Ecole Nationale des Ponts et Chaussées - Département VET
Projet d'Ingénieur en Laboratoire

Discrete Choice Models for Bike-Sharing Transportation Systems

Supplementary Material

Juste Raimbault[1 2, a]

[1] Graduate School, Ecole Polytechnique, Palaiseau
[2] LVMT, UMR-T 9403 IFSTTAR, Champs-sur-Marne

Abstract. We provide here various supplementary material to the joint paper, which should ensure total reproducibility by giving technical details about processes used.
We detail in particular the following points :
• Discrete choice questionnaire : description of variables, implementation of the questionnaire and results of modeling.
• Data on system dynamics : data collection process and pre-processing, descriptive statistics.
• Agent-based modeling : Model evaluation, Model implementation, Model parametrization, Model application, Gradient algorithm.

1 Discrete Choice Questionnaire

1.1 Data structure and description

Socio-economic variables

Chosen socio-economic variable are relatively simple and were designed to capture characteristics expected to be most significant. As the aim of the survey was to distinguish between modal choice that can strongly depend on "basic"profile (in the sense that using bike-sharing and bike in general has been shown to be typical behavior depending on simple types of socio-economic profiles [Parkin et al., 2008]), this reduced choice was already accepted as providing reasonable complexity of the resulting models, especially combined with stratified choice attributes. They are the following (with respective reasons of the choice) :

- Age of people (as bike is a physical activity) ; in years.

- Socio-economic category, stratified into student (1), active (2) and other (3) (as social origin but also type of occupation should influence modal choice) ; transformed into 4 boolean dummy variables (csp_student, csp_active, csp_other).

[a] e-mail: juste.raimbault@polytechnique.edu

- If the user has a regular bike-sharing account (although this variable may not make sense directly to explain choice of bike-sharing, since the fact to register will be more a consequence of the habit to use the service than the contrary. It can be seen therefore as a control variable, useful to check internal validity of models) ; boolean variable.

- Distance between residence and closest public transportation station (it is not clear if long distance but bike availability should favor bike choice, that variable was taken to check that point) ; in meters.

- Weather (i.e. weather in the situation of scenarii) is also considered as a variable, as it is not alternative-dependent. Stratified with dummy variables (weather_good, weather_mean, weather_hardcore).

- The same way, hour of the day is considered as a variable ; in hours.

Attributes of scenarii

Possible modal choices are bus and bike-sharing, which have common and distinct attributes, stratified into different levels used to build scenarii in a combinatorial way.

Common attributes are not stratified (continuous random values) and treated as variables, and are :

- Distance of the travel (in meters)

- Estimated travel time, assumed as the same for both mode choice, what is a reasonable hypothesis in that case.

Bike-sharing attributes describe quality of service :

- Time to find a bike ; stratified into 1, 5 and 10 minutes

- Time to find a parking place when at destination, with same stratification.

The same way, bus attributes describe level of service, including comfort :

- Bus delay, stratified into 0, 5 and 15 minutes

- Bus comfort, qualitatively decomposed into : sit available (1), standing (2), standing and overcrowded (3).

Descriptive statistics

Table 1 give basic descriptive statistics on collected data, on 288 rows of the database. Data can be directly downloaded into the biogeme format at the export address of the platform[1]. We confirm the poor quality of the data, since it is for example unbalanced in socio-economic categories.

[1] http://37.187.242.99/Questionnaire/php/utils/export.php

variable	min	1st Qrt	med	mean	3rd Qrt	max
age	19	21	22	23.95	24	50
localisation	50	60	150	161.6	200	800
csp_student	N/A	N/A	N/A	0.75	N/A	N/A
csp_active	N/A	N/A	N/A	0.14	N/A	N/A
csp_other	N/A	N/A	N/A	0.01	N/A	N/A
subscription	N/A	N/A	N/A	0.27	N/A	N/A
travel hour	0	9	15	13.51	18	23
weather_good	N/A	N/A	N/A	0.44	N/A	N/A
weather_mean	N/A	N/A	N/A	0.35	N/A	N/A
weather_hardcore	N/A	N/A	N/A	0.21	N/A	N/A

Table 1: Descriptive statistics of socio-economic variables. For dummies or booleans, quantile have no sense but mean gives naturally proportion of corresponding category.

1.2 Online Questionnaire Application Architecture and implementation

General architecture and implementation main lines were given in the paper. We provide here examples on working of the platform. Note that website is still online and source code available, for any particular concern on web architecture.

Figures 1 and 2 show the public pages available to fill the questionnaire online. These pages are the same than restricted access pages reserved to surveyors, at the exception than the latest do not need to fill security captcha at the end. Figure 3 shows the administration page, through which new questionnaire can be created (reserved to administrator). For this, total flexibility is allowed, as number of variables, attributes and choices is free, as their respective types or level in the case of attributes (one can add line by clicking on corresponding links). Once the form is filled, corresponding structures are automatically created in the SQL database, avoiding shady database administrative tasks. An capture of the database is shown in figure 4.

1.3 Modeling

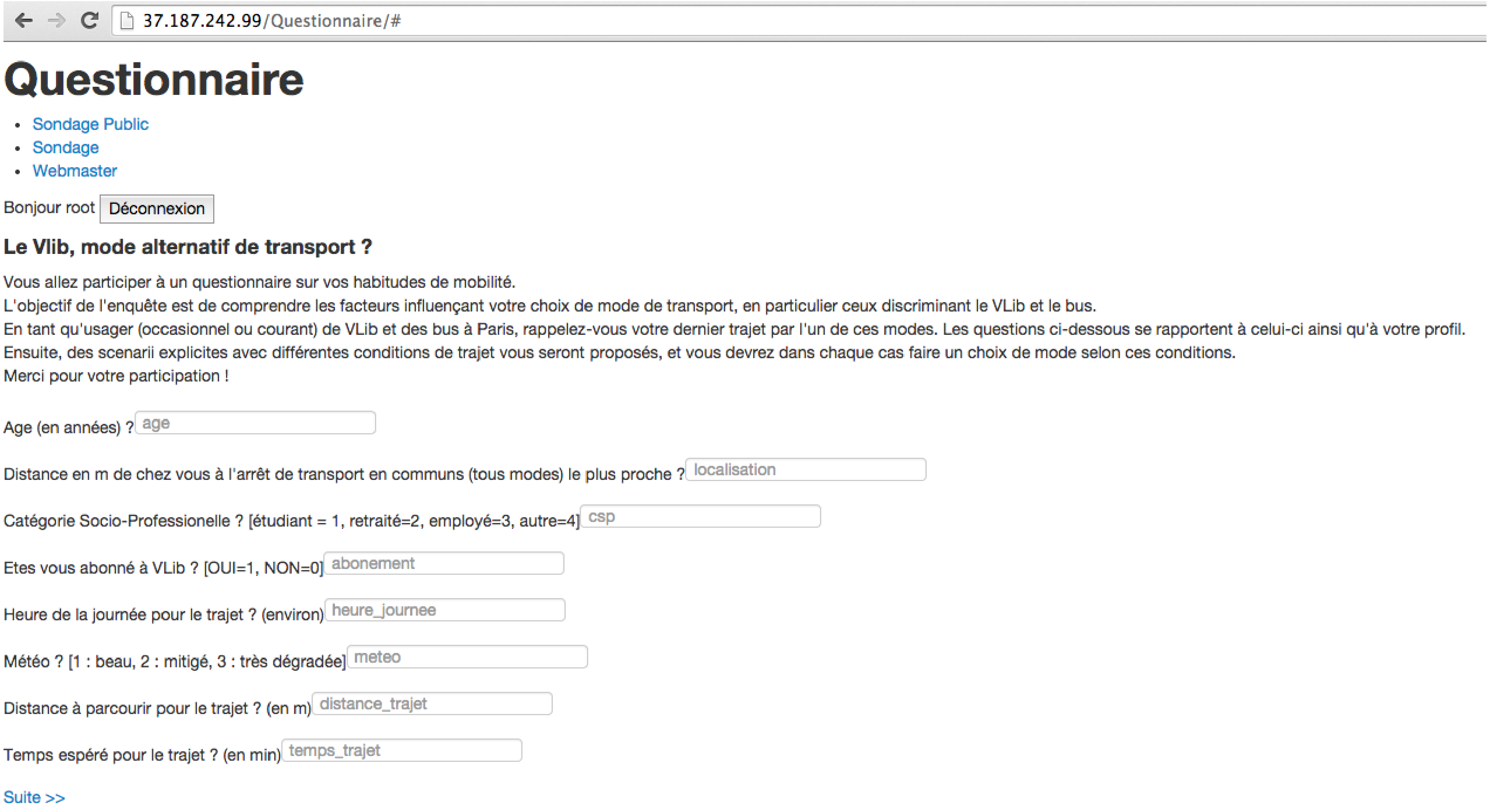
We describe here the full model, from which results have been extracted for the a priori parameter values in the calibration procedure.

We take linear utilities, and use socio-economic variables in only one choice as usual :

$$U_{bus} = K_{bus} + \sum_{i=1}^{3} \beta_{confort=i} \mathbb{1}_{confort=i} + \beta_{delay} delay + \varepsilon_{bus}$$

$$\begin{aligned} U_{bike} =& K_{bike} + \beta_{age} age + \sum_{i=1}^{3} \beta_{csp=i} \mathbb{1}_{csp=i} + \beta_d d + \beta_{loc} d_{loc} + \beta_{subscr} \mathbb{1}_{subscr} \\ &+ \sum_{i=1}^{3} \beta_{weather=i} \mathbb{1}_{weather=i} + \beta_t t + \beta_{waiting} waiting + \beta_{parking} parking + \varepsilon_{bike} \end{aligned}$$

The model is written in Biogeme and directly run with the datafile. We obtain the results presented in the following table :



37.187.242.99/Questionnaire/#

Questionnaire

- Sondage Public
- Sondage
- Webmaster

Bonjour root Déconnexion

Le Vlib, mode alternatif de transport ?

Vous allez participer à un questionnaire sur vos habitudes de mobilité.
L'objectif de l'enquête est de comprendre les facteurs influençant votre choix de mode de transport, en particulier ceux discriminant le VLib et le bus.
En tant qu'usager (occasionnel ou courant) de VLib et des bus à Paris, rappelez-vous votre dernier trajet par l'un de ces modes. Les questions ci-dessous se rapportent à celui-ci ainsi qu'à votre profil.
Ensuite, des scenarii explicites avec différentes conditions de trajet vous seront proposés, et vous devrez dans chaque cas faire un choix de mode selon ces conditions.
Merci pour votre participation !

Age (en années) ? age

Distance en m de chez vous à l'arrêt de transport en communs (tous modes) le plus proche ? localisation

Catégorie Socio-Professionelle ? [étudiant = 1, retraité=2, employé=3, autre=4] csp

Etes vous abonné à VLib ? [OUI=1, NON=0] abonement

Heure de la journée pour le trajet ? (environ) heure_journee

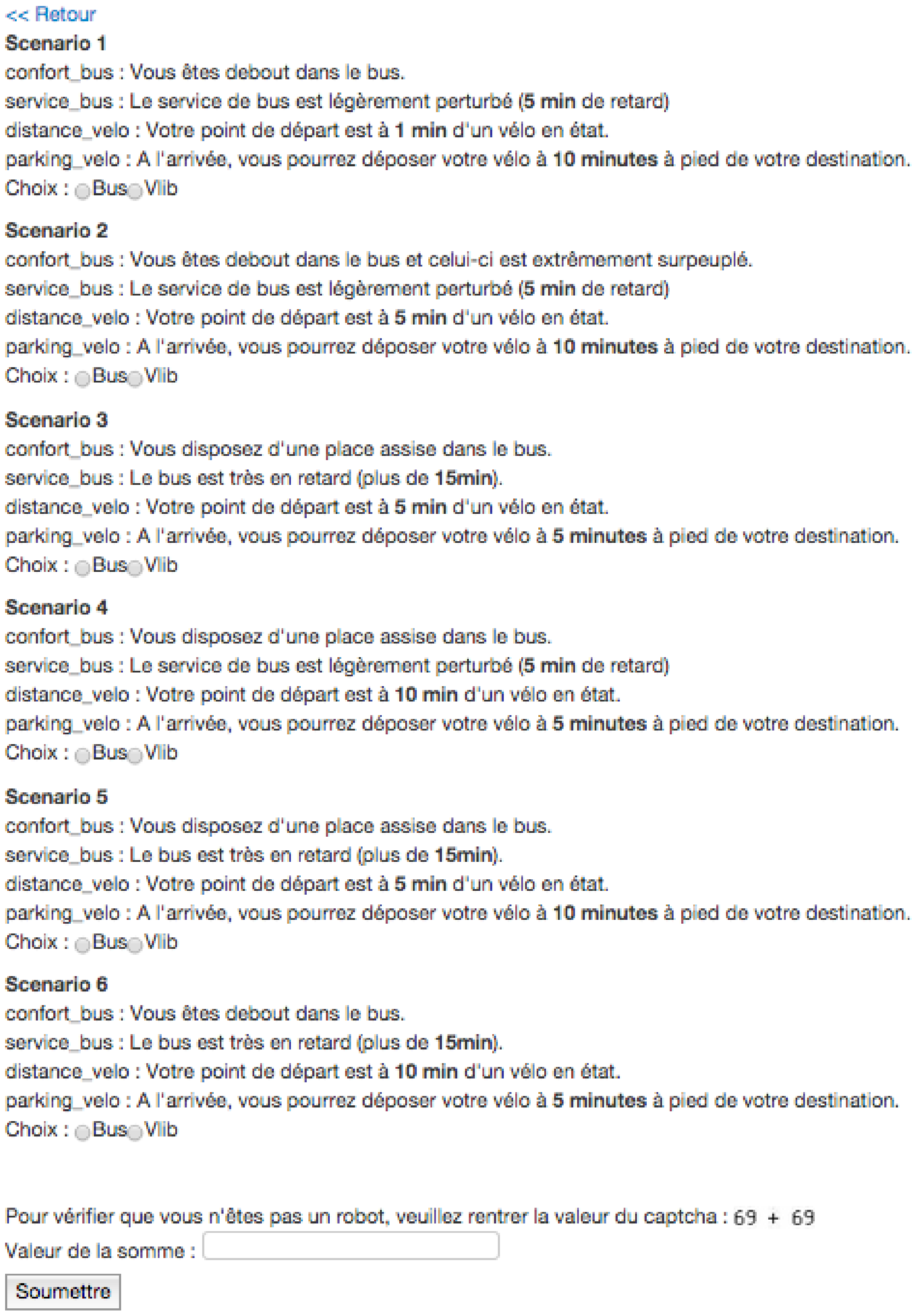
Météo ? [1 : beau, 2 : mitigé, 3 : très dégradée] meteo

Distance à parcourir pour le trajet ? (en m) distance_trajet

Temps espéré pour le trajet ? (en min) temps_trajet

Suite >>

Fig. 1: Homepage of online questionnaire (including first part of the questionnaire).



<< Retour

Scenario 1

confort_bus : Vous êtes debout dans le bus.
service_bus : Le service de bus est légèrement perturbé (**5 min** de retard)
distance_velo : Votre point de départ est à **1 min** d'un vélo en état.
parking_velo : A l'arrivée, vous pourrez déposer votre vélo à **10 minutes** à pied de votre destination.
Choix : ○Bus○Vlib

Scenario 2

confort_bus : Vous êtes debout dans le bus et celui-ci est extrêmement surpeuplé.
service_bus : Le service de bus est légèrement perturbé (**5 min** de retard)
distance_velo : Votre point de départ est à **5 min** d'un vélo en état.
parking_velo : A l'arrivée, vous pourrez déposer votre vélo à **10 minutes** à pied de votre destination.
Choix : ○Bus○Vlib

Scenario 3

confort_bus : Vous disposez d'une place assise dans le bus.
service_bus : Le bus est très en retard (plus de **15min**).
distance_velo : Votre point de départ est à **5 min** d'un vélo en état.
parking_velo : A l'arrivée, vous pourrez déposer votre vélo à **5 minutes** à pied de votre destination.
Choix : ○Bus○Vlib

Scenario 4

confort_bus : Vous disposez d'une place assise dans le bus.
service_bus : Le service de bus est légèrement perturbé (**5 min** de retard)
distance_velo : Votre point de départ est à **10 min** d'un vélo en état.
parking_velo : A l'arrivée, vous pourrez déposer votre vélo à **5 minutes** à pied de votre destination.
Choix : ○Bus○Vlib

Scenario 5

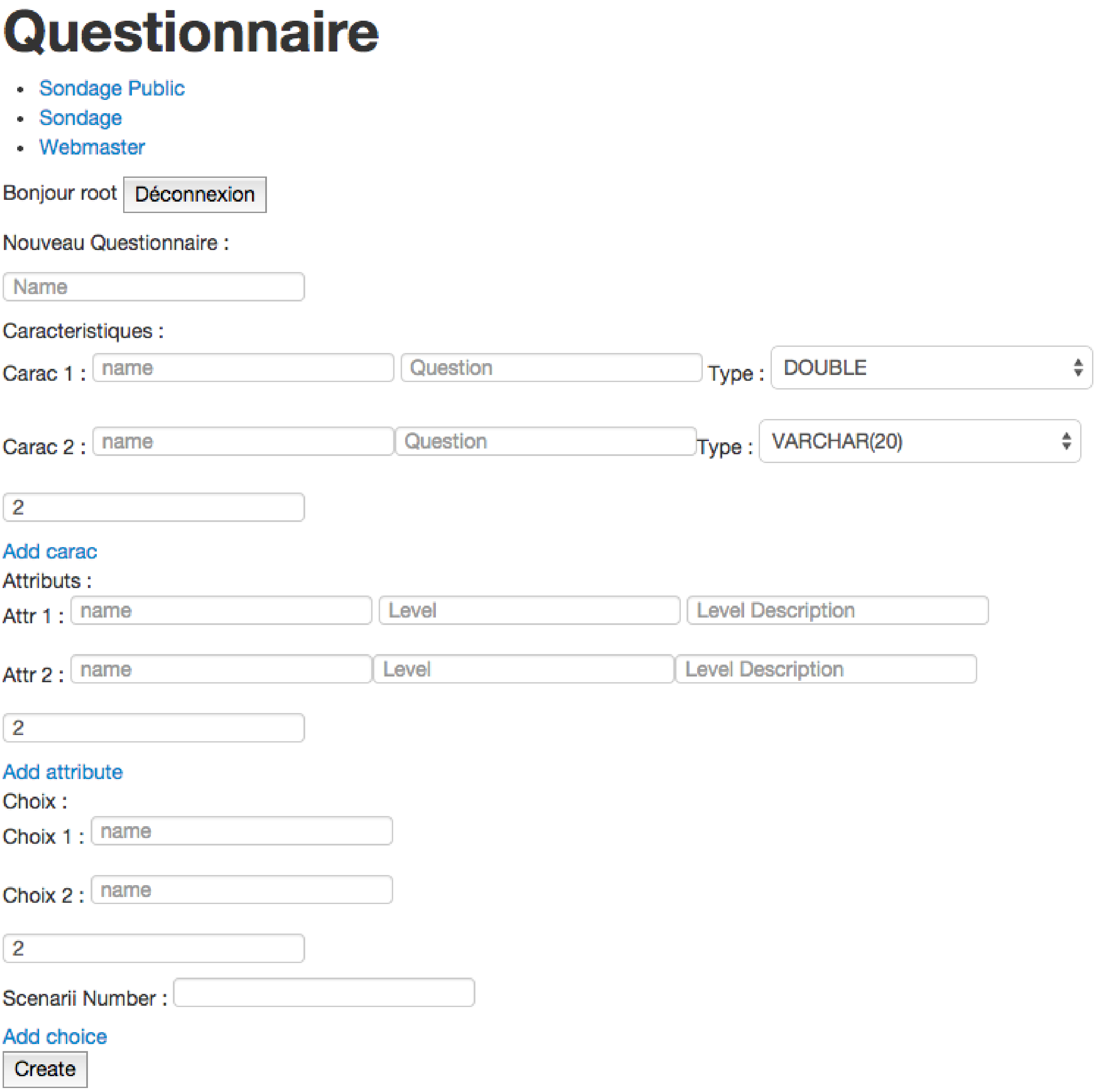
confort_bus : Vous disposez d'une place assise dans le bus.
service_bus : Le bus est très en retard (plus de **15min**).
distance_velo : Votre point de départ est à **5 min** d'un vélo en état.
parking_velo : A l'arrivée, vous pourrez déposer votre vélo à **10 minutes** à pied de votre destination.
Choix : ○Bus○Vlib

Scenario 6

confort_bus : Vous êtes debout dans le bus.
service_bus : Le bus est très en retard (plus de **15min**).
distance_velo : Votre point de départ est à **10 min** d'un vélo en état.
parking_velo : A l'arrivée, vous pourrez déposer votre vélo à **5 minutes** à pied de votre destination.
Choix : ○Bus○Vlib

Pour vérifier que vous n'êtes pas un robot, veuillez rentrer la valeur du captcha : 69 + 69
Valeur de la somme :

Fig. 2: Second part of the questionnaire : scenarii and modal choices.



Questionnaire

- Sondage Public
- Sondage
- Webmaster

Bonjour root Déconnexion

Nouveau Questionnaire :

Name

Caracteristiques :

Carac 1 : name Question Type : DOUBLE

Carac 2 : name Question Type : VARCHAR(20)

2

Add carac

Attributs :

Attr 1 : name Level Level Description

Attr 2 : name Level Level Description

2

Add attribute

Choix :

Choix 1 : name

Choix 2 : name

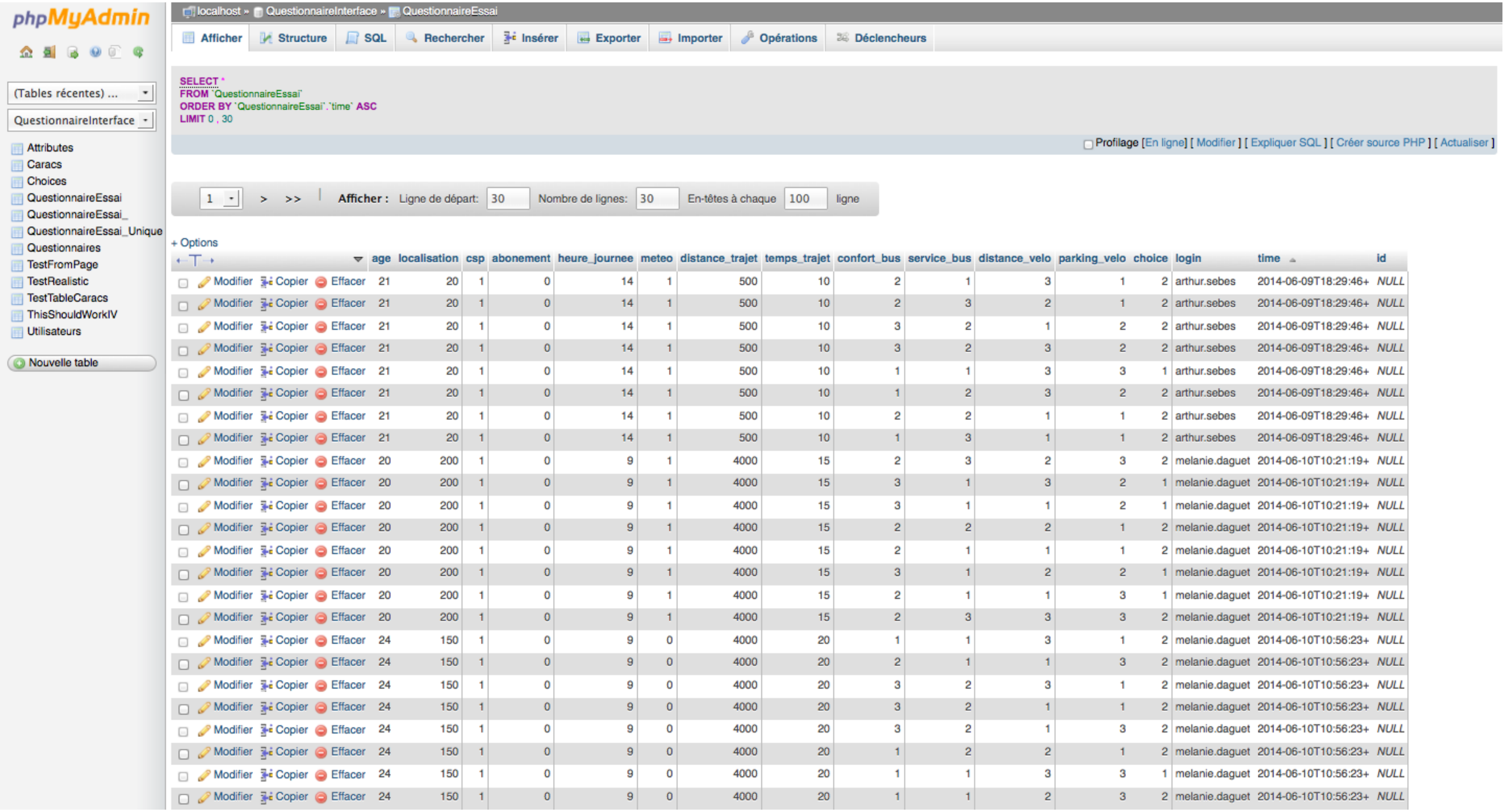
2

Scenarii Number :

Add choice

Create

Fig. 3: Management page of online questionnaire.

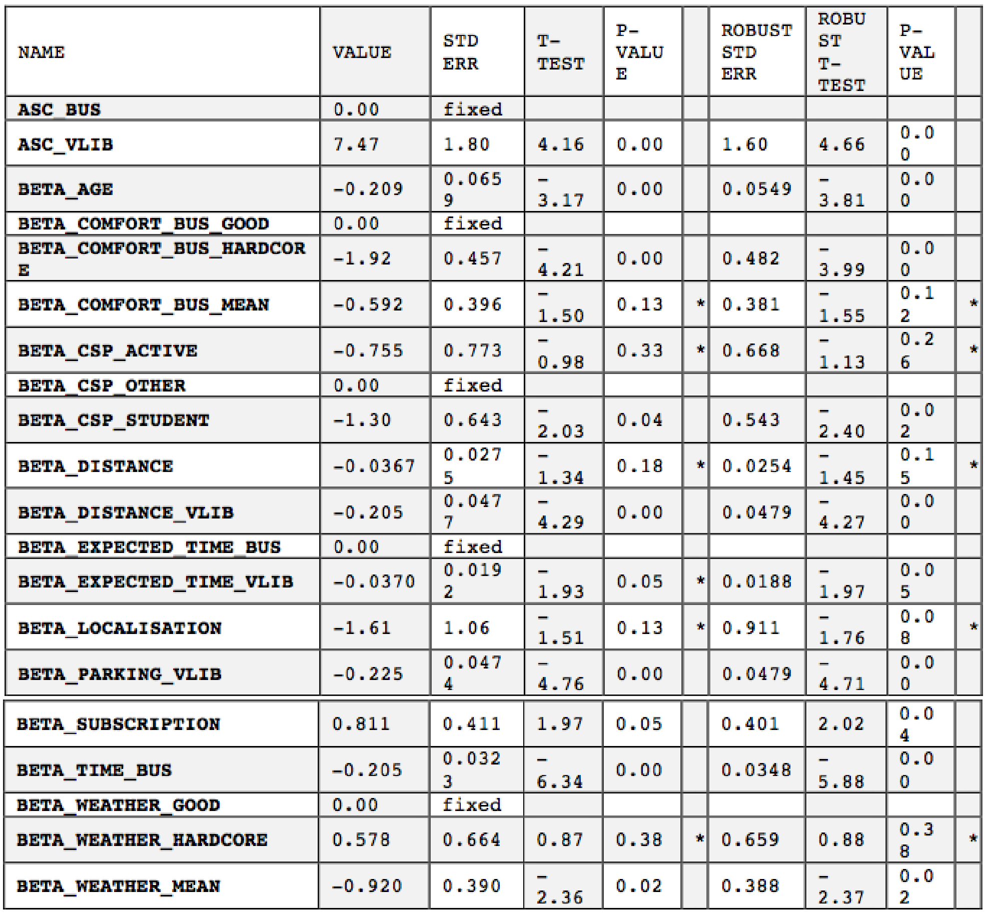


```
SELECT *
FROM `QuestionnaireEssai`
ORDER BY `QuestionnaireEssai`.`time` ASC
LIMIT 0 , 30
```

Afficher : Ligne de départ: 30 Nombre de lignes: 30 En-têtes à chaque 100 ligne

	age	localisation	csp	abonement	heure_journee	meteo	distance_trajet	temps_trajet	confort_bus	service_bus	distance_velo	parking_velo	choice	login	time	id
□ Modifier Copier Effacer	21	20	1	0	14	1	500	10	2	1	3	1	2	arthur.sebes	2014-06-09T18:29:46+	NULL
□ Modifier Copier Effacer	21	20	1	0	14	1	500	10	2	3	2	1	2	arthur.sebes	2014-06-09T18:29:46+	NULL
□ Modifier Copier Effacer	21	20	1	0	14	1	500	10	3	2	1	2	2	arthur.sebes	2014-06-09T18:29:46+	NULL
□ Modifier Copier Effacer	21	20	1	0	14	1	500	10	3	2	3	2	2	arthur.sebes	2014-06-09T18:29:46+	NULL
□ Modifier Copier Effacer	21	20	1	0	14	1	500	10	1	1	3	3	1	arthur.sebes	2014-06-09T18:29:46+	NULL
□ Modifier Copier Effacer	21	20	1	0	14	1	500	10	1	2	3	2	2	arthur.sebes	2014-06-09T18:29:46+	NULL
□ Modifier Copier Effacer	21	20	1	0	14	1	500	10	2	2	1	1	2	arthur.sebes	2014-06-09T18:29:46+	NULL
□ Modifier Copier Effacer	21	20	1	0	14	1	500	10	1	3	1	1	2	arthur.sebes	2014-06-09T18:29:46+	NULL
□ Modifier Copier Effacer	20	200	1	0	9	1	4000	15	2	3	2	3	2	melanie.daguet	2014-06-10T10:21:19+	NULL
□ Modifier Copier Effacer	20	200	1	0	9	1	4000	15	3	1	3	2	1	melanie.daguet	2014-06-10T10:21:19+	NULL
□ Modifier Copier Effacer	20	200	1	0	9	1	4000	15	3	1	1	2	1	melanie.daguet	2014-06-10T10:21:19+	NULL
□ Modifier Copier Effacer	20	200	1	0	9	1	4000	15	2	2	2	1	2	melanie.daguet	2014-06-10T10:21:19+	NULL
□ Modifier Copier Effacer	20	200	1	0	9	1	4000	15	2	1	1	1	2	melanie.daguet	2014-06-10T10:21:19+	NULL
□ Modifier Copier Effacer	20	200	1	0	9	1	4000	15	3	1	2	2	1	melanie.daguet	2014-06-10T10:21:19+	NULL
□ Modifier Copier Effacer	20	200	1	0	9	1	4000	15	2	1	1	3	1	melanie.daguet	2014-06-10T10:21:19+	NULL
□ Modifier Copier Effacer	20	200	1	0	9	1	4000	15	2	3	3	3	2	melanie.daguet	2014-06-10T10:21:19+	NULL
□ Modifier Copier Effacer	24	150	1	0	9	0	4000	20	1	1	3	1	2	melanie.daguet	2014-06-10T10:56:23+	NULL
□ Modifier Copier Effacer	24	150	1	0	9	0	4000	20	2	1	1	3	2	melanie.daguet	2014-06-10T10:56:23+	NULL
□ Modifier Copier Effacer	24	150	1	0	9	0	4000	20	3	2	3	1	2	melanie.daguet	2014-06-10T10:56:23+	NULL
□ Modifier Copier Effacer	24	150	1	0	9	0	4000	20	3	2	1	1	2	melanie.daguet	2014-06-10T10:56:23+	NULL
□ Modifier Copier Effacer	24	150	1	0	9	0	4000	20	3	2	1	3	2	melanie.daguet	2014-06-10T10:56:23+	NULL
□ Modifier Copier Effacer	24	150	1	0	9	0	4000	20	1	2	2	1	2	melanie.daguet	2014-06-10T10:56:23+	NULL
□ Modifier Copier Effacer	24	150	1	0	9	0	4000	20	1	1	3	3	1	melanie.daguet	2014-06-10T10:56:23+	NULL
□ Modifier Copier Effacer	24	150	1	0	9	0	4000	20	1	1	2	3	2	melanie.daguet	2014-06-10T10:56:23+	NULL

Fig. 4: Capture of the database (through phpMyadmin mysql manager application).



NAME	VALUE	STD ERR	T-TEST	P-VALUE		ROBUST STD ERR	ROBUST T-TEST	P-VALUE	
ASC_BUS	0.00	fixed							
ASC_VLIB	7.47	1.80	4.16	0.00		1.60	4.66	0.00	
BETA_AGE	-0.209	0.0659	–3.17	0.00		0.0549	–3.81	0.00	
BETA_COMFORT_BUS_GOOD	0.00	fixed							
BETA_COMFORT_BUS_HARDCORE	-1.92	0.457	–4.21	0.00		0.482	–3.99	0.00	
BETA_COMFORT_BUS_MEAN	-0.592	0.396	–1.50	0.13	*	0.381	–1.55	0.12	*
BETA_CSP_ACTIVE	-0.755	0.773	–0.98	0.33	*	0.668	–1.13	0.26	*
BETA_CSP_OTHER	0.00	fixed							
BETA_CSP_STUDENT	-1.30	0.643	–2.03	0.04		0.543	–2.40	0.02	
BETA_DISTANCE	-0.0367	0.0275	–1.34	0.18	*	0.0254	–1.45	0.15	*
BETA_DISTANCE_VLIB	-0.205	0.0477	–4.29	0.00		0.0479	–4.27	0.00	
BETA_EXPECTED_TIME_BUS	0.00	fixed							
BETA_EXPECTED_TIME_VLIB	-0.0370	0.0192	–1.93	0.05	*	0.0188	–1.97	0.05	
BETA_LOCALISATION	-1.61	1.06	–1.51	0.13	*	0.911	–1.76	0.08	*
BETA_PARKING_VLIB	-0.225	0.0474	–4.76	0.00		0.0479	–4.71	0.00	
BETA_SUBSCRIPTION	0.811	0.411	1.97	0.05		0.401	2.02	0.04	
BETA_TIME_BUS	-0.205	0.0323	–6.34	0.00		0.0348	–5.88	0.00	
BETA_WEATHER_GOOD	0.00	fixed							
BETA_WEATHER_HARDCORE	0.578	0.664	0.87	0.38	*	0.659	0.88	0.38	*
BETA_WEATHER_MEAN	-0.920	0.390	–2.36	0.02		0.388	–2.37	0.02	

Sense of influence are all correct, at the exception of the bad weather, what should be due to the unbalance issue in data. Indeed, some parameters have poor p-value and can be difficultly trusted, as socio-economic category, what was expected as most interrogated persons were students. We use these results to extract boundaries for the calibration procedure explained in the main paper, taking in this table parameters $\beta_{distance}$ and β_{tbus} with associated standard deviations.

```
filename=$PATHdata/‘date +‘‘\%s‘‘  ‘.json

curl -G --data 'apiKey=‘cat apiKey‘&contract=Paris'
     https://api.jcdecaux.com/vls/v1/stations > $filename
```

Table 2: Data collecting shell script.
Unique file identifier is created by epoch date, calling the `date` utility (echoing e.g. the number 1417980916 with the `%s` option). Data is then collected and stored directly in file by a simple `curl` request in GET mode (`-G` option). API website returns json file of the specified contract (Paris in our case) if the user api key is valid (key is here stored in an external file and paste by `cat`.

2 Data on system dynamics

2.1 Data Collection

Data collection was automized, by collecting at a fixed time step during a long period of time (1 year of data used in the project, but data collection is still currently running for possible future developments of the project). Therefore, the following daemons were installed on a remote server (code is given and explained in the respective tables, as for security reasons - mainly active access ids - this code was not pushed to the git repository of the project) :

- Data collection shell script. Put in crontab at a 5 minutes frequency, it collects from the operator API the real time data on stations status, provided as json data structure. See table 1 for code and explanation.

- Data archiving shell script. Json file are rapidly heavy on disk (around 300Mo for one day), it is for this reason of a crucial importance to archive the data in a compressed format. Every day, a script is automatically called to archive all data file collected during the day into a single zip file, weighting 5 times less.

2.2 Description

A raw data file consist in a text file containing json data, more precisely an array of which each entry is a docking station, coded as a json object. An example of json record representing a docking station is :

{"number":11029,"name":"11029 - MENILMONTANT OBERKAMPF","address":"137 BOULEVARD MENILMONTANT - 75011 PARIS", "position":{"lat":48.8666176586814, "lng":2.38301344041578}, "banking":true,"bonus":false,"status":"OPEN", "contract_name":"Paris","bike_stands":40, "available_bike_stands":12, "available_bikes":28,"last_update":1377639873000}

Reading the record with rjson package, we construct data objects that can easily be managed. Coordinates of docking stations allow to export through rgis their position and ids in a shapefile that is then used by the agent-based model implementation. The field `bike_stands` allows to compute the time-series of load-factors combined with `available_bikes`. As parametrization is done on typical day extracted through the clustering procedure, we do not user the field giving status of docking station.

```
#Automatic zip everyday to use 5 less memory

archivename=$PATH/archives/‘date ”+%s”‘.zip

#need to wait for a possible current collection to be finished

p=$(ps −e | grep collect | grep −v grep)

while [ ! −z $p ]
do
  sleep 1
  p=$(ps −e | grep collect | grep −v grep)
done

#now we sure no data will be lost during zipping and removing

zip −r $archivename $PATH/data
rm −r $PATH/data
mkdir $PATH/data
```

Table 3: Daily data archiving shell script.
We name the archive by the epoch date of archiving. The following loop aims to wait for any possible data collection in progress to be achieved, by listing and filtering by name the processes. We can then zip the daily data directory, remove it, and create a new empty directory for the next day.

Also, “bonus”character of a station is not taken into account in our model so we do not use that field. Furthermore, we consider for the sake of simplicity that system topology is stationary even on the larger time scale studied, i.e. we do not consider stations additions or extensions during the study period, what is a negligible approximation as less than 10 stations were extended during the period and no stations were constructed.

We show in figure 5 example of (non-normalized) load-factor curves for different docking stations, that can be dramatically different thus the need of the clustering procedure described in the main paper to isolate typical profiles and reduce dimension of the representation of the system (1230 docking stations).

We show in figure 6 the geographical representation (drawn with QGIS), for which road network of Paris and neighborhood has been extracted from OpenStreetMap, and docking stations layer has been constructed from raw data. Figure 7 provides a more detailed view of data prepared to be feed into the agent-based model implementation.

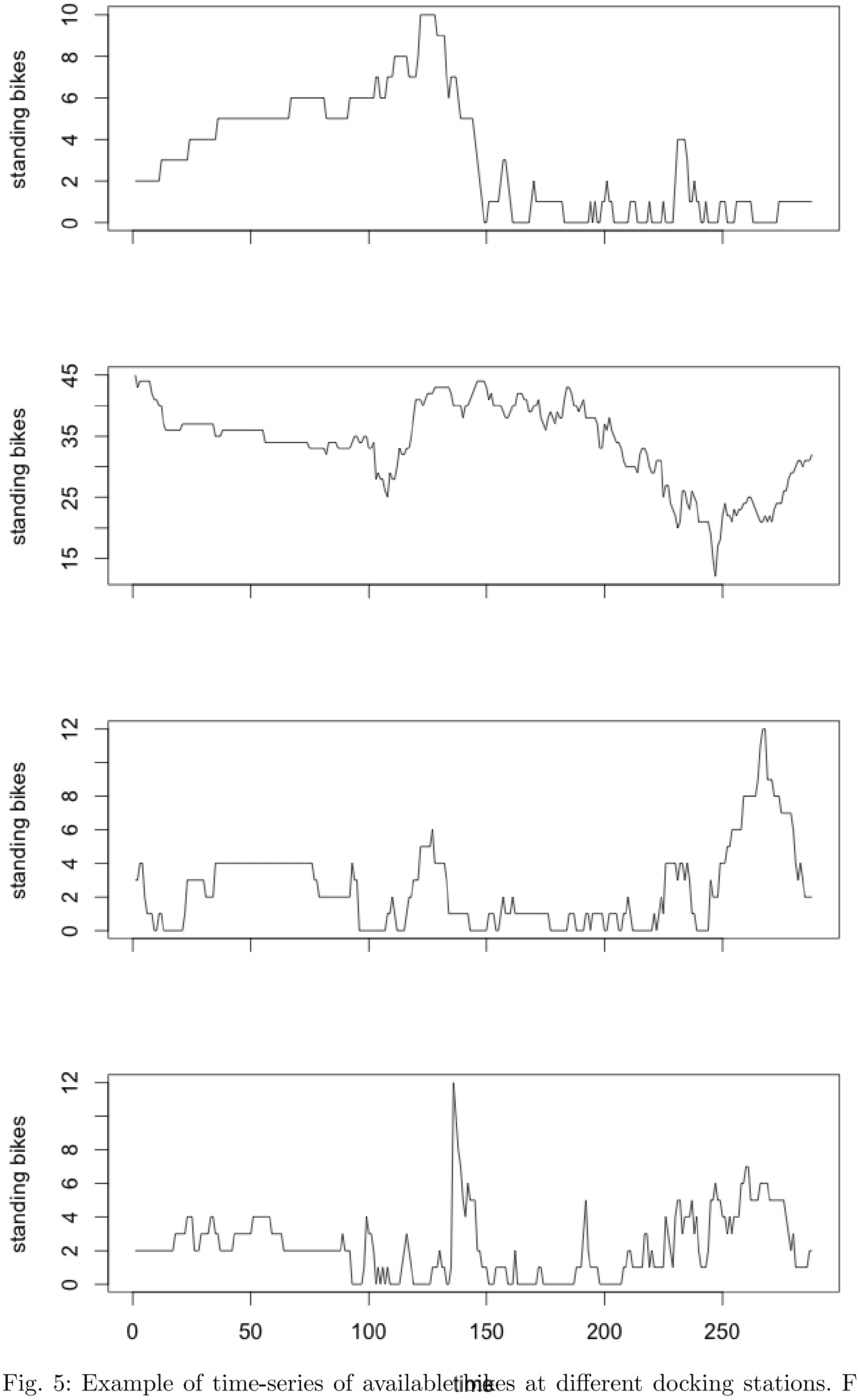


Fig. 5: Example of time-series of available bikes at different docking stations. First is a typical work area station, as number of bikes increase in the morning, to fall to quasi null in the afternoon. Second is more complicated and may be an hybrid area. The two other are zones were flux are very high, as dropped bikes directly disappear.

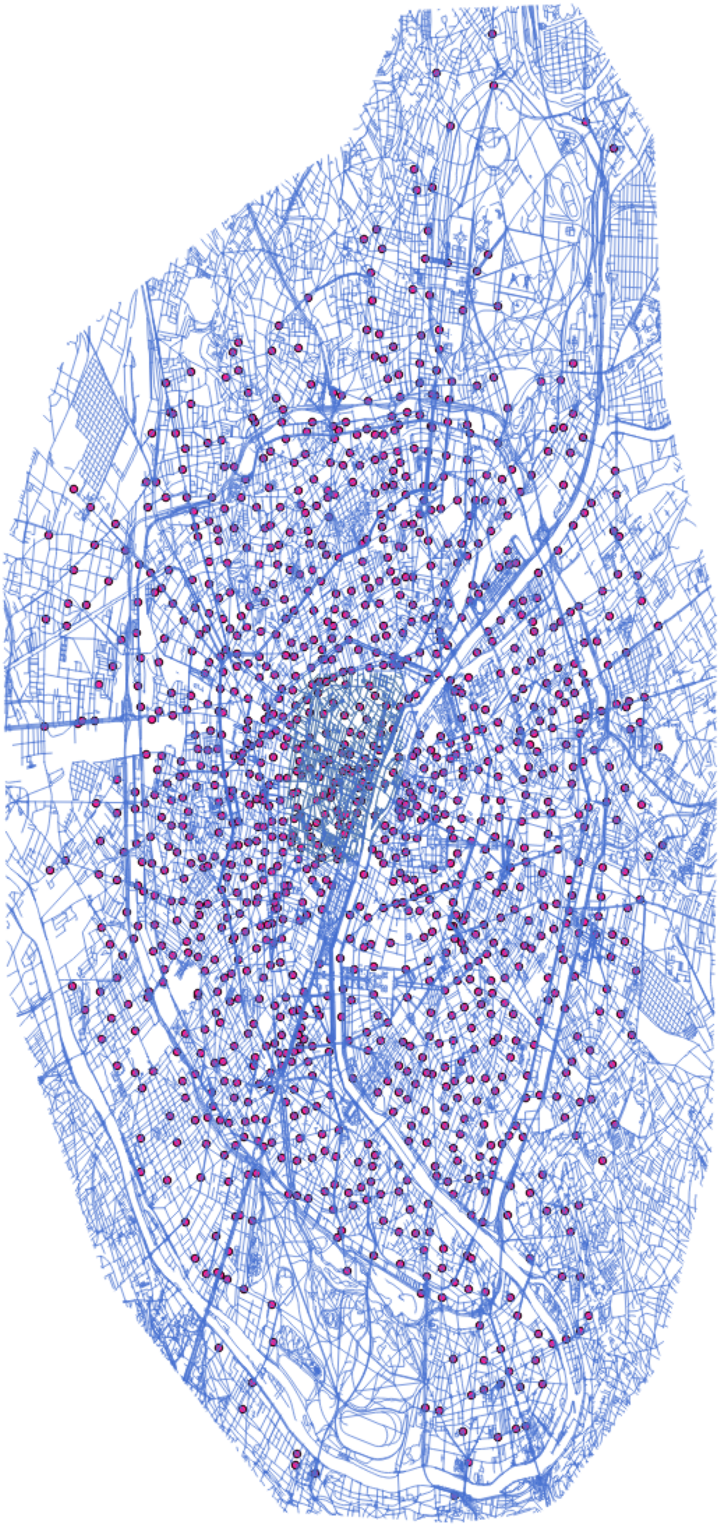


Fig. 6: Spatial representation of the system, including street network (blue) and docking stations (red).

```
#in order to deal with temporary interruption , we have a fixed local file
#with remaining zip files to download , which is updated progressively

n=$(cat remainingFiles | wc -l)

while [ $n -gt 0 ]
do
  fileName=$(head -1 remainingFiles)
  echo "Getting file "$fileName
  sshpass -p $PASSWORD scp -r $USER@$HOST:$PATH/archives/$fileName ../../Data/data
  #unzips
  echo $fileName |
      awk -F"." '{print "unzip -j -o ../../Data/data/"$1".zip -d ../../Data/data/"$1}' | sh
  #puts dir name in temp local file
  echo $fileName | awk -F"." '{print "echo "$1}' | sh >currentDir
  #calls R for file creation , that uses currentDir file and creates csv
  echo "Calling R extraction script..."
  r -f csvImport.R
  #cleaning zip and dir
  rm ../../Data/data/$fileName
  rm -R ../../Data/data/`cat currentDir`
  #putting head of files away
  tail -n `expr $n - 1` remainingFiles > tempfiles
  cat tempfiles > remainingFiles
  rm tempfiles

  n=`expr $n - 1`

done
```

Table 4: Remote data collection shell script.
This script is client-side and aims to collect and process remote raw data. File to be collected on the remote server are stored in a file (created manually for more flexibility). They are downloaded, unzipped and R script transforming heavy json into lightweight csv is called.

3 Agent-Based Modeling

3.1 Evaluation criteria

In order to quantify the performance of the system, to compare different realizations for different points in the parameter space or to evaluate the fitness of a realization towards real data, we need to define some functions of evaluation, proxies of what are considered as “qualities” of the system. Indeed, a model in itself has no use if it can not be projected in an interpretable euclidian space (in analogy to events in a probability space that take sense only through the Transfer Theorem allowing to define real probabilities).

Temporal evaluation functions

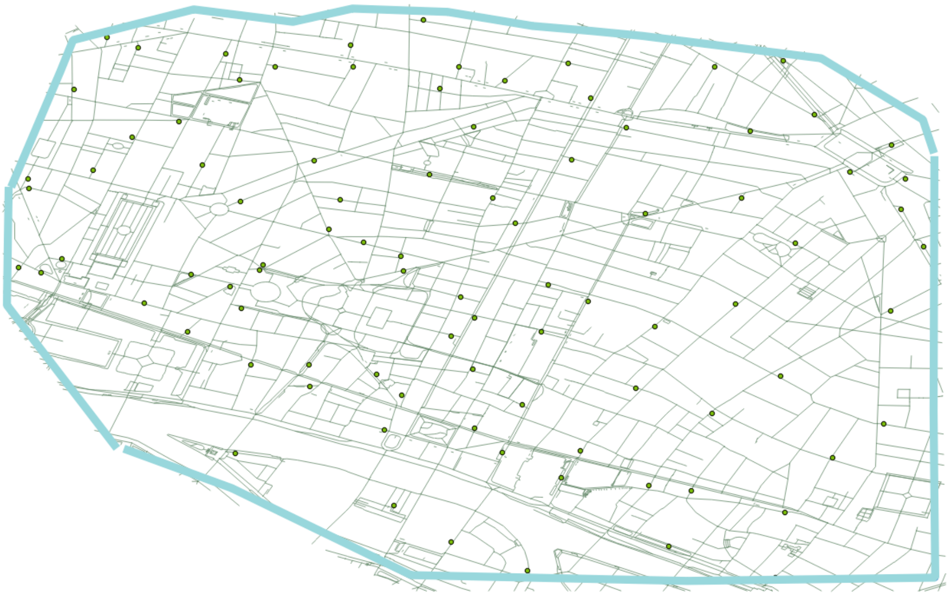


Fig. 7: Construction of GIS data for a particular district. We extract the sublayer of roads (keeping all hirearchy levels in such a size of district), docking stations, and boundaries that has to be manually constructed (it is then used to determine in- and outpoints by intersecting with street network).

These are criteria evaluated at each time step and for which the output on the all shape of the time-series will be compared.

- Mean load factor
$$\bar{l}(t) = \frac{1}{|S|}\sum_{s\in S}\frac{p_b(s)}{c(s)}$$
- Heterogeneity of bike distribution: we aggregate spatial heterogeneity of load factors on each station through a standard normalized heterogeneity indicator, defined by
$$h(t) = \frac{2}{\sum_{s\neq s'\in S}\frac{1}{d(s,s')}}\cdot\sum_{s\neq s'\in S}\frac{\left|\frac{p_b(s,t)}{c(s)} - \frac{p_b(s',t)}{c(s')}\right|}{d(s,s')}$$

Aggregated evaluation functions

These are criteria aggregated on a all day quantifying the level of service integrated on all travels. We note $\mathcal{T}$ the set of travels for a realization of the system and $\mathcal{A}$ the set of travel for which an "adverse event" occurred, i. e. for which a potential dropping station was full or a starting station was empty. For any travel $v \in \mathcal{T}$, we denote by $d_{th}(v)$ the theoretical distance (defined by the network distance between origin and initial destination) and $d_r(v)$ the effective realized distance.

- Proportion of adverse events: proportion of users for which the quality of service was doubtful.
$$A = \frac{|\mathcal{A}|}{|\mathcal{T}|}$$

– Total quantity of detours: quantification of the deviation regarding an ideal service

$$D_{tot} = \frac{1}{|\mathcal{T}|} \cdot \sum_{v \in \mathcal{T}} \frac{d_r(v)}{d_{th}(v)}$$

3.2 Model Parametrization

The model was designed in order to have real proxies for most of parameters. Most of parameters described in the model description section of the main paper are taken from the literature from comparable contexts and problematics.

Mean travel speed is taken as $\bar{v}$ =14km.h^{-1} from [Jensen et al., 2010], where data of trips where studied for the bike system of the city of Lyon, France. To simplify, we take same speed for all bikers : $v(b) = \bar{v}$. A possible extension with tiny gaussian distribution around mean speed showed in experiments to bring nothing more. It has been shown in [O'Brien et al., 2013] that profiles of use of bike systems stays approximatively the same for european cities (but can be significantly different for cities as Rio or Taipei), what justify the use of these inferred data in our case. We also use the determined mean length of travel from [Nair et al., 2013] (here that parameter should be more sensible to the topology so we prefer extract it from this second paper although it seems to have subsequent methodological bias compared to the first rigorous work on the system of Lyon), which is 2.3km, in order to determine the diameter of the area on which our approach stays consistent. Indeed the model is built in order to have emphasis on travels coming from the outside and on travels going out, internal travels have to stay a small proportion of all travels. In our case, a district of diameter 2km gives a proportion of internal travels $p_{it} \approx 20\%$. We always take districts of this size with this fixed proportion in the project.

3.3 Model Implementation

Languages

The model was implemented in NetLogo [Wilensky, 1999] including GIS data through the GIS extension. Preliminary treatment of GIS data was done with the open geographic information system QGIS [QGIS Development Team, 2009]. Statistical pretreatment of real temporal data was done in R [R Core Team, 2013], using the NL-R extension ([Thiele et al., 2012]) to import directly the data. For complete reproducibility, source code (including data collection scripts, statistical R code and NetLogo agent-based modeling code) and data (raw and processed) are available on the open git repository of the project [2].

Architecture

Following best practices typically used when developing reasonable size and complexity models with NetLogo (in the sense of more complex and complicated than a toy model but less than a fully operational data-driven model as a LUTI e.g.) we separate main model file from auxiliary source, containing agents or static procedures, in an analog way to classes in an object oriented language (being close in that sense to the GAMA language as an agent-oriented language). Model `CityBikes.nlogo` contains agent description, variables definition and includes.

We have then the following auxiliary files :

[2] at `http://github.com/JusteRaimbault/DiscreteChoiceBikeSharing`

- Setup
  - `setup.nls` Main setup function
  - `gis-setup.nls` GIS specific setup (layer loading and transformation into abstract agents)
  - `r-setup.nls` Import of parametrization data
- Core
  - `main.nls` Main runtime
  - `bikers.nls` Bikers agents
  - `depqueue.nls` Static abstract class managing walking agents
- Evaluation : `evaluation.nls`
- Display : `display.nls`
- Exploration and Calibration
  - `exploration.nls` Parameter space exploration
  - `mapping.nls` Visualization of integrated trajectories
  - `calibration.nls` Gradient calibration algorithm
- Standard utilities

3.4 Model application

We give here details on model execution and application, by first showing example of potential behaviors, then developing statistical analysis for internal validation and finally investigation on user strategies.

Figures 8 shows a capture of the interface of the model. Figure 9 and 10 show example of possible outputs of the model for different parameter values.

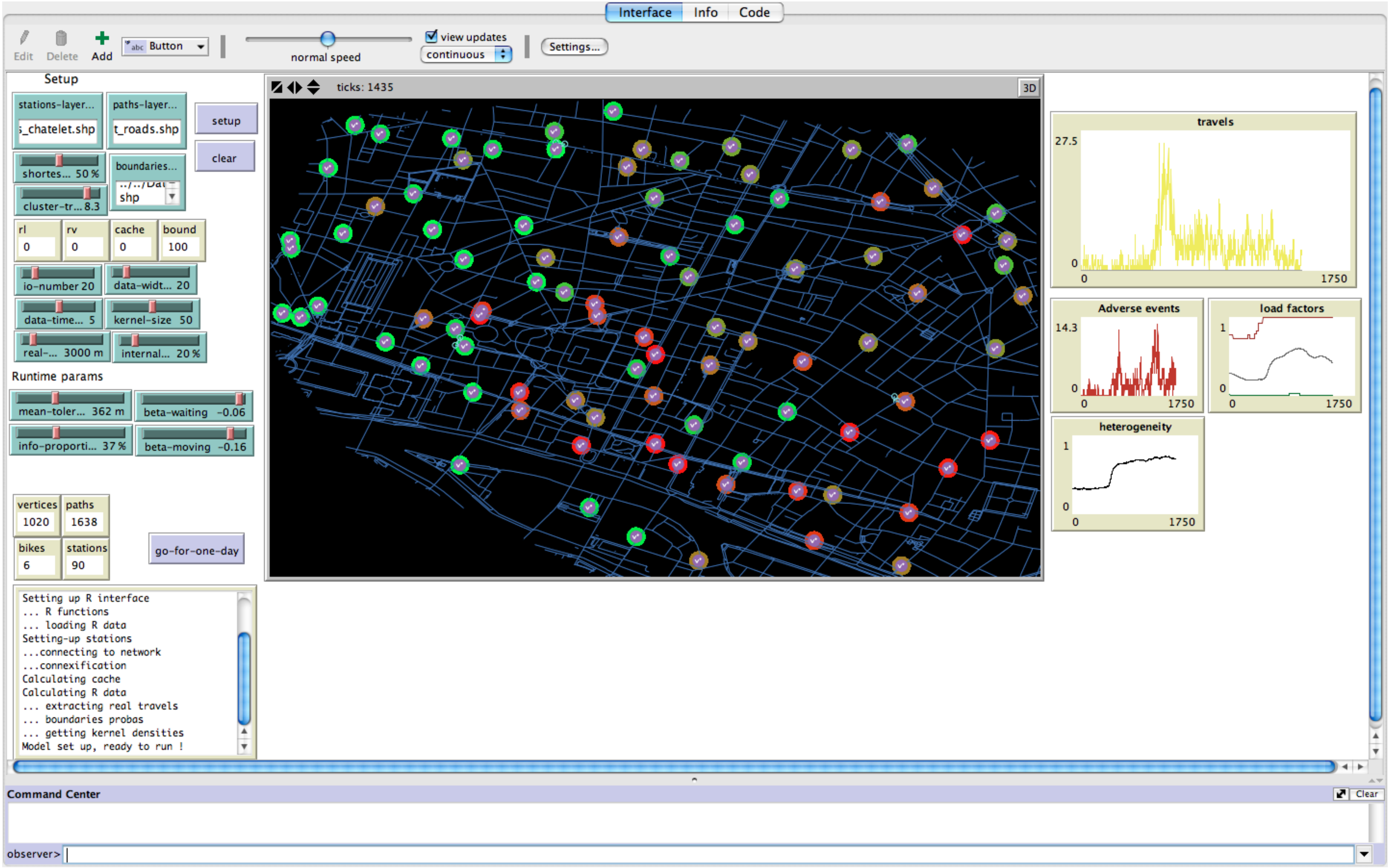
Internal consistence of the model

Before using simulations of the model to explore possible strategies, it is necessary to assess that the results produced are internally consistent, i. e. that the randomness introduced in the parametrization and in the internal rules do not lead to divergences in results. Simulations were launched on a large number of repetitions for different points in the parameter space and statistical distribution of aggregated outputs were plotted. Fig. 3 shows example of these results. The relative good gaussian fits and the small deviation of distributions confirm the internal consistence of the model. We obtain the typical number of repetitions needed to have a 95% confidence interval of length half of the standard deviation, what is around 60, and we take that number in all following experiments and applications. These experiments allowed a grid exploration of the parameter space, confirming expected behavior of indicators. In particular, the shape of MSE suggested to use the simplified calibration procedure presented in the following.

Investigation of user-based strategies

The first interesting application of the model is the investigation of user-based optimization strategies, that we detail here.

Influence of walking radius Taking for kernel-size and quantity of information the values given by the calibration, we can test the influence of walking radius on the performance of the system. Note that we make a strong assumption, that is that the



☑ view updates

continuous

ticks: 1435

stations-layer... _chatelet.shp

paths-layer... t_roads.shp

boundaries... ../../Dat shp

shortes... 50 %

cluster-tr... 8.3

rl 0

rv 0

cache 0

bound 100

io-number 20

data-widt... 20

data-time... 5

kernel-size 50

real-... 3000 m

internal... 20 %

mean-toler... 362 m

beta-waiting -0.06

info-proporti... 37 %

beta-moving -0.16

vertices 1020

paths 1638

bikes 6

stations 90

```
Setting up R interface
... R functions
... loading R data
Setting-up stations
...connecting to network
...connexification
Calculating cache
Calculating R data
... extracting real travels
... boundaries probas
... getting kernel densities
Model set up, ready to run !
```

observer>

Fig. 8: Capture of the NetLogo model interface.

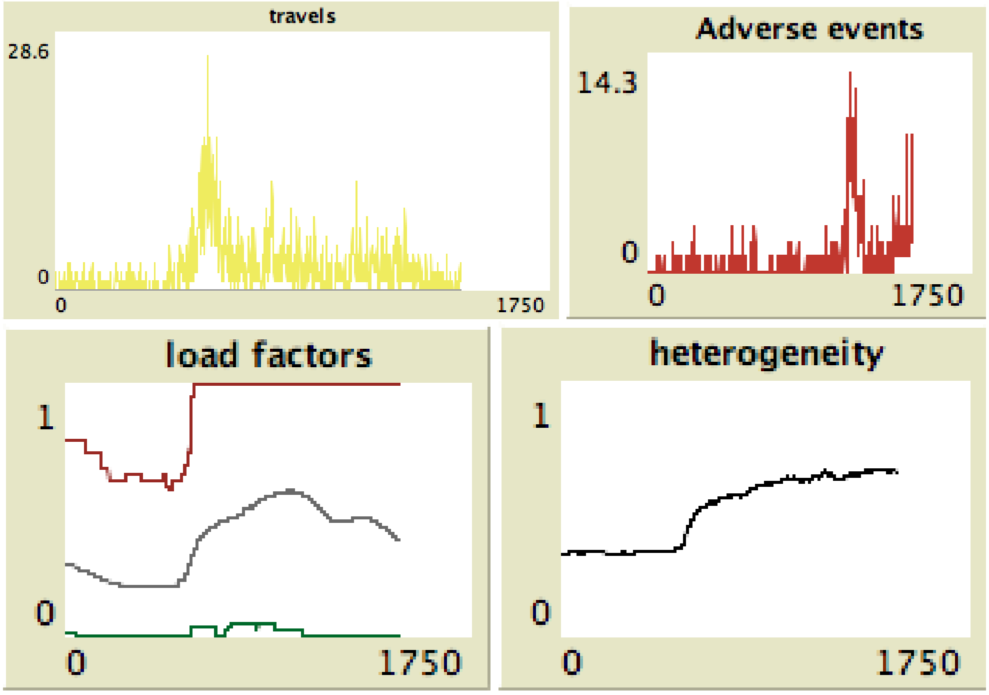


Fig. 9: Example of output time series for evaluation functions of the model, on a 24 hours single run. Here, tunable parameters are typically low ($p_{info} = 10, r = 50$) and discrete choice parameters are fixed.

calibration stay valid for different values of the radius. As we stand previously, this stays true as soon as we stay in a reasonable range of values (we obtained 300m to 600m) for the radius. Discrete Choice parameters stay also fixed to standard value. The influence of variations of walking radius on indicators were tested. Most interesting results are shown in figure 12. Concerning the indicators evaluated on time-series (h and $\bar{l}(t)$), it is hard to have a significant conclusion since the small difference that one can observe between curves lies inside errors bars of all curves. For A, we see a decreasing of the indicator after a certain value (300m), what is significant if we consider that radius under that value are not realistic, since a random place in the city should be at least in mean over 300m from a bike station. However, the results concerning the radius are not so concluding, what could be due to the presence of periodic negative feedbacks: when the mean distance between two stations is reached, repartitions concerns neighbor stations as expected, but the relation is not necessarily positive, depending on the current status of the other station. A deeper understanding and exploration of the behavior of the model regarding radius should be the object of further work.

Influence of information For the quantity of information, we are on the contrary able to draw significant conclusions. Again, behavior of indicators were studied according to variations of p_{info}. Most significant are shown on figure 13. Results from time-series are also not concluding, but concerning aggregated indicators, we have a constant and regular decrease for each and for different values of the radius. We are able to quantify a critical value of the information for which most of the progress concerning

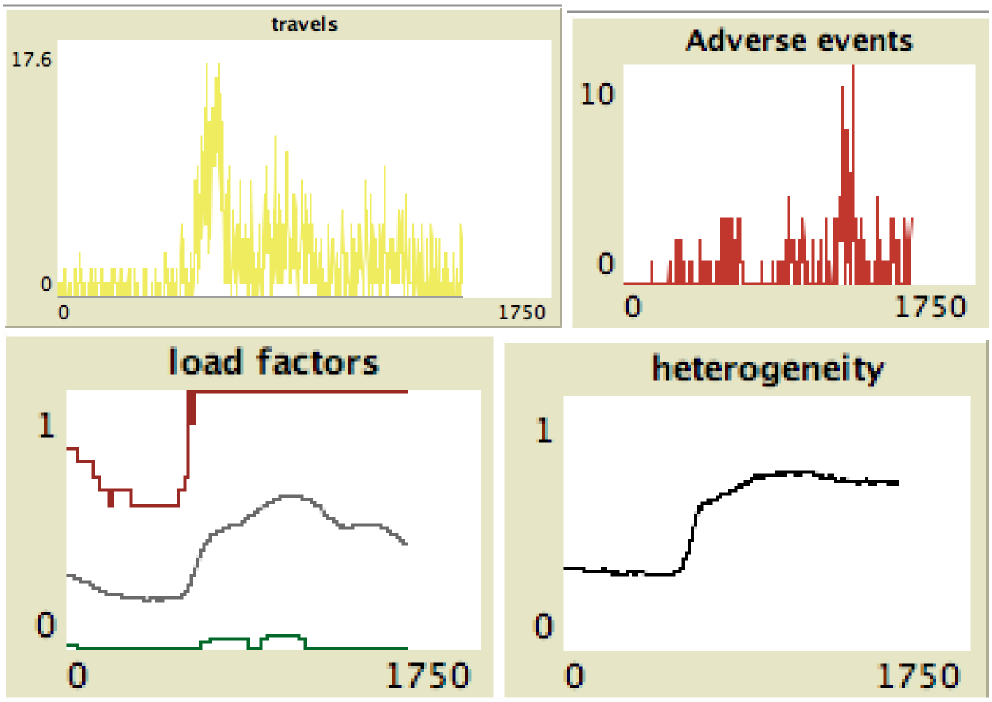


Fig. 10: Other example of run with higher values of parameters ($p_{info} = 40, r = 150$). We observe differences but of small magnitude. Final heterogeneity is smaller, meaning this case is more performant in that sense. Also, more travel are done since people are more about to walk to start a travel, what means also less adverse events, what we can see on the evening period. The order of magnitude of differences confirms that the data-driven character of the model strongly governs its behavior, since patterns depend on the first order of spatial geometry and temporal evolution of parametrization time-series.

indicator A (adverse events) is done, that is around 35%. We observe for this value an amelioration of 4% in the quantity of adverse events, that is interesting when compared to the total number of bikers. Regarding the management strategy for an increase in the level of service, that implies an increase of the penetration rate of online information tools (mobile application e. g.) if that rate is below 50%. If it is over that value, we have shown that efforts for an increase of penetration rate would be pointless.

3.5 Model Calibration

Runtime for one run of the model is about 45 seconds, and gradient simplex algorithm required around 50 iterations for convergent cases (it was set as maximal iteration number). We were able to run 100 times the algorithm with a total CPU time of 95 hours (almost 4 days), and found 76 runs satisfying convergence criteria. This result stays statistically poor as it is difficult to draw a conclusion on so few repetitions, whereas the dispersion is quite large in the large dimension parameter space. We plot in figure 14 confidence interval for β_d and the estimated (mean) value, computed on convergent runs, what may be not asymptotically valid depending on the convergence rate (that we can not now, lacking of computation power).

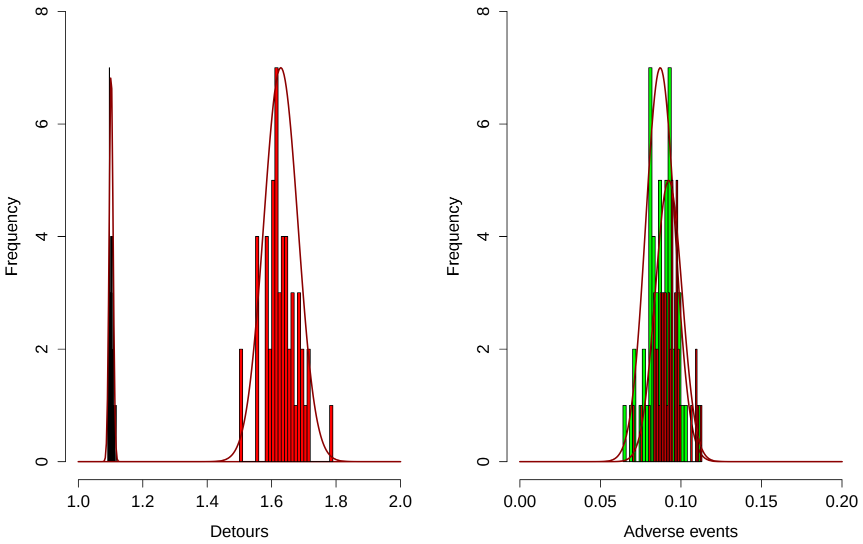
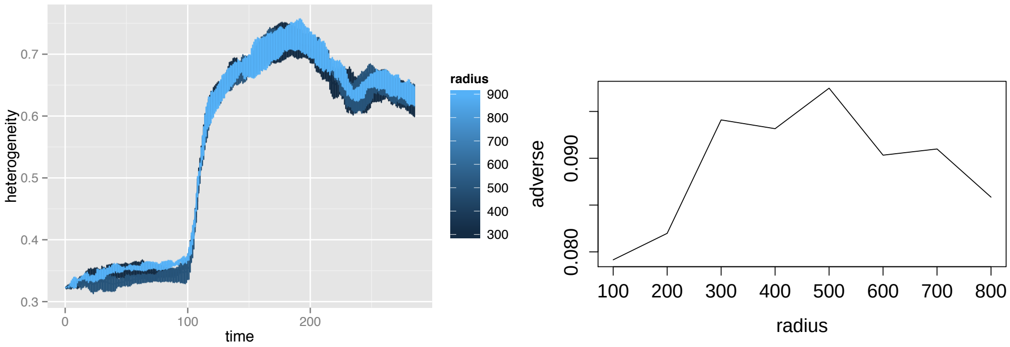


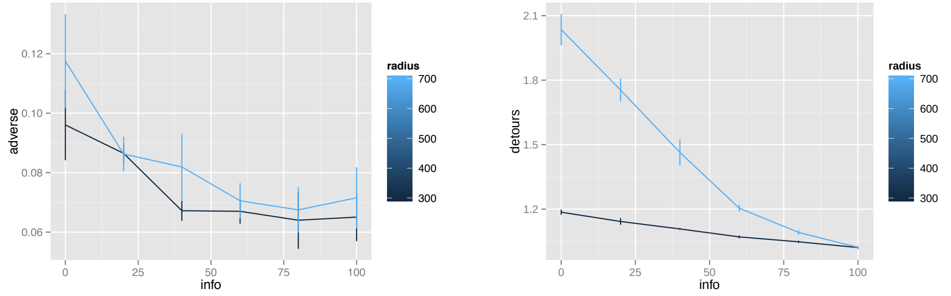
Fig. 11: Statistical analysis of outputs.
For some aggregated outputs (here the overall quantity of detours and the proportion of adverse events), we plotted histograms of the statistical distribution of the functions on many realizations of the model for a point in the parameter space. Two points of the parameter space, corresponding to $(r = 300, p_{info} = 50, \sigma = 80)$ (green histogram) and $(r = 700, p_{info} = 50, \sigma = 80)$ (red) are plotted here as examples. Gaussian fits are also drawn. The relative good fit shows the internal consistence of the model and we are able to quantify the typical number of repetitions needed when applying the model : supposing normal distributions for the indicator and its mean, a 95% confidence interval of size $\sigma/2$ is obtained with $n = (2 \cdot 2\sigma \cdot 1.96/\sigma)^2 \approx 60$



(a) Time series of heterogeneity indicator $h(t)$ for different values of walking radius. Small differences between means could mislead to a positive effect of radius on heterogeneity, but the error bars of all curves recover themselves, what makes any conclusion non-significant.

(b) Influence of walking radius on the quantity of adverse events A. After 400m, we observe a relative decrease of the proportion. However, values under 300-400m should be ignored since these are smaller than the mean distance of a random point to a station.

Fig. 12: Results on the influence of walking radius.



(a) Influence of proportion of information on adverse events A for two different values of walking radius. We can conclude significantly that the information has a positive influence. Quantitatively, we extract the threshold of 35% that corresponds to the majority of decrease, that means that more is not necessarily needed.

(b) Influence of information on quantity of detours D_{tot}. Curves for $r = 300m$ and $r = 700m$ are shown (scale color). Here also, the influence is positive. The effect is more significant for high values of walking radius. The inflection is around 50% of users informed, what is more than for adverse events.

Fig. 13: Results on the influence of proportion of information.

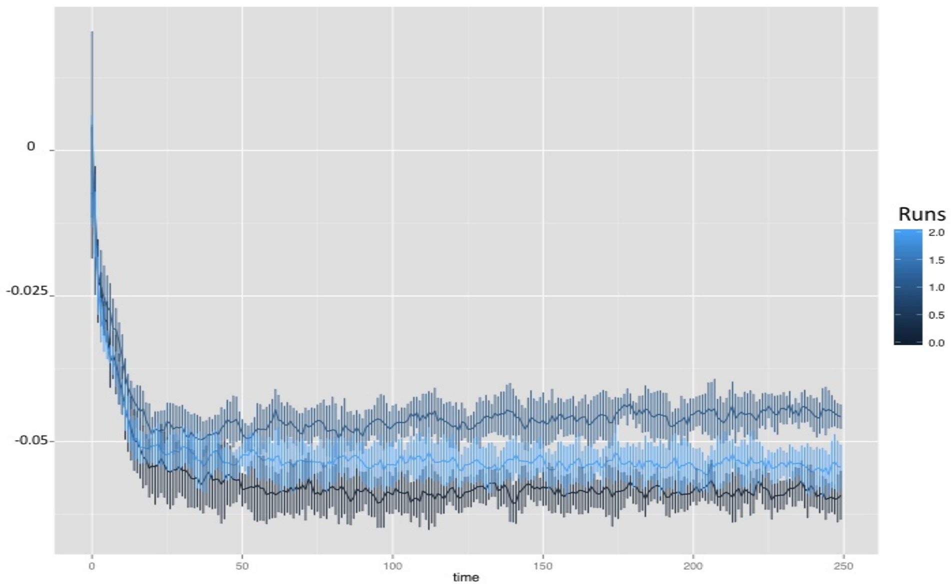


Fig. 14: Convergence of β_d for the confidence interval and estimated value, computed on convergent runs (may be therefore biased).

References

Jensen, P., Rouquier, J.-B., Ovtracht, N., and Robardet, C. (2010). Characterizing the speed and paths of shared bicycle use in lyon. *Transportation research part D: transport and environment*, 15(8):522–524.

Nair, R., Miller-Hooks, E., Hampshire, R. C., and Bušić, A. (2013). Large-scale vehicle sharing systems: Analysis of vélib'. *International Journal of Sustainable Transportation*, 7(1):85–106.

O'Brien, O., Cheshire, J., and Batty, M. (2013). Mining bicycle sharing data for generating insights into sustainable transport systems. *Journal of Transport Geography*.

Parkin, J., Wardman, M., and Page, M. (2008). Estimation of the determinants of bicycle mode share for the journey to work using census data. *Transportation*, 35(1):93–109.

QGIS Development Team (2009). *QGIS Geographic Information System*. Open Source Geospatial Foundation.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Thiele, J. C., Kurth, W., and Grimm, V. (2012). Agent-based modelling: Tools for linking netlogo and r. *Journal of Artificial Societies and Social Simulation*, 15(3):8.

Wilensky, U. (1999). Netlogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.