Travel demand modeling with smart card tap-in information

EIDEIC

Borja Alonso Oreña
Director

Juan Benavente Ponce
PhD student

José Luis Moura Berodia
Director

Department of Transportation and Projects and Processes Technology
University of Cantabria

Santander. May 22, 2020

Public transport demand modeling with smart card tap-in information

1. Introduction
2. Objectives
   - Data cleansing and imputation
   - Trip chaining
   - Public transport demand models
3. Dataset analysis
   - Bus stops
   - AFC
   - AVL
4. Methodology
   - Pre-processing
   - Trip chaining
   - Demand modeling
5. Results sample
6. Skills and Capacities Assessment

Public transport demand modeling with smart card tap-in information

1. Introduction
2. Objectives
    - Data cleansing and imputation
    - Trip chaining
    - Public transport demand models
3. Dataset analysis
    - Bus stops
    - AFC
    - AVL
4. Methodology
    - Pre-processing
    - Trip chaining
    - Demand modeling
5. Results sample
6. Skills and Capacities Assessment

Introduction

Smart Card data in public transport analysis

Strengths

- Low cost per datum (byproduct of access control).
- Adequate to create key performance indicators of a public transport system.

Opportunities

- Real time data.
- Regional or nation-wide integration.
- Multimodal transport.
- Data fusion.

Threats

- Free, open public transport.
- More stringent privacy requirements.

Weaknesses

- Sample bias.
- Data problems.
- No tap-out.

Public transport demand modeling with smart card tap-in information

1 Introduction

2 Objectives

- Data cleansing and imputation
- Trip chaining
- Public transport demand models

3 Dataset analysis

- Bus stops
- AFC
- AVL

4 Methodology

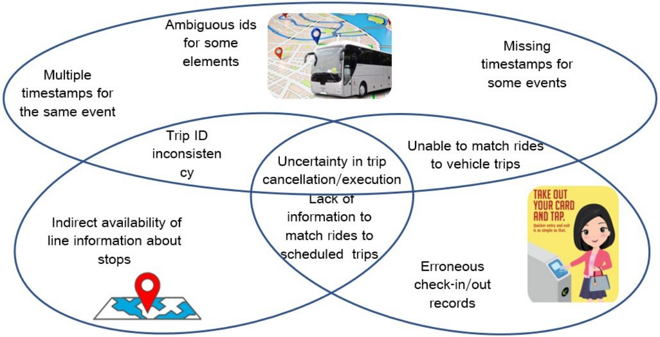
- Pre-processing
- Trip chaining
- Demand modeling

5 Results sample

6 Skills and Capacities Assessment

Objectives

Data cleansing and imputation



Ambiguous ids for some elements

Missing timestamps for some events

Multiple timestamps for the same event

Trip ID inconsistency

Uncertainty in trip cancellation/execution

Unable to match rides to vehicle trips

Lack of information to match rides to scheduled trips

Indirect availability of line information about stops

Erroneous check-in/out records

TAKE OUT YOUR CARD AND TAP.
Quicker entry and exit is as simple as that.

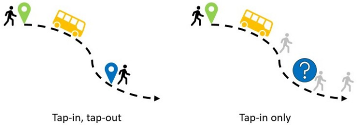
Common smart card data issues

Goal

Lay out a new methodology to infer services and load profiles from the ticketing and bus position records in public transport systems.

Objectives

Trip chaining



Tap-in, tap-out

Tap-in only

Goal

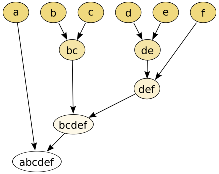
Apply and improve the trip chaining method to infer each leg of a passenger's journey. Use the results to:

- Build vehicle load profiles.
- Create origin-destination matrices.

Objectives

Demand models comparison

Hierarchical agglomerative clustering



Dendogram, *(Stathis Sideris, 2005)*

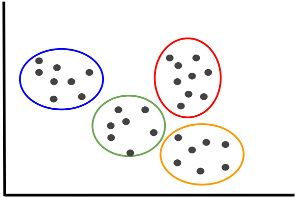
$$F\left(A,B\right)=\sqrt{\mathrm{Tr}\left(\left(A-B\right)^{T}\left(A-B\right)\right)}$$

$$d_1\left(A,B\right)=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|a_{ij}-b_{ij}\right|}$$

$$d_2\left(A,B\right)=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(a_{ij}-b_{ij}\right)^2}$$

Measurement of the similarity between matrices

Unsupervised machine learning clustering



Set of data points, already labeled (Google problem framing course)

Goal

Compare the patterns found by each method in the evolution of public transportation demand with each other, and with the planning strategies of the service operator.

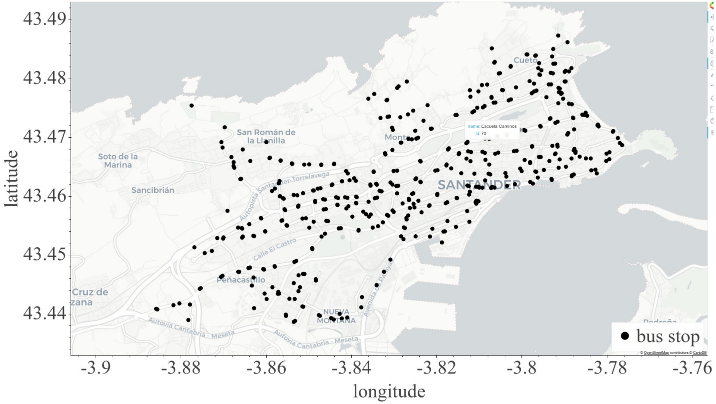
Public transport demand modeling with smart card tap-in information

1. Introduction
2. Objectives
   - Data cleansing and imputation
   - Trip chaining
   - Public transport demand models
3. Dataset analysis
   - Bus stops
   - AFC
   - AVL
4. Methodology
   - Pre-processing
   - Trip chaining
   - Demand modeling
5. Results sample
6. Skills and Capacities Assessment

Dataset analysis

Santander

The city



450 Bus stops

Dataset analysis

Automated fare collection (AFC)

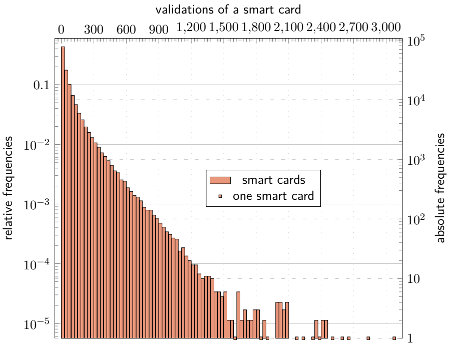
Column	***Description***
IDSesion	Data transfer session id.
IDViaje	AFC service id.
Instante	Instant the passenger's card was read.
NodoOrg	Id of the bus where the user boarded the bus.
Título	Ticket type: 1: Youth; 2 and 5: Standard; … 3: Elderly: 8: Large Family; 9 & 10: metálico; …).
Tarifa	Rate (0: free; 9: all others).
Transbordo	Whether the ride is considered a transfer.
Viajeros	Number of passengers validated in this operation.
Tarjeta	Card ID.
TipoTarjeta	Card type: 0: Ti10, no card nmbr; 1: most of the cards; 2: old contact smart; 8: Ten-numbers cards; Ti2: <null> (cash payments); ….
Saldo	Balance (€).
Linea	Line id.
Coche	Vehicle id.
Sublinea	Subline id.
Ruta	Direction.

raw AFC information

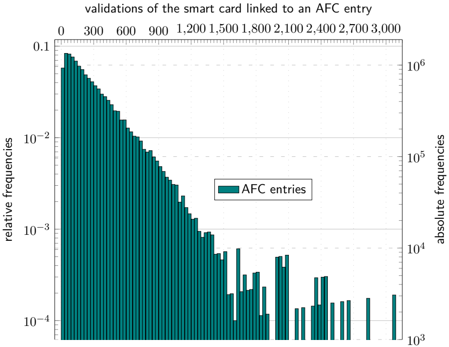
- 1 year data.
- 178 247 smart cards.
- 17 470 526 records ⇒17 729 665 legs:
  - 9 % cash payments.
  - 91 % smart card validations. Of these:
    - 12 % elderly.
    - 6 % youth.
- Vehicle id not linked to AVL.
- More reliable than AVL.

Dataset analysis

Automated Fare Collection



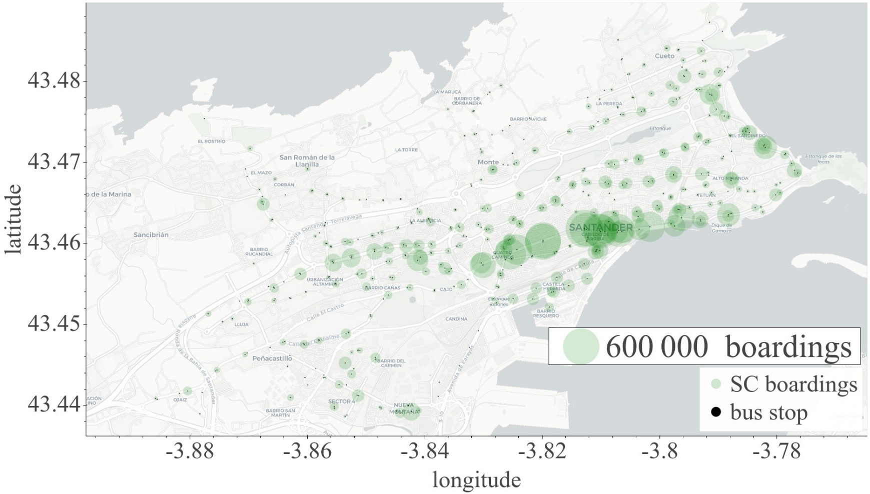
Distribution of the uses of a smart card during a year



Distribution of the uses through the year of the smart card utilized in a transaction

Dataset analysis

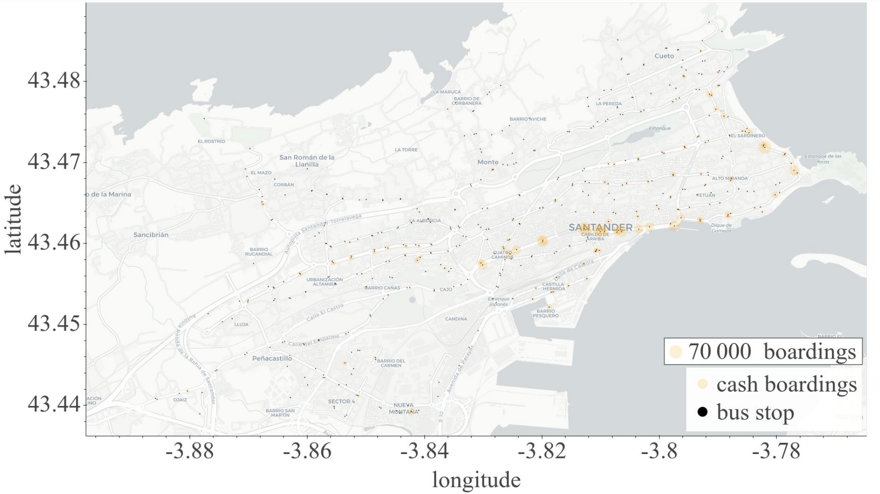
Automated Fare Collection



Smart card ingresses

Dataset analysis

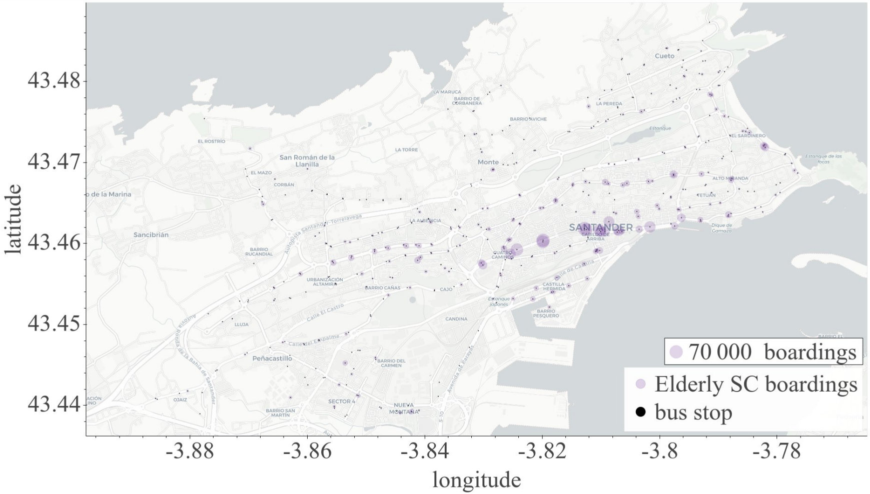
Automated Fare Collection



Cash paying ingresses

Dataset analysis

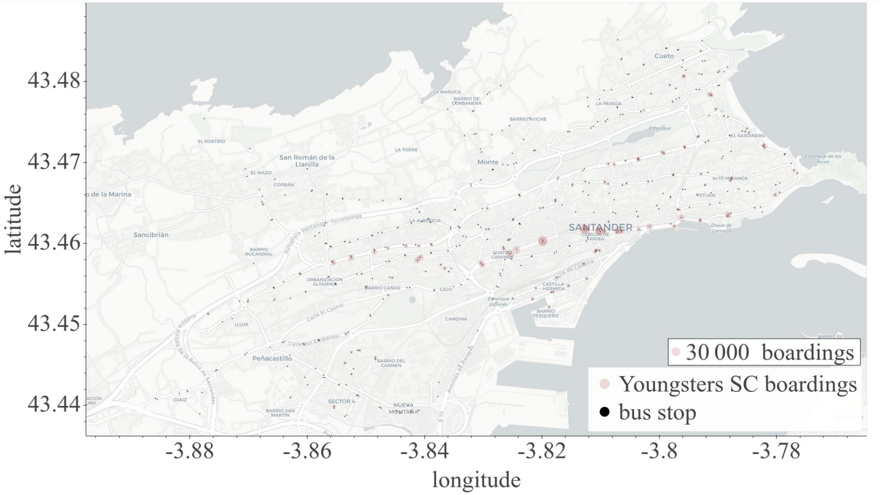
Automated Fare Collection



Elderly boardings

Dataset analysis

Automated Fare Collection



Youth boardings

Dataset analysis

Automatic Vehicle Location (AVL)

Column	***Description***
fecha	The date of the PT schedule in effect.
línea	Line id.
sublínea	Subline id.
coche	Vehicle id.
viaje	Service id during the PT schedule in effect.
instante	Instant when doors were opened at a stop.
parada	Bus stop id.
tparada	How long the doors where open (seconds).
psuben	Number of boarding passengers.
pbajan	Number of alighting passengers.

raw AVL information

- 1 year data.
- 112 buses, though 60 account for 93 % of all entries.
- 12 412 884 rows.
- 35 distinct line ids.
- One reading each time the bus opens its doors or passes by the stop.
- Prevalent issues:
  - Missing or erroneous entries.
  - Unreliable boarding and alighting.
  - Multiple readings linked to same visit.
  - Duplicate rows.

1. Introduction
2. Objectives
   - Data cleansing and imputation
   - Trip chaining
   - Public transport demand models
3. Dataset analysis
   - Bus stops
   - AFC
   - AVL
4. Methodology
   - Pre-processing
   - Trip chaining
   - Demand modeling
5. Results sample
6. Skills and Capacities Assessment

Pre-processing

Automated Fare Colection

Goal

AFC data is used to complete missing AVL entries. In order to do so:

1 row per payment or validation → 1 row per bus visit

Obstacle

Grouping criteria based on just one aspect of the data are not good enough:

- Time: There is not a clear-cut threshold value of the interval between validations of the same boarding group.
- Following each vehicle as it performs services (line, subline, vehicle, IDViaje): Well defined, but IDViaje is not reliable (sometimes it does not change between one run and the next).

Pre-processing

Automated Fare Colection

Chosen approach

Combine both perspectives:

- Group entries by *line*, *subline* and *vehicle* (a, u, v); and identify successive entries linked to the same *bus stop* (b).
- Break those groups where the interval between entries is higher than a threshold parameter (h). (e.g. $h = 30\,\text{min}$).

stop	***Idx. over each (a, u, v) ordered by n***	**-**	***Idx. over each (a, u, v, b) ordered by n***	**=**	***Visit group number***	***stop***	***validation instant (n)***	***Δn***	***¿Δn ≤ 30 min?***	***group change***	***group id***	***stop***
						...						
A	1	-	1	=	0	A	03-24 12:31:23	<null>	✗	1	987	A
B	2	-	1	=	1	B	03-24 12:33:56	00:02:31	✓	0	987	B
C	3	-	1	=	2	C	03-24 12:36:14	00:02:18	✓	0	987	C
C	4	-	2	=	2		03-24 12:36:16	00:00:02	✓	0	987	
D	5	-	1	=	4	D	03-24 12:37:24	00:01:10	✓	0	987	D
E	6	-	1	=	5	E	03-24 12:39:44	00:01:20	✓	0	987	E
						...						
B	45	-	2	=	43	B	03-24 13:49:28	00:01:44	✓	0	987	B
C	46	-	3	=	43	C	03-24 13:51:33	00:02:03	✓	0	987	C
D	47	-	2	=	45	D	03-24 13:54:02	00:02:29	✓	0	987	D
D	48	-	3	=	45		03-24 13:44:03	00:00:01	✓	0	987	
D	49	-	4	=	45		03-24 23:05:49	09:11:46	✗	1	988	D
E	50	-	2	=	48	E	03-24 23:05:49	09:11:46	✓	0	988	E
						...						

AFC pre-processing procedure

Pre-processing

Automated Fare Colection

Column	***Type***	***Description***
first_ticket	timestamp w/o TZ	Instant of the earlist tap-in.
last_ticket	timestamp w/o TZ	Instant of the latest tap-in.
id	integer w/o TZ	Id.
avl_real	integer	Id of linked avl_coalesced entry.
avl_inferred	integer	Id of linked avl_inferred entry.
bus_stop	integer	Bus stop id.
line	smallint	Line id.

Boarding groups characterization

This table has 5 210 074 rows, 30 % of *afc* table.

Pre-processing

Automatic Vehicle Location

Goal

Define the services that have been offered by the operator.

Obstacles

- Multiple entries linked to a single visit of a bus to a stop.
- Missing information.
- Incorrect entries.

Procedure

1. Define raw trajectories.
2. Coalesce *AVL* entries and split raw trajectories.
3. Template sequences.
4. Re-create transportation offer.

AVL pre-processing

Defining raw trajectories

Goal

Identify each service, as recorded by the AVL system (imperfect).

How

AVL records that share sched_day, line, subline, vehicle, and service number values (s, a, u, v, c) are part of the same raw trajectory (r):

With:	r	:	*raw trajectory* id	$r \in [1 \ldots 424\,428]$
	s	:	*schedule* date	time, 1 day resolution
	u	:	*subline* number	$u \in \mathbb{Z}_{\geq 0}$
	c	:	*service number*	$c \in \mathbb{Z}_{\geq 0}$
	i, j	:	*avl* row *ids*	$i, j \in [1 \ldots 12\,412\,856]$

Then: $r_i = r_j \iff s_i = s_j \wedge a_i = a_j \wedge u_i = u_j \wedge v_i = v_j \wedge c_i = c_j$

AVL pre-processing

Defining raw trajectories

Extra information

The beginning and end of each raw trajectory is also stored:

$(q_r)_k$	:	instant	$(q_r)_k \in Q_r$. The time the bus doors opened.
k	:	index	$k \in [1 \ldots \lvert Q_r \rvert]$
Q_r	:	instants	Set of door opening timestamps for raw trajectory r.
β_r	:	instant	First door opening event for r. $\beta_{r_x} = \min\left(Q_{r_x}\right) = (q_r)_1$
γ_r	:	instant	Last door opening event for r. $\gamma_{r_x} = \max\left(Q_{r_x}\right) = (q_r)_{\lvert Q_r \rvert}$

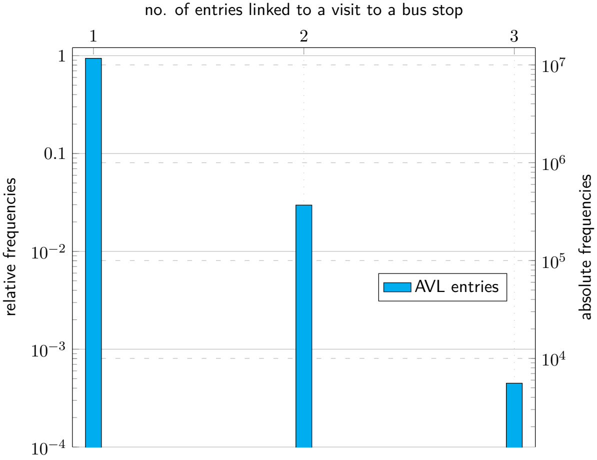
Column	***Type***	***Description***
duration	interval	How long the trip took.
start	timestamp w/o TZ	Arrival at first stop.
conclusion	timestamp w/o TZ	Last instant at the final stop.
sched_day	date	Date of the active PT schedule.
id	integer	Id.
sequence	integer	Stops sequence id.
line	smallint	Line id.
subline	smallint	Subline id.
vehicle	smallint	Vehicle id.
service_number	smallint	Service id during the PT schedule.
bus_stops	smallint	No. of stops made.

Raw trajectories table

AVL pre-processing: coalesce entries and raw trajectories splitting

The issues

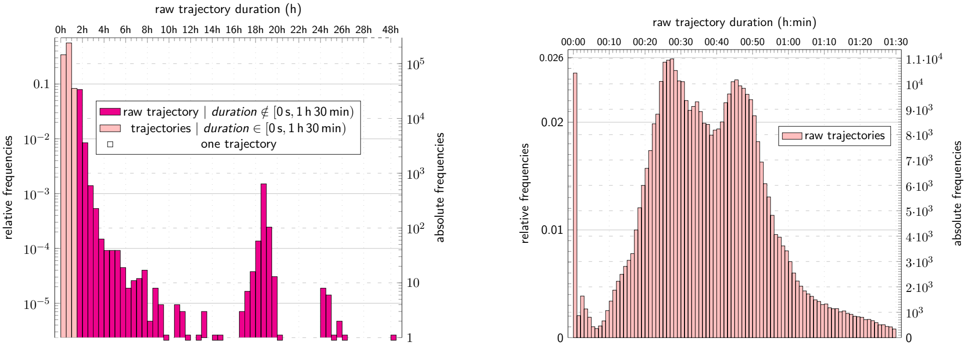
- A significant part of AVL entries correspond to 2 or 3 consecutive readings from a bus at a stop (5 % registering different information, probably due to opening the doors more than once; and 1 % of perfect duplicates).
- Sometimes the system does not properly identify when a service ends and the next one begins.



How many AVL rows represent a visit to a stop

AVL pre-processing

Coalesce entries and raw trajectories splitting



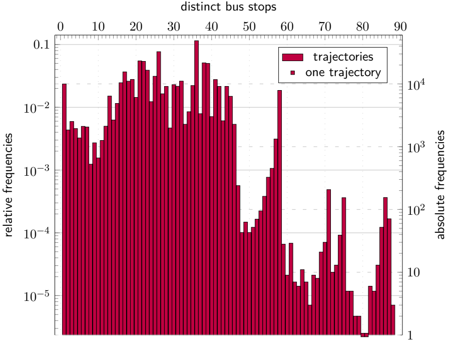
Whole *duration* range

Durations between 0 s and 1 h 30 min

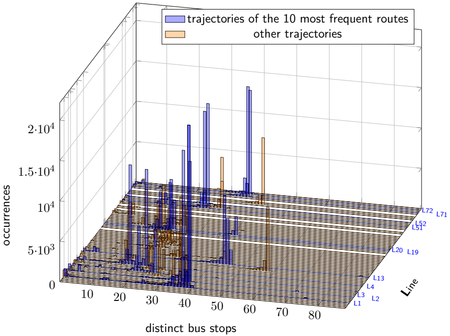
Duration of the raw trajectories

AVL pre-processing

Coalesce entries and raw trajectories splitting



Distinct stops visited by a raw trajectory



Distinct stops of a raw trajectory, by route

AVL pre-processing

Coalesce entries and raw trajectories splitting

Goals

- Condense the data from a single visit from several rows into a single one.
- Split raw trajectories where the system failed to identify the end of a service.

How

- Identify consecutive stops of a trajectory, and split when the gap between entries is too long. Similar to what was done for AFC, but:
  - If the logging system works flawlessly, all bus stops of the offered services appear.
  - Each entry provides two temporal values:
    - The instant the doors were opened or when the bus passed by.
    - How long the doors of the bus were open.
  - Different threshold value (15 min).
- Combine their information.

AVL pre-processing

Coalesce entries and raw trajectories splitting

stop	*Idx. over each r ordered by n*	-	*Idx. over each (r, b) ordered by n*	=	*Visit group number*	*stop*	*doors opening instant (n)*	*stop duration (β)*	*last instant*	*Δn*	*¿Δn ≤ 15 min?*	*group change*	*group id*	*stop*
...														
A	1	-	1	=	0	A	03-24 12:31:23	00:00:55	12:32:16	<null>	✗	1	311	A
B	2	-	1	=	1	B	03-24 12:33:56	<null>	12:33:56	00:01:40	✓	0	311	B
C	3	-	1	=	2	C	03-24 12:36:14	00:00:15	12:36:29	00:02:18	✓	0	311	C
C	4	-	2	=	2		03-24 12:36:31	00:00:09	12:36:40	00:00:02	✓	0	311	
D	5	-	1	=	4	D	03-24 12:37:24	00:00:31	12:37:55	00:00:44	✓	0	311	D
E	6	-	1	=	5	E	03-24 12:39:44	00:00:07	12:39:51	00:01:49	✓	0	311	E
...														
B	45	-	2	=	43	B	03-24 13:49:28	00:00:12	13:49:40	00:01:44	✓	0	311	B
C	46	-	3	=	43	C	03-24 13:51:33	00:00:07	13:51:40	00:01:53	✓	0	311	C
D	47	-	2	=	45	D	03-24 13:54:02	00:00:11	13:54:13	00:02:22	✓	0	311	D
D	48	-	3	=	45		03-24 13:44:03	00:00:07	13:54:10	-00:00:03	✓	0	311	
D	49	-	4	=	45		03-24 23:05:49	00:00:25	23:06:14	09:21:39	✗	1	312	D
E	50	-	2	=	48	E	03-24 23:06:50	00:00:18	23:06:07	00:0:36	✓	0	312	E
...														

Splitting raw trajectories and finding AVL entries linked to same visit to a stop

instant	*duration*	*stop*	*boardings*	*alightings*
...				
2015-09-26 12:50:29	0	151	<null>	0
2015-09-26 12:51:11	19	135	<null>	0
2015-09-26 12:51:18.117	1098	135	255	255
2015-09-26 12:51:18.117	<null>	135	255	255
2015-09-26 12:52:37	14	134	<null>	0
...				

⇒

arrival	*departure*	*stop*	*boardings*	*alightings*
...				
2015-09-26 12:50:29	2015-09-26 12:50:29	151	<null>	0
2015-09-26 12:51:11	2015-09-26 13:09:36.117	135	<null>	0
2015-09-26 12:52:37	2015-09-26 12:52:51	134	<null>	0
...				

avl

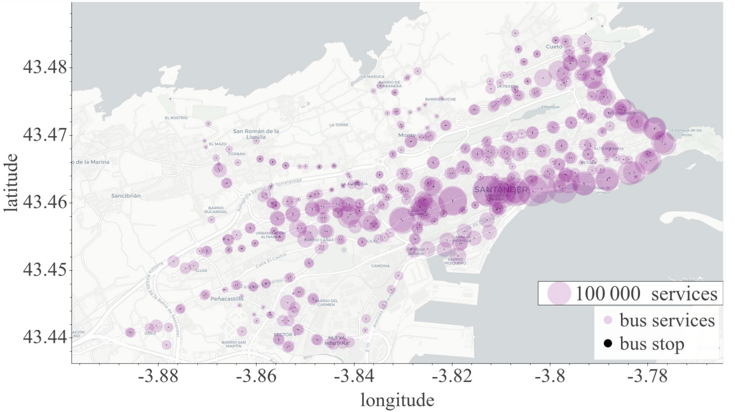
avl_coalesced

Condensing data from several rows linked to a single visit

AVL pre-processing

Coalesce entries and raw trajectories splitting

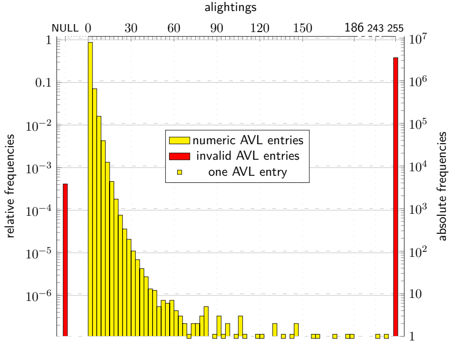
As part of this task has removed redundant and duplicate entries from the AVL information, it is interesting to visualize the resulting data.



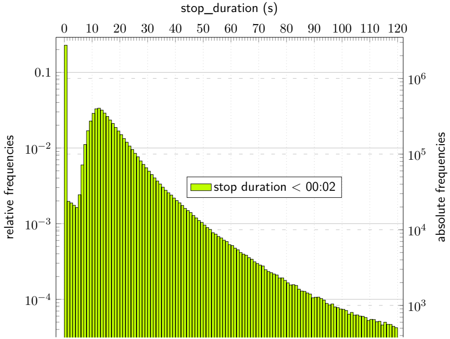
Distribution of bus visits

AVL pre-processing

Coalesce entries and raw trajectories splitting



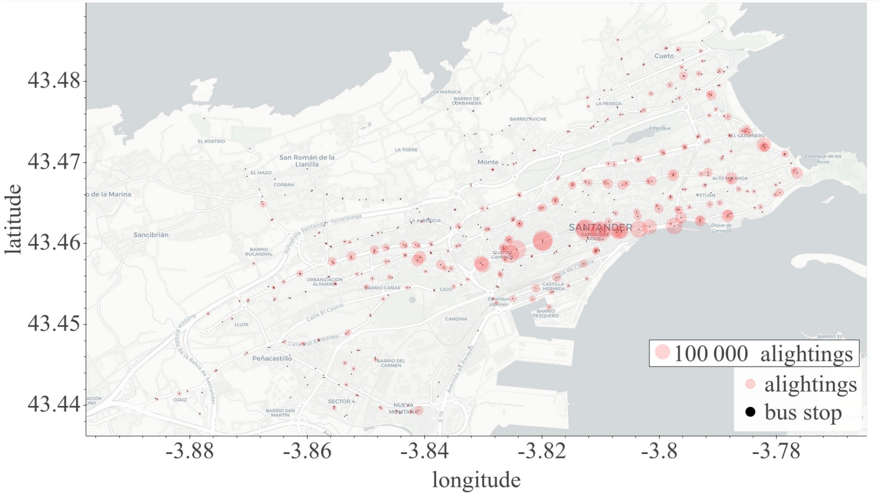
Number of alightings each time a bus arrives at a stop



Bus stop spans shorter than 2 min

AVL pre-processing

Coalesce entries and raw trajectories splitting



Distribution of bus visits

AVL pre-processing

Template sequences

Goal

Identify the sequences of bus stops that characterize flawlessly registered services.

How

- Extract the sequences of bus stops of each trajectory, and group them by line.
- Study the distribution of each distinct sequence of stops for each line.
- Manually identify the template sequence(s).

AVL pre-processing

Template sequences

occurrences	*frequency*	*sequence of bus stops*
7877	0.263	488,487,454,338,343,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,196,197,198,199,200,333,326,327,328
7219	0.241	328,330,200,201,167,168,169,28,453,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,342,341,455,489,488
3071	0.103	338,343,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,196,197,198,199,200,333,326,327,328
753	0.025	328,330,200,201,167,168,169,28,453,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,342,341,455,489
641	0.021	488
591	0.020	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,196,197,198,199,200,333,326,327,328
586	0.020	328,330,200,201,167,168,169,28,453,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,342,341
349	0.012	328,330,200,201,167,168,169,28,453,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,342,341,455
297	0.010	328,330,200,201,167,168,169,28,453,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53
263	0.009	328,330,200,201,167,168,169,28,453,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,342,341,455,338
...	...	...
222	0.007	328
...	...	...
58	0.002	330,201,167,168,169,28,453,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,342,341,455,489,488,487,454,338,343,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,196,197,198,199,200,333,326,327,328

Linear route (1)

occurrences	*frequency*	*sequence of bus stops*
9526	0.714	482,483,47,214,215,216,217,218,26,219,73,74,75,100,101,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,220,7,223,179,482
978	0.073	483,47,214,215,216,217,218,26,219,73,74,75,100,101,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,220,7,223,179,482
553	0.041	482,483,47,214,215,216,217,218,26,219,73,74,75,100,101,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,220,7,223,179
324	0.024	482
245	0.018	42,43,44,45,220,7,223,179,482
156	0.012	31,482,483,47,214,215,216,217,218,26,219,73,74,75,100,101,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,220,7,223,179
...	...	...
78	0.005	488

Circular route (7, ↻)

sequences samples

AVL pre-processing

Transportation offer definition

Goal

Re-create the services that constitute the transportation offer in the city.

How

1. Trajectories that correspond to template sequences: directly interpreted as services. ✓
2. Trajectories longer that the templates: those sub-sequences that are a math to the templates are identified as services. ✓
3. The rest of the data: sequences that partially match the templates. The missing visits to stops are inferred. ¿?

AVL pre-processing

Transportation offer definition

Imputation

Sequences of visits that do not conform a full service are analyzed, from longer to shorter, and the missing data is obtained by the first of these sources that yields a match with a sufficient confidence level:

1. Another compatible sub-sequence (AVL).
2. A boarding group (AFC).
3. Most likely value (AVL templates).

Hypotheses

The distributions of how long a bus stays at each stop during a service (d); and how long it takes to drive from one bus stop to the next (t) can be approximated as Normal.

AVL pre-processing

Transportation offer definition

Example #1/3

When the next stop in a service has not been recorded by the AVL system, the information regarding the arrival and departure of the bus is missing. Estimating a confidence interval for these values requires:

- Departure time from the previous stop (n).
- Distribution of the travel time (t).
- Distribution of the duration of the stop (d).

Implementation details #1/2

In order to improve the accuracy of the estimations, days are segregated according to how the operator plans the transportation supply:

Weekly categories: working days, Saturdays, and holidays (including Sundays).

Yearly categories: summer and the rest of the year.

AVL pre-processing

Transportation offer definition

Implementation details #2/2

Temporal bins are established, dividing each day in a series of intervals. Their width is established for each line, looking for a balance between the advantage of using shorter intervals (the predictions will be based on data in more similar conditions) and keeping a significant sample size.

Formulation

$t_{m,d,f,b,o}$	:	travel times between bus stops	time
$d_{m,d,f,b,o}$	:	duration of stops	time

Where:

m	:	template sequence id	$m \in T$
T	:	set of template sequences ids	$T \subseteq E$
E	:	set of ids of seqs. w/o overlap	$E \subset \mathbb{Z}_{>0}$
d	:	type of day	$d \in D$
D	:	set of day types	$D = \{$‘working day’, ‘saturday’, ‘holiday’$\}$
f	:	part of the year	$f \in F$
F	:	set of parts of the year	$F = \{$‘summer’, ‘rest of the year’$\}$
b	:	bin number	$b \in \mathbb{Z}_{>0}, 1 \leq bn \leq \lvert B_m \rvert$
l	:	length of one day	$l = 24\,\mathrm{h}$
g_m	:	length of a bin	time, $g_m \mid l$
B_m	:	set of time bins	$B_m = \{\, [00{:}\,00{:}\,00, g_m) \ldots [24{:}\,00{:}\,00 - g_m, 24{:}\,00{:}\,00) \,\}$
o	:	$\begin{cases} \text{for } t: \text{ number of the } 1^{\text{st}} \text{ stop} \\ \text{for } d: \text{ stop number} \end{cases}$	$o \in \mathbb{Z}_{>0} \begin{cases} \text{for } t: \ 1 \leq o \leq \lvert O_m \rvert - 1 \\ \text{for } d: \ 1 \leq o \leq \lvert O_m \rvert \end{cases}$
O_m	:	sequence of bus stops	$O_m = [1 \ldots \lvert O_m \rvert]$

AVL pre-processing

Transportation offer definition

Example #2/3

Template $m = 63$ has $|O_{63}| = 39$ stops, and characterizes the services of line 1 between Odriozola and Pctcan termini. If a bin length of $g_m = 30$ min is chosen, creating $|B_{63}| = 48$ bins along a day, the total number of data points available would be:

$$\begin{aligned} |D| \cdot |F| \cdot |B_{63}| \cdot (|O_{63}| - 1) &= 10\,944 \text{ data points} \quad \text{for } t_{63,d,f,b,o} \\ |D| \cdot |F| \cdot |B_{63}| \cdot |O_{63}| &= 11\,232 \text{ data points} \quad \text{for } d_{63,d,f,b,o} \end{aligned}$$

AVL pre-processing

Transportation offer definition

Example #3/3

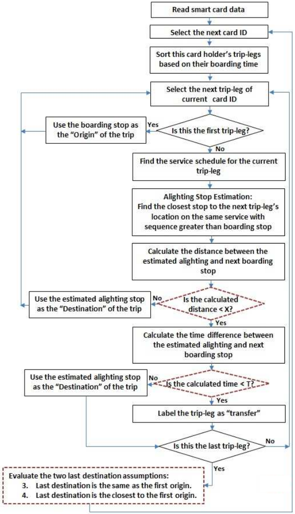
If the missing entry has occurred the 24th of March (a working day of the "winter schedule" of the operator), at 11:15:00, at Correos stop (no. 40), the statistical parameters from the appropiate bins are:

```
[local]:5432 postgres@pt_toolbox=# SELECT * FROM get_travel_time_inference_info(template => 63,
instant => '2015-03-24␣11:15:00', final_bus_stop_number => 40);
      mean       |   ps_std_dev    | sample_size
-----------------+-----------------+-------------
 00:01:21.176056 | 00:00:22.848791 |         284


[local]:5432 postgres@pt_toolbox=# SELECT * FROM get_stop_time_inference_info(template => 63,
bus_stop_number => 40::smallint, arrival => '2015-03-24␣11:15:00');
      mean       |   ps_std_dev    | sample_size
-----------------+-----------------+-------------
 00:00:32.632075 | 00:00:11.875678 |         212
```

Trip chaining

Model definition



Read smart card data

Select the next card ID

Sort this card holder's trip-legs based on their boarding time

Select the next trip-leg of current card ID

Is this the first trip-leg?

Yes

Use the boarding stop as the "Origin" of the trip

No

Find the service schedule for the current trip-leg

Alighting Stop Estimation: Find the closest stop to the next trip-leg's location on the same service with sequence greater than boarding stop

Calculate the distance between the estimated alighting and next boarding stop

Is the calculated distance < X?

No

Use the estimated alighting stop as the "Destination" of the trip

Yes

Calculate the time difference between the estimated alighting and next boarding stop

Is the calculated time < T?

No

Use the estimated alighting stop as the "Destination" of the trip

Yes

Label the trip-leg as "transfer"

Is this the last trip-leg?

No

Yes

Evaluate the two last destination assumptions:

3. Last destination is the same as the first origin.
4. Last destination is the closest to the first origin.

Trip chaining method, *(A. Alsger, B. Eng; 2016)*

Parameters

mdp: max. dist. between stops during a transfer (800 m).

mtt: maximum time since passengers leave a bus until they board the next while transferring (40 min).

mtr: maximum duration of a chain of trips in the city.

ws: walking speed $\left(1.4\,\frac{\mathrm{m}}{\mathrm{s}}\right)$.

Output

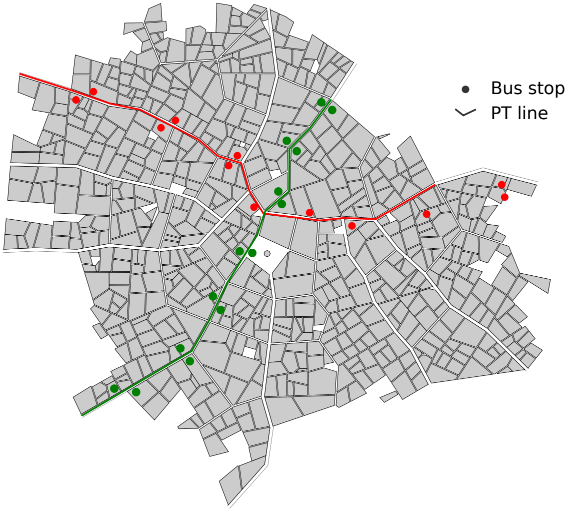
The trip chaining model reveals the journeys of public transport users, and the legs that compose them.

The former provide, after temporal and spatial aggregation, OD matrices during the period under analysis.

The latter, materialized by boarding and alighting different bus services, provide insight in the load profile of the vehicles.

Trip chaining

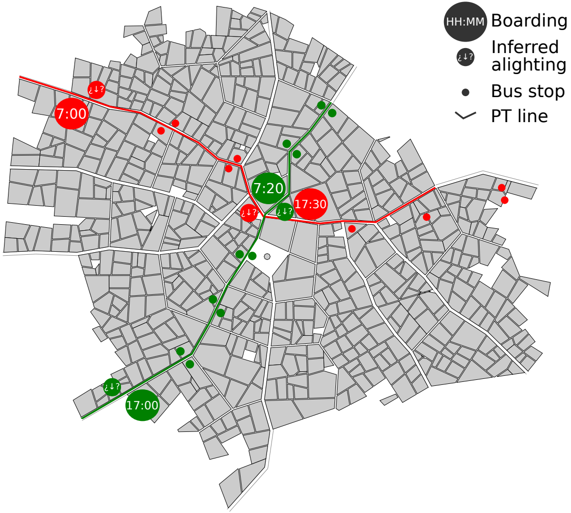
Example



Sample city generated with https://watabou.itch.io

Trip chaining

Example

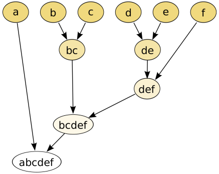


Inferred trip chains from 4 validations of a SC in a day

Demand modeling

Models comparison

Hierarchical agglomerative clustering



Dendogram, *(Stathis Sideris, 2005)*

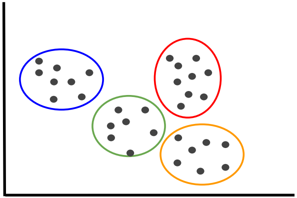
$$\mathrm{F}\left(A,B\right)=\sqrt{\mathrm{Tr}\left(\left(A-B\right)^{T}\left(A-B\right)\right)}$$

$$\mathrm{d}_1\left(A,B\right)=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|a_{ij}-b_{ij}\right|}$$

$$\mathrm{d}_2\left(A,B\right)=\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(a_{ij}-b_{ij}\right)^2}$$

Measurement of the similarity between matrices

Unsupervised machine learning clustering



Set of data points, already labeled (Google problem framing course)

Expected results

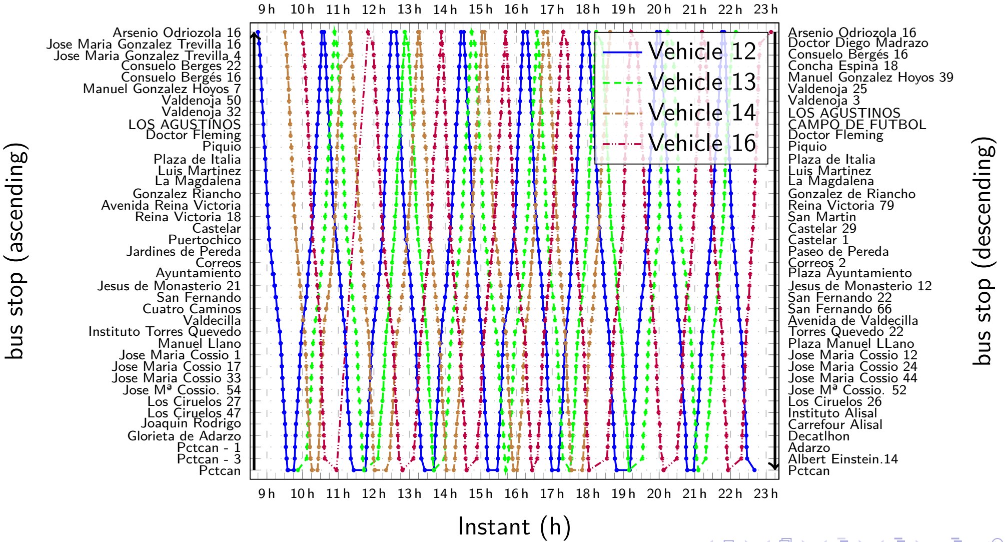
Comparison of the patterns found by each method in the evolution of public transportation demand with each other, and with the planning strategies of the service operator.

Public transport demand modeling with smart card tap-in information

1. Introduction
2. Objectives
   - Data cleansing and imputation
   - Trip chaining
   - Public transport demand models
3. Dataset analysis
   - Bus stops
   - AFC
   - AVL
4. Methodology
   - Pre-processing
   - Trip chaining
   - Demand modeling
5. **Results sample**
6. Skills and Capacities Assessment

Recreation of the transportation offer

Services provided by 4 of the vehicles that run Line 1 the 24th of March



1. Introduction
2. Objectives
   - Data cleansing and imputation
   - Trip chaining
   - Public transport demand models
3. Dataset analysis
   - Bus stops
   - AFC
   - AVL
4. Methodology
   - Pre-processing
   - Trip chaining
   - Demand modeling
5. Results sample
6. Skills and Capacities Assessment

Skills and Capacities Assessment

Basic Skills

CB11: I have had access to the bibliography and tools I needed. Also, I have attended many helpful courses and seminars.

CB12: I need to publish my results. Also, I still struggle to follow a work plan. During my doctoral studies I have had the opportunity to collaborate at TU Delft for three months.

CB13: I need to publish my results.

CB14: I have had to assess and utilize new ideas to carry out my studies.

CB15: I really need to publish my results. As part of my activities, I have had to communicate with researchers, scientists, and workers from many european organizations.

CB16: I have met and worked with rigorous and ethical researchers, and tried to learn from them.

Skills and Capacities Assessment

Capacities and personal abilities

CA01: I have learned this ability during my research.

CA02: I need to publish my results.

CA03: I am working on improving my ability to devise and follow a realistic work plan. Part of my multidisciplinary training from the EDUC (and other seminars I have attended to) deal with research funding.

CA04: I am able to work individually or as part of multidisciplinary or international groups.

CA05: My ability to evaluate complex problems with limited information has improved greatly.

CA06: The day-to-day interactions with my colleges includes the discussion of different topics related to our research.

Travel demand modeling with smart card tap-in information

Thanks

email: juan.benavente@unican.es

Borja Alonso Oreña

Director

Juan Benavente Ponce

PhD student

José Luis Moura Berodia

Director