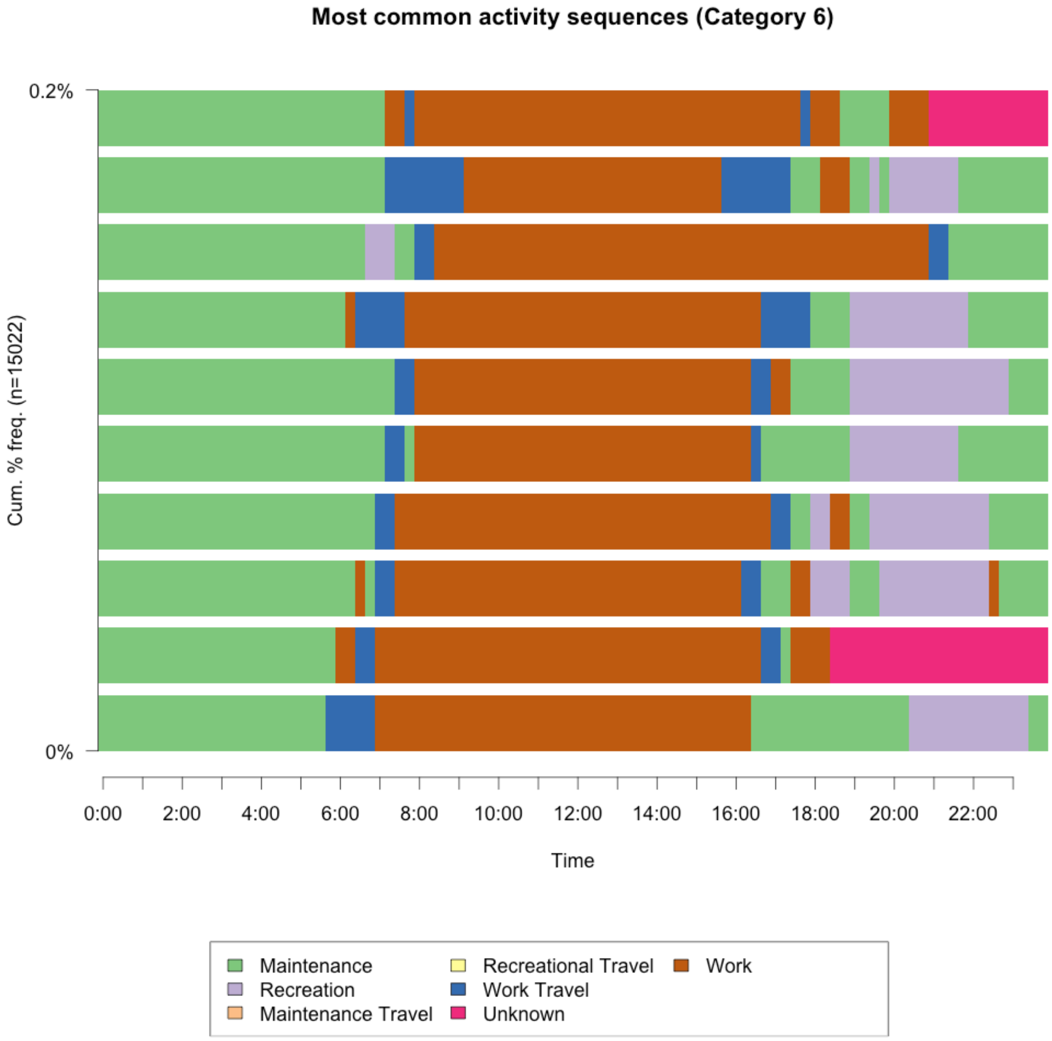


Figure 22: Most common activity sequences (Category 6)

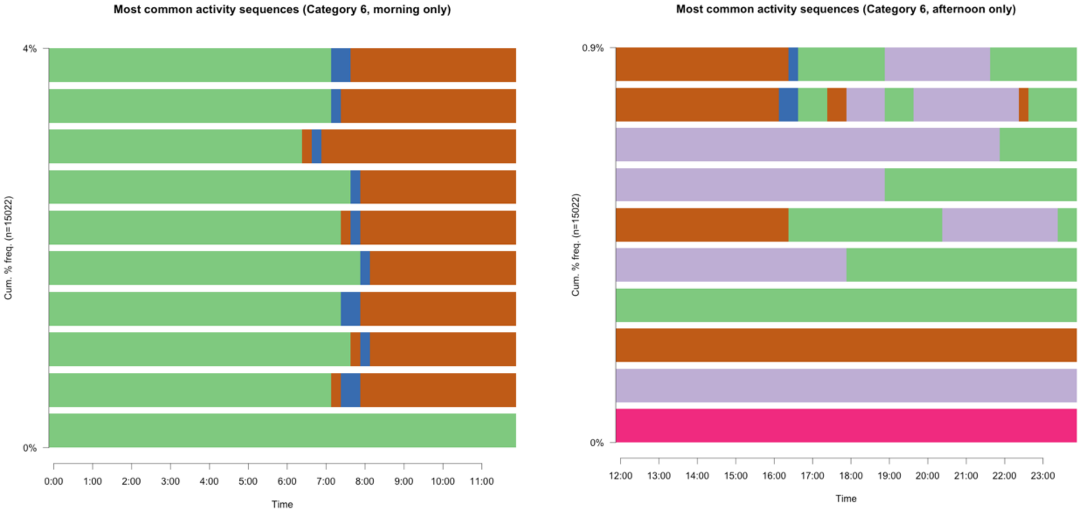


Figure 23: Most common half-day activity sequences (Category 6)

For category 6 (*WealthIndex* = 2, *UrbanIndex* = 2, *AgeIndex* = 2), we analyzed a total of 15'022 diaries, making it the largest category within our dataset. We observe very irregular patterns with high numbers of different activities during the day. Notable observations for these graphs include:

1) We observe regular, long work episodes which are almost always preceded and succeeded by work travel episodes with medium to long durations.
2) The patterns exhibit relatively low amounts of recreational activities, found almost exclusively towards the evening.
3) Much like in Categories 2 and 4, the start- and end-times of common *work travel* → *work* → *work travel* sequences are very similar for the most common patterns.

7.3.7 Lifestyle Category 7: "Younger Suburban Middle-Class"

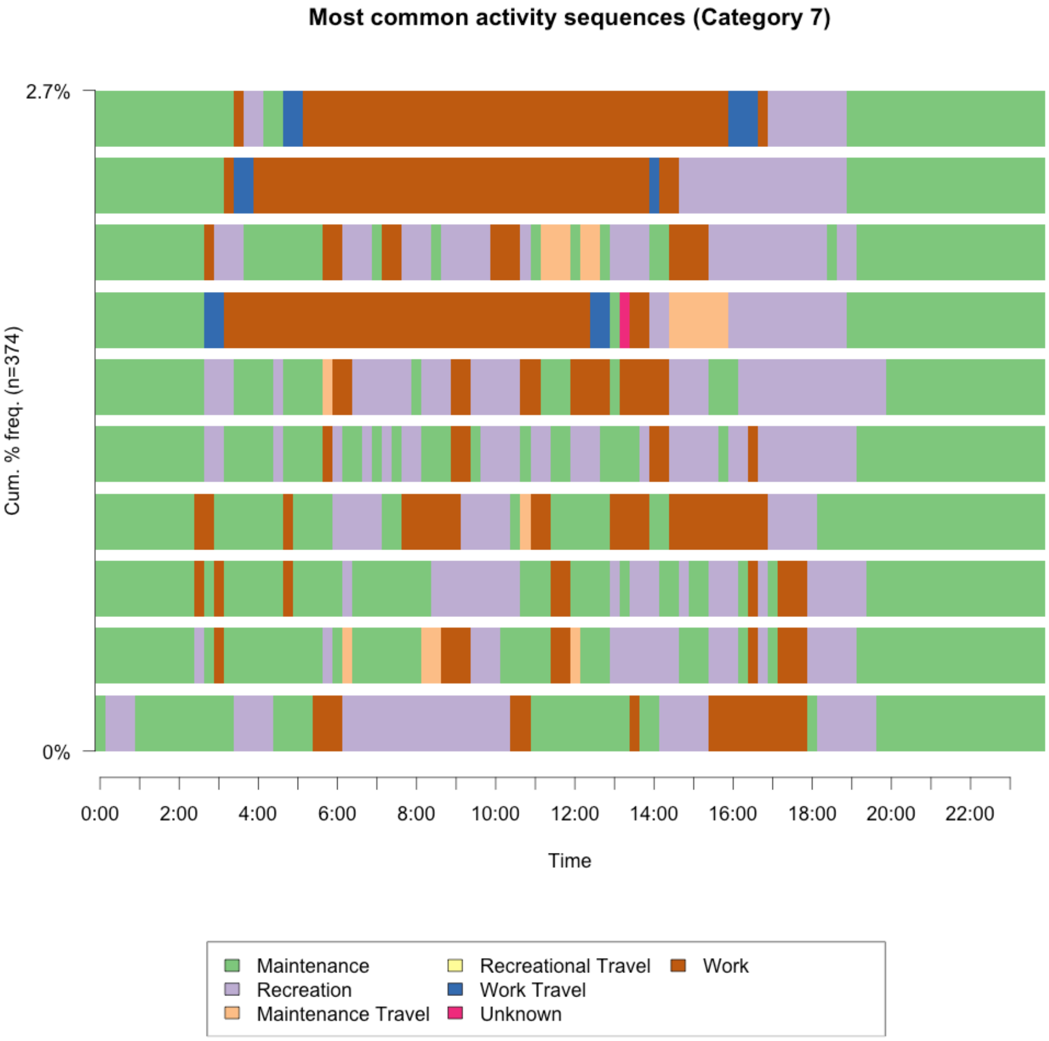


Figure 24: Most common activity sequences (Category 7)

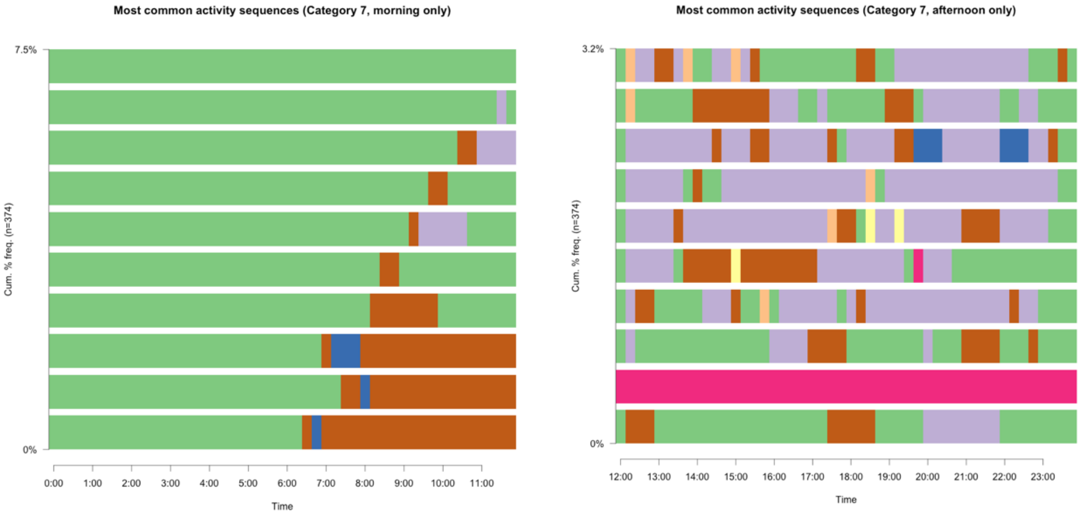


Figure 25: Most common half-day activity sequences (Category 7)

For category 7 (*WealthIndex* = 2, *UrbanIndex* = 1, *AgeIndex* = 1), we analyzed a total of 374 diaries, making it one of the smallest categories. We observe a great variety in patterns, with some being composed of mostly long activities, while others are composed of a high number of short activities. Notable observations for these graphs include:

1) For the most common full-day sequences, we observe typical sequences of *work travel* → *work* → *work travel*, wherein the *work travel* episodes are of medium duration.
2) The patterns containing low amounts of work hours exhibit relatively high amounts of recreational activities, coupled with scattered short episodes of work.
3) In some afternoon activity sequences, we find very short episodes of maintenance travel and recreational travel which are arranged into very irregular patterns.

7.3.8 Lifestyle Category 8: "Urban Lower-Class"

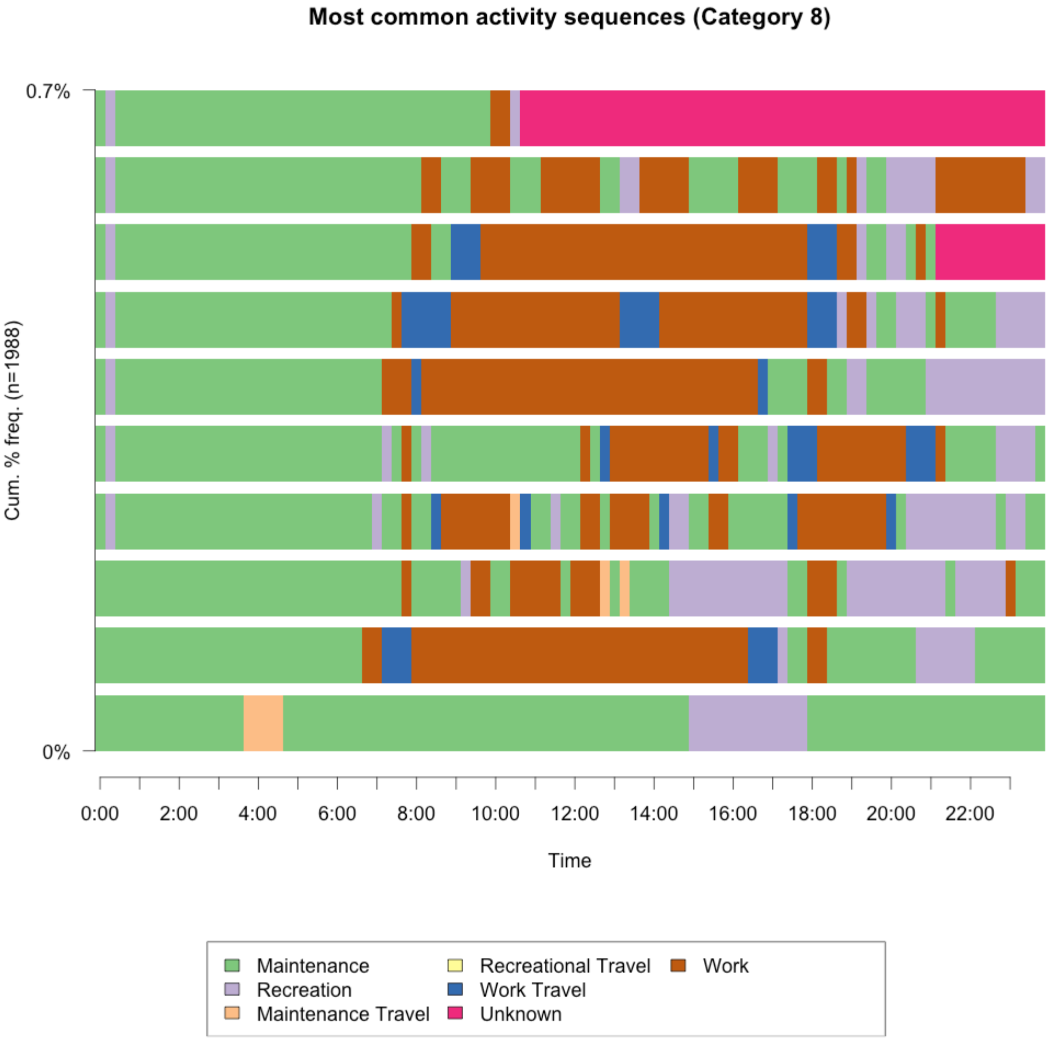


Figure 26: Most common activity sequences (Category 8)

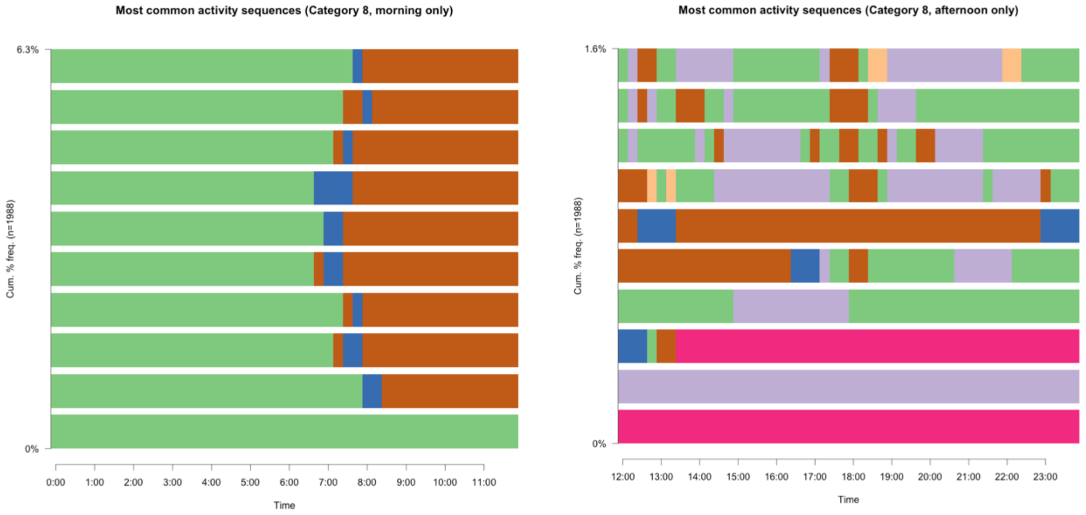


Figure 27: Most common half-day activity sequences (Category 8)

For category 8 (*WealthIndex* = 1, *UrbanIndex* = 2, *AgeIndex* = 1 & 2), we analyzed a total of 1988 diaries. We observe a low to medium variety in patterns, as well as a high amount of work hours for the most common morning activity sequences. Notable observations for these graphs include:

1) Most longer episodes of work activities are preceded and succeeded by relatively long episodes of work travel.
2) The patterns exhibit medium amounts of recreational activities, occurring almost exclusively in the afternoon.
3) While the most common morning activity sequences are mostly very similar (*maintenance* → *work* travel → *work*), we observe a variety that is around four times higher among afternoon sequences.
4) We find low to medium amounts of maintenance travel, which are arranged into very irregular patterns.

7.3.9 Lifestyle Category 9: "Older Suburban Lower-Class"

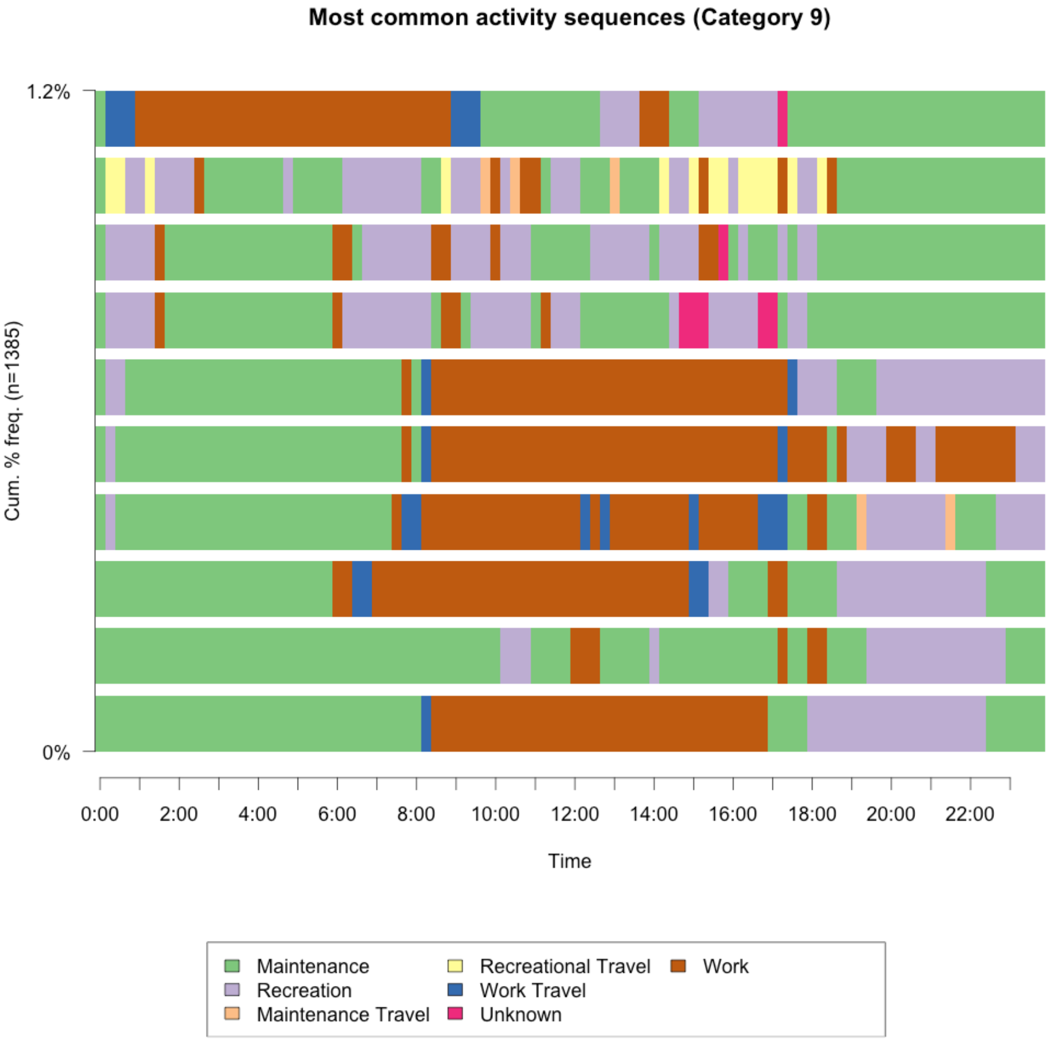


Figure 28: Most common activity sequences (Category 9)

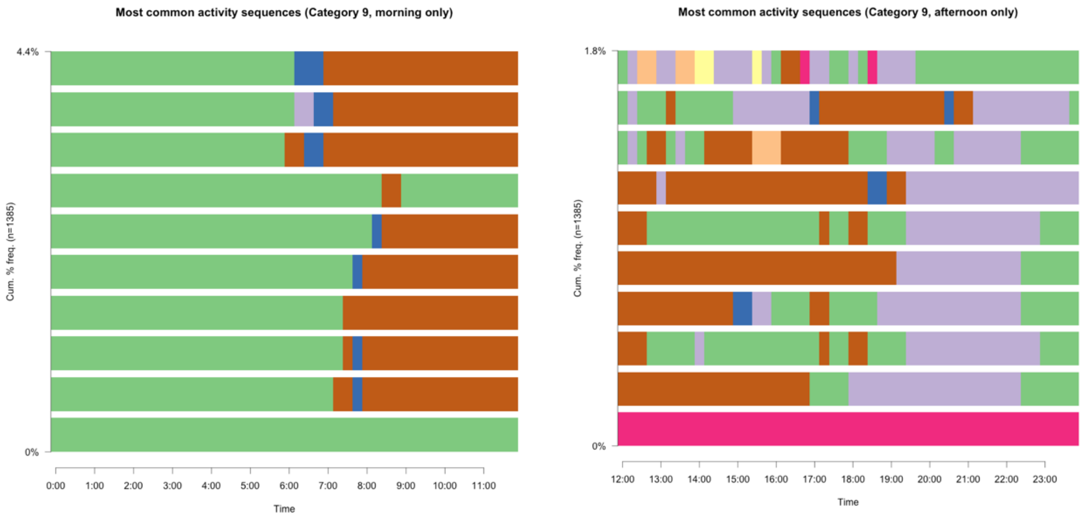


Figure 29: Most common half-day activity sequences (Category 9)

For category 9 (*WealthIndex* = 1, *UrbanIndex* = 1, *AgeIndex* = 2), we analyzed a total of 1385 diaries. For full-day activity sequences, we observe a high number of short episodes arranged in irregular patterns. Notable observations for these graphs include:

1) While we observe a high variety among the full-day activity sequences for category 9, examining only morning sequences shows very regular patterns consisting mostly of *maintenance* → *work travel* → *work* sequences.
2) Maintenance travel mostly occurs in medium duration episodes, but not very frequently.
3) For many of the most common patterns, small work episodes spread throughout the day (apart from one longer work episode during the day) can be observed.
4) Recreational activities occur mostly in the evening, but are also found frequently during the morning for common full-day activity sequences.

7.3.10 Lifestyle Category 10: “Younger Suburban Lower-Class”

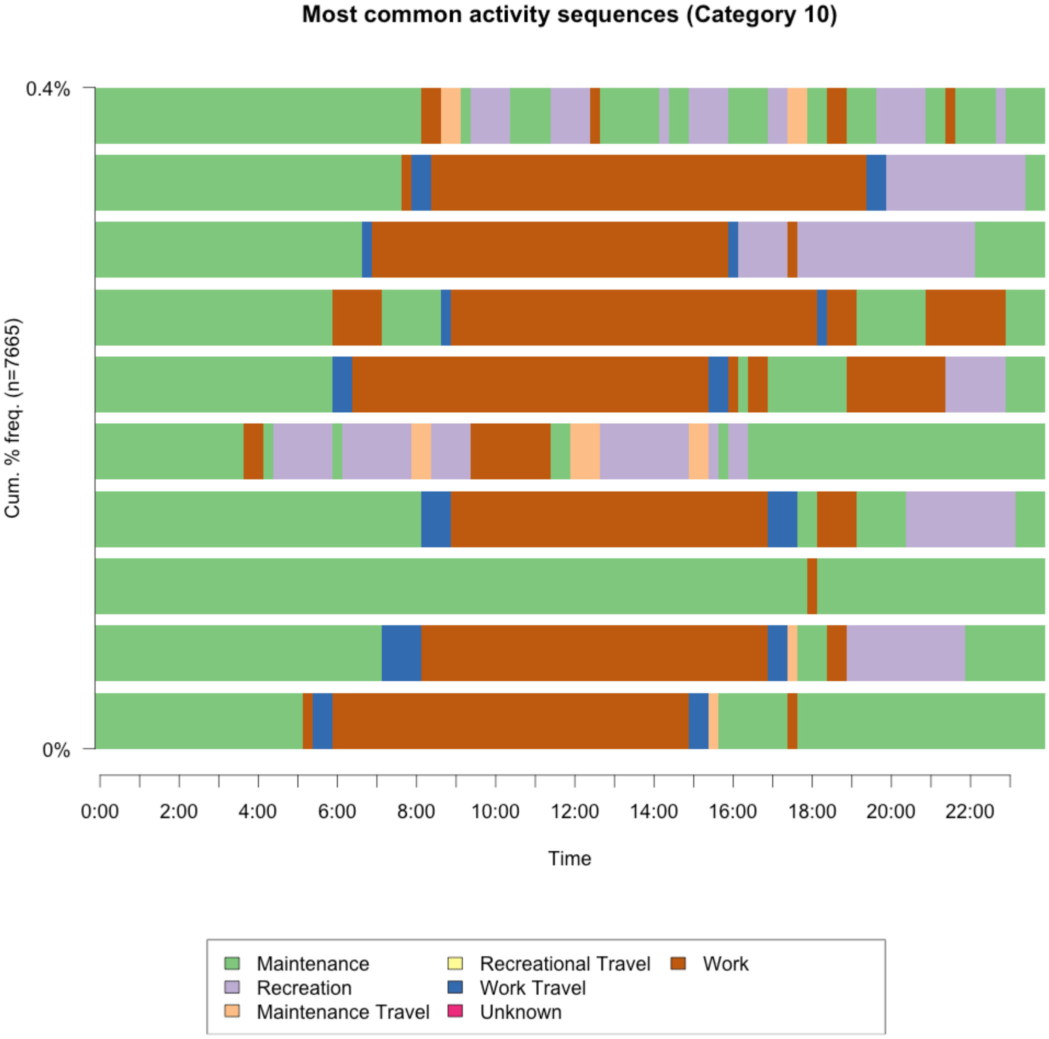


Figure 30: Most common activity sequences (Category 10)

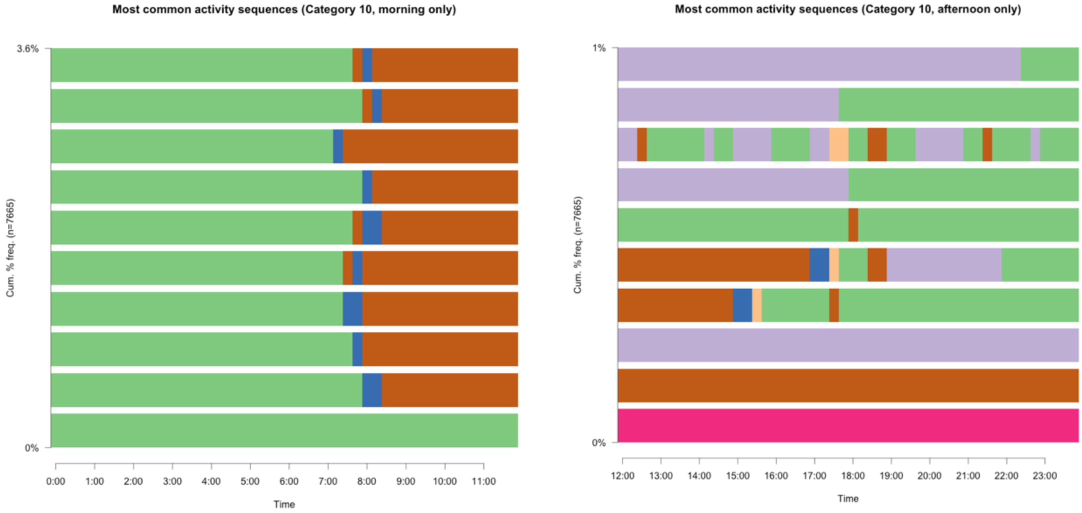


Figure 31: Most common half-day activity sequences (Category 10)

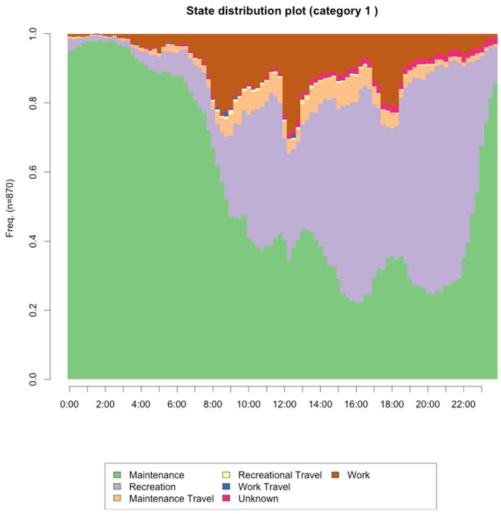
For category 10 (*WealthIndex* = 1, *UrbanIndex* = 1, *AgeIndex* = 1), we analyzed a total of 7665 diaries. Contrary to most other groups with an *AgeIndex* of 1, we observe mostly typical work-day structures similar to those found in younger Categories. Notable observations for these graphs include:

1) Work episodes are arranged very similarly among diarists and are almost always preceded and succeeded by work travel of short to medium duration.
2) The most common afternoon activity patterns exhibit high amounts of recreational activities, as well as low amounts of maintenance travel.
3) For the most common morning activity sequences, we observe very similar starting times for *work travel* → *work* sequences.
4) Diarists within this group seem to fall within two groups: one that works regular hours and another that exhibits high amounts of time spent on recreational activities.

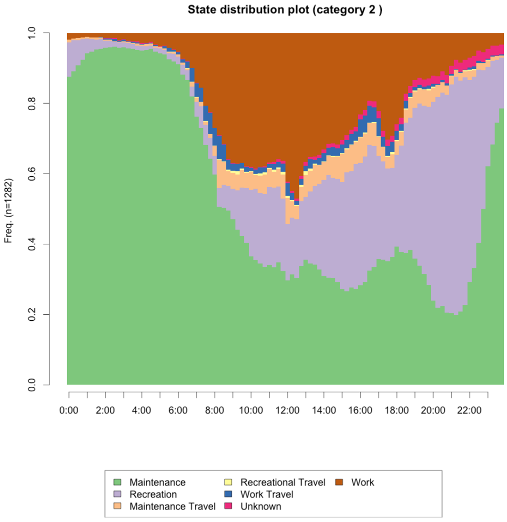
7.4 State Distribution by Lifestyle Category

In addition to analyzing the most common activity sequences for all lifestyle categories, we also plot the activity state distribution (as described in chapter 7.2) for each lifestyle category individually. The following graph shows the state distribution plots for all categories:

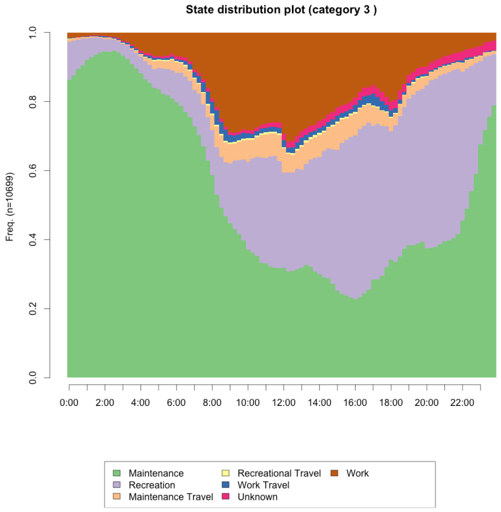
x



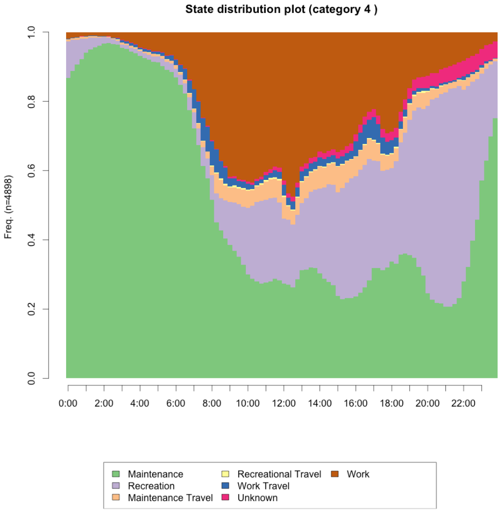
State distribution plot (category 1)



State distribution plot (category 2)



State distribution plot (category 3)



State distribution plot (category 4)

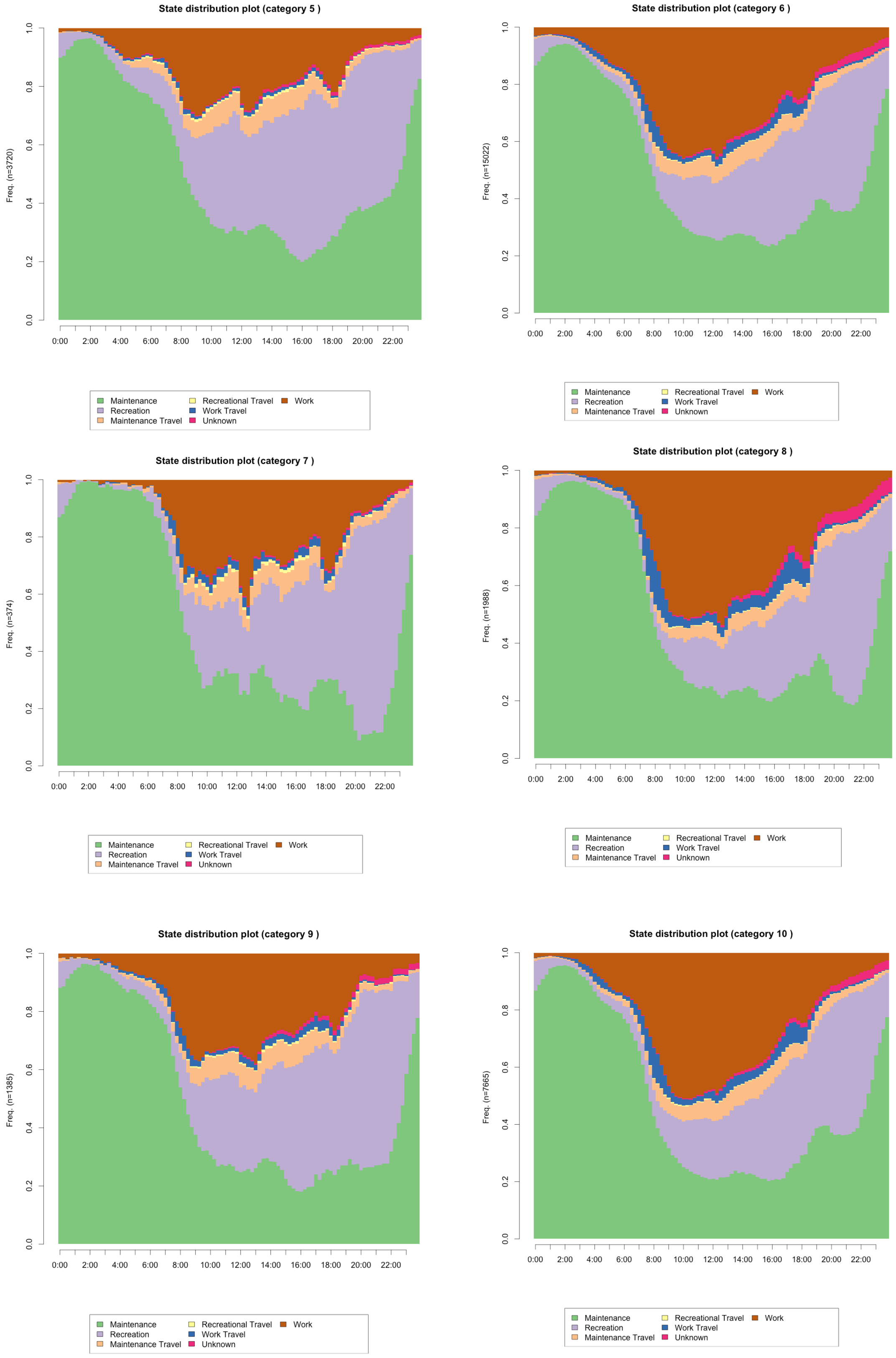


Figure 32: State distribution plots per Lifestyle Category

From these graphs, we observe similar trends as described in Chapter 7.3. One major point is the fact that the daily distribution of maintenance travel looks very similar in both amounts and distribution among all categories. This would mean similar amounts of maintenance travel for all categories; more on this topic will be discussed in chapter 7.5.

For some categories with high varieties of activity sequences (like categories 1 and 7), we observe more erratic state distribution patterns, whereas categories with high regularity in diaries (like categories 6 and 10) exhibit relatively smooth state distributions during the day.

Within categories with higher amounts of work travel, we also typically observe peaks in work travel activities around typical commuter times, such as between 7am and 8am as well as between 5pm and 6pm.

7.5 Quantitative Analysis by Lifestyle Category

We also analyzed how each lifestyle category distributed their time per day on average. First, this was done for all six activity types (as well as "*unkn*", which was included to maintain completeness). This distribution looks as follows:

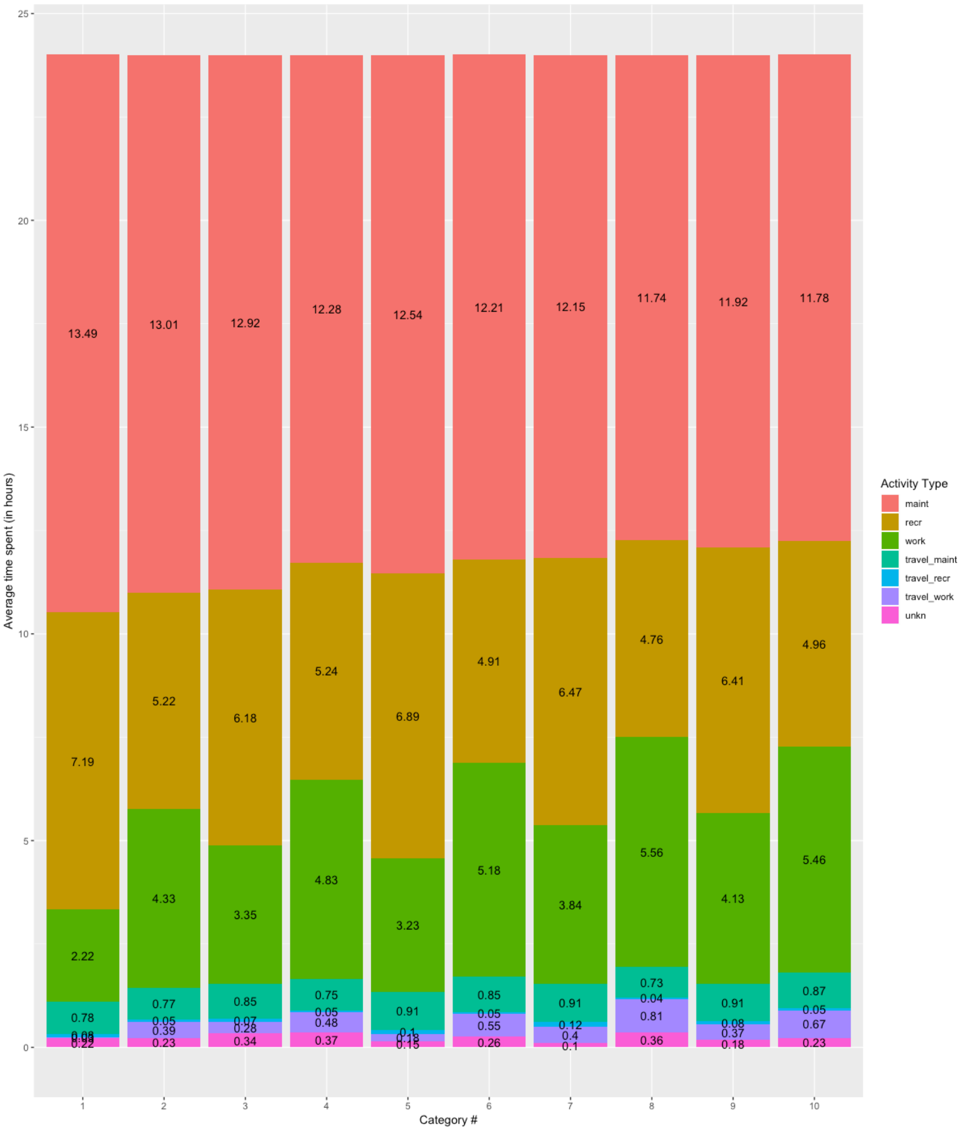


Figure 33: Time distribution by category and activity type

From this graph, we draw the following conclusions:

- Time spent on recreation is moderately dependent on lifestyle category.
- Time spent on work is very category-dependent, as its values range from 2.22 hours per day to 5.56 hours per day.
- Time spent on maintenance seems largely uncorrelated to lifestyle category, with a mean of 12.38 hours and a variance of only 0.34.

However, as our focus lies on travel behavior, we further break down this time distribution graph by omitting all non-travel activities (*"work"*, *"recr"*, *"maint"*, *"unkn"*). This results in the following time distribution per category:

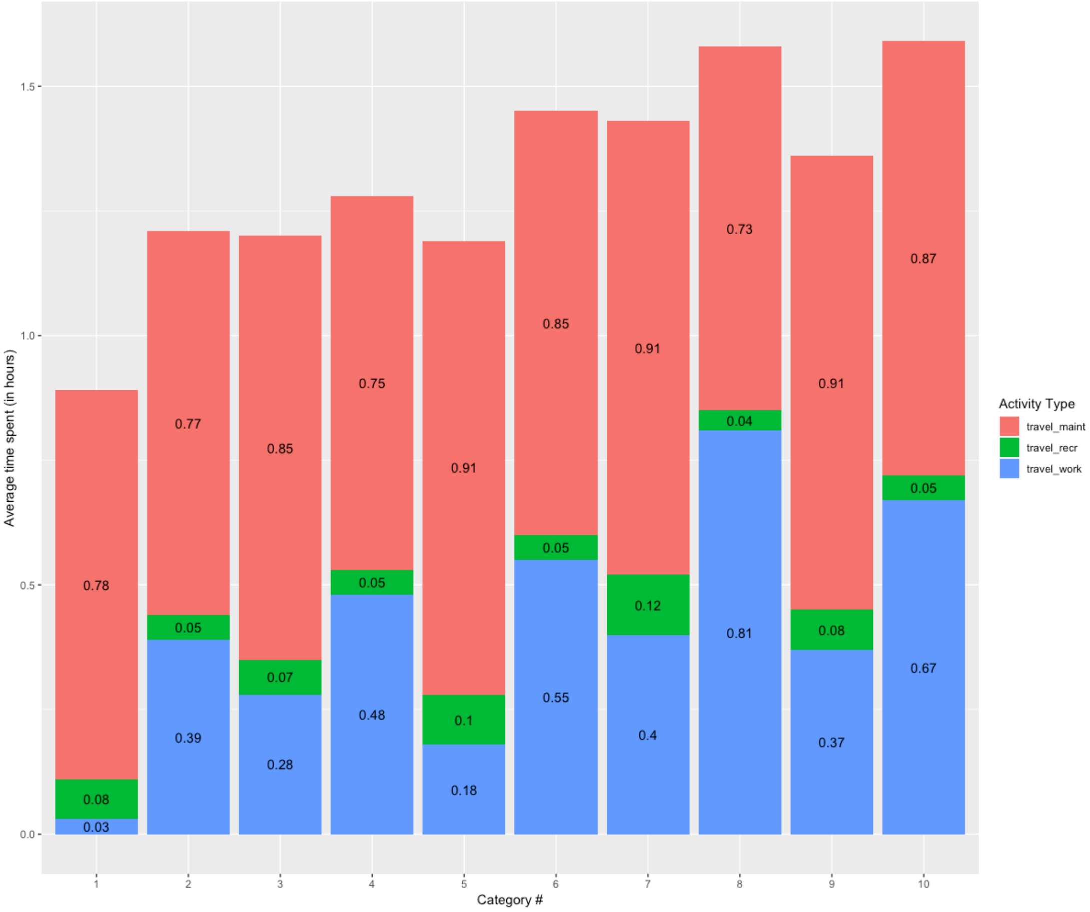
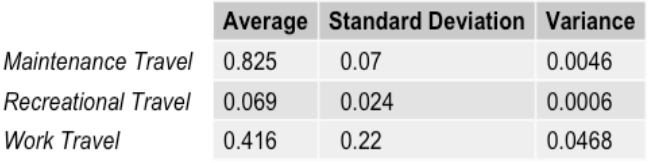


Figure 34: Travel type distribution by Lifestyle Category

From this graph, the differences in travel times between the different lifestyle categories become much more apparent. Calculating the average duration (in hours), standard deviation and variance across all categories yields the following results:



	Average	Standard Deviation	Variance
Maintenance Travel	0.825	0.07	0.0046
Recreational Travel	0.069	0.024	0.0006
Work Travel	0.416	0.22	0.0468

Figure 35: Statistical values by Travel Type

While we observe relatively high standard deviations for Work Travel and Recreational Travel, Maintenance Travel seems to be largely unaffected by the criteria that went into our lifestyle categorization, meaning we see very similar values across all lifestyle categories.

We observe the highest amounts of maintenance travel for lifestyle categories 5, 7 and 9 with 54.6 minutes of average daily maintenance travel for both. However, the amounts are very similar across all lifestyle categories, with the minimum being found for lifestyle category 8 with a daily average of 43.8 minutes of maintenance travel. Recreational travel is found most often for lifestyle category 7 ("Younger Suburban Middle-Class"), with a daily average of 0.12 hours (7.2 minutes). Work travel is the most varied travel category among the diarists: While lifestyle category 1 ("Older Wealthy Urbanites") only reported a daily average of 1.8 minutes, lifestyle category 8 ("Urban Lower-Class") reported a daily average work travel duration of 48.6 minutes.

7.6 Results

The analyses conducted in this chapter result in several conclusions about the diarists within the MTUS HEF dataset.

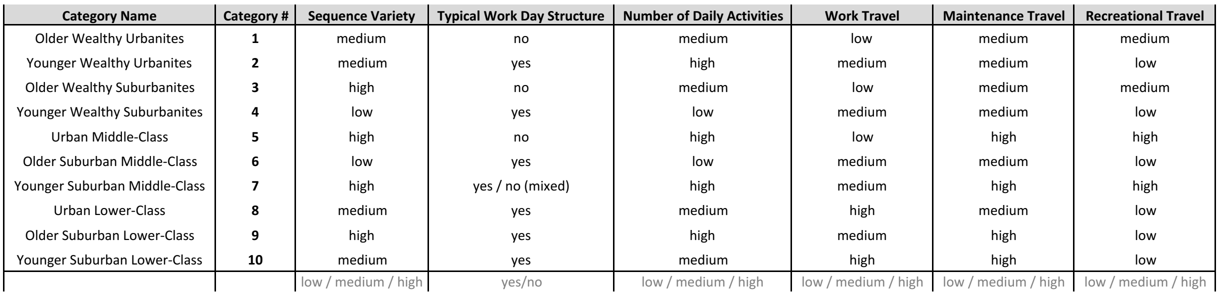
In the sequence analyses described in chapter 7.3, we observed mostly two different groups: Firstly, those with regular, scheduled days consisting mostly of work, work travel and maintenance activities. Secondly, there were groups of diarists that commonly had very irregular daily schedules, consisting mostly of maintenance and recreational activities, coupled with low working hours. This is also backed by chapter 7.4, wherein we saw notably more irregular state distribution patterns for the groups showing irregular daily schedules. As shown in both chapter 7.3 and chapter 7.4, older diarists (*ageIndex* = 1, meaning an age of more than 55 years) exhibit more erratic and thus unpredictable activity sequence patterns. This also coincides with the fact that we found older lifestyle categories to have a lower average amount of work hours, thus showing an inverse correlation between the diarists' amount of work hours and the irregularity in their activity sequences.

Regarding travel times, chapter 7.5 shows that the lifestyle categorization we applied to the MTUS dataset is most suitable for determining work travel frequency, as this has proven to be most correlated to the different lifestyle categories we determined.

Recreational travel shows moderate correlation to our lifestyle categories, but since we only observed low amounts of recreational travel among our dataset, we cannot make any definitive claims about the correlation between this type of travel and our lifestyle categorization.

Contrary to this, maintenance travel seems largely unaffected by the different lifestyle categories and varies only very slightly among them, making these lifestyle categories unsuitable for analyzing and comparing maintenance travel.

The following table summarizes our findings across all lifestyle categories with regard to sequence variety within the category, whether a typical work day structure is present, the number of daily activities, as well as the average amount of time spent on each travel type:



Category Name	Category #	Sequence Variety	Typical Work Day Structure	Number of Daily Activities	Work Travel	Maintenance Travel	Recreational Travel
Older Wealthy Urbanites	**1**	medium	no	medium	low	medium	medium
Younger Wealthy Urbanites	**2**	medium	yes	high	medium	medium	low
Older Wealthy Suburbanites	**3**	high	no	medium	low	medium	medium
Younger Wealthy Suburbanites	**4**	low	yes	low	medium	medium	low
Urban Middle-Class	**5**	high	no	high	low	high	high
Older Suburban Middle-Class	**6**	low	yes	low	medium	medium	low
Younger Suburban Middle-Class	**7**	high	yes / no (mixed)	high	medium	high	high
Urban Lower-Class	**8**	medium	yes	medium	high	medium	low
Older Suburban Lower-Class	**9**	high	yes	high	medium	high	low
Younger Suburban Lower-Class	**10**	medium	yes	medium	high	high	low
		low / medium / high	yes/no	low / medium / high	low / medium / high	low / medium / high	low / medium / high

Figure 36: Summarized Activities per Lifestyle Category

Based on this table, we can summarize our findings on how the different lifestyle groups behave as follows:

- We find higher variety among the daily activity sequences of diarists within older lifestyle categories.
- Typical work day structures are almost exclusively found among younger diarists.
- The number of daily activities, e.g. the amount of state transitions during a day, is higher for diarists with higher ages.
- Work travel is found most often in lifestyle categories with lower income as well as urbanicity.
- Maintenance travel is found more commonly among lifestyle categories with lower incomes.
- We observe that recreational travel is found mostly in lifestyle categories with relatively high wealth.

8 Discussion and Conclusion

8.1 Discussion

We have found a great variation of behavior patterns among diarists within our dataset, from which we were able to work out significant differences between our predetermined lifestyle categories. Sequence analysis on MTUS survey data showed correlations between various factors that went into our lifestyle categorization and the diarists' daily activity schedules.

However, while these results are applicable to diarists within our dataset, we cannot claim that they hold for any given population. As various factors outside the scope of our definition of a lifestyle also determine a person's behavior, the results might vary significantly for diarists in other cultures, countries and societal structures. We also have no method of verifying whether the diarists within MTUS survey data accurately represent a random sample of a country's population; thus, any results based on MTUS data might be skewed due to the fact that certain types of people might be more likely to participate in a statistical survey.

8.2 Conclusion

As described in chapter 7.6, we observe very different correlations between lifestyle categories and each different kind of travel within our dataset. We noticed that work travel is heavily dependent on the diarist's lifestyle category, while both other kinds of travel exhibit much lower deviations across the dataset in both frequency and duration. This fact makes predicting recreational travel and maintenance travel harder across a varied dataset, as we have not found any clear correlation between our chosen categorization variables and the observed travel behavior.

Regarding the hypotheses described in chapter 2.2, we come to the following conclusions:

1) While we see significant behavioral differences between lifestyle categories, the type of lifestyle alone has proven to not be accurate for completely determining a diarist's behavior. As other parameters outside the scope of our lifestyle categorization influence diarists' behavior, further refinement of the way we categorize lifestyles would be needed. Thus, we can neither fully accept nor reject hypothesis 1).
2) Our data analysis shows a significant influence of the variables that were chosen as a base for lifestyle categorization in our adapted version of Claritas PRIZM premier; as we saw both quantitative and behavioral differences between different lifestyle categories that were

established based on the variables wealth, urbanicity and age, we consider this to be a strong indicator of these variables carrying significant weight in a person's travel behavior. Thus, we consider hypothesis 2) to be confirmed.

8.3 Further Improvements and Research Topics

To conclude, this chapter describes further research possibilities as well as improvements based on the data and techniques used in this thesis. One such possibility for expanding upon the topic of sequence analysis in the MTUS HEF dataset is to analyze the correlation between the number of state transitions during a day (within a lifestyle category's typical activity sequences) and their respective travel behavior. However, in order to pursue this topic further, a finer granularity in the categorization of activities might be necessary.

An improvement worth considering would be exclusively analyzing any sub-sequences containing or leading to travel times; this could serve in determining the activities which contribute most significantly towards the generation of travel times. Thus, it might be possible to more accurately determine the leading causes for avoidable travel activities.

Lastly, a very interesting albeit complex approach towards the MTUS episode dataset would be to analyze the transitional probabilities between different activity types as well as their dependencies on other parameters (such as the parameters we used to categorize lifestyles). Based on this, one could then attempt to generate a number of behavioral models through the use of clustering techniques on these transitional probabilities.

9 Bibliography

Behrens, R., & Mistro, R. D. (2010). Shocking habits: Methodological issues in analyzing changing personal travel behavior over time. *International Journal of Sustainable Transportation*, 4(5), 253-271.

Bieser, J. C., & Hilty, L. M. (2018). An approach to assess indirect environmental effects of digitalization based on a time-use perspective. *Advances and New Trends in Environmental Informatics* (pp. 67-78). Springer, Cham.

Brown, M. S. (2014). Data Mining For Dummies. New Jersey: John Wiley & Sons, Inc.

Brzinsky-Fay, C., Kohler, U., & Luniak, M. (2006). Sequence analysis with Stata. *The Stata Journal*, *6*(4), 435-460.

Claritas Inc. *Claritas PRIZM Premier Segment Narratives 2016* (2016). Retrieved May 14, 2019 from http://pages.srds.com/rs/259-INB-778/images/NielsenPRIZMPremierSegmentNarratives2015.pdf

Frias-Martinez, V., Soguero, C., & Frias-Martinez, E. (2012). Estimation of urban commuting patterns using cellphone network data. *Proceedings of the ACM SIGKDD international workshop on urban computing* (pp. 9-16). ACM.

Froemelt, A., Dürrenmatt, D. J., & Hellweg, S. (2018). Using Data Mining To Assess Environmental Impacts of Household Consumption Behaviors. *Environmental science & technology*, *52*(15), 8467-8478.

Gabadinho, A., G. Ritschard, M. Studer, & N. S. Müller (2010). Mining sequence data in R with the TraMineR package: A user's guide. University of Geneva.

Gangrade, S. A. C. H. I. N., Pendyala, R. M., & McCullough, R. G. (2002). A nested logit model of commuters' activity schedules. *Journal of Transportation and Statistics*, 5(2/3), 19-36.

Geopath. *Market Segmentation*. Retrieved May 14, 2019 from https://support.geopath.io/hc/en-us/sections/360000917931-Market-Segmentation

Lin, H-Z., Lo, H-P., & Chen, X-J. (2009). Lifestyle classifications with and without activity-travel patterns, *Transportation Research Part A: Policy and Practice*, 43(6), 626-638.

MTUS User Guide. Retrieved May 14, 2019 from https://www.timeuse.org/MTUS-User-Guide

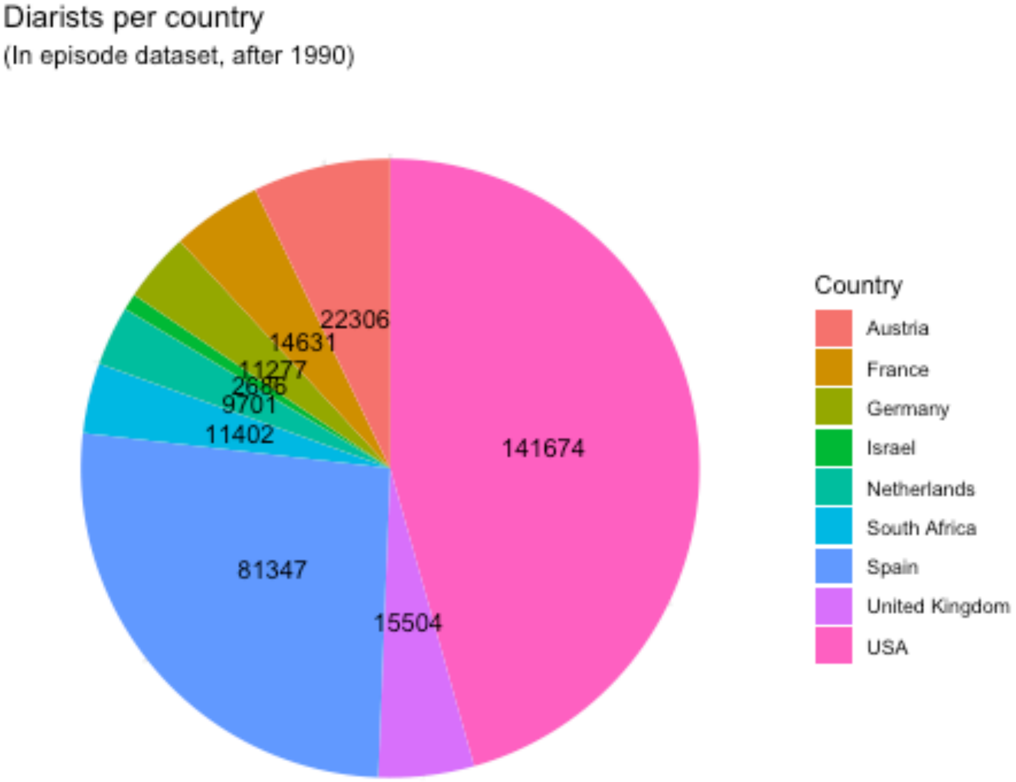
Nilsson, M., & Küller, R. (2000). Travel behaviour and environmental concern. *Transportation Research Part D: Transport and Environment*, 5(3), 211-234.

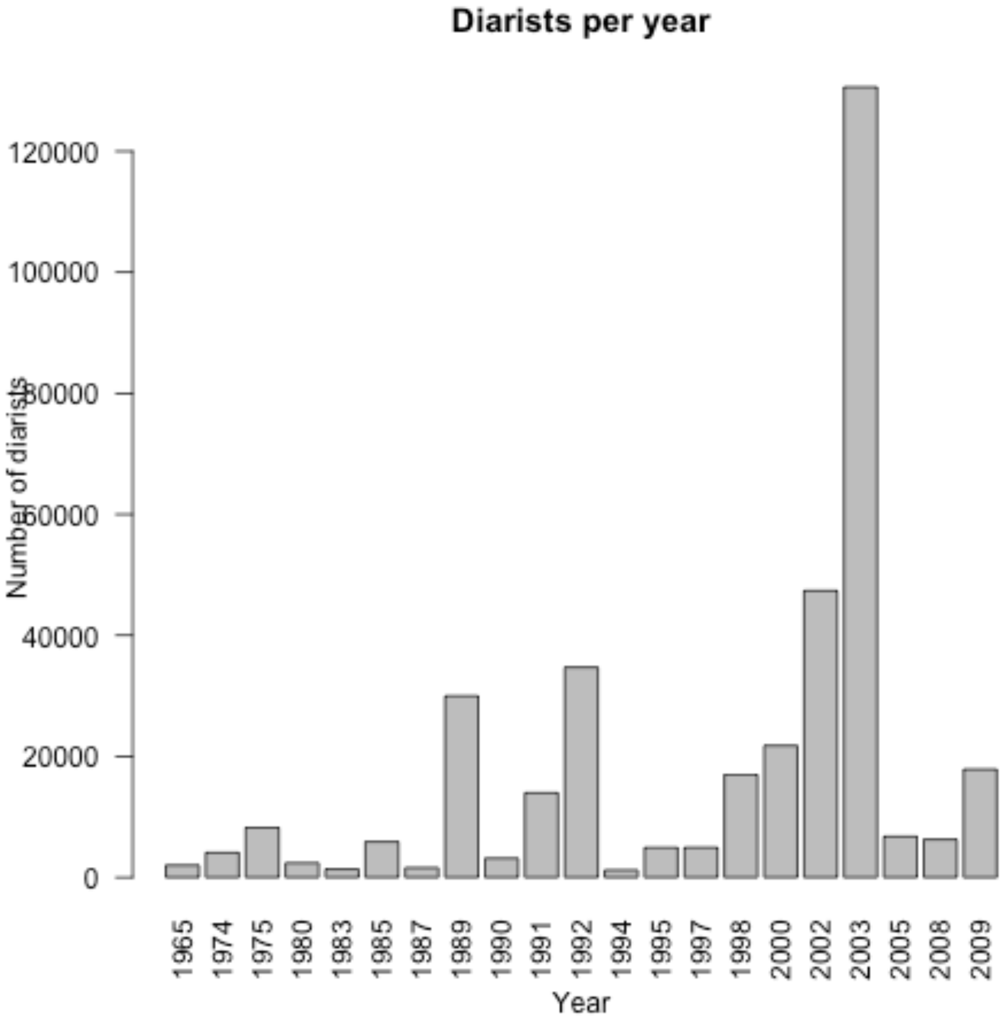
Susilo, Y. O., & Dijst, M. (2010). Behavioural decisions of travel-time ratios for work, maintenance and leisure activities in the Netherlands. *Transportation planning and technology*, *33*(1), 19-34.

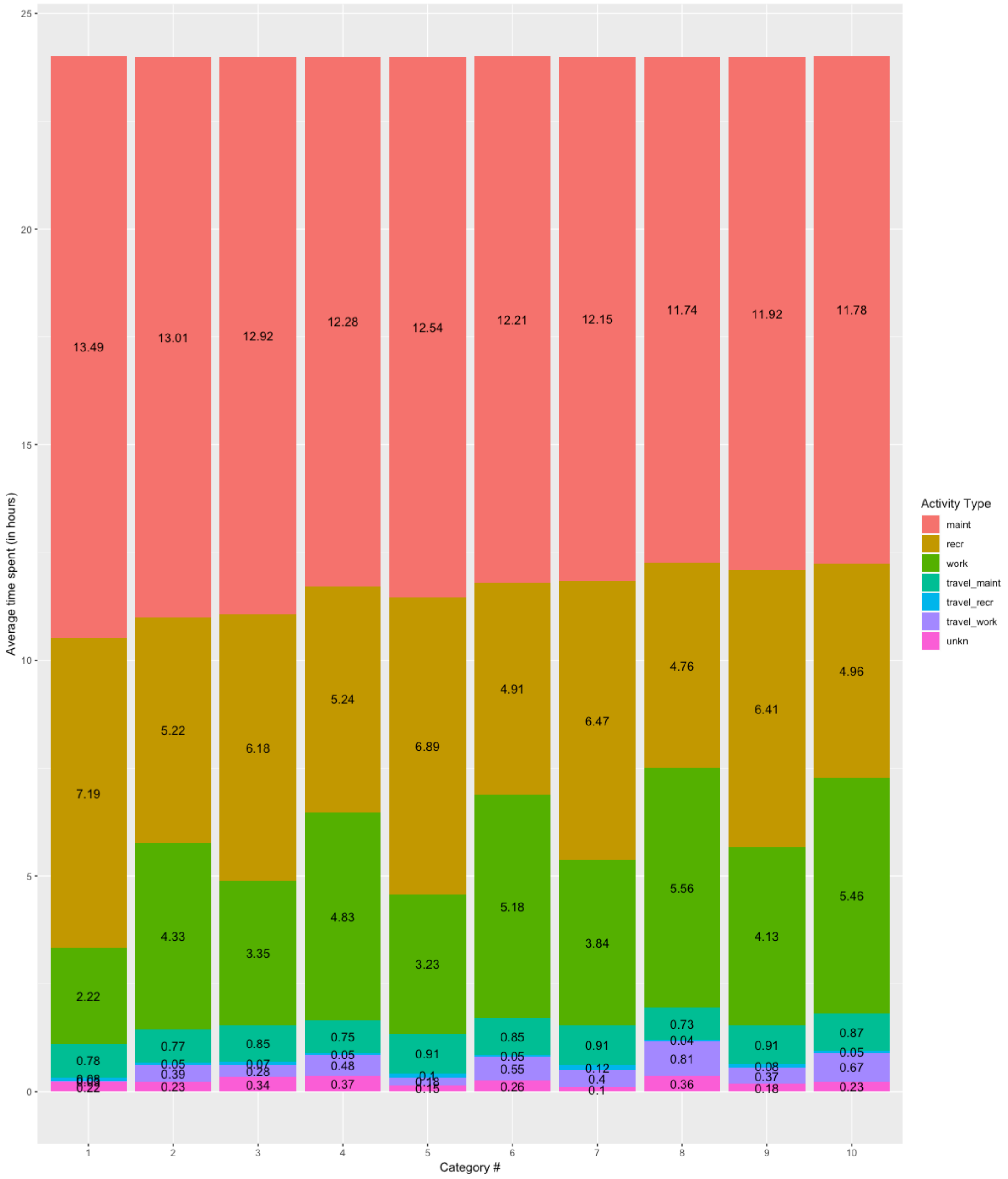
TraMineR User Guide. Retrieved May 14, 2019 from http://mephisto.unige.ch/pub/TraMineR/doc/TraMineR-Users-Guide.pdf

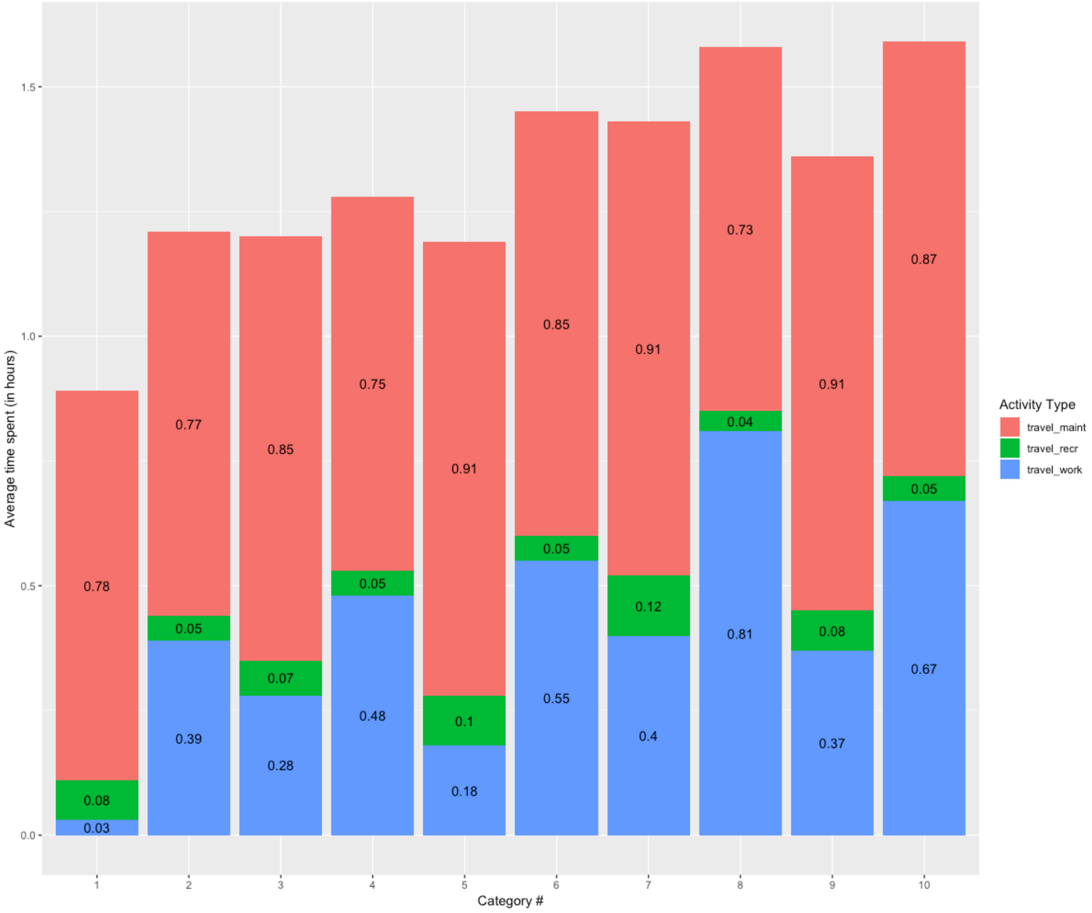
Zhao, Y. (2015). R and Data Mining: Examples and Case Studies. Elsevier.

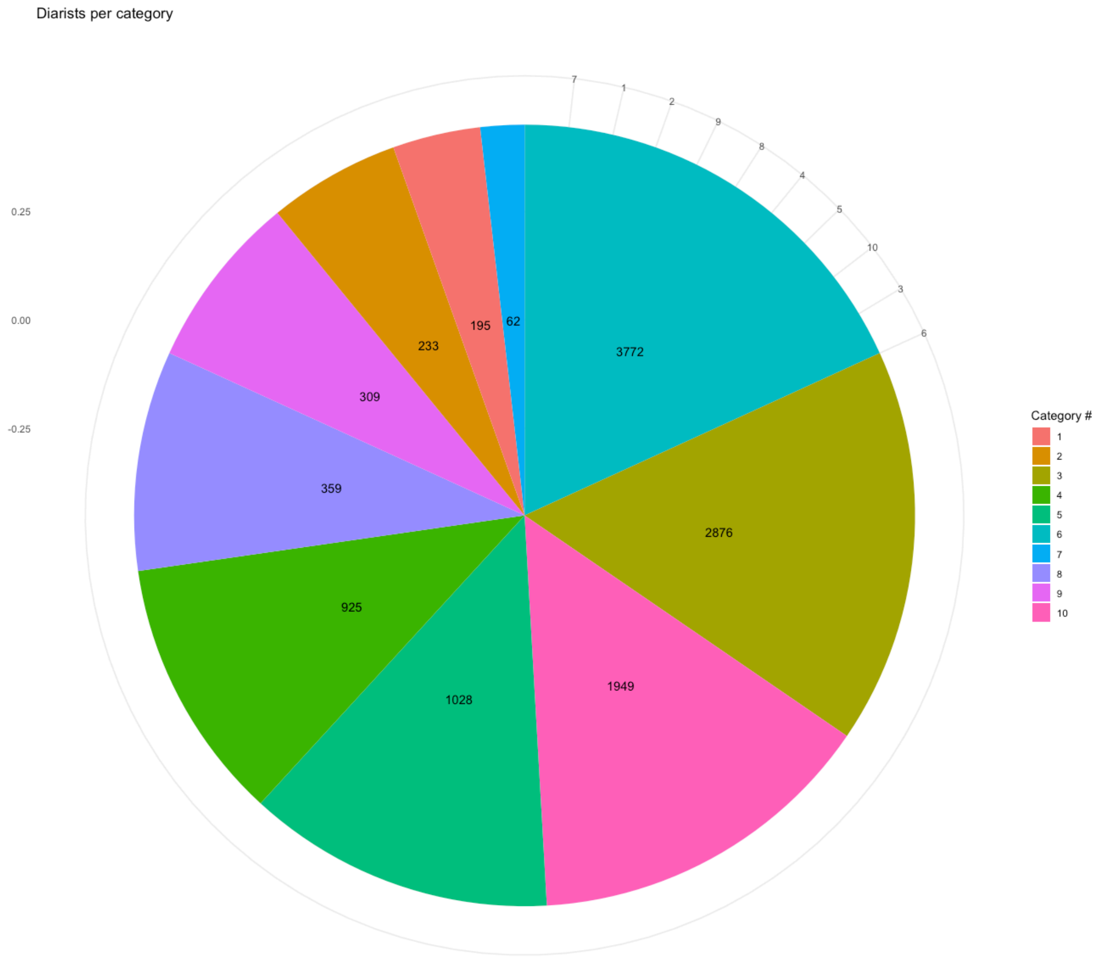
10 Appendix

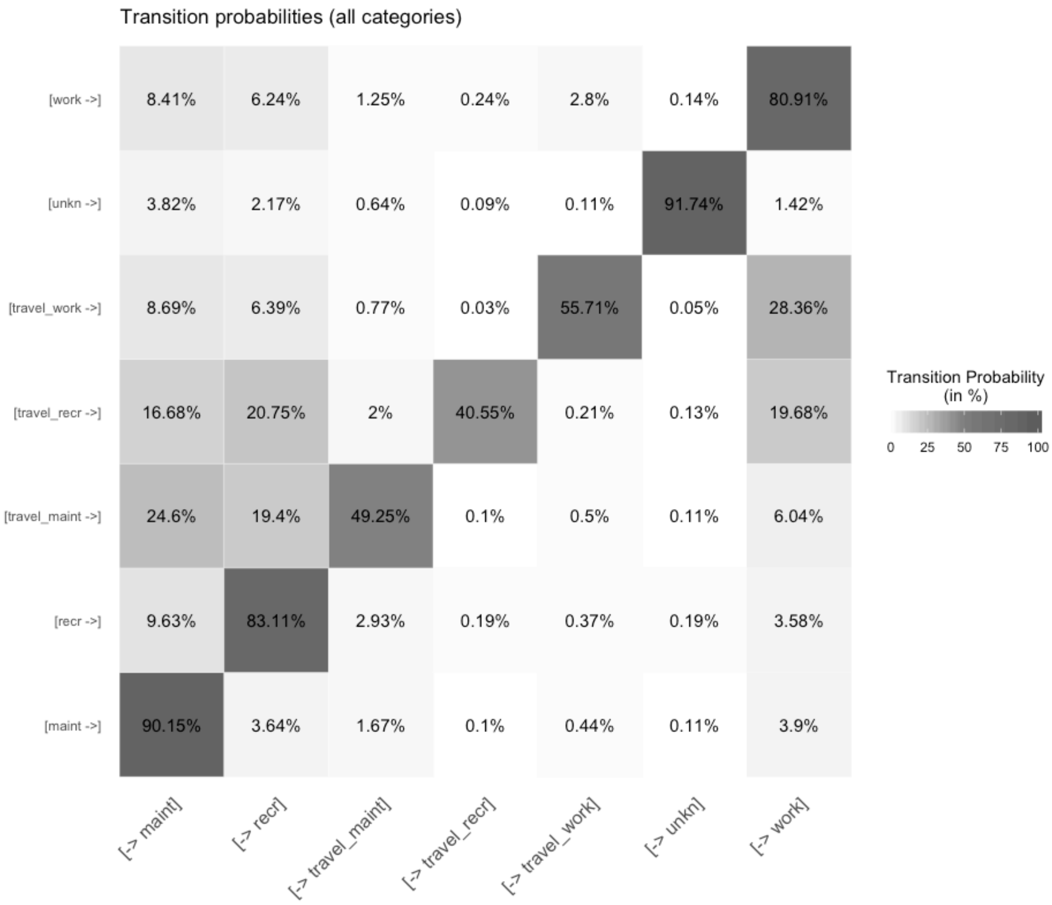

Diarists per country
(In episode dataset, after 1990)


Diarists per year
Number of diarists
0
20000
40000
60000
80000
100000
120000
1965
1974
1975
1980
1983
1985
1987
1989
1990
1991
1992
1994
1995
1997
1998
2000
2002
2003
2005
2008
2009
Year


Average time spent (in hours)
25
20
15
10
5
0
13.49
7.19
2.22
0.78
0.08
0.22
13.01
5.22
4.33
0.77
0.05
0.39
0.23
12.92
6.18
3.35
0.85
0.07
0.28
0.34
12.28
5.24
4.83
0.75
0.05
0.48
0.37
12.54
6.89
3.23
0.91
0.1
0.18
0.15
12.21
4.91
5.18
0.85
0.05
0.55
0.26
12.15
6.47
3.84
0.91
0.12
0.4
0.1
11.74
4.76
5.56
0.73
0.04
0.81
0.36
11.92
6.41
4.13
0.91
0.08
0.37
0.18
11.78
4.96
5.46
0.87
0.05
0.67
0.23
1
2
3
4
5
6
7
8
9
10
Category #
Activity Type
maint
recr
work
travel_maint
travel_recr
travel_work
unkn

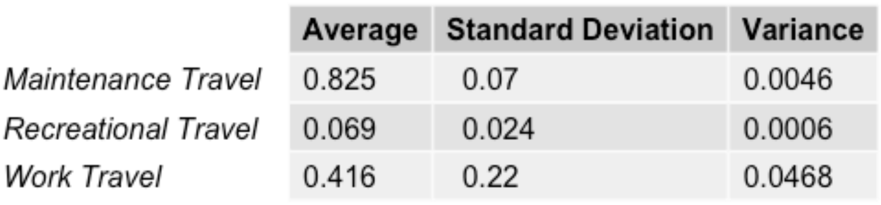

Average time spent (in hours)
1.5
1.0
0.5
0.0
Category #
1
2
3
4
5
6
7
8
9
10
0.78
0.08
0.03
0.77
0.05
0.39
0.85
0.07
0.28
0.75
0.05
0.48
0.91
0.1
0.18
0.85
0.05
0.55
0.91
0.12
0.4
0.73
0.04
0.81
0.91
0.08
0.37
0.87
0.05
0.67
Activity Type
🟥 travel_maint
🟩 travel_recr
🟦 travel_work


Diarists per category
0.25
0.00
-0.25
7
1
2
9
8
4
5
10
3
6
3772
2876
1949
1028
925
359
309
233
195
62
Category #
1
2
3
4
5
6
7
8
9
10

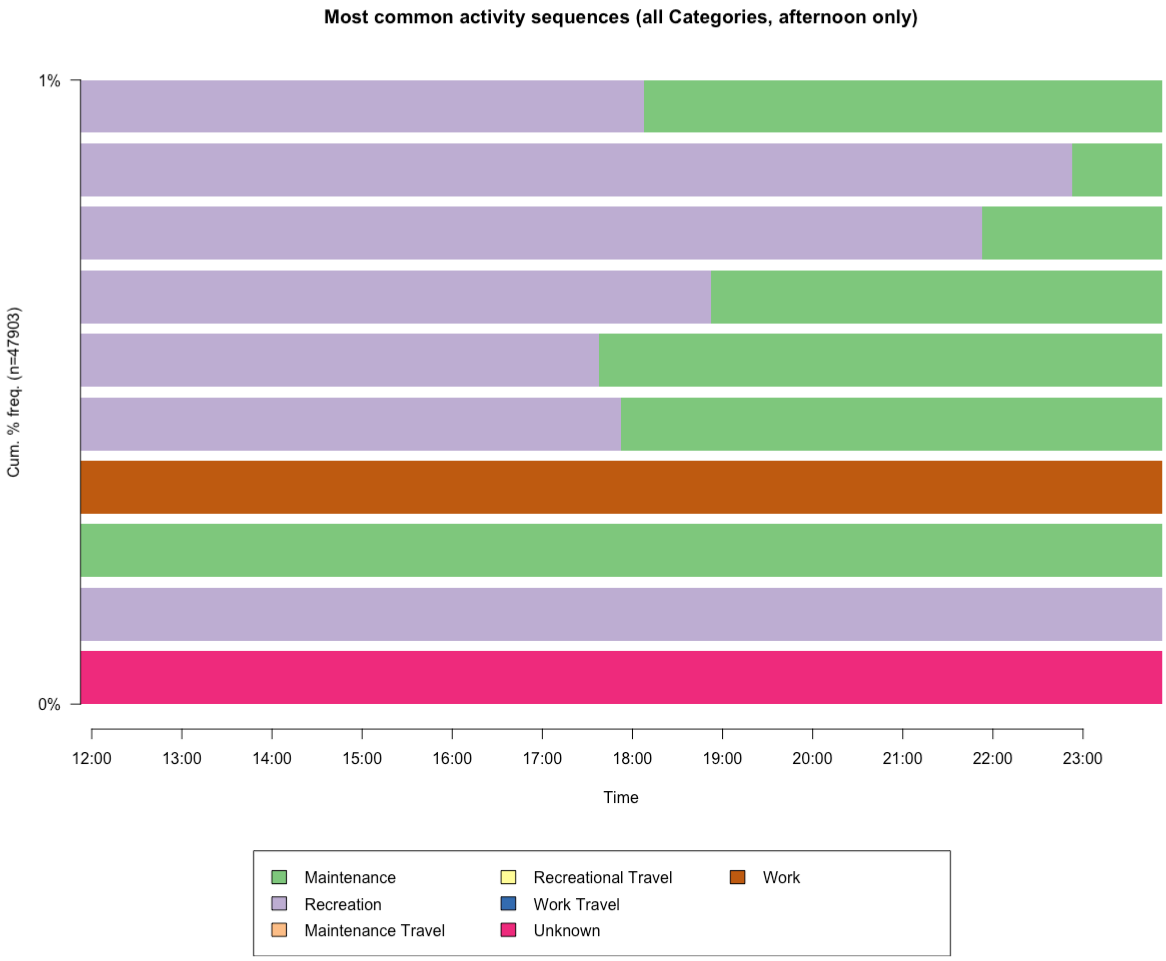


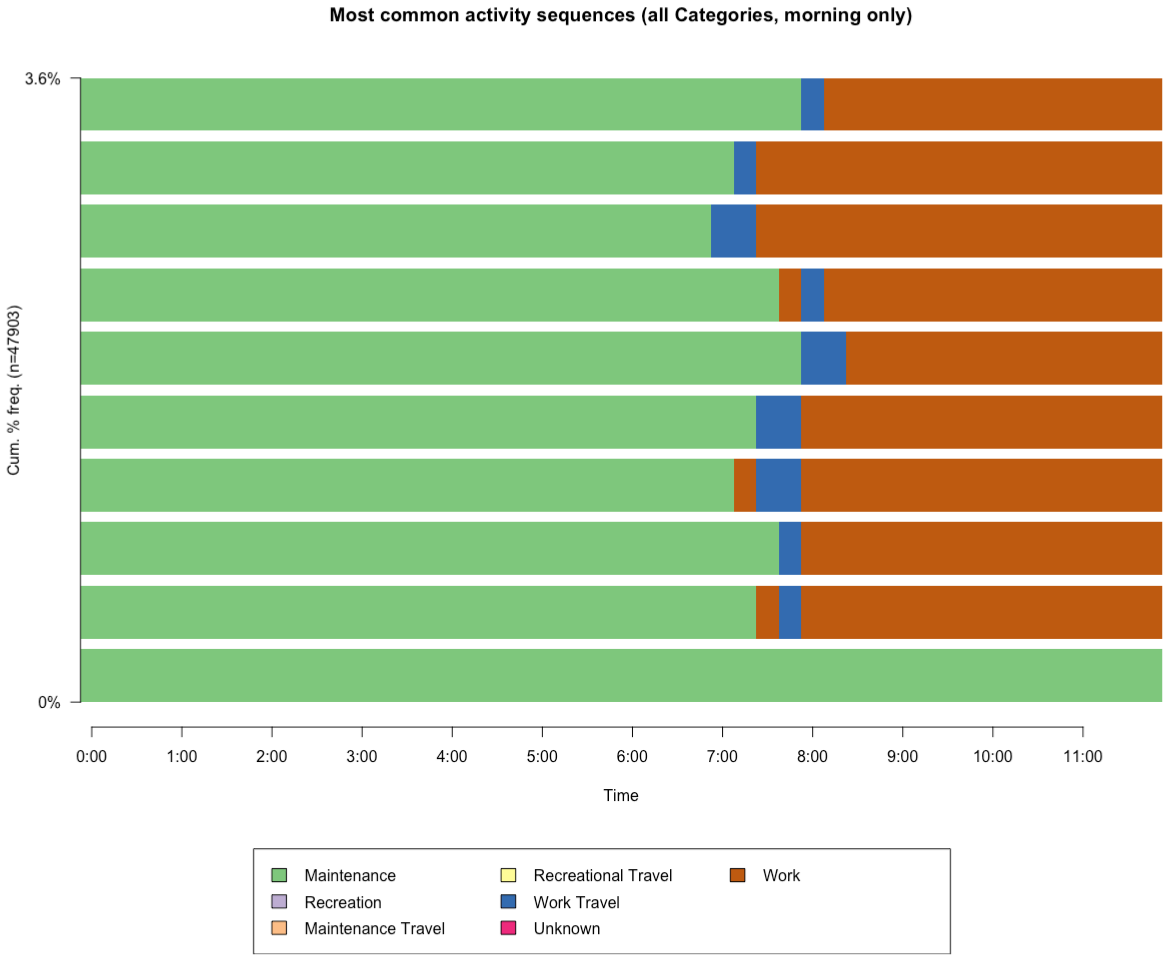
Transition probabilities (all categories)

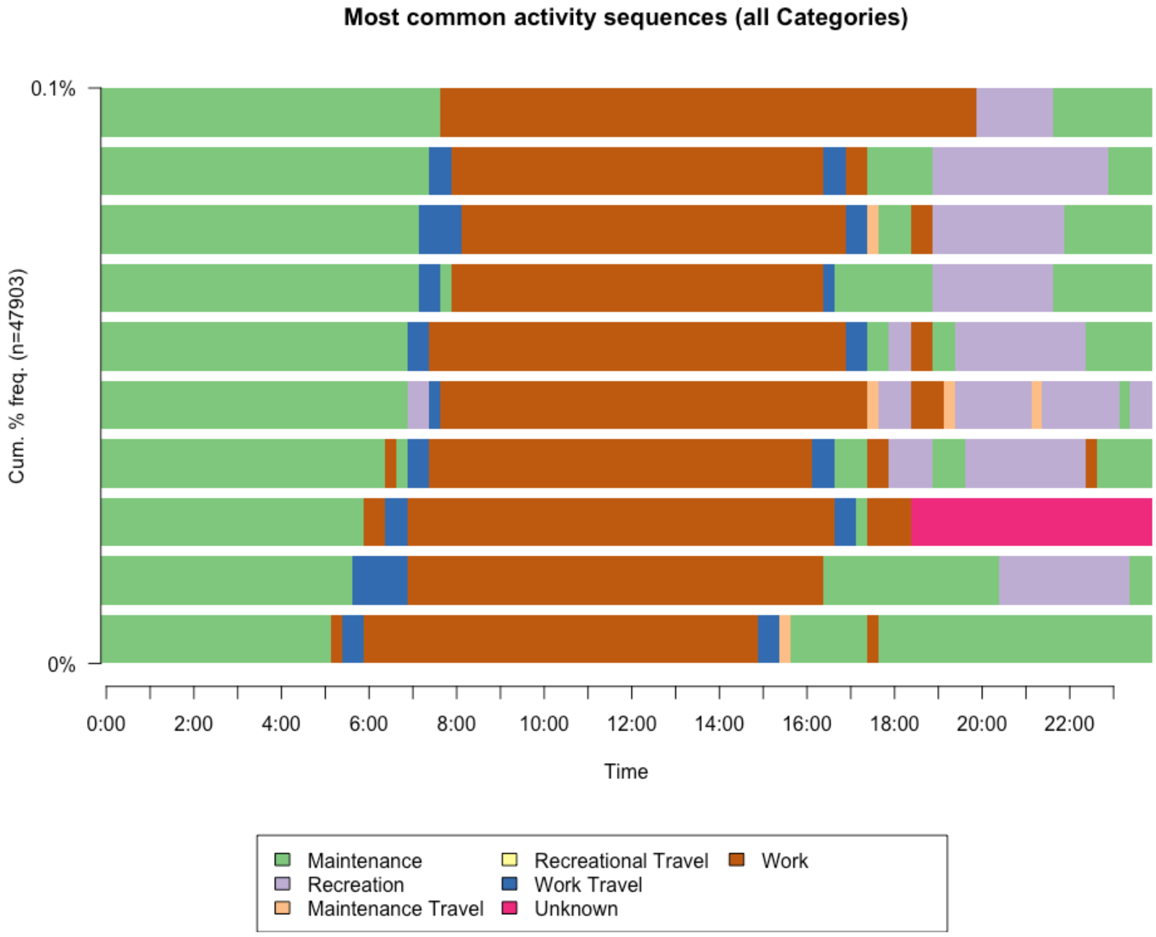
	[-> maint]	[-> recr]	[-> travel_maint]	[-> travel_recr]	[-> travel_work]	[-> unkn]	[-> work]
[work ->]	8.41%	6.24%	1.25%	0.24%	2.8%	0.14%	80.91%
[unkn ->]	3.82%	2.17%	0.64%	0.09%	0.11%	91.74%	1.42%
[travel_work ->]	8.69%	6.39%	0.77%	0.03%	55.71%	0.05%	28.36%
[travel_recr ->]	16.68%	20.75%	2%	40.55%	0.21%	0.13%	19.68%
[travel_maint ->]	24.6%	19.4%	49.25%	0.1%	0.5%	0.11%	6.04%
[recr ->]	9.63%	83.11%	2.93%	0.19%	0.37%	0.19%	3.58%
[maint ->]	90.15%	3.64%	1.67%	0.1%	0.44%	0.11%	3.9%

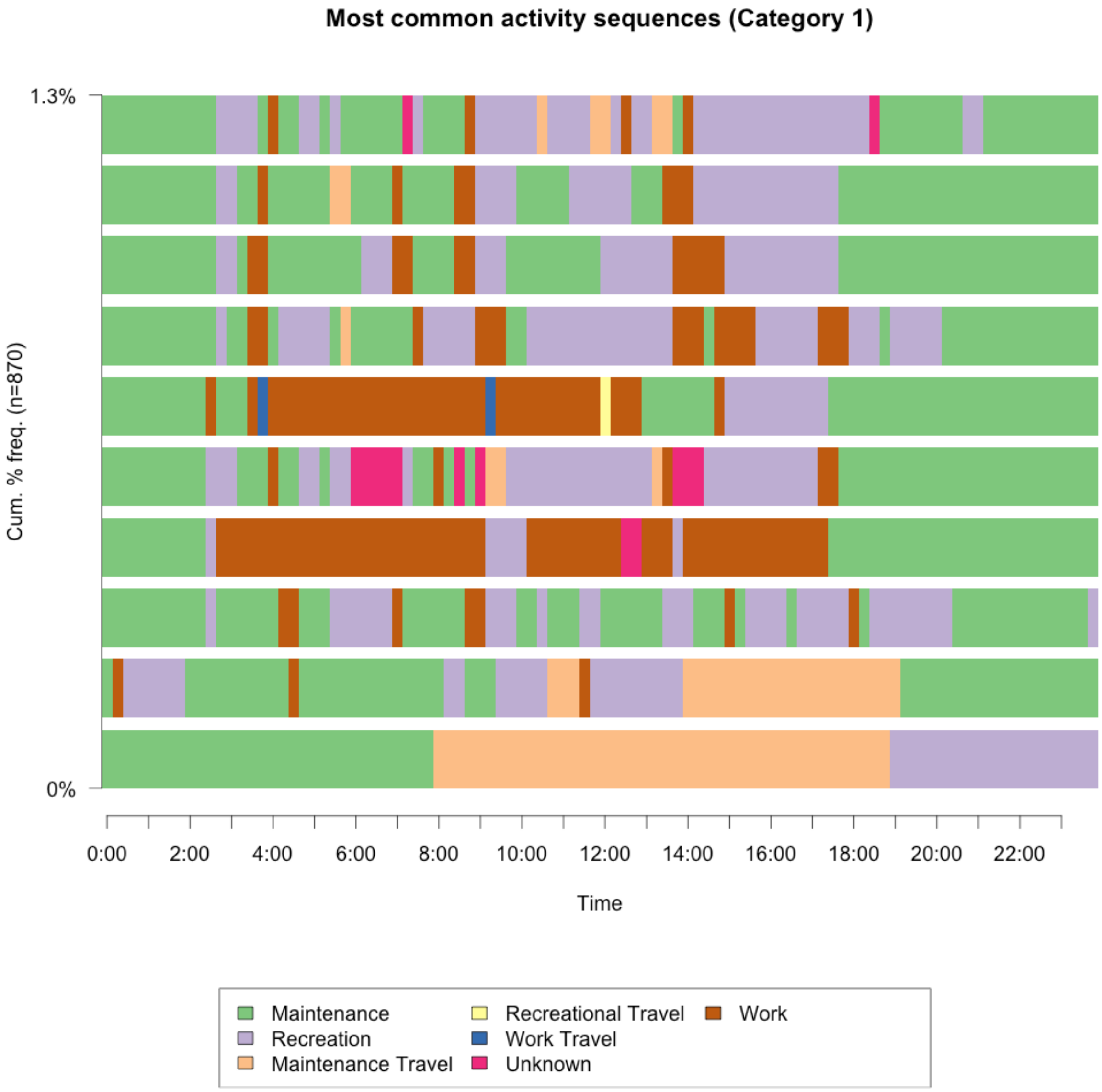


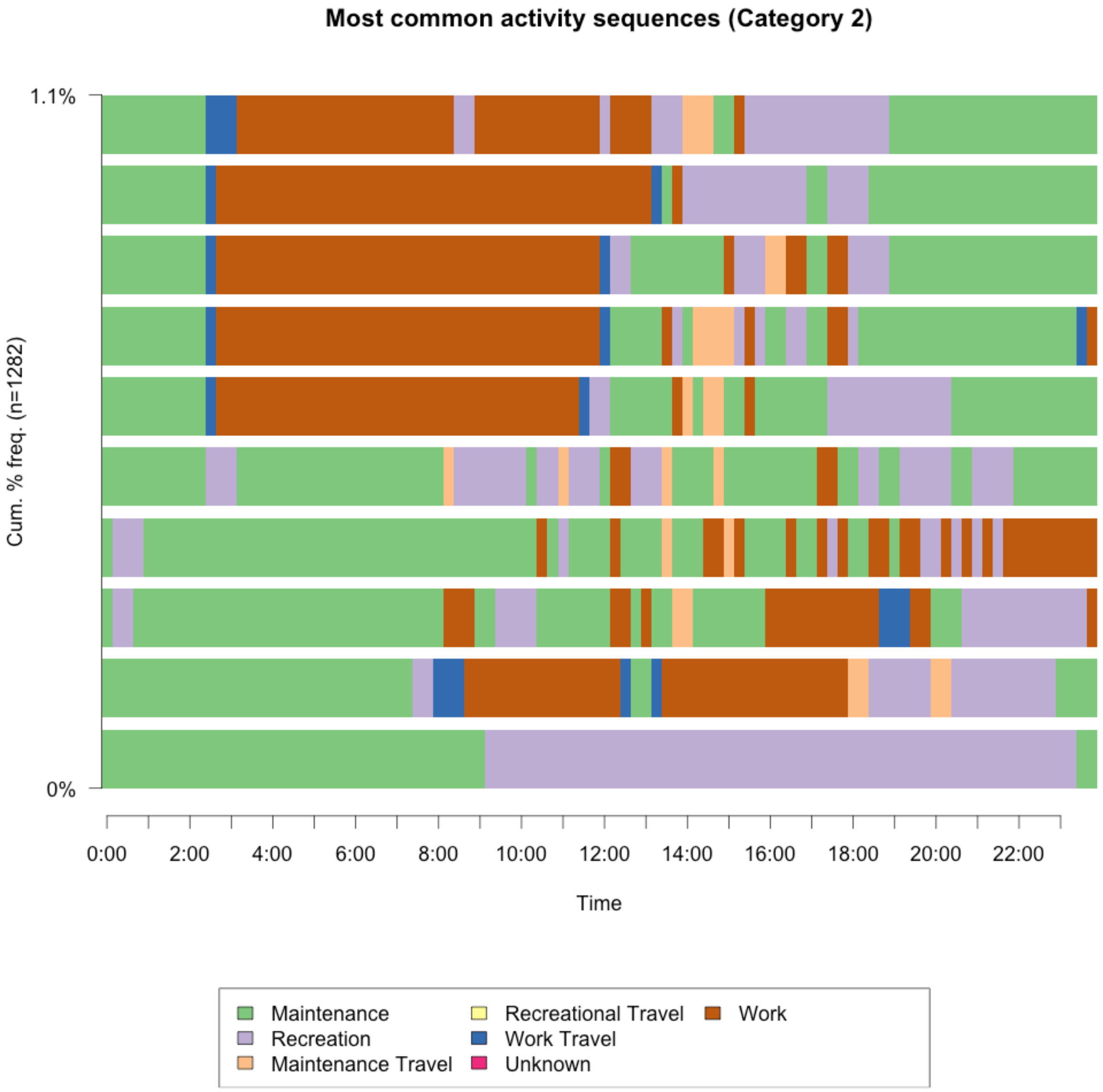
	Average	Standard Deviation	Variance
Maintenance Travel	0.825	0.07	0.0046
Recreational Travel	0.069	0.024	0.0006
Work Travel	0.416	0.22	0.0468

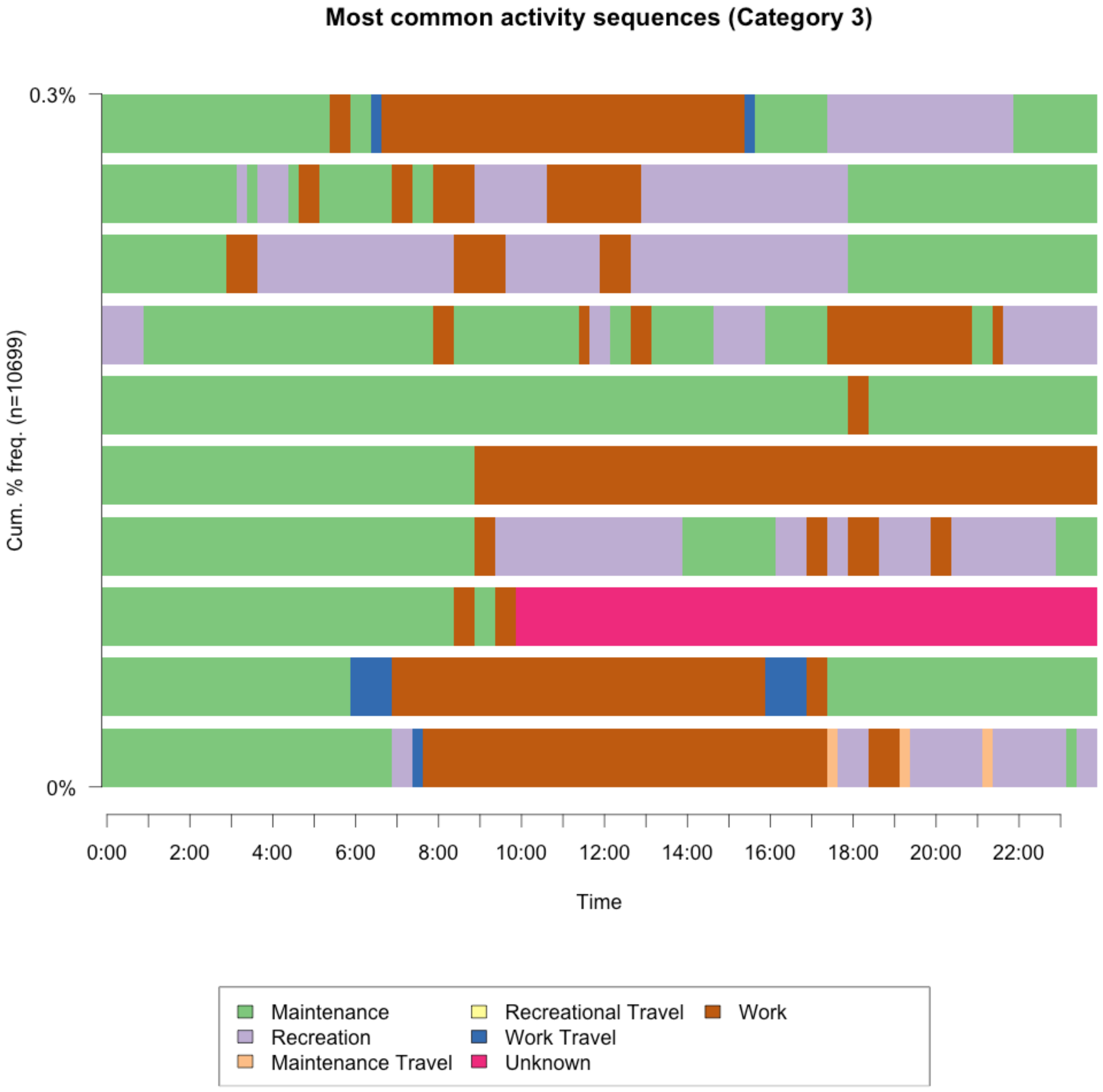

Most common activity sequences (all Categories, afternoon only)
1%
0%
Cum. % freq. (n=47903)
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Time
🟩 Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

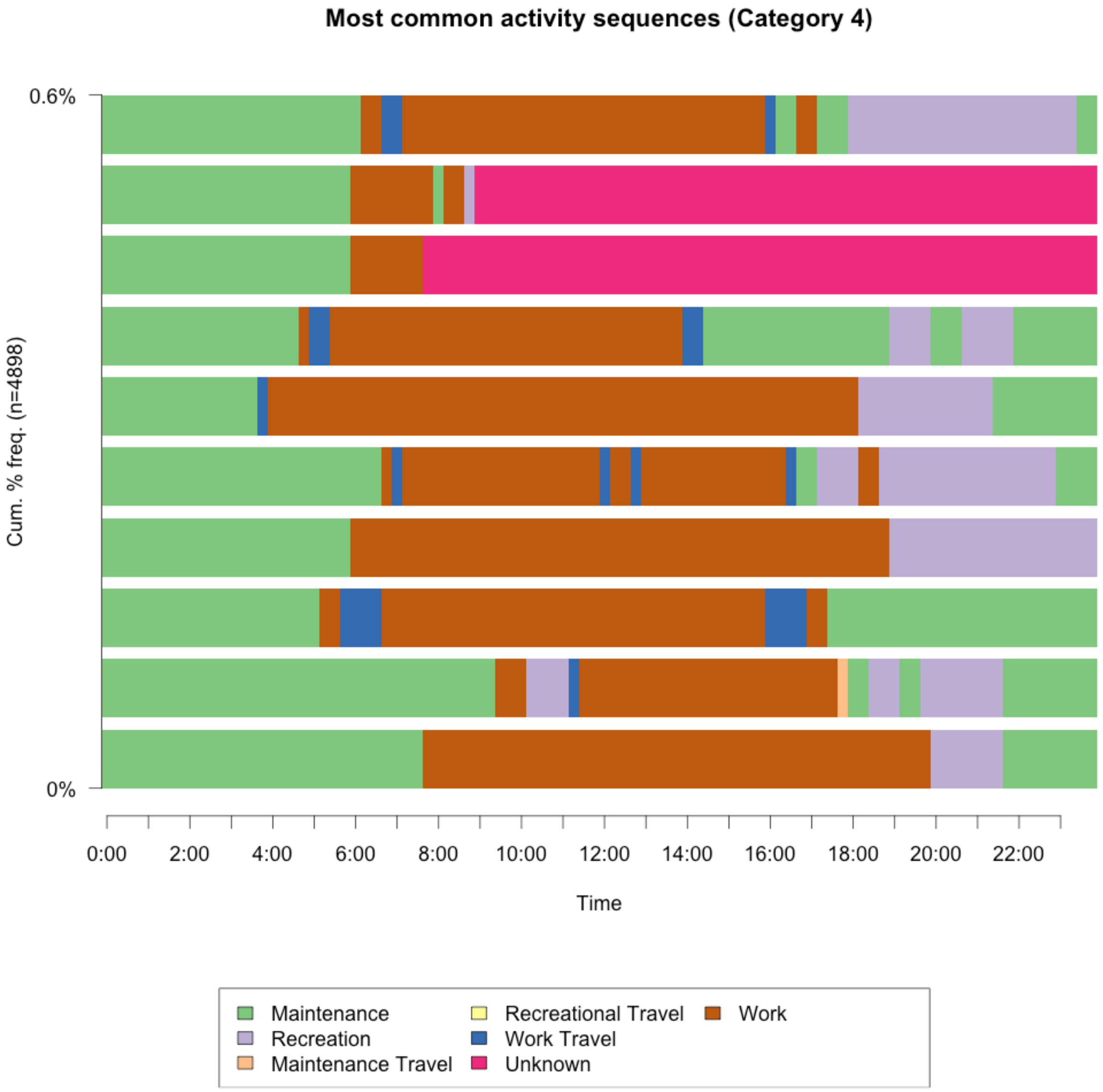

Most common activity sequences (all Categories, morning only)
3.6%
0%
Cum. % freq. (n=47903)
0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:00
Time
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

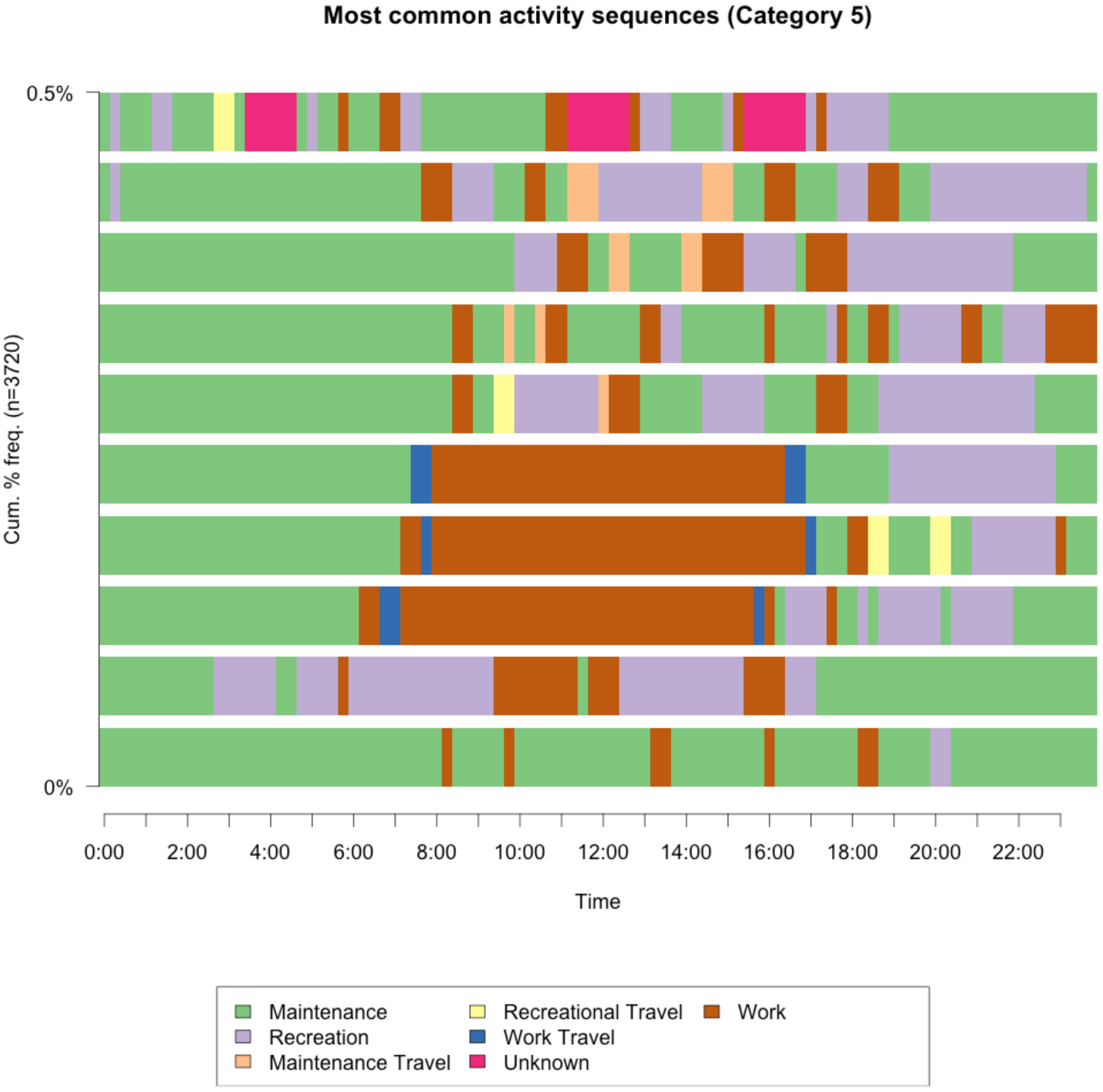

Most common activity sequences (all Categories)
0.1%
0%
Cum. % freq. (n=47903)
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Time
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

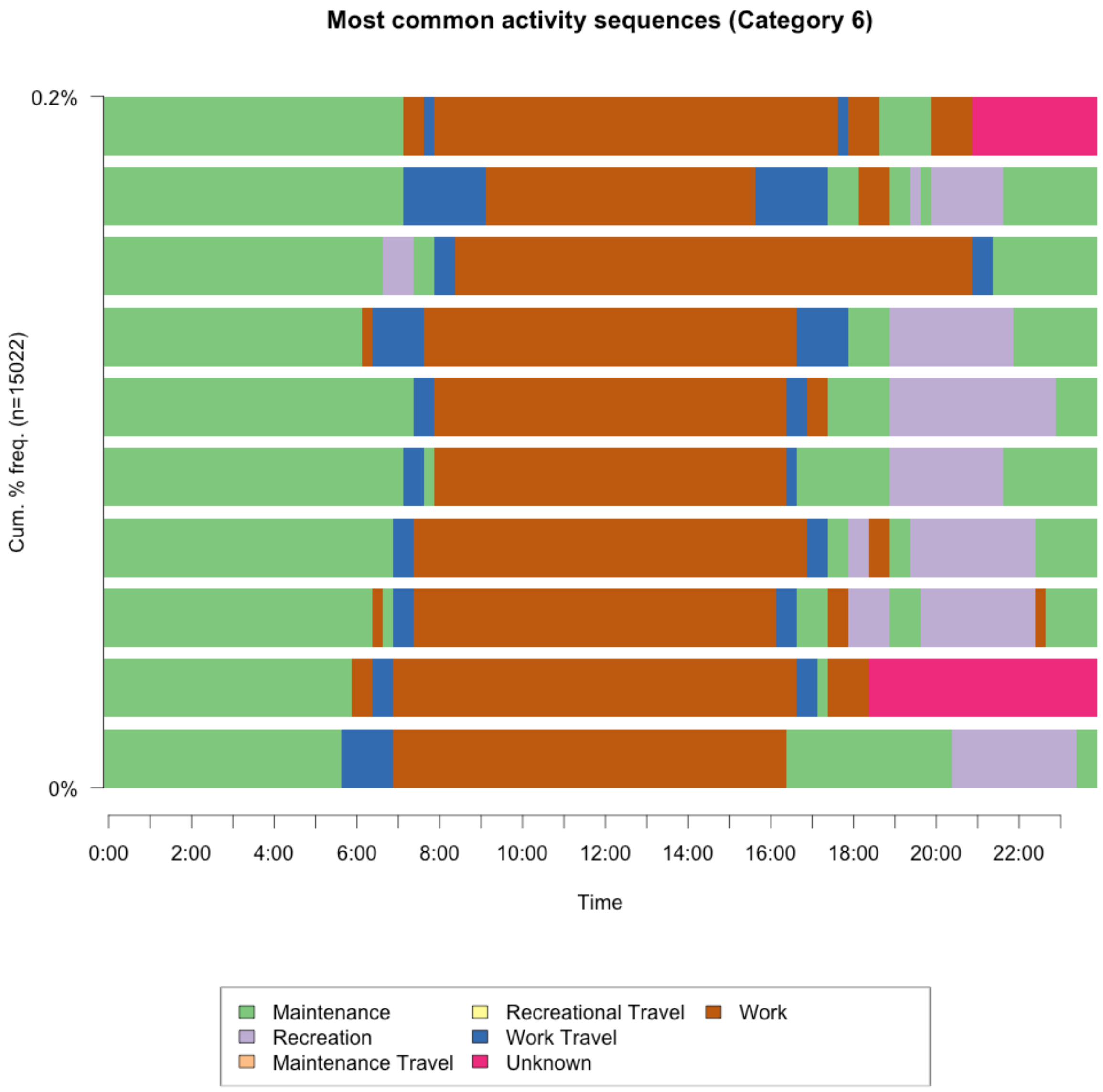

Most common activity sequences (Category 1)
1.3%
0%
Cum. % freq. (n=870)
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Time
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

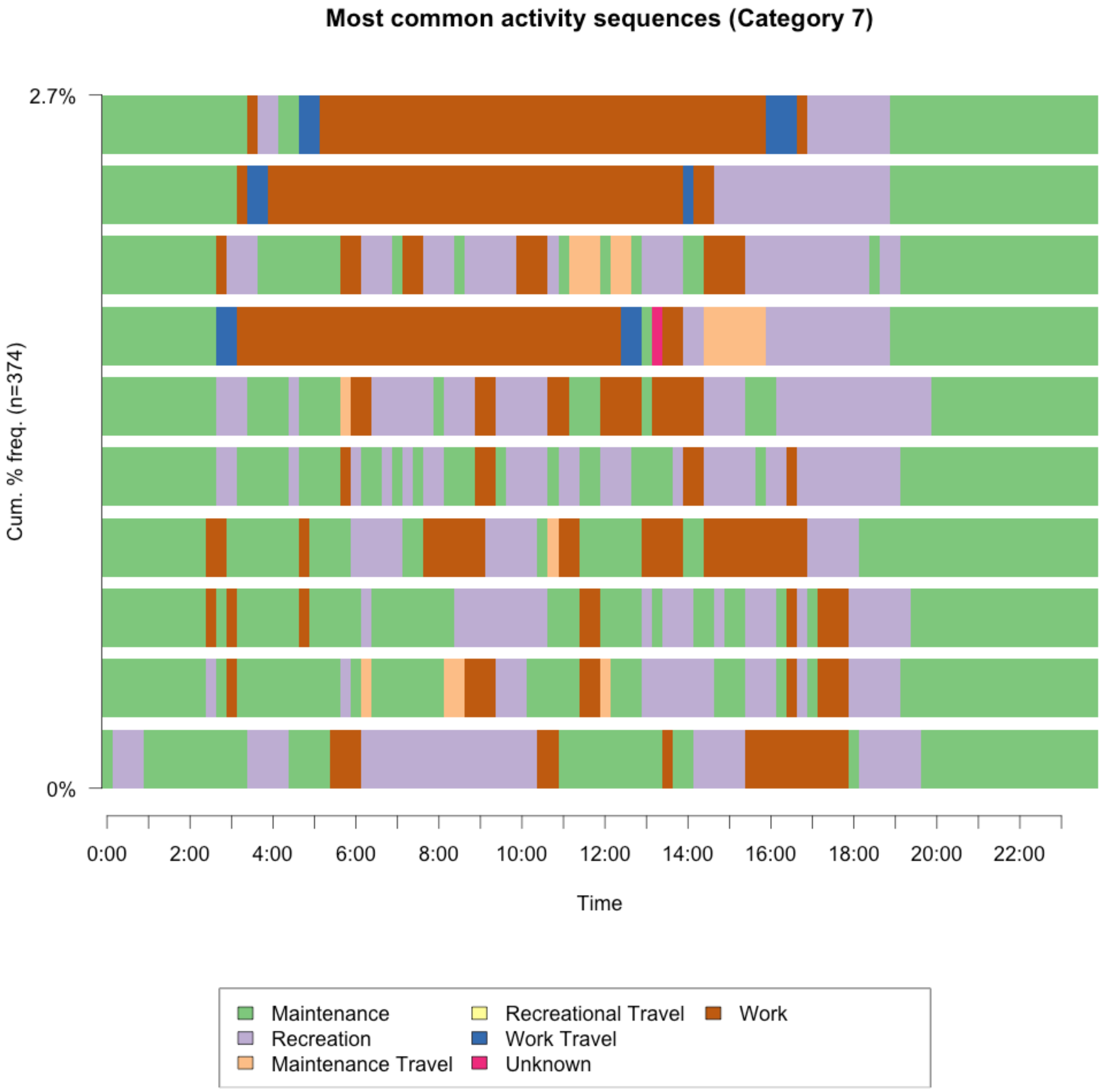

Most common activity sequences (Category 2)
Cum. % freq. (n=1282)
1.1%
0%
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Time
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

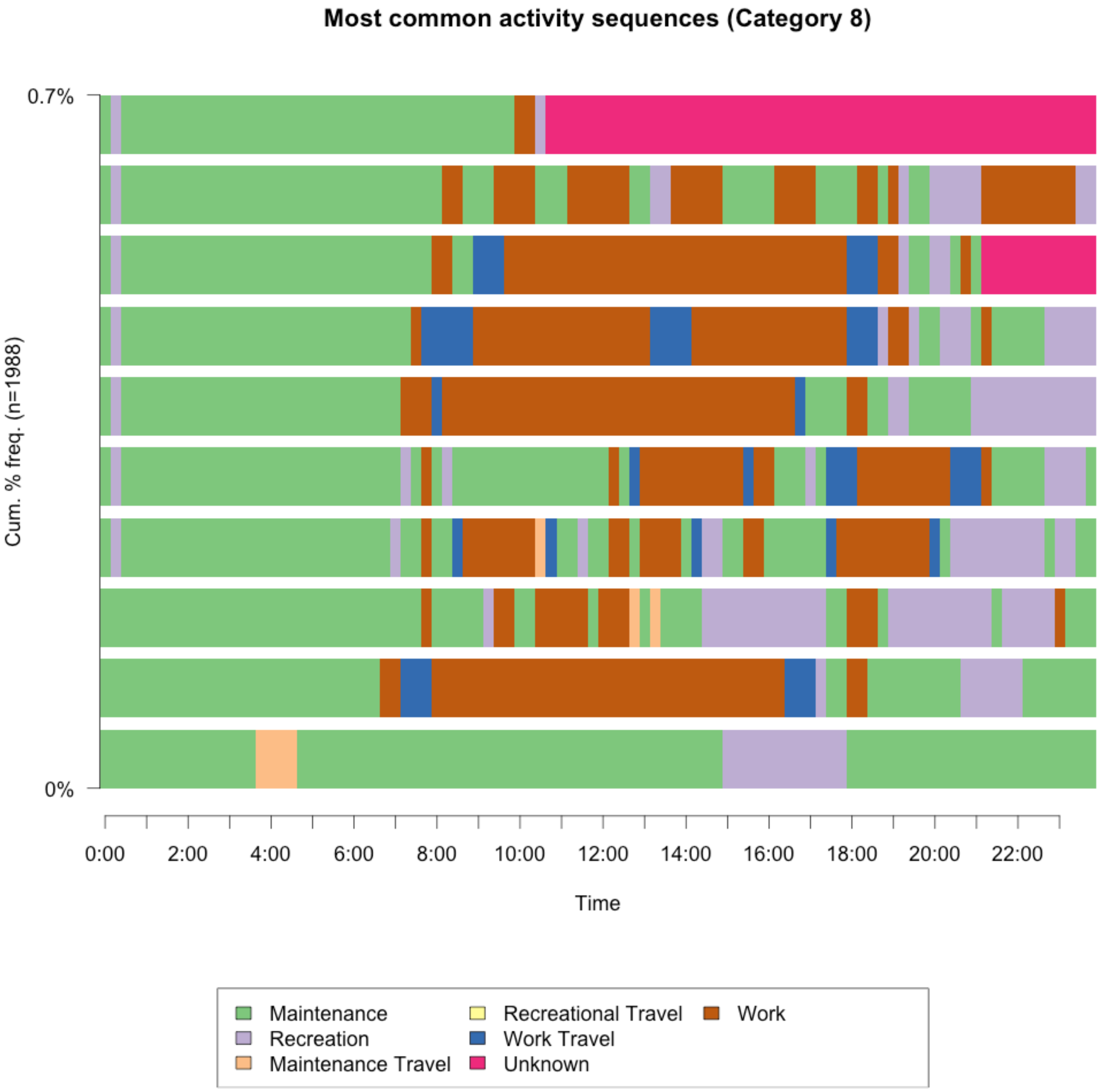

Most common activity sequences (Category 3)
Cum. % freq. (n=10699)
0.3%
0%
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Time
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

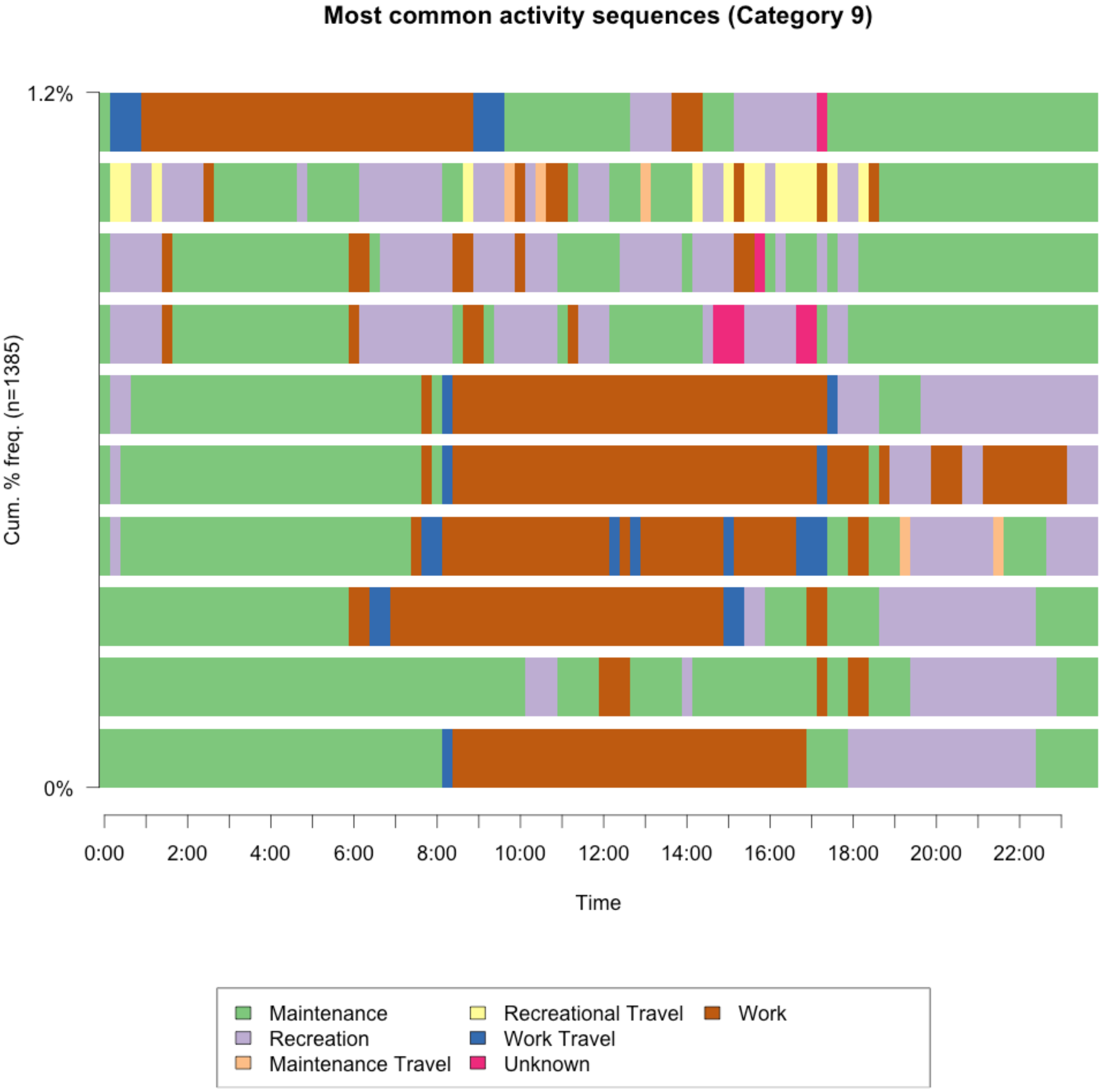

Most common activity sequences (Category 4)
0.6%
0%
Cum. % freq. (n=4898)
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Time
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

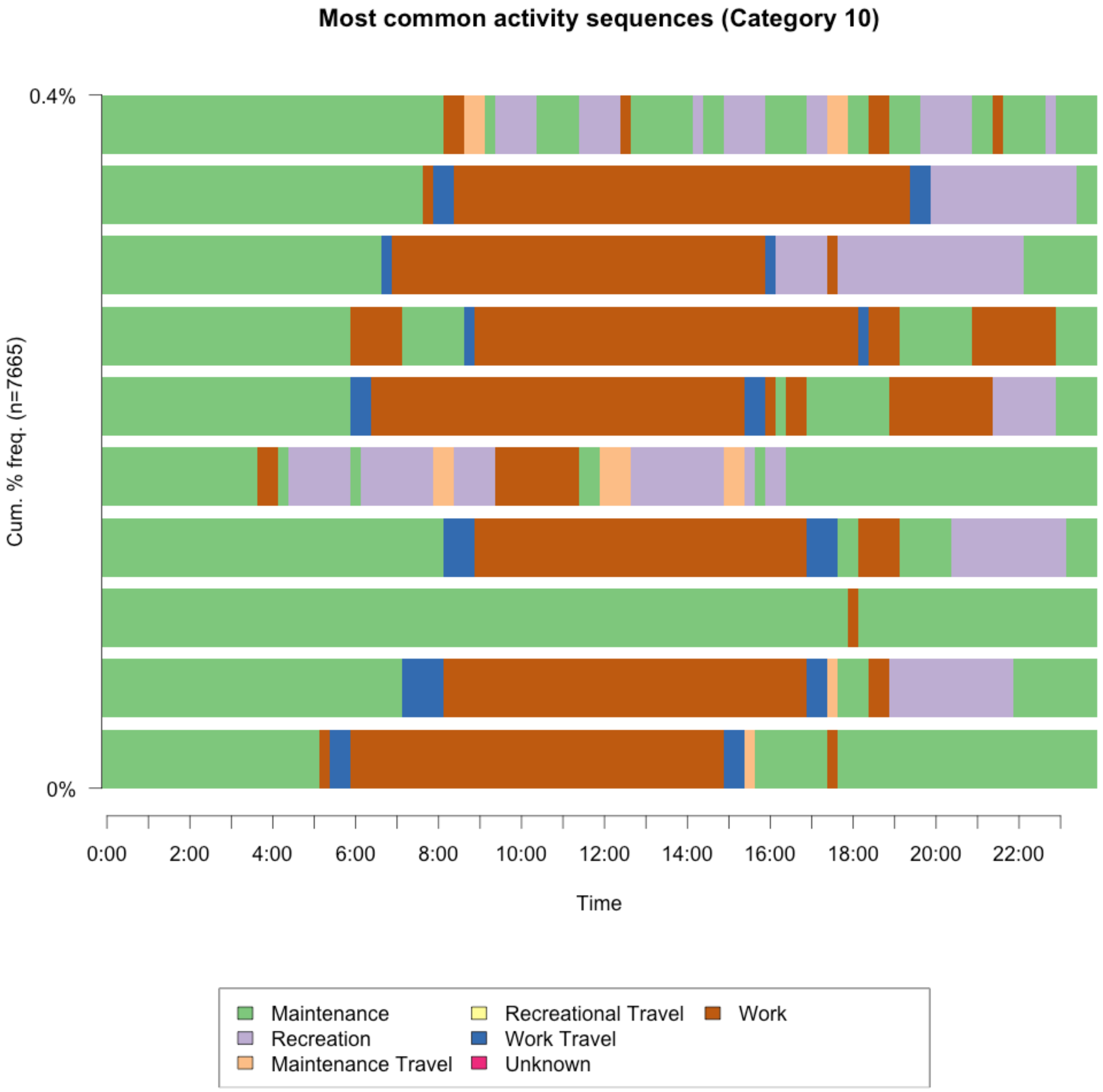

Most common activity sequences (Category 5)
0.5%
0%
Cum. % freq. (n=3720)
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Time
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

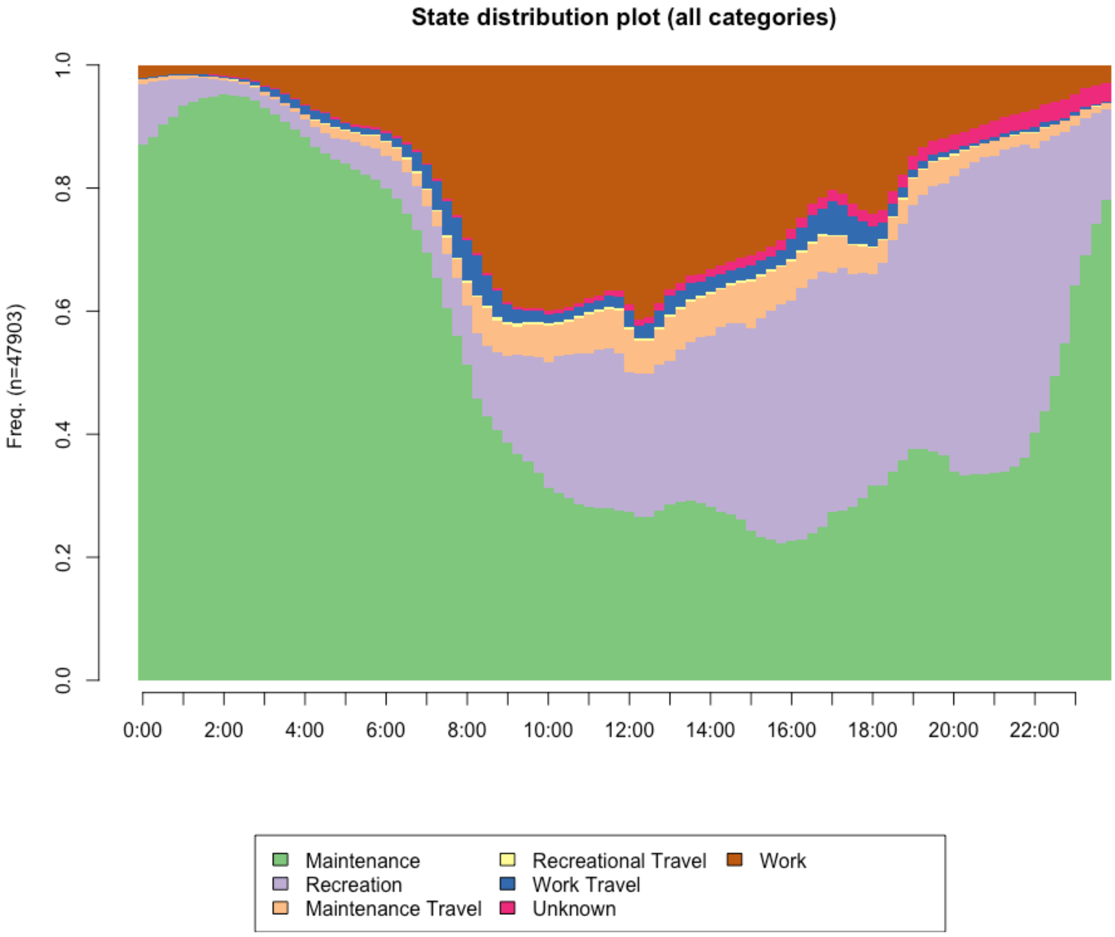

Most common activity sequences (Category 6)
Cum. % freq. (n=15022)
0.2%
0%
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Time
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

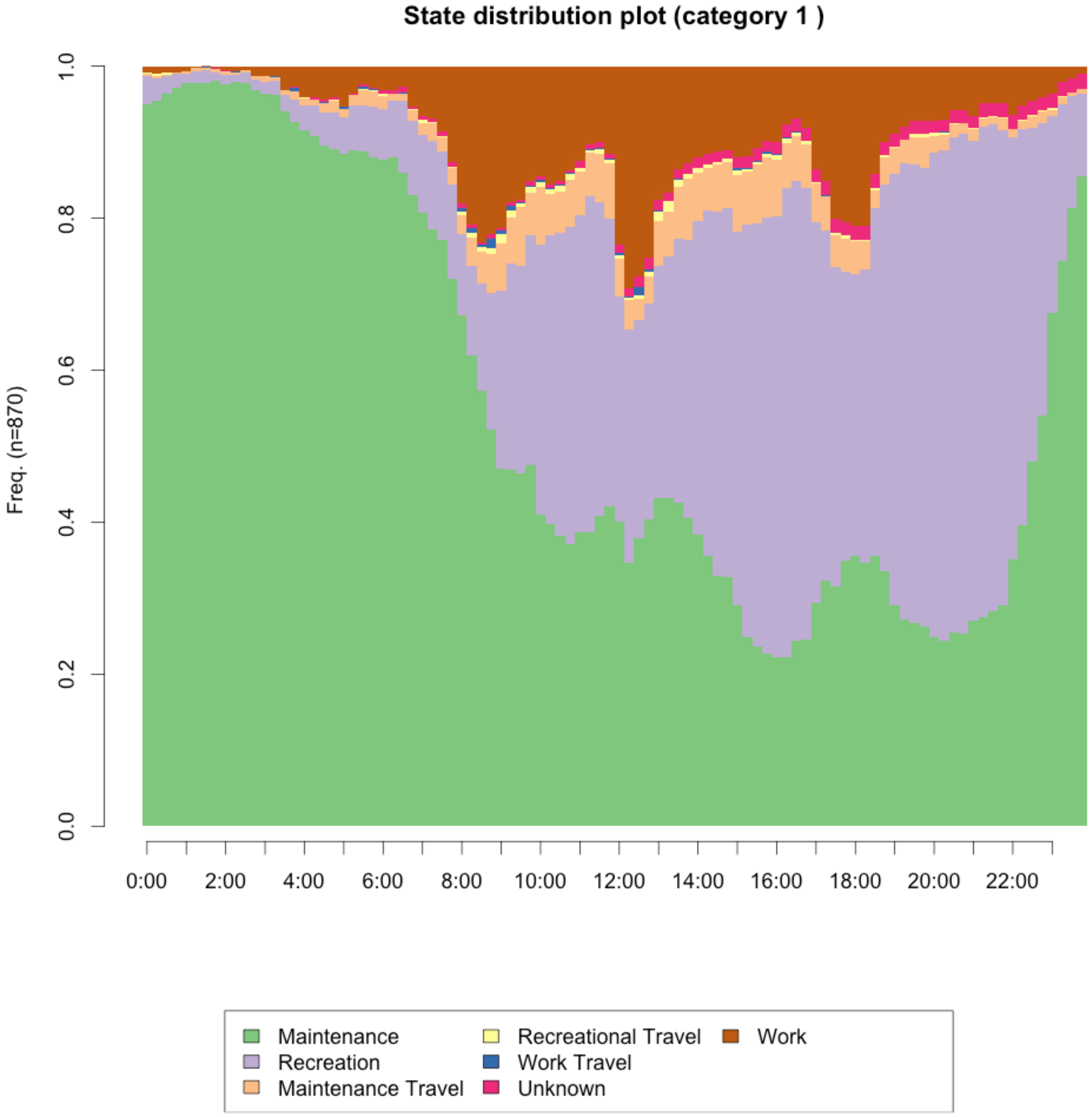

Most common activity sequences (Category 7)
2.7%
0%
Cum. % freq. (n=374)
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Time
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

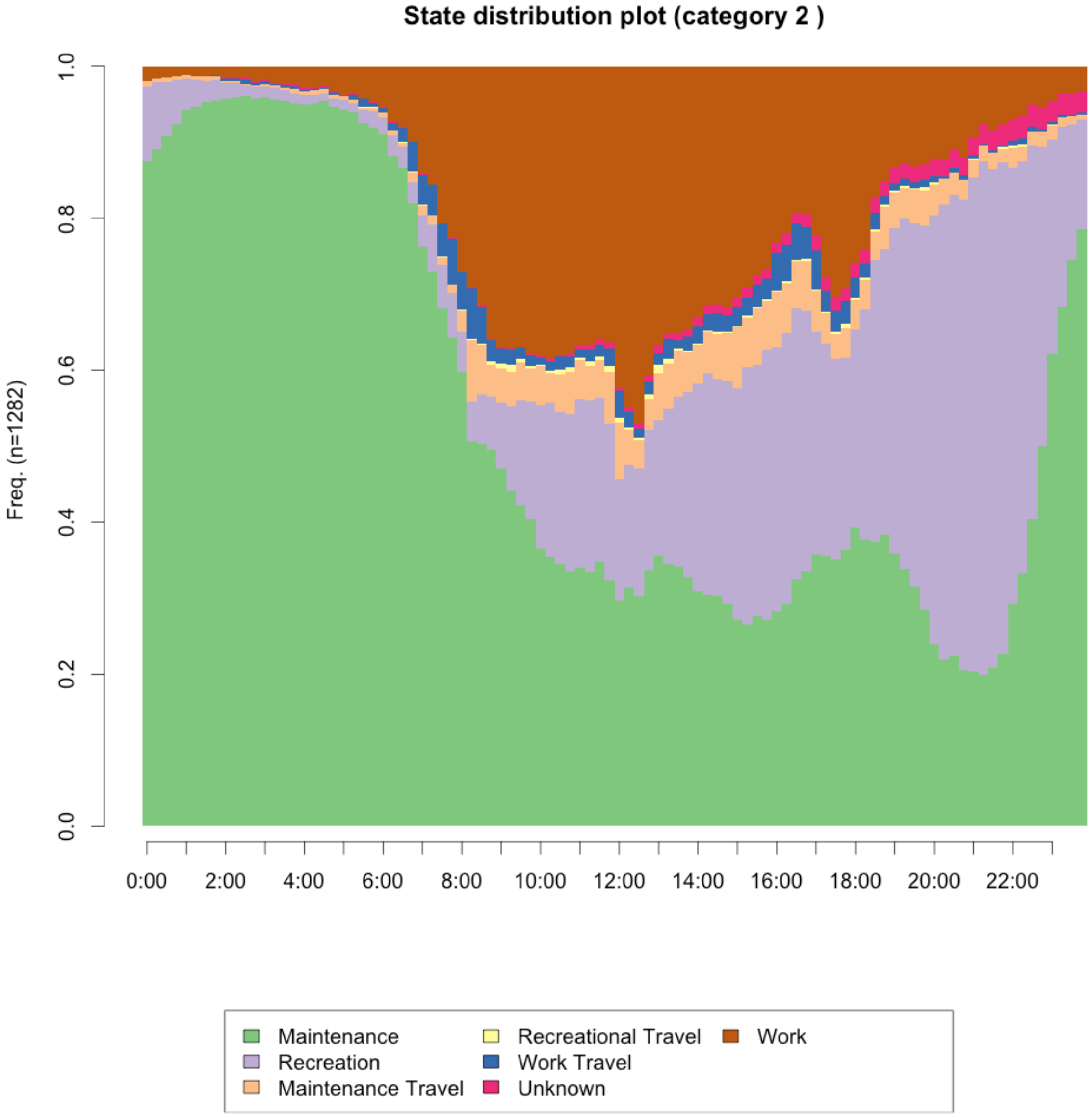

Most common activity sequences (Category 8)
0.7%
0%
Cum. % freq. (n=1988)
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Time
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

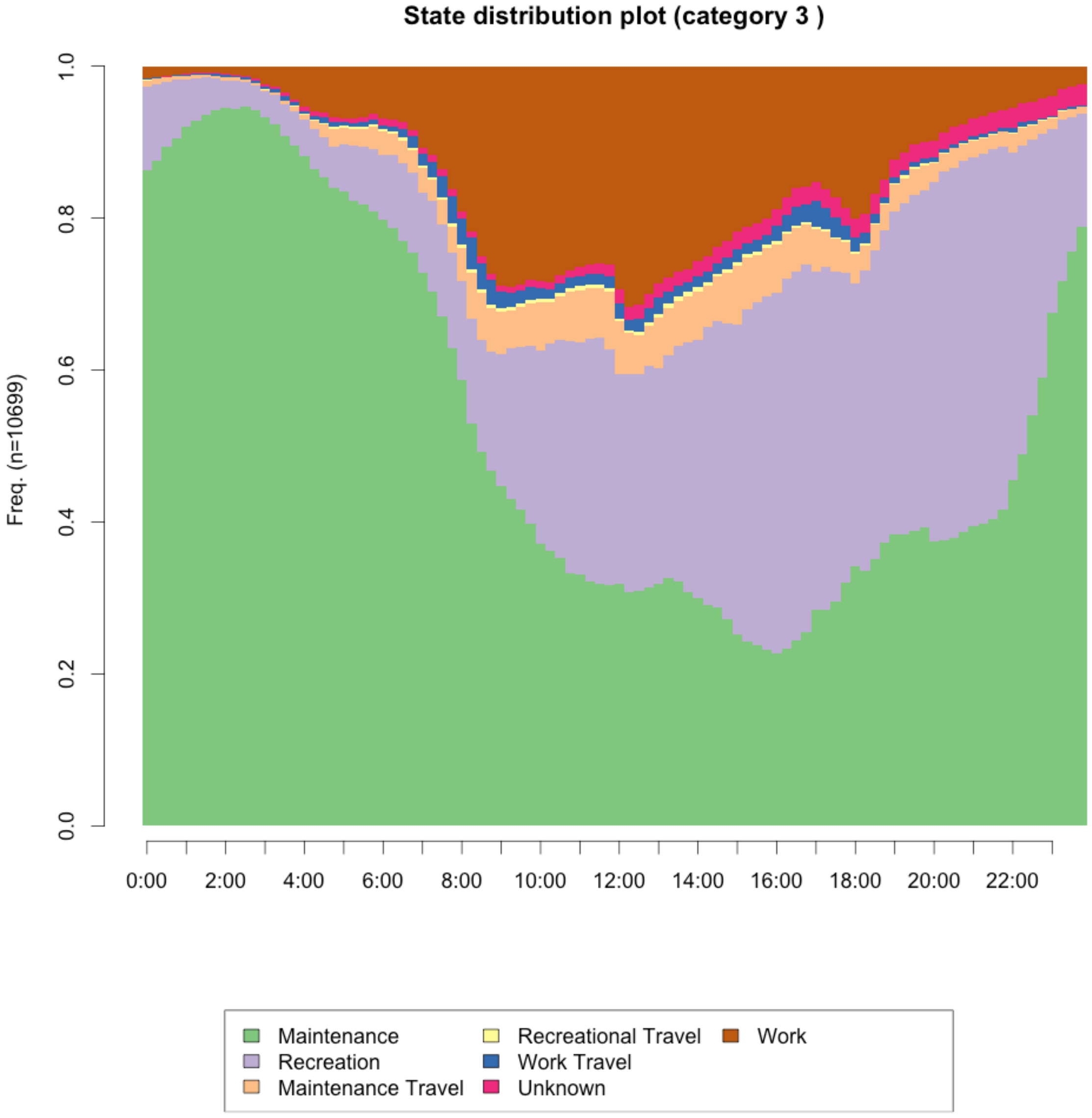

Most common activity sequences (Category 9)
1.2%
0%
Cum. % freq. (n=1385)
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Time
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

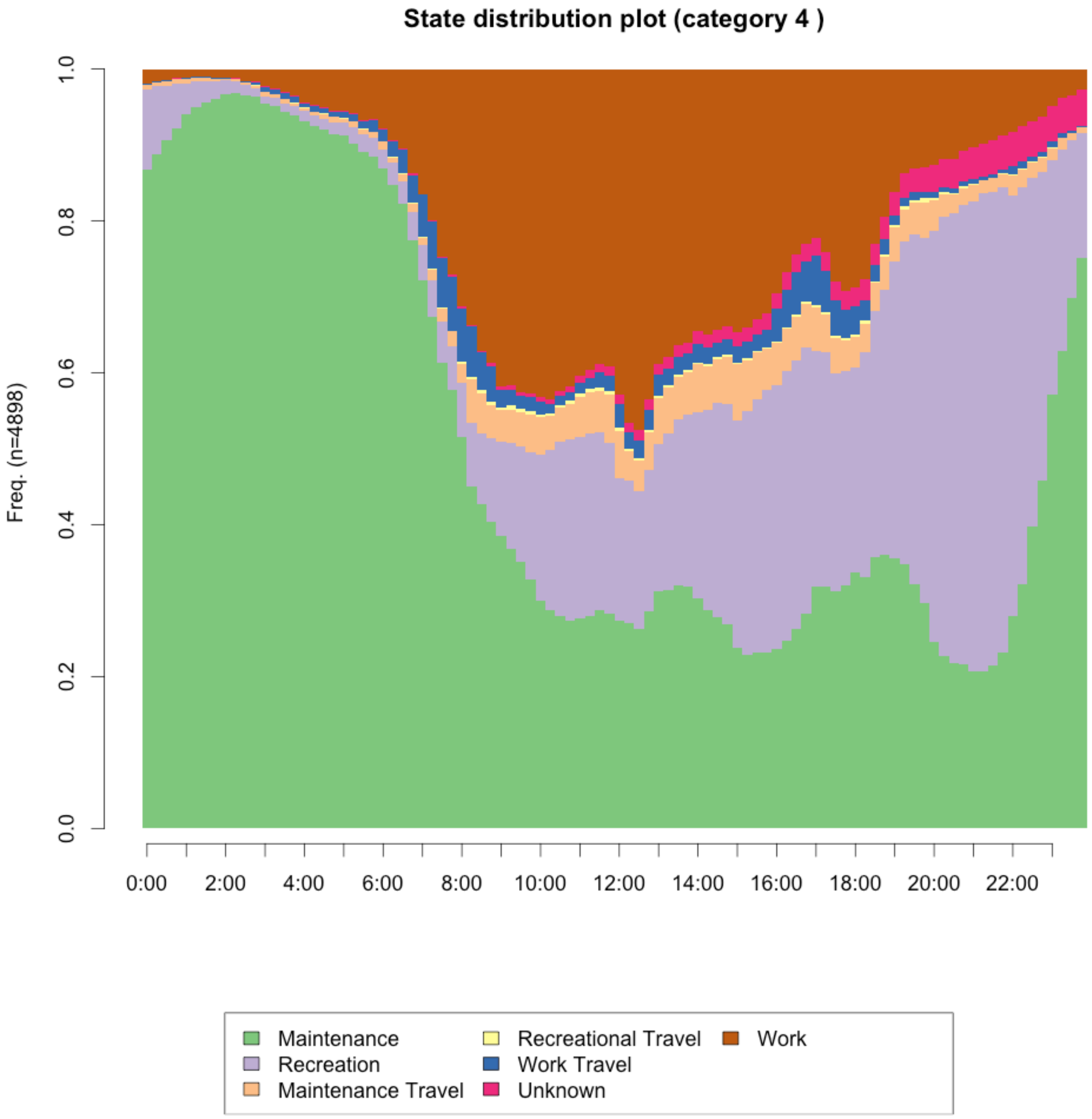

Most common activity sequences (Category 10)
0.4%
0%
Cum. % freq. (n=7665)
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Time
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

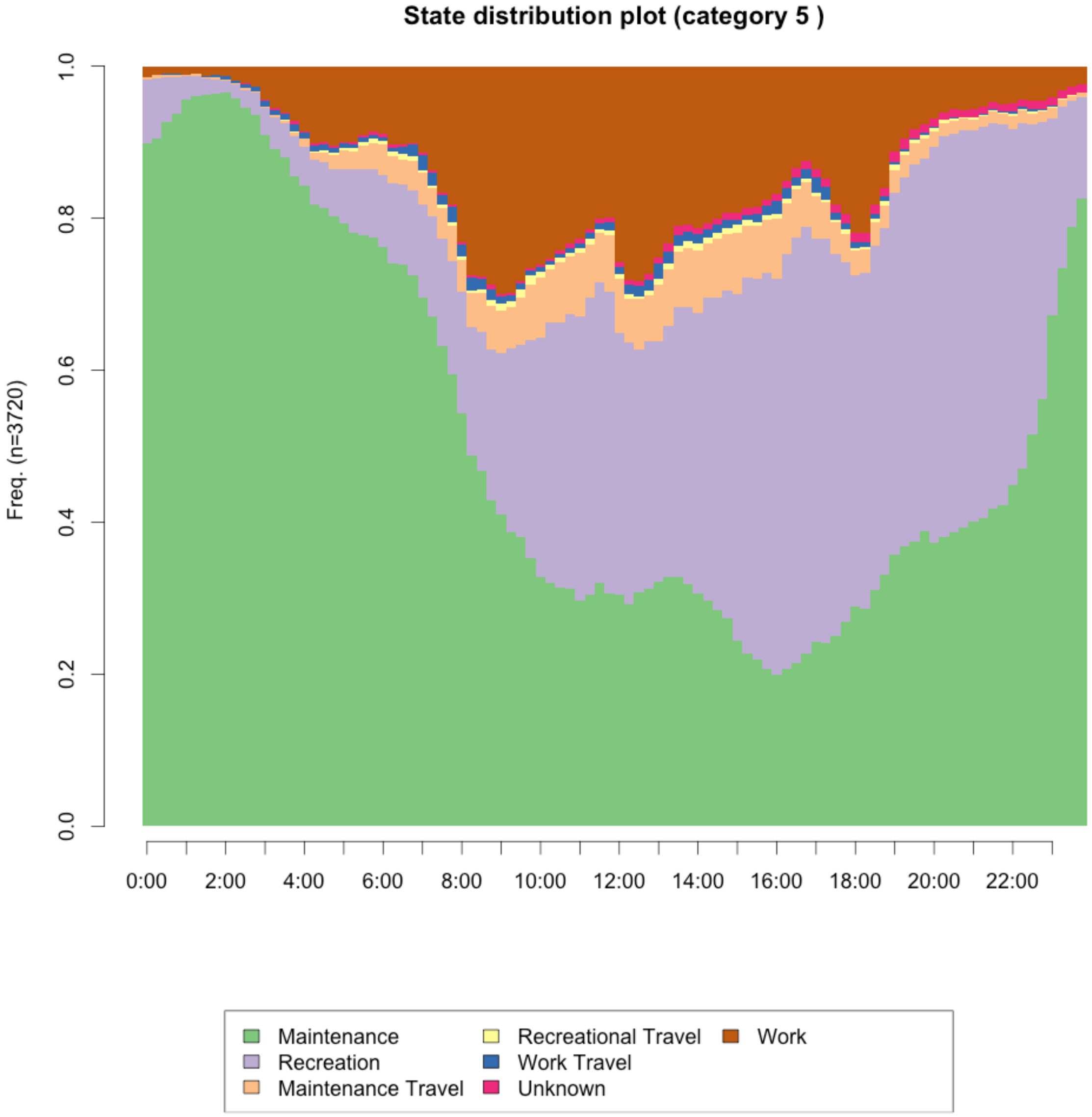

State distribution plot (all categories)
Freq. (n=47903)
0.0
0.2
0.4
0.6
0.8
1.0
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
🟩 Maintenance
▨ Recreation
▨ Maintenance Travel
▨ Recreational Travel
🟦 Work Travel
🟥 Unknown
🟫 Work

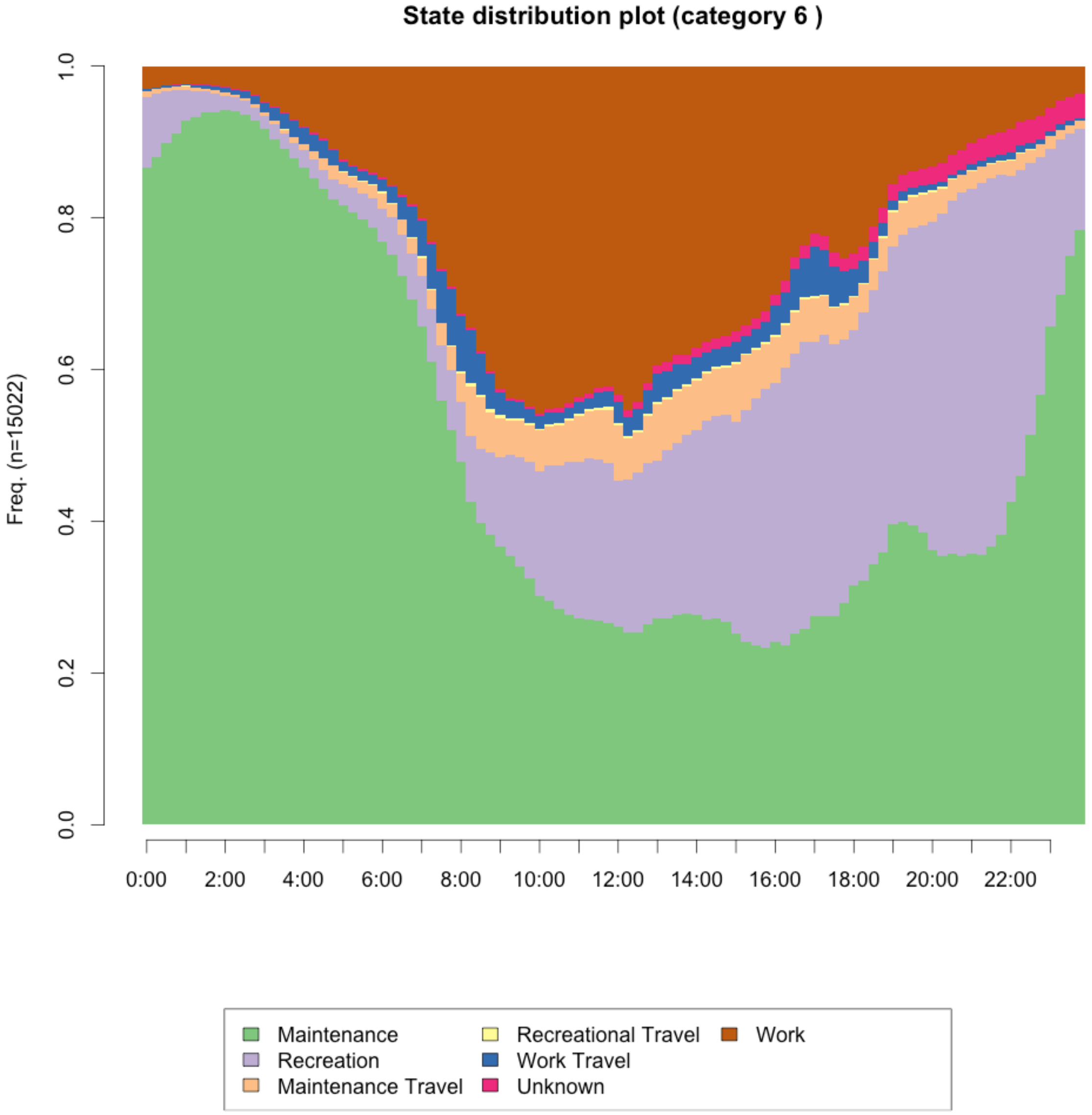

State distribution plot (category 1)
Freq. (n=870)
0.0
0.2
0.4
0.6
0.8
1.0
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

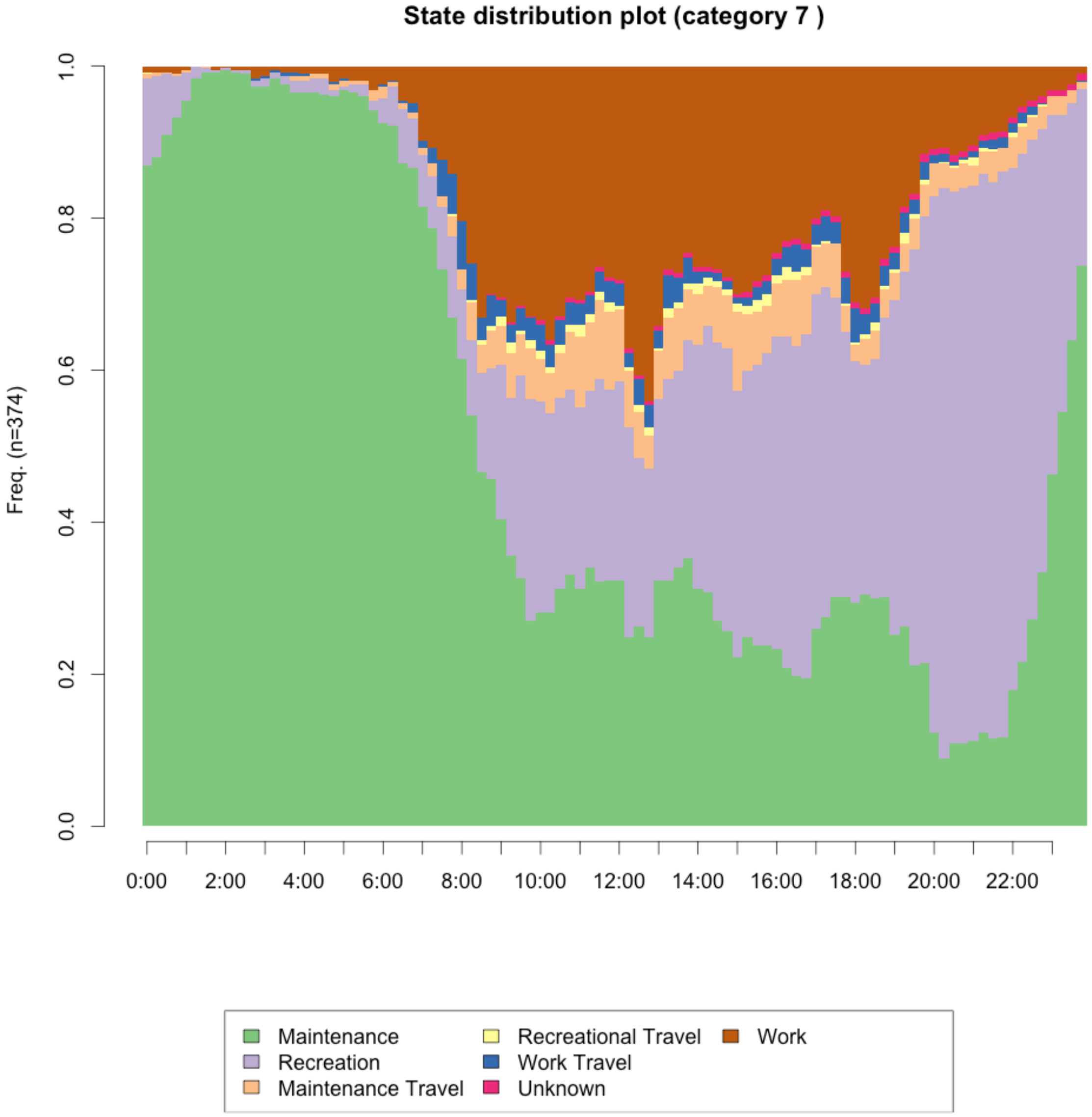

State distribution plot (category 2)
Freq. (n=1282)
0.0
0.2
0.4
0.6
0.8
1.0
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

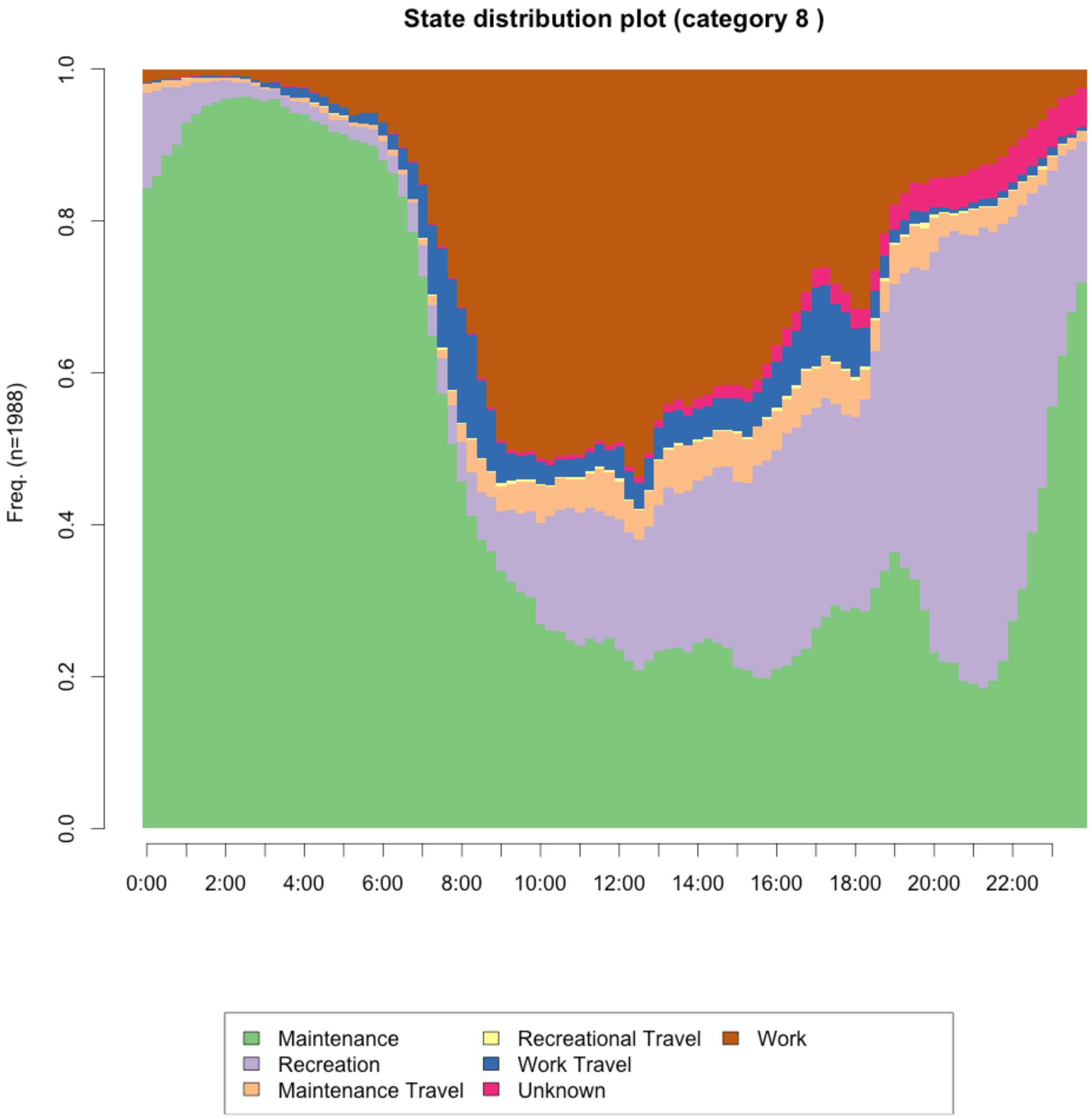

State distribution plot (category 3)
Freq. (n=10699)
0.0
0.2
0.4
0.6
0.8
1.0
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

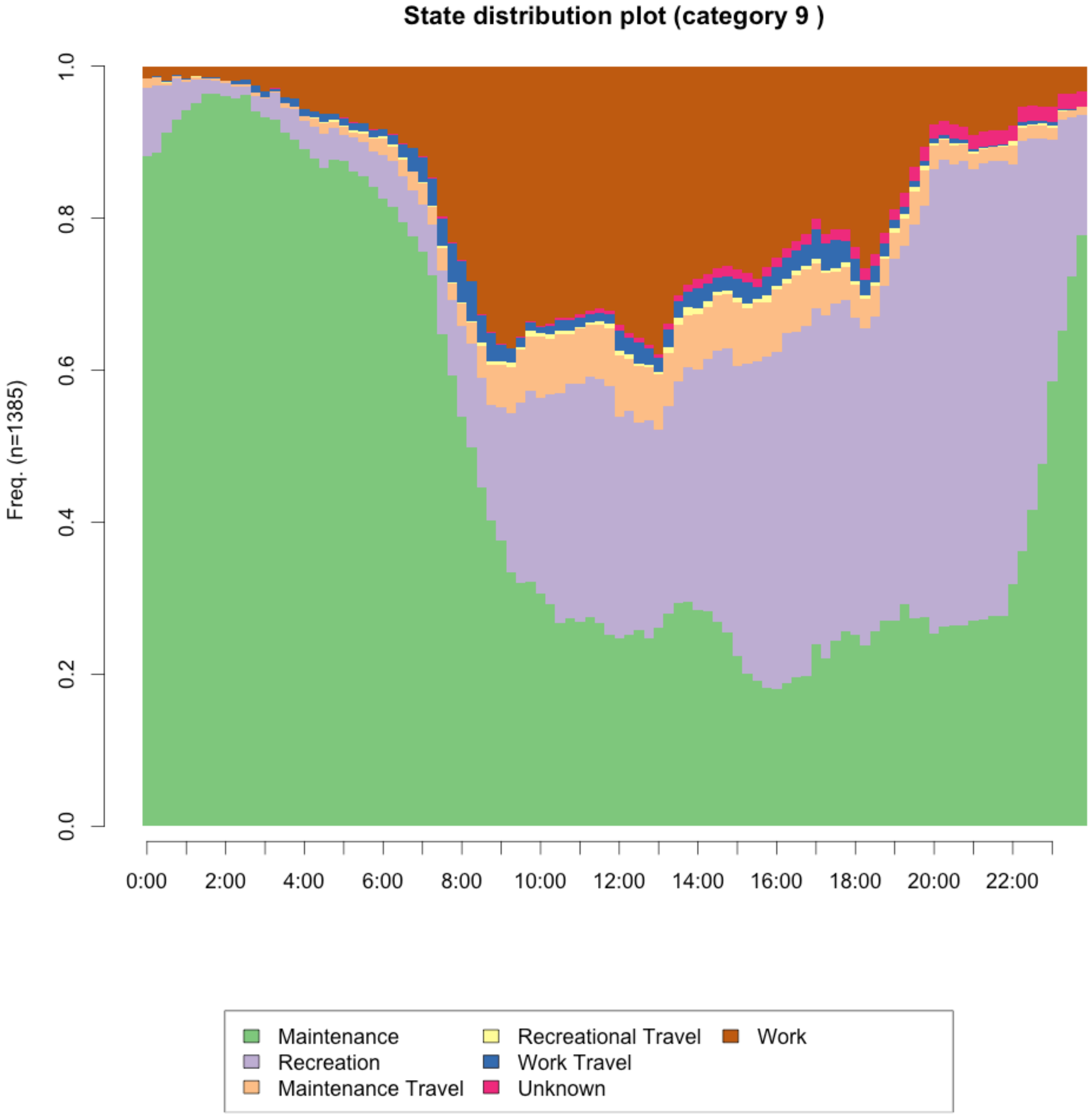

State distribution plot (category 4)
Freq. (n=4898)
0.0
0.2
0.4
0.6
0.8
1.0
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

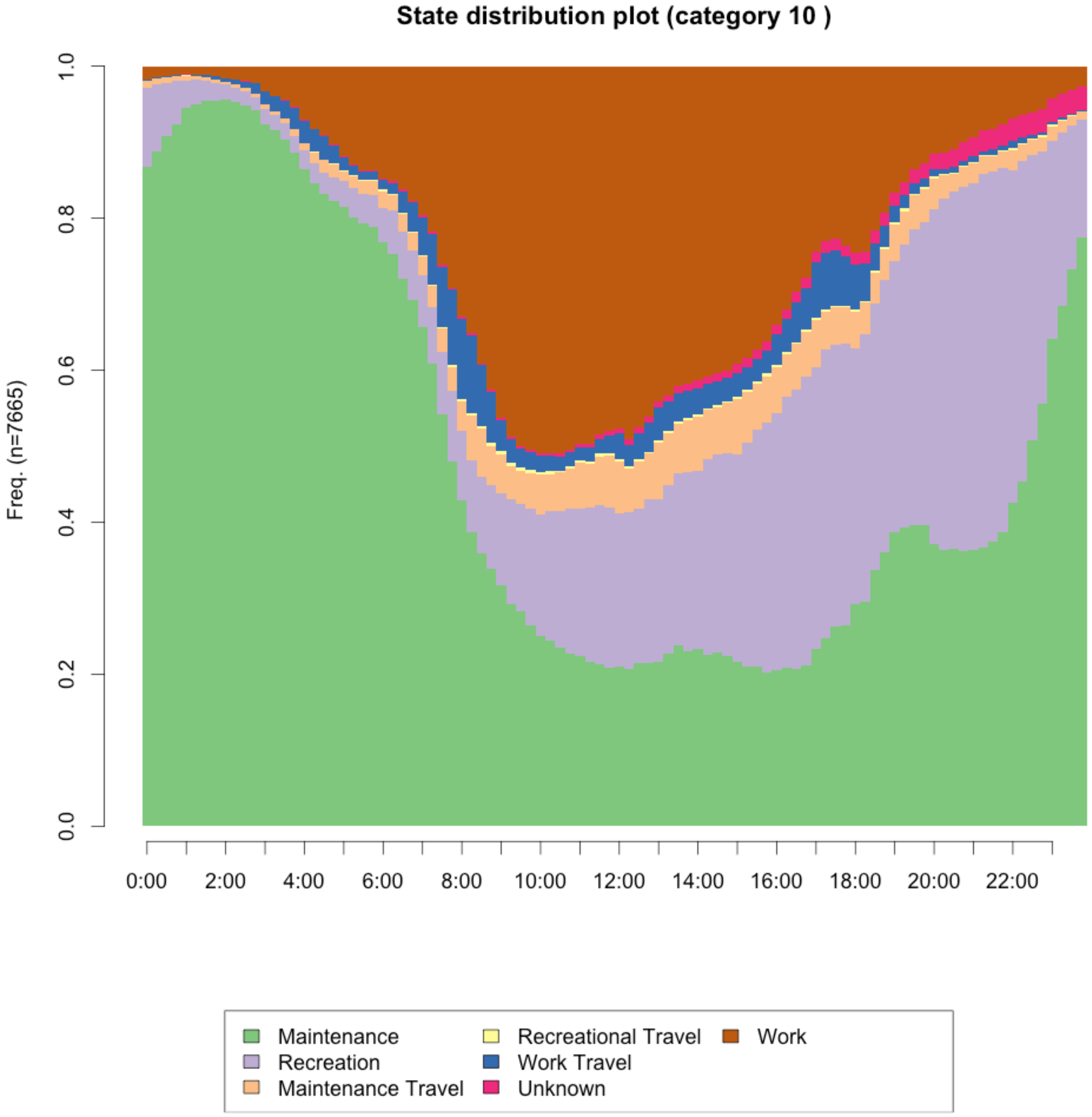

State distribution plot (category 5)
Freq. (n=3720)
0.0
0.2
0.4
0.6
0.8
1.0
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

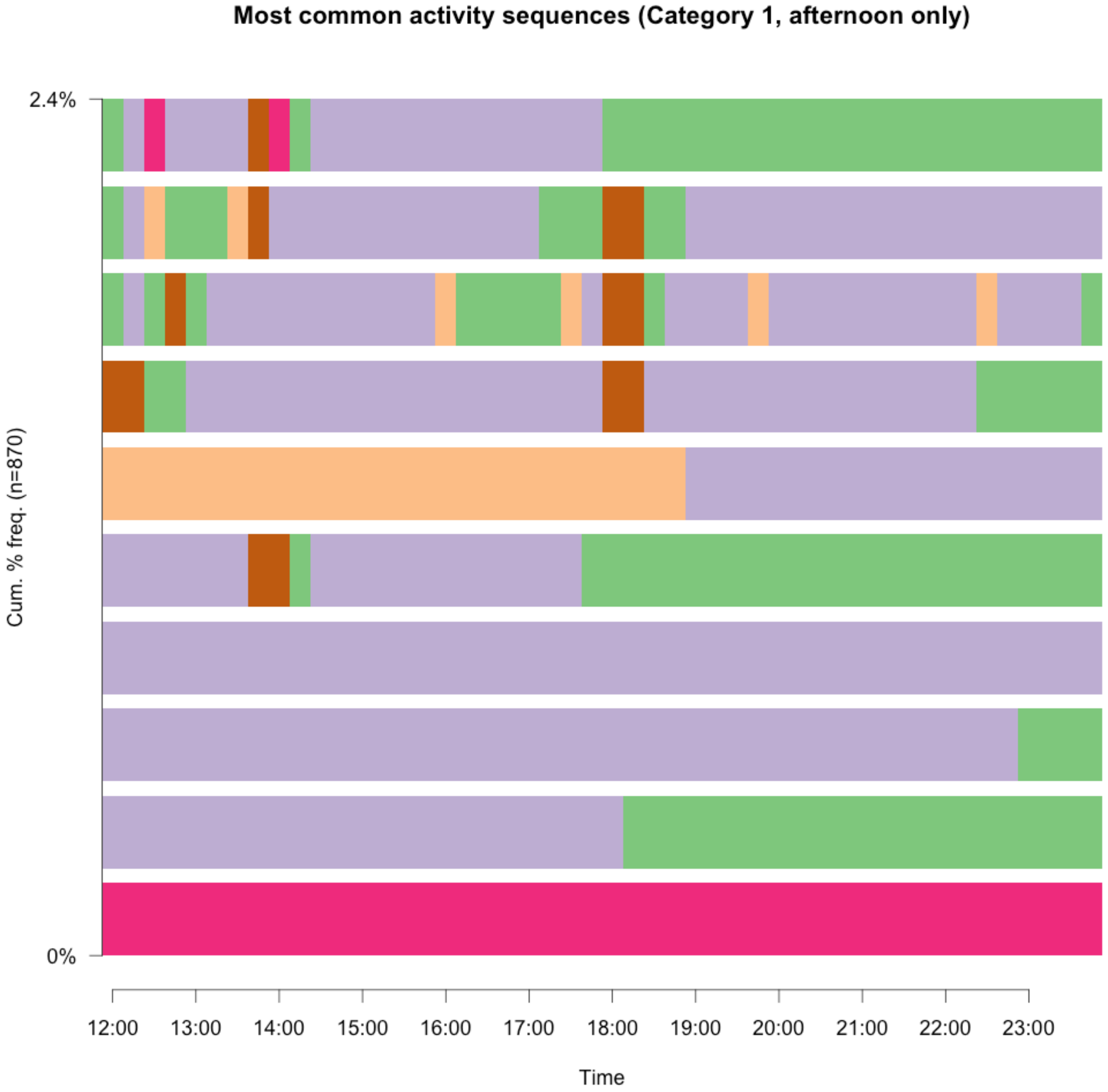

State distribution plot (category 6)
Freq. (n=15022)
0.0
0.2
0.4
0.6
0.8
1.0
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

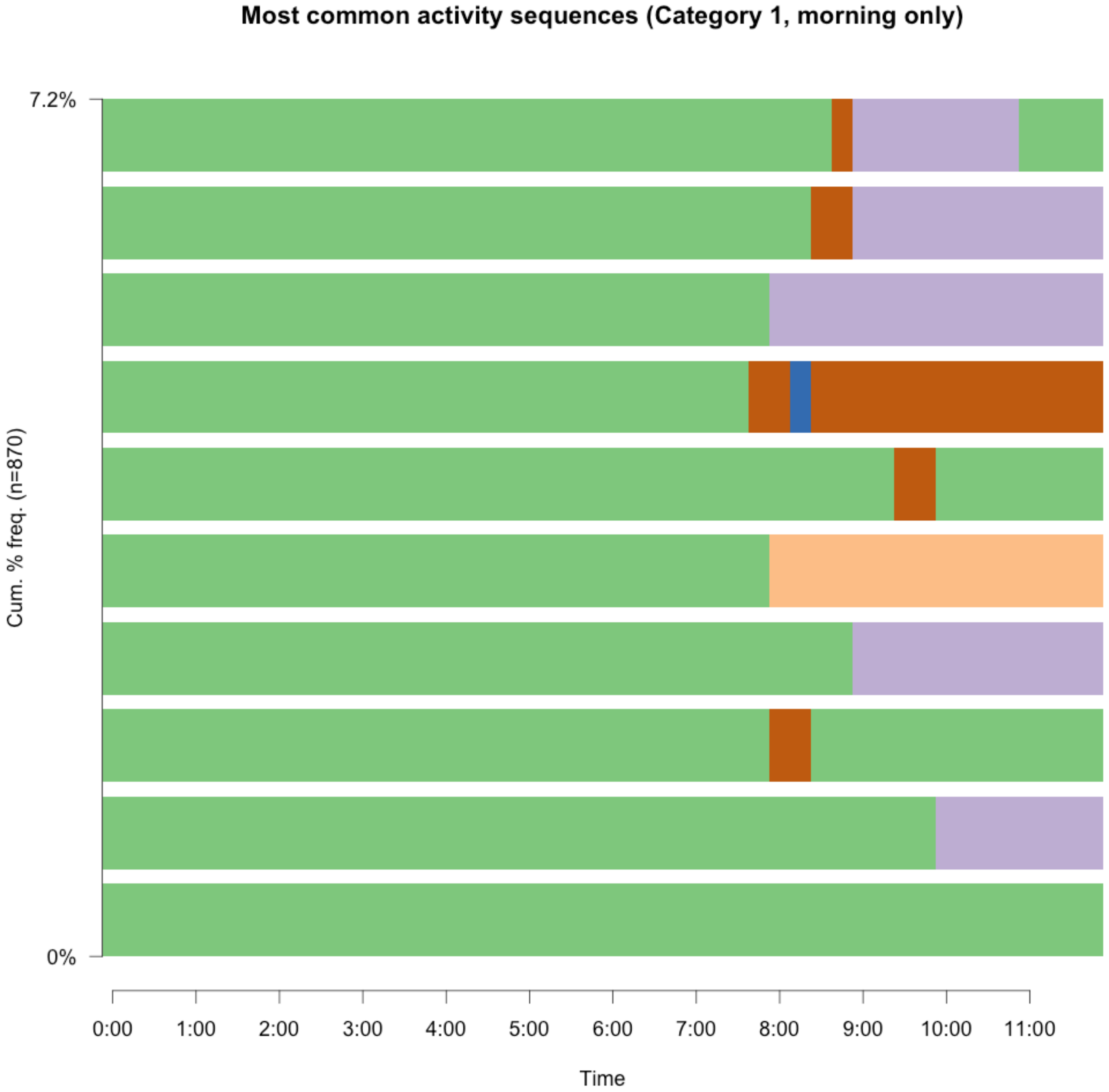

State distribution plot (category 7)
Freq. (n=374)
0.0
0.2
0.4
0.6
0.8
1.0
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

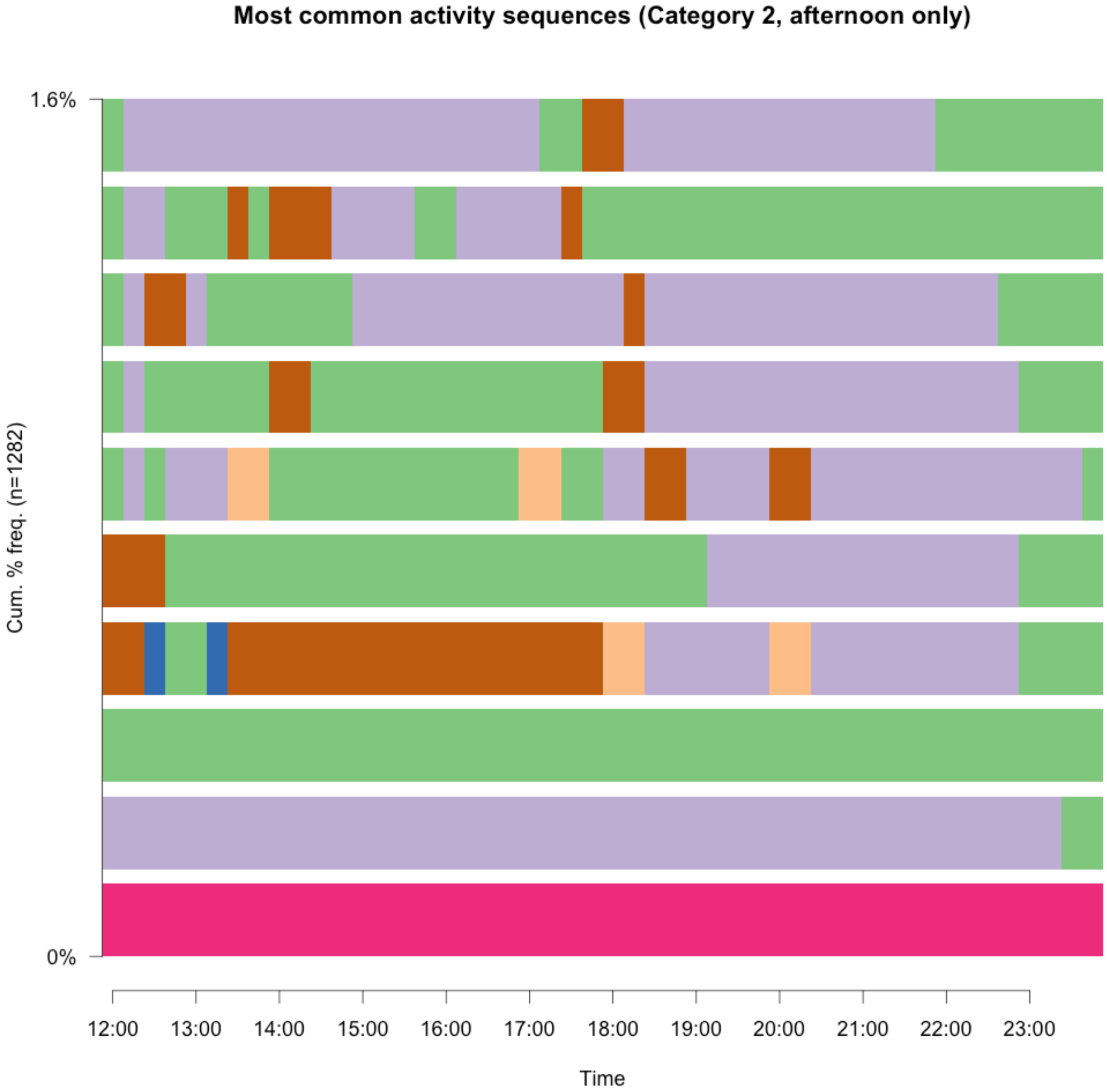

State distribution plot (category 8)
Freq. (n=1988)
0.0
0.2
0.4
0.6
0.8
1.0
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

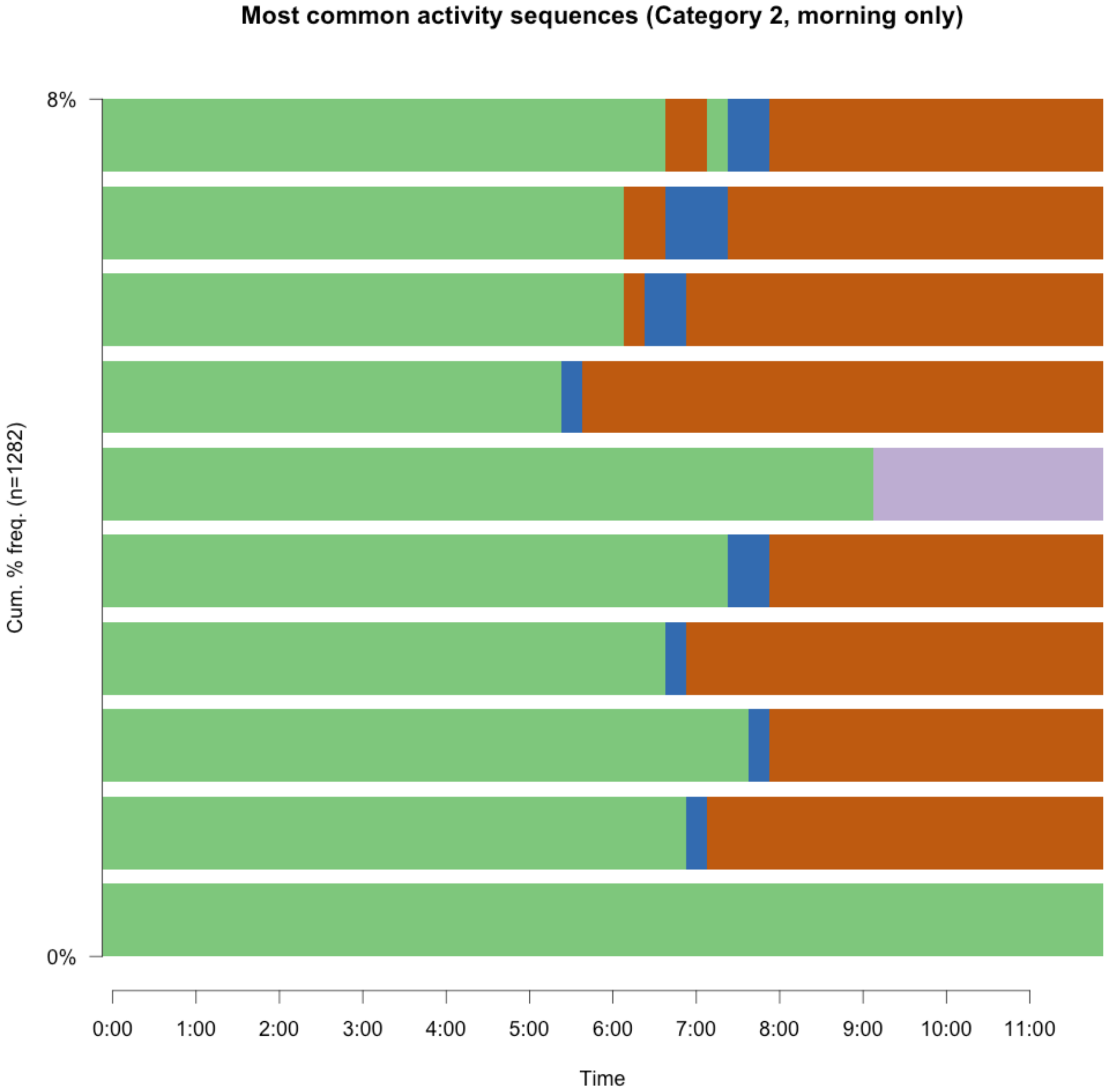

State distribution plot (category 9)
Freq. (n=1385)
0.0
0.2
0.4
0.6
0.8
1.0
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
🟩 Maintenance
🟪 Recreation
🟧 Maintenance Travel
🟨 Recreational Travel
🟦 Work Travel
🟥 Unknown
🟫 Work

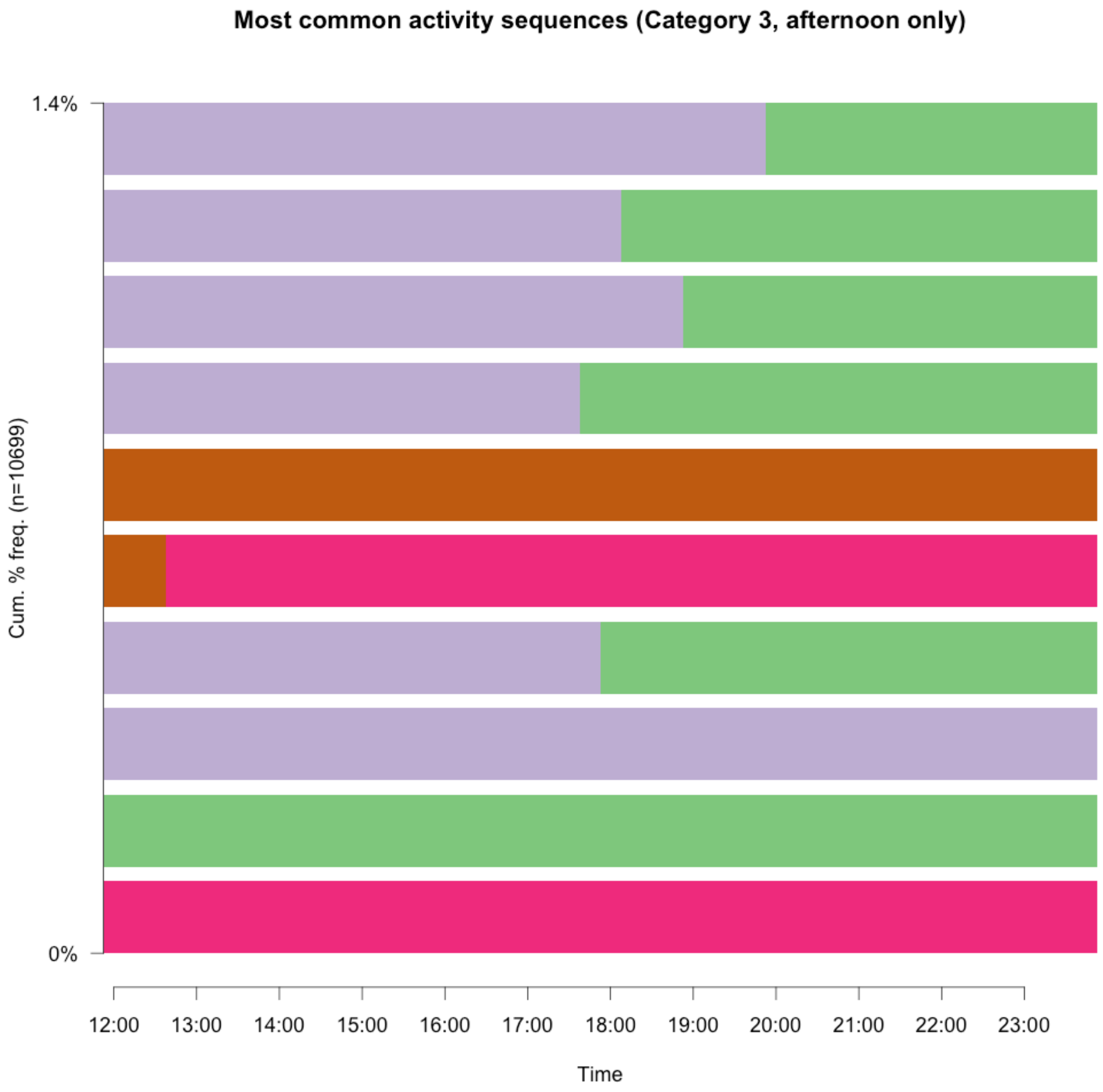

State distribution plot (category 10)
Freq. (n=7665)
0.0
0.2
0.4
0.6
0.8
1.0
0:00
2:00
4:00
6:00
8:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
Maintenance
Recreation
Maintenance Travel
Recreational Travel
Work Travel
Unknown
Work

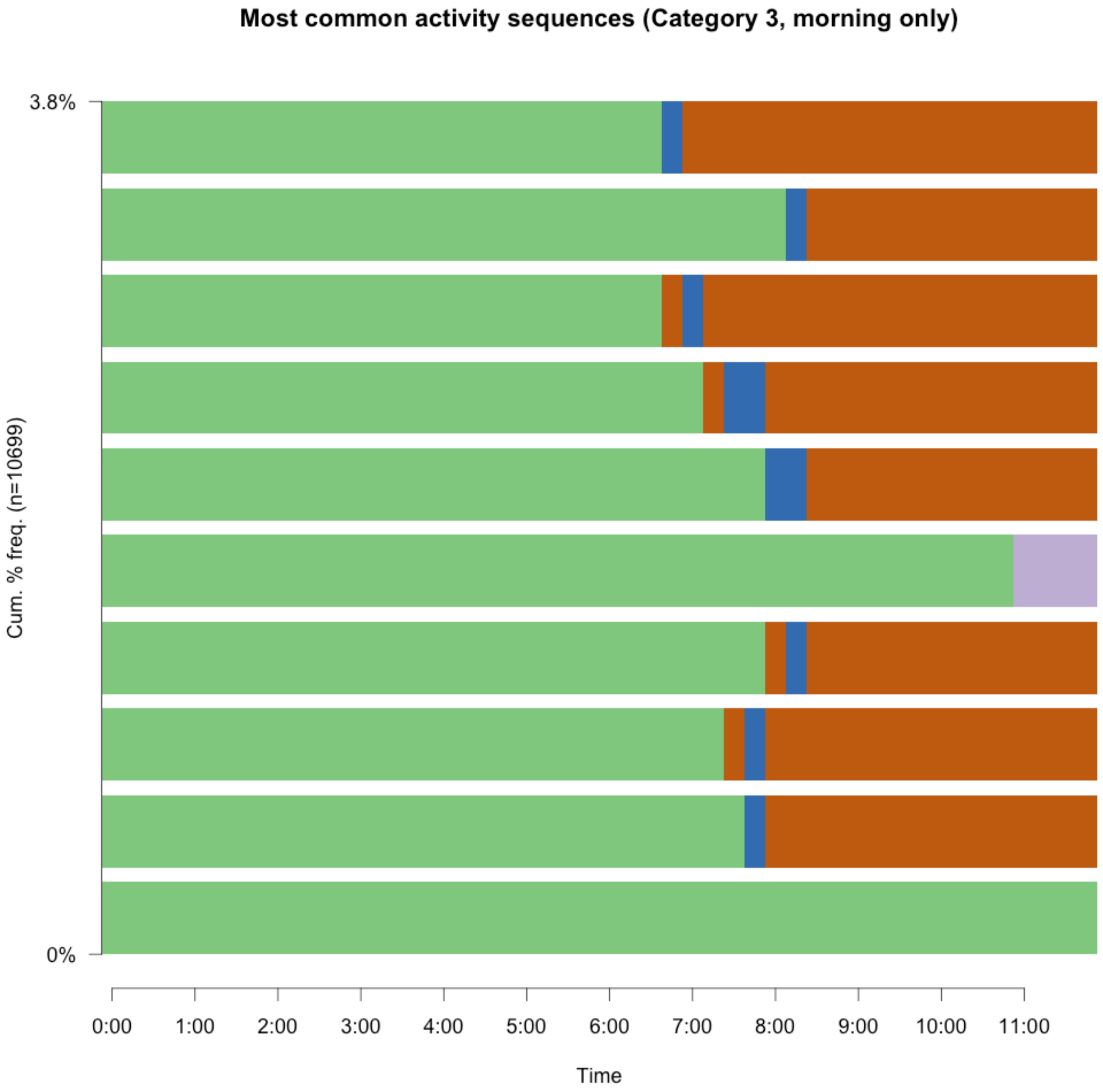

Most common activity sequences (Category 1, afternoon only)
2.4%
0%
Cum. % freq. (n=870)
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Time

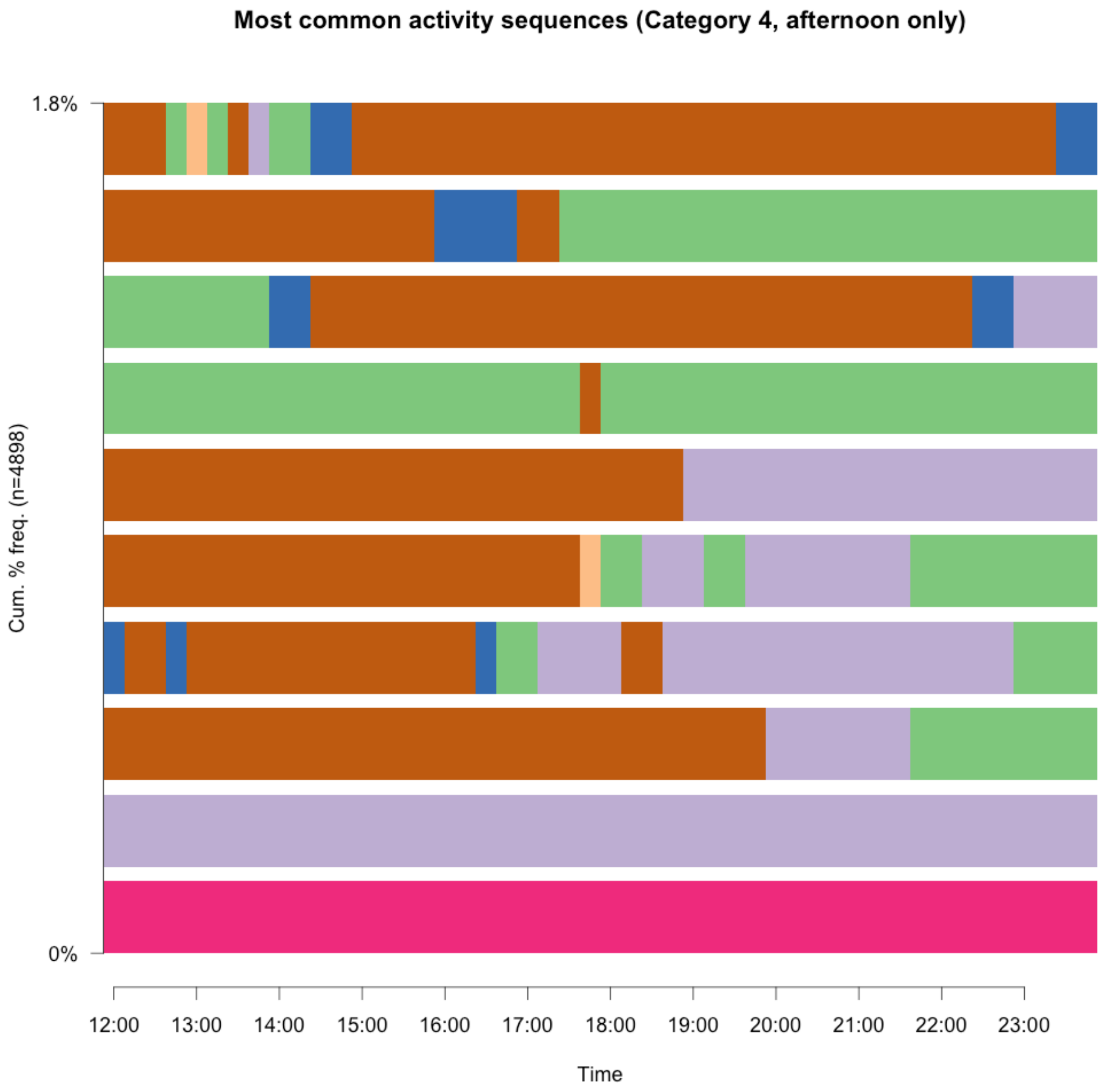

Most common activity sequences (Category 1, morning only)
7.2%
0%
Cum. % freq. (n=870)
0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:00
Time

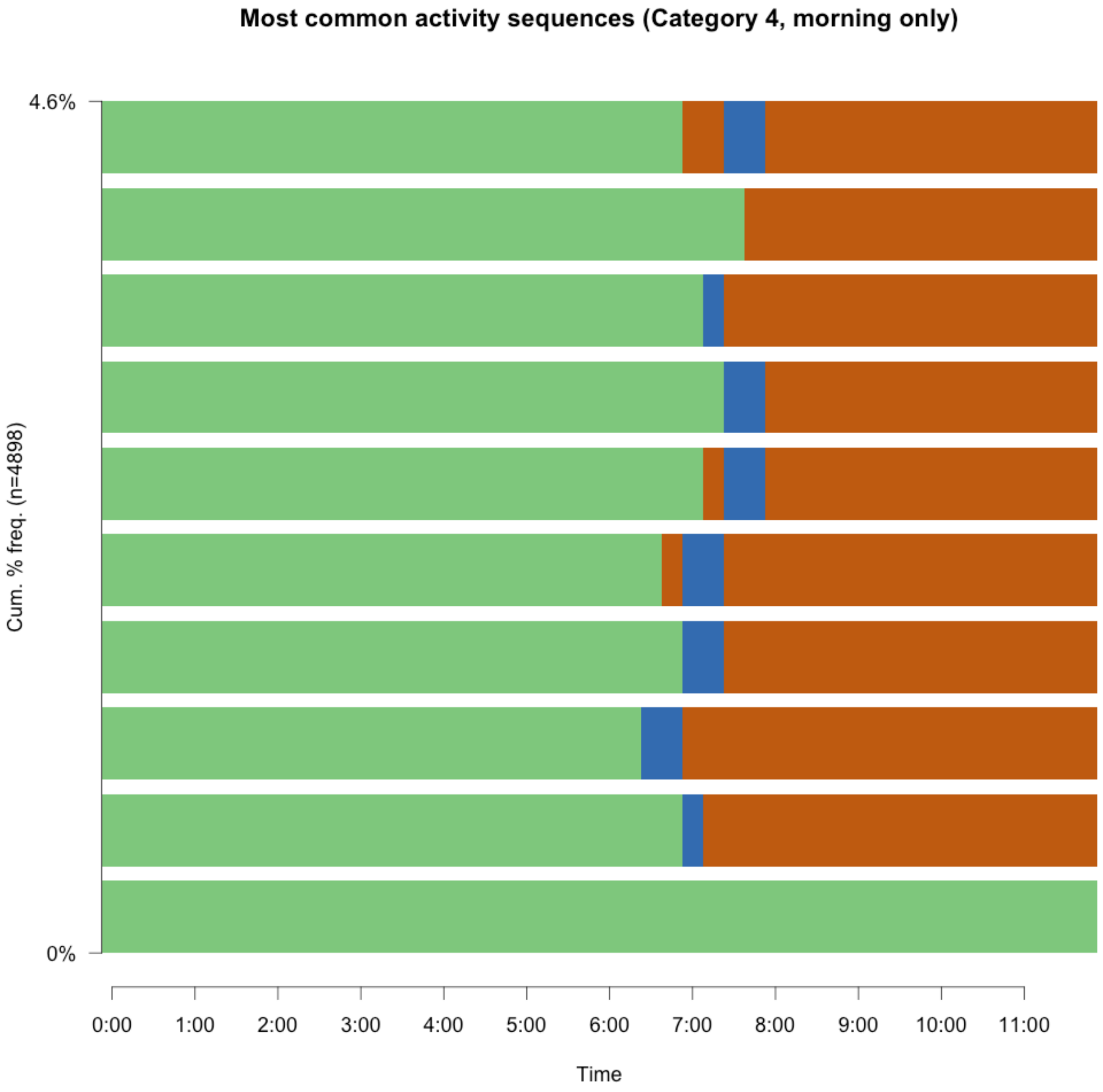

Most common activity sequences (Category 2, afternoon only)
1.6%
0%
Cum. % freq. (n=1282)
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Time

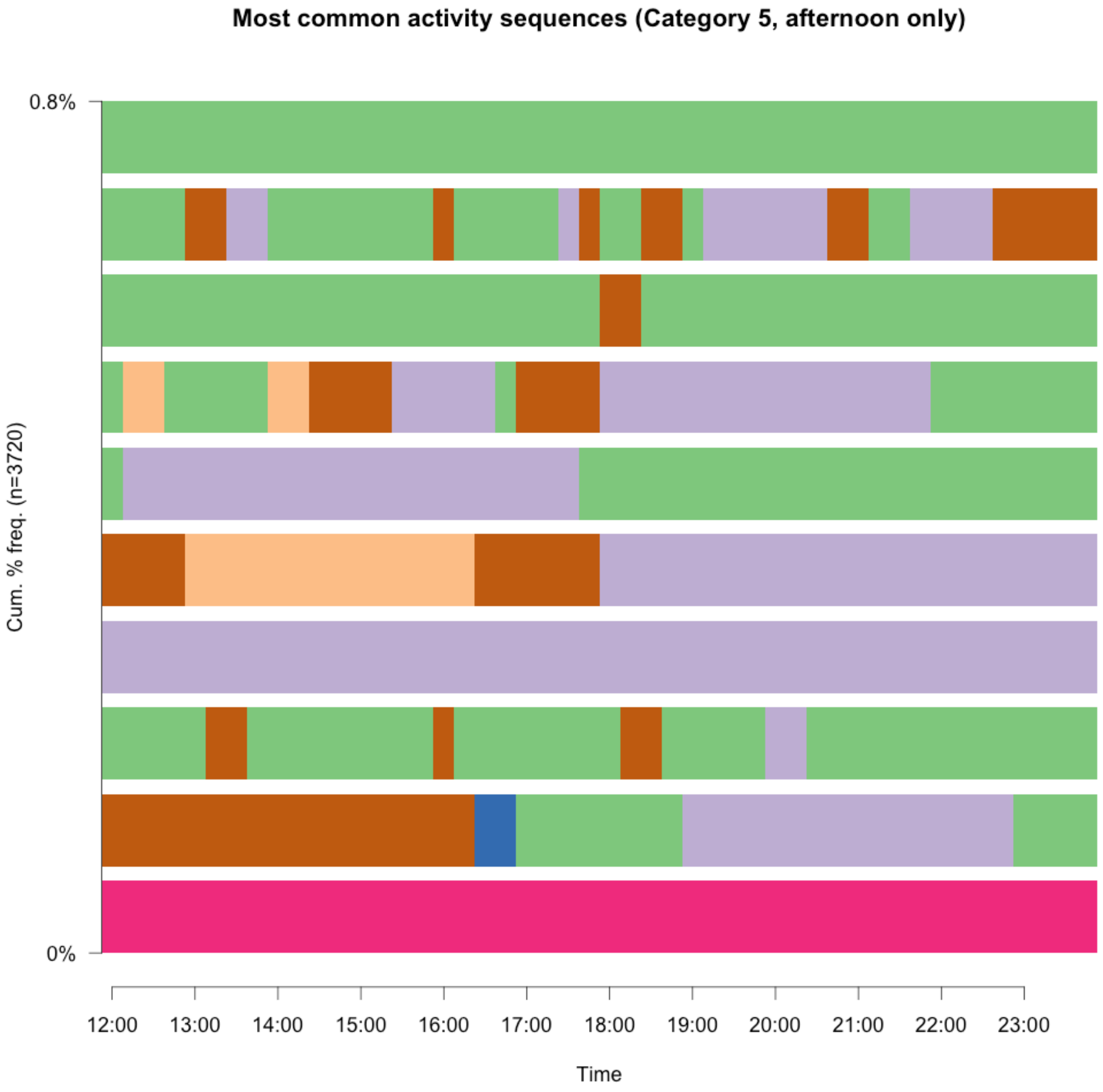

Most common activity sequences (Category 2, morning only)
8%
0%
Cum. % freq. (n=1282)
0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:00
Time

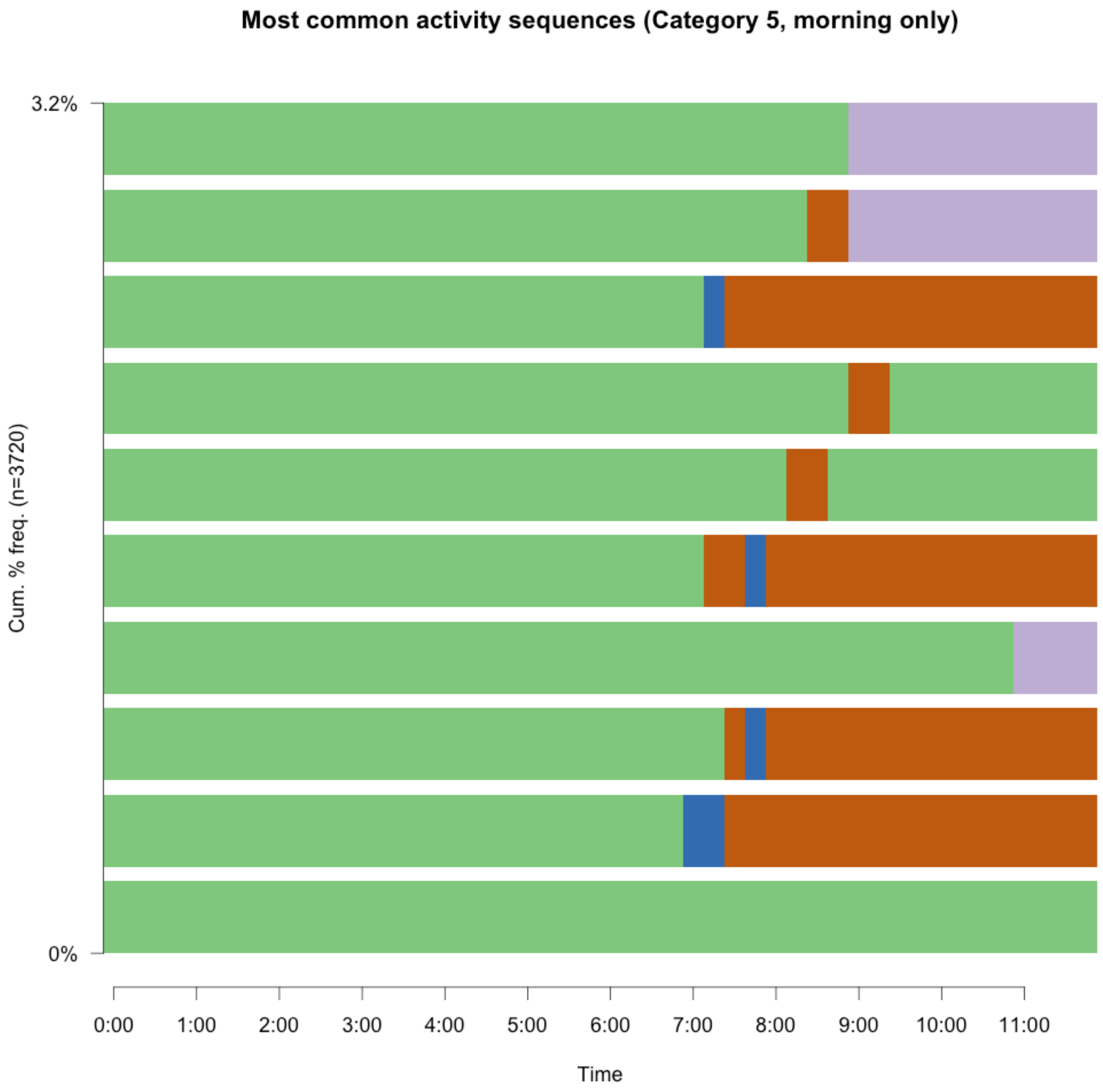

Most common activity sequences (Category 3, afternoon only)
1.4%
0%
Cum. % freq. (n=10699)
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Time

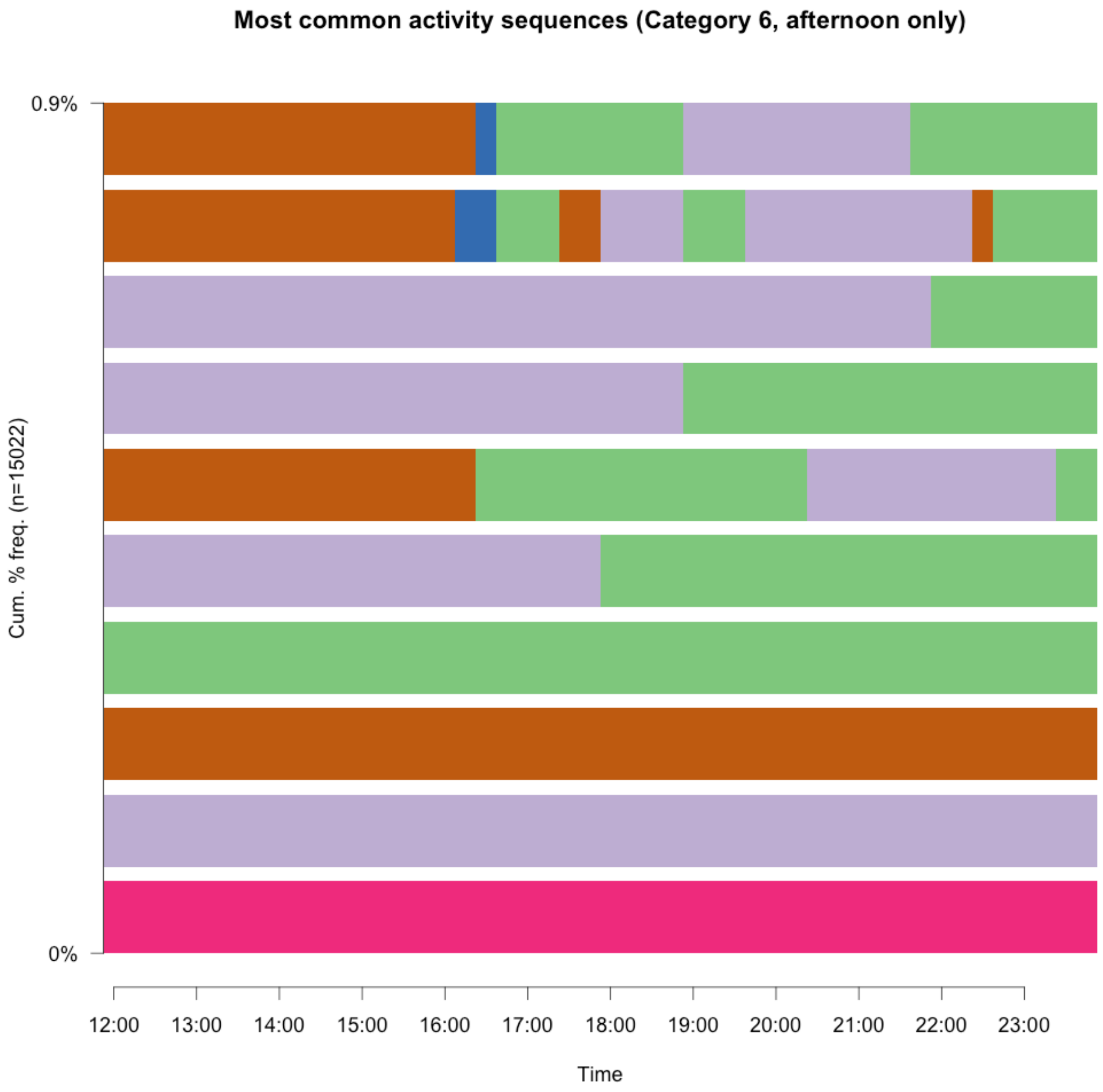

Most common activity sequences (Category 3, morning only)
3.8%
0%
Cum. % freq. (n=10699)
0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:00
Time

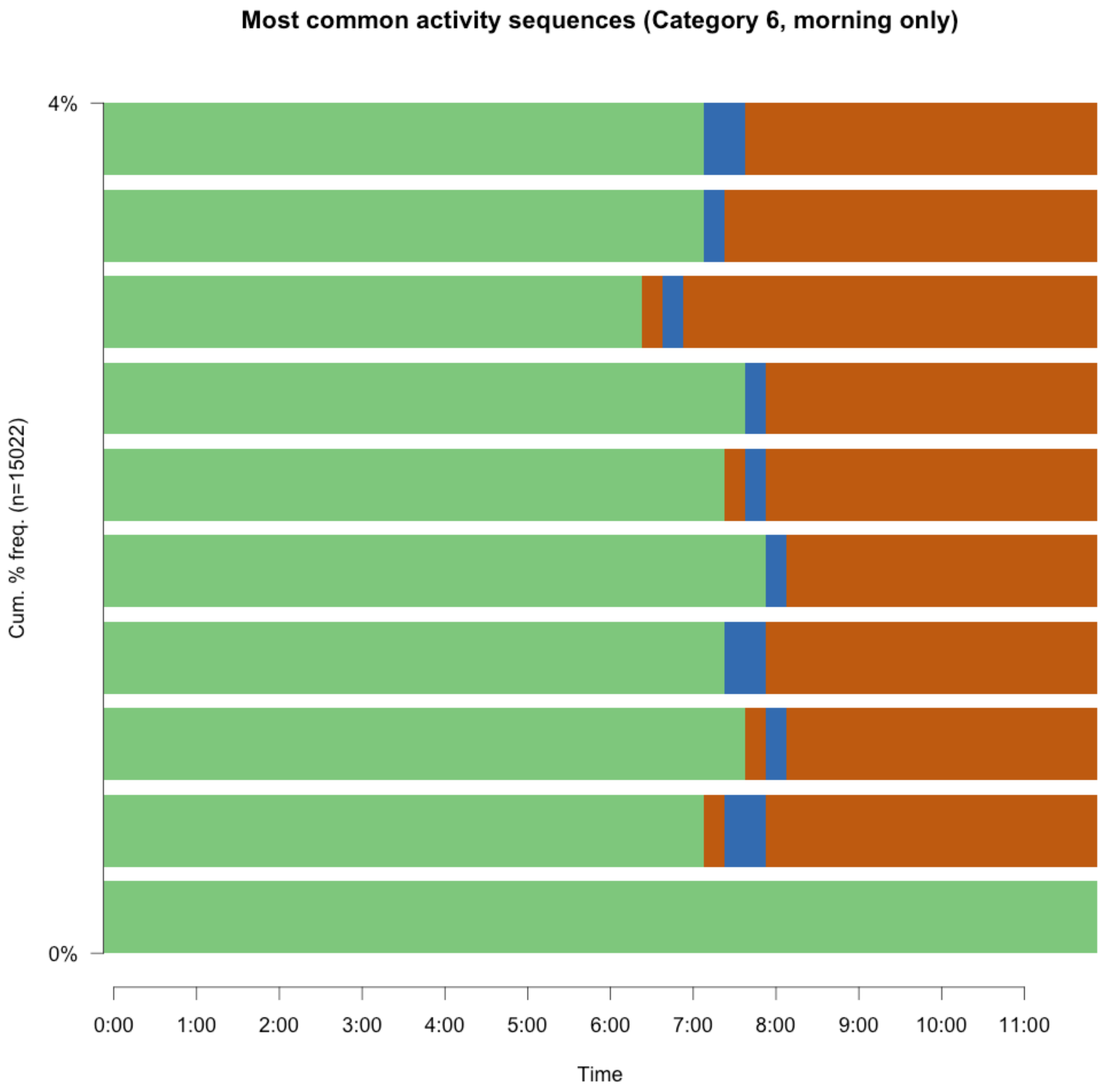

Most common activity sequences (Category 4, afternoon only)
1.8%
0%
Cum. % freq. (n=4898)
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Time

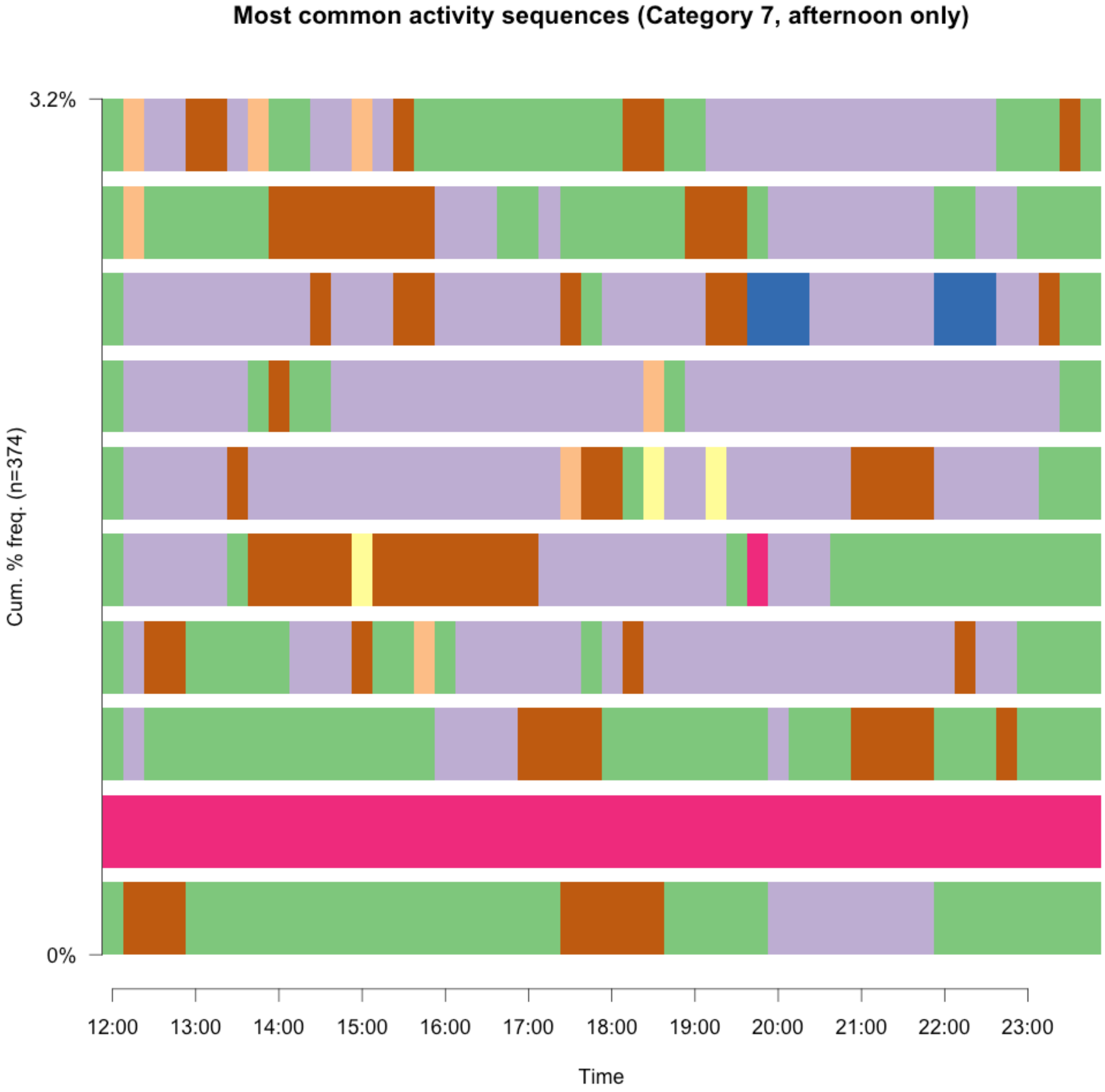

Most common activity sequences (Category 4, morning only)
4.6%
0%
Cum. % freq. (n=4898)
0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:00
Time

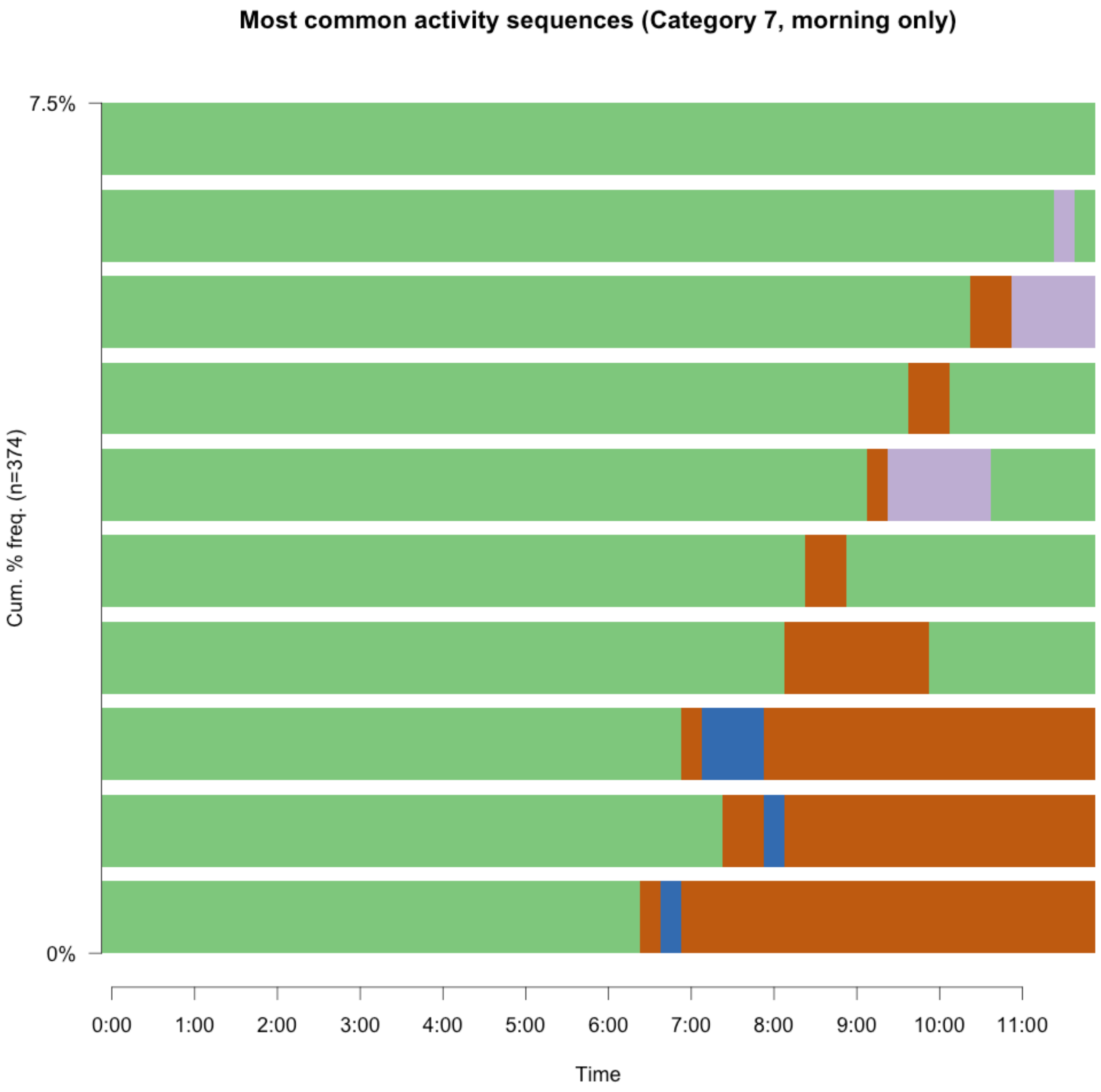

Most common activity sequences (Category 5, afternoon only)
0.8%
0%
Cum. % freq. (n=3720)
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Time

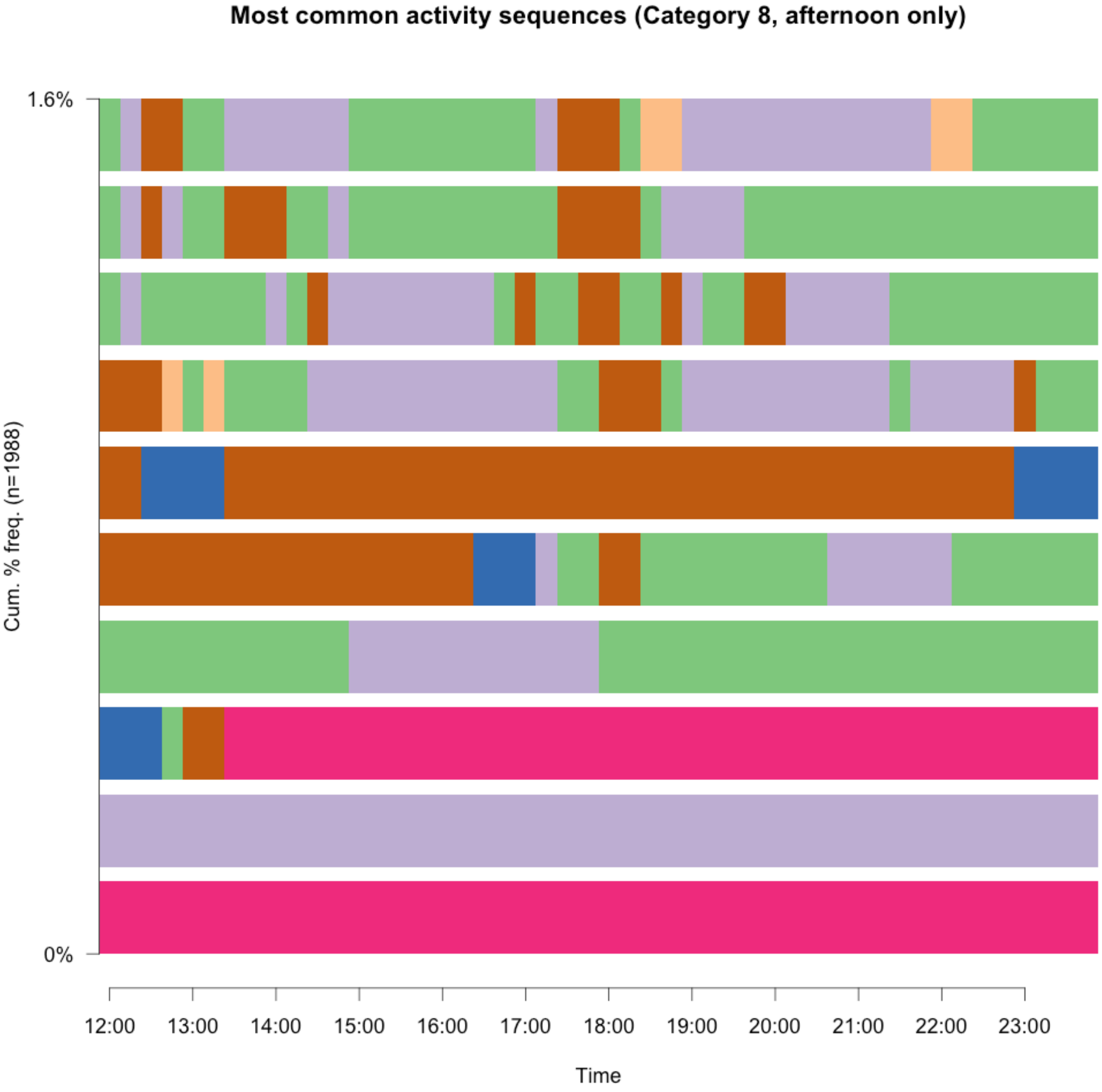

Most common activity sequences (Category 5, morning only)
3.2%
0%
Cum. % freq. (n=3720)
0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:00
Time

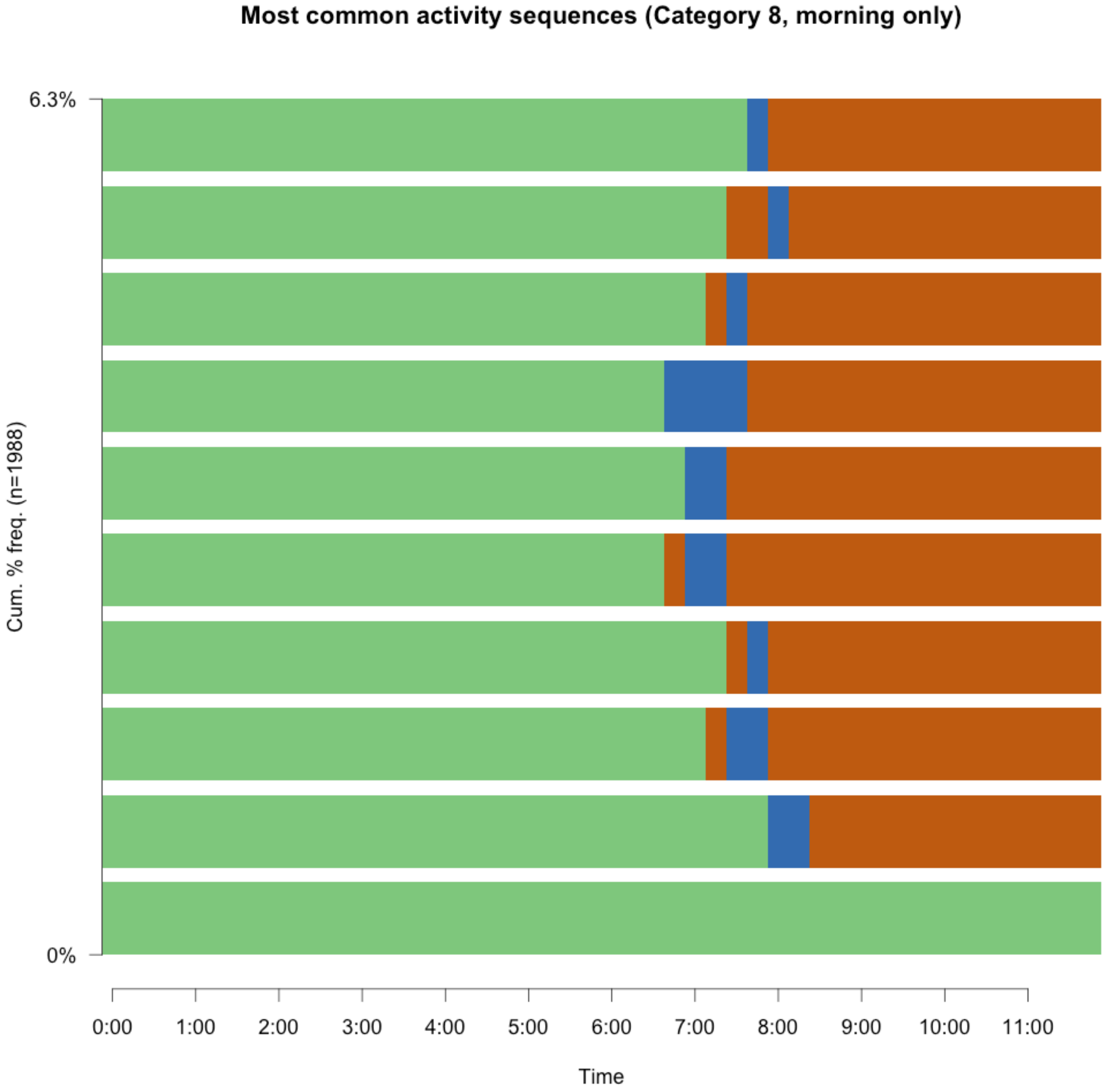

Most common activity sequences (Category 6, afternoon only)
0.9%
0%
Cum. % freq. (n=15022)
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Time

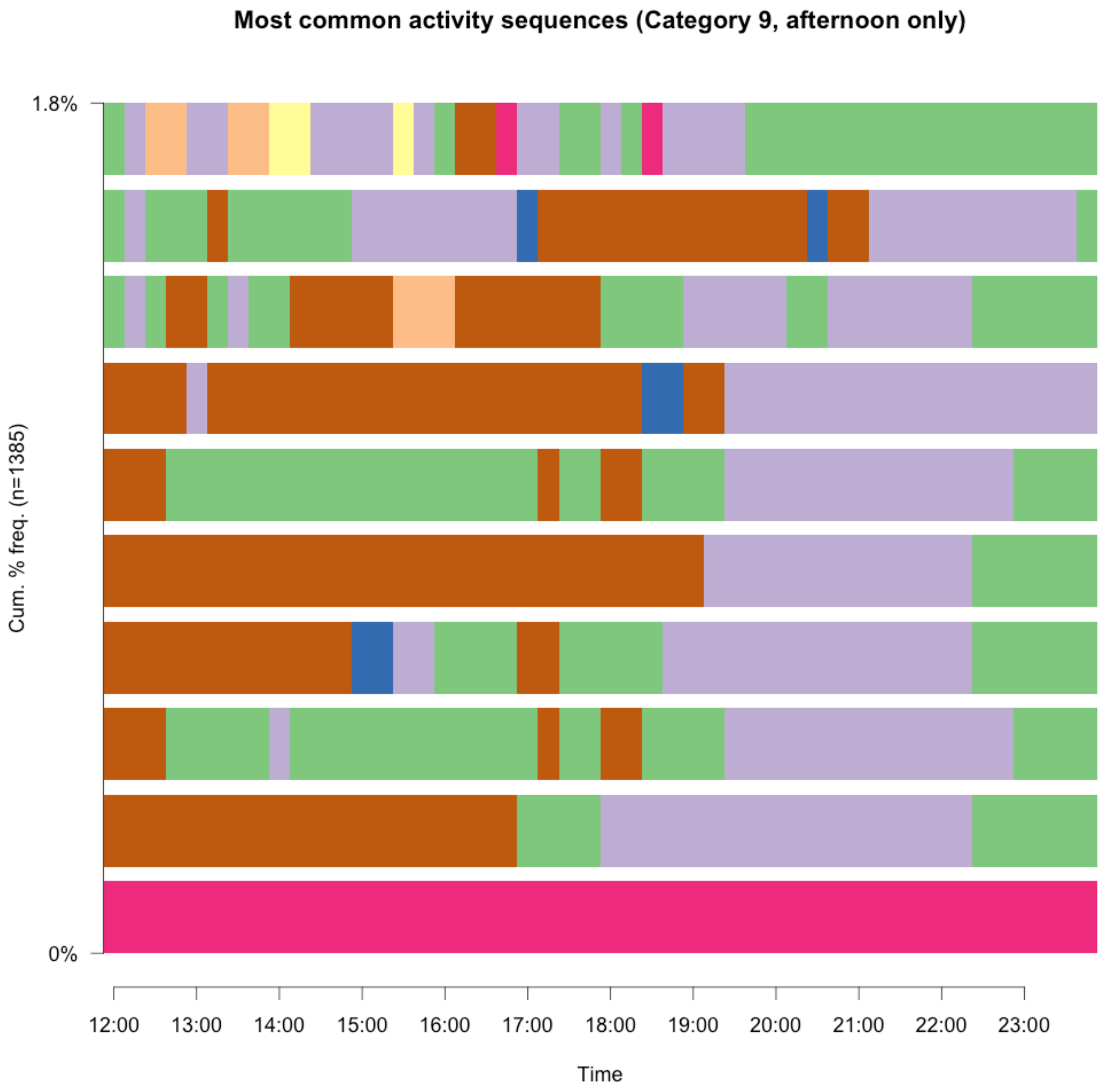

Most common activity sequences (Category 6, morning only)
4%
0%
Cum. % freq. (n=15022)
0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:00
Time

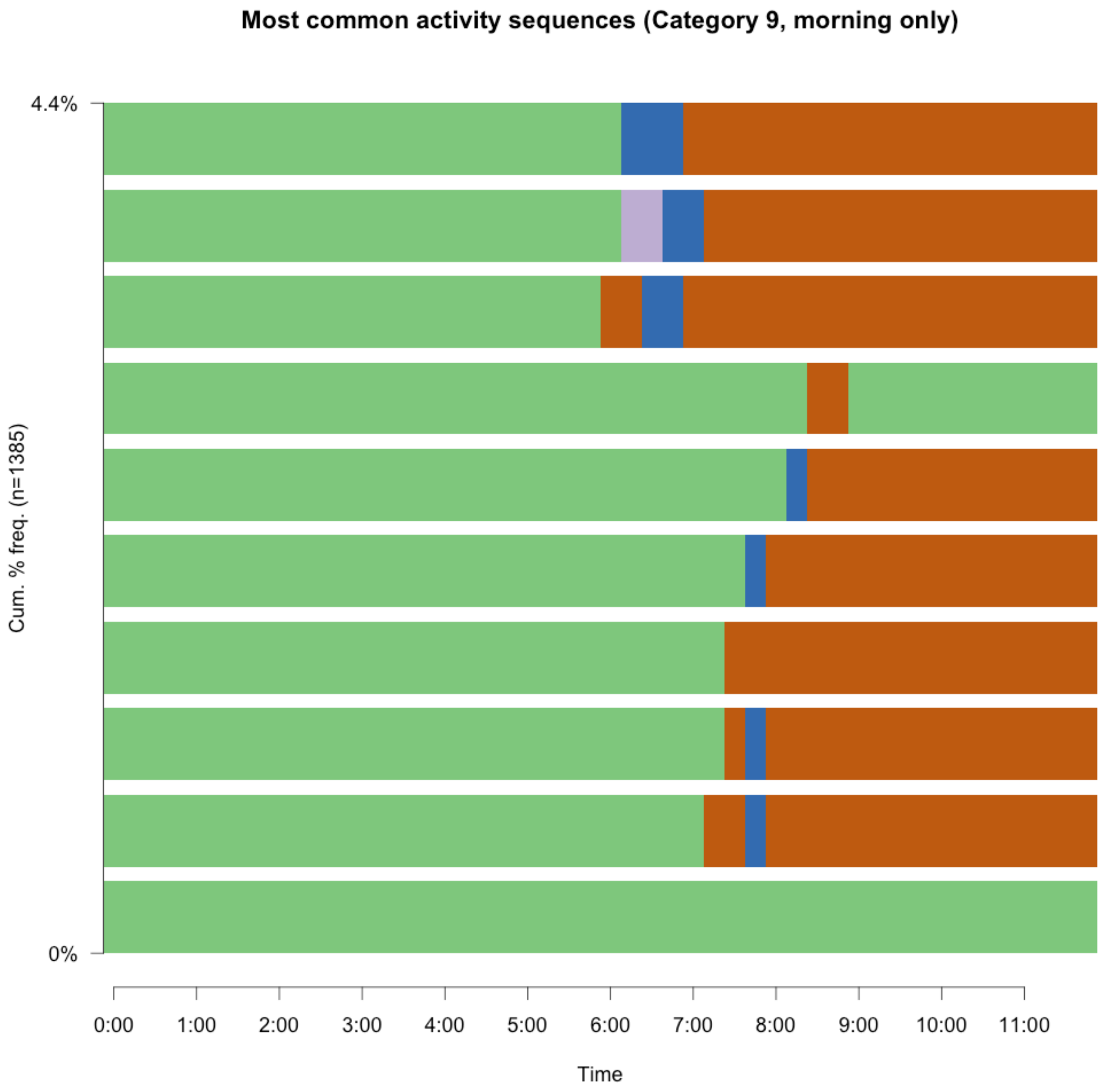

Most common activity sequences (Category 7, afternoon only)
3.2%
0%
Cum. % freq. (n=374)
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Time

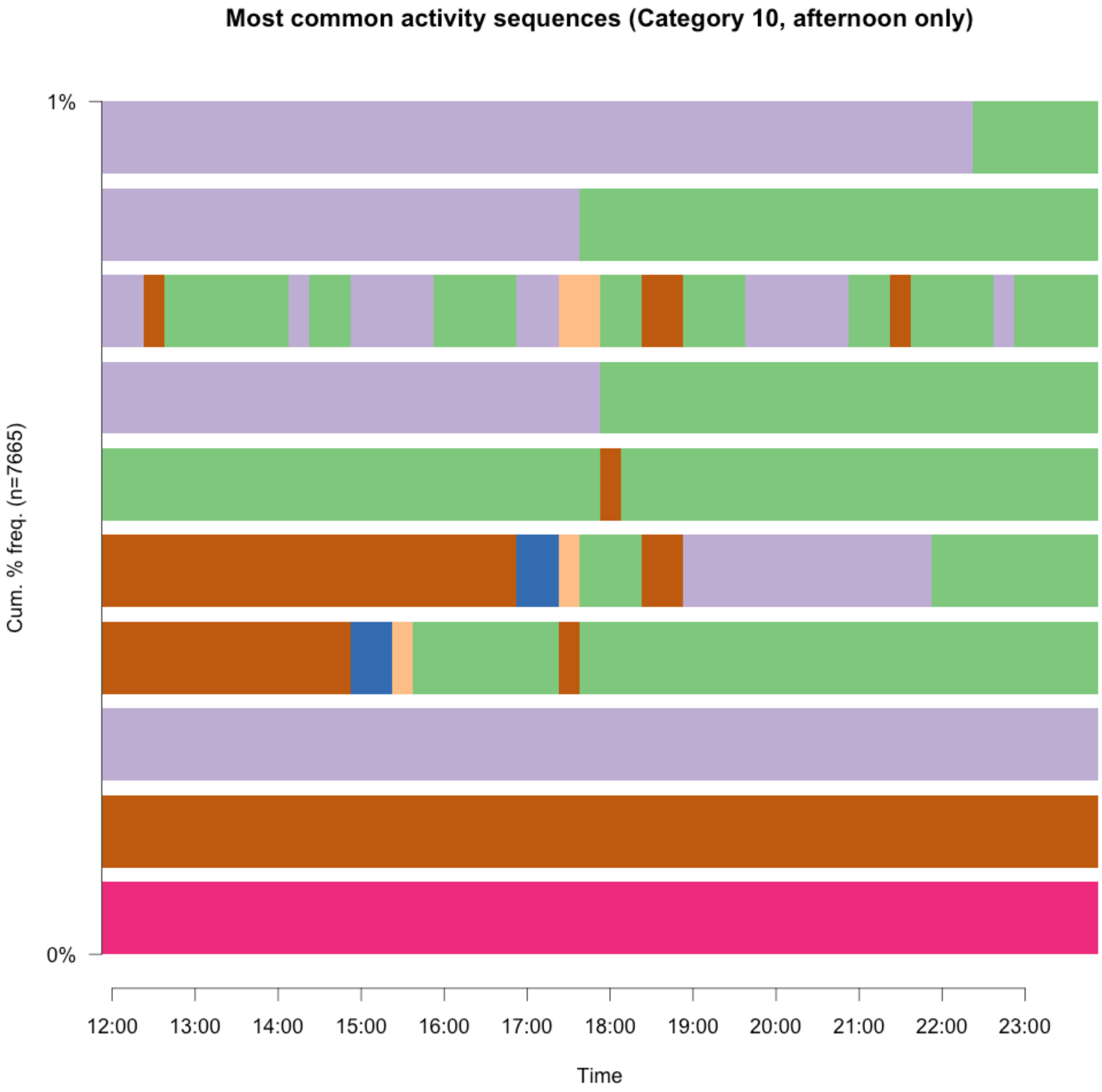

Most common activity sequences (Category 7, morning only)
7.5%
0%
Cum. % freq. (n=374)
0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:00
Time

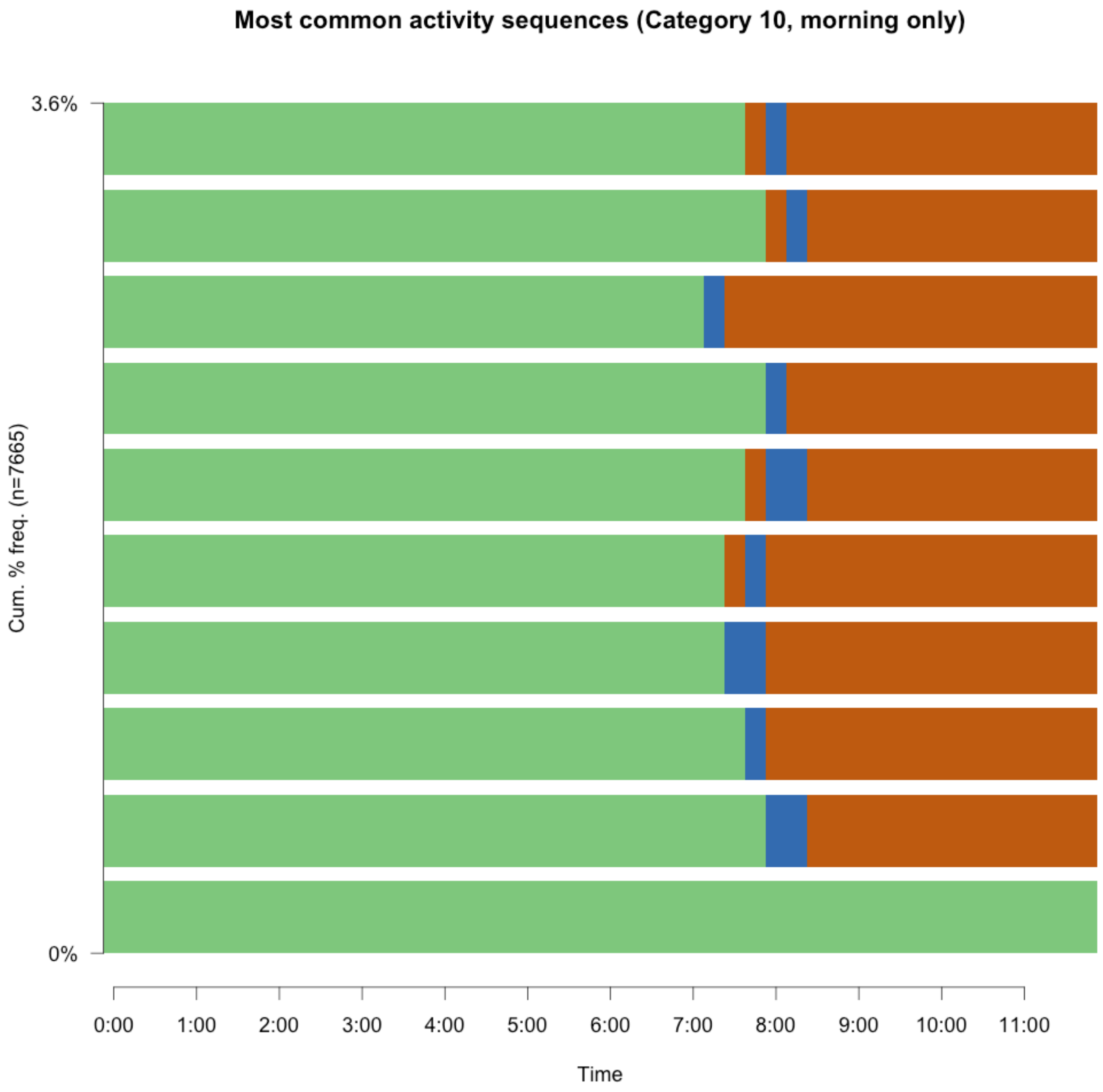

Most common activity sequences (Category 8, afternoon only)
1.6%
0%
Cum. % freq. (n=1988)
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Time


Most common activity sequences (Category 8, morning only)
6.3%
0%
Cum. % freq. (n=1988)
0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:00
Time


Most common activity sequences (Category 9, afternoon only)
1.8%
0%
Cum. % freq. (n=1385)
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Time


Most common activity sequences (Category 9, morning only)
4.4%
0%
Cum. % freq. (n=1385)
0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:00
Time


Most common activity sequences (Category 10, afternoon only)
1%
0%
Cum. % freq. (n=7665)
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Time


Most common activity sequences (Category 10, morning only)
3.6%
0%
Cum. % freq. (n=7665)
0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:00
Time

```
library(haven)
episode_full <-
read_sav("~/Dropbox/Bachelorarbeit/MTUS/Datensätze/MTUS-adult-
episode.sav")
aggregate_full <-
read_sav("~/Dropbox/Bachelorarbeit/MTUS/Datensätze/MTUS-adult-
aggregate.sav")
episode_nl <- episode_full[episode_full$countrya==22,]
episode_uk <- episode_full[episode_full$countrya==37,]


episode_uk_nl <- merge(episode_nl, episode_uk, all=TRUE)

#Filter by year
episode_uk_nl <- episode_uk_nl[episode_uk_nl$survey > 1990,]
library(dplyr)


idmaker = function(vec){
  return(paste(vec, collapse="_"))
}

#In this step, we prepare the aggregate Dataset for merging with the
Episode dataset (to gain additional information about the diarists.)
#In order to do this, we select the most relevant variables to add to
the episode dataset

# Parameters we want to use: (note: Sex & Age are already contained in
HEF dataset)
# --------------------------
# PERSID: unique diarist ID, used for joining te datasets
# HHTYPE: household type
# HHLDSIZE: household number of people
# HLDID: household ID
# INCORIG: household income -> not harmonised, might not use
# INCOME: total household income
# OWNHOME: own/rent/other for household accomodation
# URBAN: Urban/suburban or rural/semi-rural (most likely only consider
cases where this is 1)
# VEHICLE: household vehicle ownership
# FAMSTAT: family status
# CITIZEN: yes/no, might be relevant later
# EMPSTAT: employment status
# EMP: true/false, is employed
# UNEMP: true/false, is unemployed
# STUDENT: true/false, is student
# RETIRED: true/false, might be relevant to correlate with age/wealth
etc.
# WORKHRS: work hours
# OCCUPO: occupation
# SECTOR: public/private employment sector
# EDCAT: harmonised education level
# HEALTH: health status (0 to 4), might be used later

# NUMBER OF CHILDREN TODO
# COHAB TODO
# SPOUSE EMPLOYMENT TODO

aggregate_selected <- aggregate_full %>% select(persid, hldid,
countrya, survey, swave, msamp, hhtype, hhldsize, incorig, income,
ownhome, urban, vehicle, famstat, citizen, empstat, emp, unemp,
student, retired, workhrs, occup, sector, edcat, health)



#Get all unique persid/hldid combinations in the episode dataset
episode_ids <- episode_uk_nl %>% select(persid, hldid, countrya,
survey, swave, msamp)
episode_ids <- unique(episode_ids[,c("persid", "hldid", "countrya",
"survey", "swave", "msamp")])

#Assign an ID to every unique combination of persid & hldid & other
relevant attributes for single-person identification in the episode ID
set

uniqueid = apply(as.matrix(episode_ids[, c("persid", "hldid",
"countrya", "survey", "swave", "msamp")]), 1, idmaker)
episode_ids = cbind(uniqueid, episode_ids)


#Assign an ID to every unique combination of persid & hldid in the
aggregate set
uniqueid = apply(as.matrix(aggregate_selected[, c("persid", "hldid",
"countrya", "survey", "swave", "msamp")]), 1, idmaker)
aggregate_selected = cbind(uniqueid, aggregate_selected)


#Filter aggregate data to only contain unique IDs that appear in the
episode dataset - this means we have exactly as many diarists in both
datasets
aggregate_selected <- subset(aggregate_selected, (uniqueid %in%
episode_ids$uniqueid))

#Remove unused dataframe
rm(episode_ids)

#Remove duplicates caused by multiple diaries of the same person
aggregate_selected <- unique(aggregate_selected)

#Remove duplicate entries for unique IDs (Not needed, but might prevent
errors in rare cases)
aggregate_selected =
aggregate_selected[!duplicated(aggregate_selected$uniqueid),]

#Remove identification variables from aggregate data frame for merging
aggregate_selected <- subset(aggregate_selected, select= -c(persid,
hldid, survey, swave, msamp))



library(plyr)
```

```
#This merges the attributes from aggregate_selected to the
episode_uk_nl data frame.


#Assign an ID to every unique combination of persid & hldid in the
episode set
#(unique ID is used for merging with aggregate set)
uniqueid = apply(as.matrix(episode_uk_nl[, c("persid", "hldid",
"countrya", "survey", "swave", "msamp")]), 1, idmaker)
episode_uk_nl = cbind(uniqueid, episode_uk_nl)

rm(uniqueid)

#Join the two datasets on uniqueid
joined_full <- merge(episode_uk_nl, aggregate_selected, by="uniqueid")

#rename countrya
names(joined_full)[names(joined_full)=="countrya.x"] <- "countrya"


#In this step we prepare the merged data set for analysis.


classificationFinder = function(vec){
  return(paste(vec, collapse=""))
}

classifier <- function(x){
  su=sort(unique(x))
  for (i in 1:length(su)) x[x==su[i]] = i
  return(x)
}



#Filter by age > 18, also removes missing values
joined_filtered <- joined_full[joined_full$age >= 18,]

#Filter by income: no missing values allowed (eliminates approx. 400k
entries)
joined_filtered <- joined_filtered[joined_filtered$income >= 1,]

#Filter by urban (no missing values) --> leaves us with ~11.5k diarists
joined_filtered <- joined_filtered[joined_filtered$urban >= 1,]

#Filter by main (no missing values)
joined_filtered <- joined_filtered[joined_filtered$main >= 0,]

#---- Add indices for grouping ----

#Age Index: >=55 years = 1, otherwise 2
ageIndex = ifelse(joined_filtered$age >= 55, 1, 2)
joined_filtered <- cbind(ageIndex, joined_filtered)

#Urbanicity Index: 1 or 2 (1 = semi-urban, 2 = urban) --> inverse of
the variable "URBAN" in the MTUS dataset
```

```
urbanIndex = ifelse(joined_filtered$urban == 1, 2, 1)
joined_filtered <- cbind(urbanIndex, joined_filtered)

#Wealth index: directly take INCOME variable
wealthIndex = joined_filtered$income
joined_filtered <- cbind(wealthIndex, joined_filtered)

#Remove indices
rm(urbanIndex)
rm(wealthIndex)
rm(ageIndex)


#Assign category ID (1-10)

classificationId = apply(as.matrix(joined_filtered[, c("wealthIndex",
"urbanIndex", "ageIndex")]), 1, classificationFinder)

joined_filtered <- cbind(classificationId, joined_filtered)

#Convert to numeric
joined_filtered$classificationId <-
as.numeric(as.character(joined_filtered$classificationId))

#Remove old stuff
rm(classificationId)

#Handle cases 6 & 7 where we merge categories:
#Replace 212 with 211, replace 122 with 121
joined_filtered$classificationId[joined_filtered$classificationId==212]
<- 211
joined_filtered$classificationId[joined_filtered$classificationId==122]
<- 121

#Rank classification ID: we now have a column with classification from
1 to 10
classification <- classifier(joined_filtered$classificationId)
joined_filtered <- cbind(classification, joined_filtered)

rm(classification)
joined_filtered$classificationId <- NULL
#---- Dataset division ----

#Divide data into separate UK and NL datasets
joined_nl <- joined_filtered[joined_filtered$countrya == 22,]
joined_uk <- joined_filtered[joined_filtered$countrya == 37,]

#In this step, we categorize all 69 main activities into the following
buckets:

#---- Categories: ----

# (1) Work Travel
# (2) Maintenance Travel
# (3) Recreational Travel
# (4) Work
# (5) Maintenance
# (6) Recreation
# UNDEFINED: activity 69

# Short names are as follows:
# (1) travel_work
# (2) travel_maint
# (3) travel_recr
# (4) work
# (5) maint
# (6) recr

typeId <- c(1,1,2,2,3,4,4,4,5,5,6,6,7)
typeName <- c('Work Travel', 'Work Travel', 'Maintenance
Travel','Maintenance Travel', 'Recreational Travel', 'Work','Work',
'Work', 'Maintenance','Maintenance', 'Recreation','Recreation',
'Unknown')
typeShort <- c('travel_work', 'travel_work',
'travel_maint','travel_maint', 'travel_recr', 'work', 'work','work',
'maint','maint','recr', 'recr', 'unkn')

numStart <- c(11,63,62,66,65, 5,12,33,1,18,17,34,69)
numEnd <-   c(11,64,62,68,65,10,16,33,4,32,17,61,69)

#taskNums <-c(
#  c(11,63,64),
#  c(62, 66, 67, 68),
#  c(65),
#  c(5,6,7,8,9,10,12,13,14,15,16,33),
#  c(1,2,3,4,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32),
#
c(17,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,
56,57,58,59,60,61),
#  c(69)
#  )


#Create task type assignment array
taskTypes <- data.frame(typeId, typeName, typeShort, numStart, numEnd)


rm(typeId)
rm(typeName)
rm(typeShort)
rm(numStart)
rm(numEnd)

#Task classification array
taskTypes <- taskTypes[order(taskTypes$numStart),]

#Apply Task types to filtered dataset
data_tasks_classified <- transform(joined_filtered,
typeId=taskTypes$typeId[findInterval(main,
taskTypes$numStart)],typeShort=taskTypes$typeShort[findInterval(main,
taskTypes$numStart)], typeName=taskTypes$typeName[findInterval(main,
taskTypes$numStart)])
```

```
rm(taskTypes)


#---- Activity numbers by category: ----

# Work Travel (travel_work):
# 11
# 63
# 64


# Maintenance Travel (travel_maint):
# 62
# 66
# 67
# 68 (assumed)


# Recreational Travel (travel_recr)
# 65


# Work (work):
# 5
# 6
# 7
# 8
# 9
# 10
# 12
# 13
# 14
# 15
# 16
# 33

# Maintenance (maint)
# 1
# 2
# 3
# 4
# 18-32

#Recreation (recr)
# 17
# 34-61
library(plyr)
library(tibble)

diaryId = apply(as.matrix(data_tasks_classified[, c("uniqueid",
"diary")]), 1, idmaker)

#apply diary IDs to data_tasks_classified (for matching)
data_tasks_classified <- cbind(diaryId, data_tasks_classified)
rm(diaryId)
```

```
episode_individual_diaries <- data_tasks_classified

#Remove variables to create set of unique diaries (one row per diary -
content is added afterwards)
episode_individual_diaries <- subset(
  episode_individual_diaries,
  select= -c(time, clockst, start, end, epnum, main, sec, av, inout,
eloc, ict, mtrav, alone, child, sppart, oad, id, typeId, typeShort,
typeName)
)

episode_individual_diaries <- unique(episode_individual_diaries)


#Add columns for each fifteen minutes
#Format: time_0.00, time_0.25 (= 0:15 o'clock), ..., time_23.75

#Create vector for all times and add to dataset as columns
prefix <- 'time'
suffix <- seq(0, 23.75, length.out=96)
timevector <- paste(prefix, suffix, sep='_')
rm(suffix)

#prefill with "unknown"
episode_individual_diaries[ , timevector] <- 'unkn'
rm(timevector)


#prepare data_tasks_classified for conversion

#round "start" variable to 15min increments
start_rounded <- data_tasks_classified$start / 60
start_rounded <- round_any(start_rounded, 0.25, f=round)


#round "end" variable to 15min increments
end_rounded <- data_tasks_classified$end / 60
end_rounded <- round_any(end_rounded, 0.25, f=round) -0.25


data_tasks_classified <- cbind(data_tasks_classified, start_rounded)
data_tasks_classified <- cbind(data_tasks_classified, end_rounded)
rm(start_rounded)
rm(end_rounded)


#Prepare merging data frame
episode_premerge <- data_tasks_classified[, c("diaryId",
"start_rounded", "end_rounded", "typeId", "typeShort", "typeName")]
#Rename columns
names(episode_premerge)[names(episode_premerge)=="start_rounded"] <-
"start"
names(episode_premerge)[names(episode_premerge)=="end_rounded"] <-
"end"
```

```
#add strings for start/end time
episode_premerge$startTime <- paste(prefix, episode_premerge$start,
sep='_')
episode_premerge$endTime <- paste(prefix, episode_premerge$end,
sep='_')

rm(prefix)

#remove incomplete cases
episode_premerge <- episode_premerge[complete.cases(episode_premerge),
]
episode_individual_diaries <-
episode_individual_diaries[complete.cases(episode_individual_diaries),
]

#backup
eid_backup <- episode_individual_diaries

#Export CSVs
write.csv(
      episode_individual_diaries,

file='~/Dropbox/Bachelorarbeit/Code/Data/episode_individual_diaries.csv
',
      row.names = FALSE
  )
write.csv(
  data_tasks_classified,
  file='~/Dropbox/Bachelorarbeit/Code/Data/data_tasks_classified.csv',
  row.names = FALSE
)
write.csv(
  episode_premerge,
  file='~/Dropbox/Bachelorarbeit/Code/Data/episode_premerge.csv',
  row.names = FALSE
)
write.csv(
  aggregate_selected,
  file='~/Dropbox/Bachelorarbeit/Code/Data/aggregate_selected.csv',
  row.names = FALSE
)

#---- Actual merging logic ----
#Arranges data sequentially in the final dataset
(episode_individual_diaries)
# WARNING: MIGHT TAKE A VERY LONG TIME depending on your computer!

#Do for each row (e.g. episode)
from <- 1
incr <- 992
to <- incr
max <- nrow(episode_individual_diaries)
iter <- 1

#create empty data frame
```

```
finalDiaries <- episode_individual_diaries[FALSE,]

#Iterate 1000 at a time, then save
for(k in 1:iter){
  start_time <- Sys.time()
  diaries <- foreach(i=from:to, .combine=rbind) %do%{

    #Prepare single row to be filled as data frame
    row <- (data.frame(episode_individual_diaries[i, ]))
    id <- toString(row$diaryId)


    #Get rows for this diary in premerge set
    entries <- episode_premerge[episode_premerge$diaryId == id,]
    #Order by start time
    entries <- entries[order(entries$start),]

    #For each activity, enter into row
    for(i in 1:nrow(entries)){
      #Start
      entry <- (data.frame(entries[i, c('diaryId', 'start', 'end',
'startTime', 'endTime','typeShort')]))
      st <- entry$startTime
      et <- entry$endTime
      row[[st]] <- entry$typeShort
      row[[et]] <- entry$typeShort

      s <- as.numeric(entry$start)
      e <- as.numeric(entry$end)

      #longer activity: fill length
      if(e >= s+0.5){
        startindex <- which(colnames(row)==st)
        endindex <- which(colnames(row)==et)
        for(j in ((startindex+1):(endindex-1))){
          row[,j] <- entry$typeShort
        }
      }
    }
    #add row to finalDiaries
    row
  }

  end_time <- Sys.time()
  print(k)
  print(end_time - start_time)

  from <- to+1
  to <- to+incr

  finalDiaries <- rbind(finalDiaries, diaries)
}

rm(i)
rm(d)
rm(e)
rm(end)
rm(end_time)
rm(endindex)
rm(et)
rm(from)
rm(id)
rm(incr)
rm(iter)
rm(j)
rm(k)
rm(max)
rm(s)
rm(st)
rm(start)
rm(start_time)
rm(startindex)
rm(to)
rm(type)
rm(row)
rm(entry)
rm(entries)

#Export CSV
write.csv(
  finalDiaries,
  file='~/Dropbox/Bachelorarbeit/Code/Data/finalDiaries.csv',
  row.names = FALSE
)

library(TraMineR)
library(dplyr)
library(ggplot2)
library(foreach)
library(data.table)
library(reshape2)
library(gridExtra)
library(grid)

cleanup=theme(panel.grid.major = element_blank(),
              panel.grid.minor = element_blank(),
              panel.background = element_blank(),
              axis.line = element_line((color="black")))

dataset_analysis <- finalDiaries

#Preparation:
#Exclude diaries that contain only maintenance or unknown
dataset_analysis <-
dataset_analysis[rowSums(sapply(dataset_analysis[41:136], '%in%',
c('recr', 'travel_maint', 'travel_recr', 'travel_work', 'work'))) > 0,]

#Define labels
dataset_analysis.labels <- c(
  "Maintenance",
  "Recreation",
  "Maintenance Travel",
  "Recreational Travel",
  "Work Travel",
  "Unknown",
  "Work"
)

#Define codes
dataset_analysis.scode <- c(
  "maint",
  "recr",
  "travel_maint",
  "travel_recr",
  "travel_work",
  "unkn",
  "work"
)

#Define sequences
dataset_analysis.seq <- seqdef(dataset_analysis, 41:136, states =
dataset_analysis.scode, labels = dataset_analysis.labels, xtstep = 7)


#Automatically generate category datasets
#Define sequences in sub-datasets
for(i in 1:10){
  assign(paste("dataset_category_", i, sep = '', '.seq'),
seqdef(dataset_analysis[dataset_analysis$classification == i,], 41:136,
states = dataset_analysis.scode, labels = dataset_analysis.labels,
xtstep = 7))
}

categoryFrames = list(
  dataset_category_1.seq,
  dataset_category_2.seq,
  dataset_category_3.seq,
  dataset_category_4.seq,
  dataset_category_5.seq,
  dataset_category_6.seq,
  dataset_category_7.seq,
  dataset_category_8.seq,
  dataset_category_9.seq,
  dataset_category_10.seq
  )
#10 Most frequent sequences - all categories
seqfplot(dataset_analysis.seq, with.legend = T, border = NA, main =
"Most common activity sequences (all Categories)", pbarw=FALSE)

dataset_analysis.seq_morning <- dataset_analysis.seq[,1:48]
dataset_analysis.seq_afternoon <- dataset_analysis.seq[,49:96]

seqfplot(dataset_analysis.seq_afternoon, with.legend = T, border = NA,
main = "Most common activity sequences (all Categories, afternoon
only)", pbarw=FALSE)
seqfplot(dataset_analysis.seq_morning, with.legend = T, border = NA,
main = "Most common activity sequences (all Categories, morning only)",
pbarw=FALSE)
```

```
#Visualize transition rates / probabilities
transitionmatrix <- seqtrate(dataset_analysis.seq, sel.states = NULL,
time.varying = FALSE, weighted = TRUE,
         lag = 1, with.missing = FALSE, count = FALSE)
transitionmatrix <- round(transitionmatrix, 4)
transitionmatrix <- transitionmatrix * 100
melted_transitionmatrix <- melt(transitionmatrix)

ggplot(data = melted_transitionmatrix, aes(Var2, Var1, fill = value))+
  labs(fill = NULL, x = NULL, y = NULL, title = '  Transition
probabilities (all categories)') +
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "white", high = "gray37", mid =
"gray50",midpoint = 50, limit = c(0,100), space =
"Lab",name="Transition Probability\n(in %)") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, vjust = 1,size = 12,
hjust = 1))+
  coord_fixed()+
  geom_text(aes(Var2, Var1, label = paste(value, '%', sep='')), color =
"black", size = 4) +
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    axis.ticks = element_blank(),
    legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                               title.position = "top", title.hjust =
0.5))
rm(transitionmatrix)
rm(melted_transitionmatrix)

#generate 10 Most frequent by category (and export)
for(i in 1:10){
  png(paste('Graphs/mostcommon_category_', i, '.png', sep=''),
width=1000, height=1000, pointsize = 18)
  seqfplot(categoryFrames[[i]], with.legend = T, border = NA, main =
paste("Most common activity sequences (Category ", i, ")", sep=""),
pbarw=FALSE)
  dev.off()

  #Morning sequences
  png(paste('Graphs/split/mostcommon_category_', i, '_morning.png',
sep=''), width=1000, height=1000, pointsize = 18)
  seqfplot(categoryFrames[[i]][,1:48], with.legend = F, border = NA,
main = paste("Most common activity sequences (Category ", i, ", morning
only)", sep=""), pbarw=FALSE)
  dev.off()

  #Afternoon sequences
  png(paste('Graphs/split/mostcommon_category_', i, '_afternoon.png',
sep=''), width=1000, height=1000, pointsize = 18)
  seqfplot(categoryFrames[[i]][,49:96], with.legend = F, border = NA,
main = paste("Most common activity sequences (Category ", i, ",
afternoon only)", sep=""), pbarw=FALSE)
  dev.off()
}

#State distribution over a day
seqdplot(dataset_analysis.seq, with.legend = T, border = NA, main =
"State distribution plot (all categories)")

for(i in 1:10){
  png(paste('Graphs/statedistribution_category_', i, '.png', sep=''),
width=1000, height=1000, pointsize = 18)
  seqdplot(categoryFrames[[i]], with.legend = T, border = NA, main =
paste("State distribution plot (category", i, ')'))
  dev.off()
}

#TEST: generate avg day for a category TODO remove
test <- {}
for(i in 1:96){

  test[i] <- tail(names(sort(table(dataset_category_6.seq[[i]]))), 1)
}
View(test)

#Create data frame with duration for each activity type for each
category (rounded to .01 hours)
travelTimes <- foreach(i=1:10, .combine=rbind) %do%{
  category <- categoryFrames[[i]]

  res <- data.frame(table(category[[1]]))
  res$sum <- res[[2]]
  for(j in 2:96){
    res <- merge(res,table(category[[j]]),by='Var1')
    res$sum <- res$sum + res[[j+1]]
  }
  res$sum <- round(as.numeric(res$sum / nrow(category) / 4), 2)
  res <- res[c("Var1", "sum")]
  res <- res[res$Var1 != "*",]
  res <- res[res$Var1 != "%",]
  res <- transpose(res)
  colnames(res) <- as.character(unlist(res[1,]))
  res = res[-1, ]

  res
}
rm(res)
rownames(travelTimes) <- c(1:nrow(travelTimes))
travelTimes <- data.frame(travelTimes)
travelTimes <- travelTimes[,c(1,2,7,3,4,5,6)]
travelTimes$category <- c(1:10)

#Generate stacked bar plot: all activities
```

```
travelTimes.all <- melt(travelTimes,id.vars = "category")
ggplot(travelTimes.all, aes(y=as.numeric(value), x=factor(category),
fill=factor(variable))) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = value), position = position_stack(vjust = 0.5))
+
  ylab("Average time spent (in hours)") +
  scale_fill_discrete(name="Activity Type") +
  xlab("Category #")

#Generate stacked bar plot: travel only
travelTimes.travelOnly <- travelTimes[,c(4,5,6,8)]
travelTimes.travelOnly <- melt(travelTimes.travelOnly,id.vars =
"category")
ggplot(travelTimes.travelOnly, aes(y=as.numeric(value),
x=factor(category), fill=factor(variable))) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = value), position = position_stack(vjust = 0.5))
+
  ylab("Average time spent (in hours)") +
  scale_fill_discrete(name="Activity Type") +
  xlab("Category #")


#Generate & visualize matrix for values (calculated externally)
type <- c('travel_maint', 'travel_recr', 'travel_work')
avg <- c(0.825, 0.069, 0.416)
sd <- c(0.07, 0.024, 0.22)
var <- c(0.0046, 0.0006, 0.0468)
travelInfo <- data.frame(type, avg, sd, var)
rownames(travelInfo) <- c('Maintenance Travel', 'Recreational Travel',
'Work Travel')
colnames(travelInfo) <- c('Average', 'Standard Deviation', 'Variance')
travelInfo$type <- NULL
options("scipen"=100, "digits"=4)

grid.table(
  travelInfo, theme=ttheme_default(core=list(fg_params=list(hjust=0,
x=0.1)),
                                   rowhead=list(fg_params=list(hjust=0,
x=0)))
  )

rm(travelInfo)


#=== DATA OVERVIEW ====

#Diarists per category
diarists <- finalDiaries %>% select(2,6)
diarists <- unique(diarists)
diaristCount <- count(diarists, classification)
diaristCount <- diaristCount[order(diaristCount$n),]
rm(diarists)
```

```
ggplot(data = diaristCount, aes(x = 0, y = reorder(classification,n),
fill = factor(classification))) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = n), position = position_stack(vjust = 0.5)) +
  scale_x_continuous(expand = c(0,0)) +
  labs(fill = 'Category #', x = NULL, y = NULL, title = 'Diarists per
category') +
  coord_polar(theta = "y") +
  theme_minimal()
```

```
library(plyr)
library(ggplot2)

#Apply unique person IDs to entire episode dataset
uniqueid = apply(as.matrix(episode_full[, c("persid", "hldid",
"countrya", "survey", "swave", "msamp")]), 1, idmaker)
episode_analysis <-cbind(uniqueid, episode_full)

#Remove uneeded identification variables from episode data frame
episode_analysis <- subset(episode_analysis, select= -c(persid, hldid,
swave, msamp))
episode_analysis_recent <- episode_analysis[episode_analysis$survey >=
1990,]

#Create data frame with only one entry per person
episode_analysis_unique_recent =
episode_analysis_recent[!duplicated(episode_analysis_recent$uniqueid),]
episode_analysis_unique =
episode_analysis[!duplicated(episode_analysis$uniqueid),]

rm(episode_analysis)
rm(episode_analysis_recent)

#---- ANALYSIS: DIARISTS PER COUNTRY -----

#Data frame counting which country has how many entries
country_count <- count(episode_analysis_unique_recent$countrya)

#Add country name column
country_labels <- c("Armenia", "Australia", "Austria", "Belgium",
"Brazil", "Bulgaria", "Canada", "China", "Denmark", "Estonia",
"Finland", "France", "Germany", "Hungary", "India", "Ireland",
"Israel", "Italy", "Japan", "Latvia", "Lithuania", "Netherlands", "New
Zealand", "Norway", "Pakistan", "Poland", "Portugal", "Republic of
Korea", "Romania", "Serbia/Yugoslavia", "Slovak
Republic/Czechoslovakia", "Slovenia/Yugoslavia", "South Africa",
"Spain", "Sweden", "Turkey", "United Kingdom", "USA")
country_names <- factor(country_count$x,1:38,country_labels)

country_count <- cbind(country_names, country_count)

#Rename Columns
names(country_count)[names(country_count)=="country_names"] <-
"country"
```

```
names(country_count)[names(country_count)=="x"] <- "code"
names(country_count)[names(country_count)=="freq"] <- "diarists"

#Remove unneeded stuff
rm(country_labels)
rm(country_names)

#Change order
country_count <- country_count[c("code", "country", "diarists")]

#Sort descending
country_count <- country_count[order(country_count$diarists),]

#Add percentage column (accurate to .1%)
country_count$perc <- (round(1000 * country_count$diarists /
sum(country_count$diarists)))/10

#Create pie chart by country & number of diarists

png('Data Overview/diarists_per_country.png')
ggplot(data = country_count, aes(x = 0, y = diarists, fill = country))
+
  geom_bar(stat = "identity") +
  geom_text(aes(label = diarists), position = position_stack(vjust =
0.5)) +
  scale_x_continuous(expand = c(0,0)) +
  labs(fill = 'Country', x = NULL, y = NULL, title = 'Diarists per
country', subtitle = '(In episode dataset, after 1990)') +
  coord_polar(theta = "y") +
  theme_minimal()
dev.off()

#---- ANALYSIS: DIARISTS PER YEAR ----

#Data frame counting which year has how many entries
year_count <- count(episode_analysis_unique$survey)

#Rename Columns
names(year_count)[names(year_count)=="x"] <- "year"
names(year_count)[names(year_count)=="freq"] <- "diarists"

png('Data Overview/diarists_per_year.png')
#Create bar plot per year / number of diarists
barplot(year_count$diarists,
        main="Diarists per year",
        xlab="Year",
        las = 2,
        ylab="Number of diarists",
        names.arg=year_count$year,
        )

dev.off()
```