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

Discrete Choice Models for Bike-Sharing Transportation Systems

Inference of Discrete-Choice parameters by coupling Statistical Analysis and Agent-based Modeling

Juste Raimbault[1,2,a]

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

Abstract. The realization of developed user surveys, generally used to elaborate and calibrate transportation systems models with an aim of improvement of the level of service, is heavy to implement and has a certain cost. We propose an alternative method to obtain similar results, combining poor quality survey data and raw data on system evolution, fed into an hybrid statistical and agent-based model. The method is described on the particular case of bike-sharing transportation system for which surveys are destined to parametrize discrete choice models. We work on the real system of Paris (Vélib) with 1 year of recent data on system dynamics. It should be easily generalized to other kind of systems and research questions, as the core principle of hybrid modeling has already been well documented in the literature (see e.g. recent works in archeology [Crabtree and Kohler, 2012]). We find that we can retrieve discrete choice parameters indirectly by parametrizing and calibrating an agent-based model including discrete choice models at finest granularity, and propose a novel method to assess the robustness of the inference.

1 Introduction

1.1 Context

Discrete Choice Modeling has been widely used in Transportation Science, showing both a powerful theoretical framework in formalizing crucial but fuzzy components of a social systems (i.e. choices of irrational agents) and mature applications in transportation systems management with dedicated frameworks (such as Biogeme [Bierlaire, 2006] software) including e.g. mode share or route choice predictions, adapted pricing solutions, more accurate planning of public transport, or other application detailed in [Ben-Akiva and Bierlaire, 1999].

With the rise of novel transportation modes in cities, a precise knowledge of potential user behavior and pattern of system use is necessary during conception and

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

management phases. Therefore the use of discrete choice models can be a powerful asset. We are particularly interested in Bike-sharing transportation systems, which have been presented as an ecological and user-friendly transportation mode and performs a good complementary with classic public transportation modes ([Midgley, 2009]). Although the direct positive impact on public health is not systematically proved and the modal share report does not necessarily occurs from polluting modes as the recent review in [Fishman et al., 2013] highlighted, we follow [O'Brien et al., 2013] on the fact that the success of such systems in Europe and its current transposition in other regions of the world such as China ([Liu et al., 2012,Geng et al., 2009]) or the United States ([Gifford and Campus, 2004]) is an acceptable reason for helping its development by the study of its mechanisms and possible measures to improve its level of service. Our context is therefore the use of discrete choice models at small granularity (user) to investigate behaviors and mechanisms in a given bike-sharing system .We focus on the Paris case, and qualitative conclusions do depend on cultural and geographical frame [O'Brien et al., 2013], but their extension or generalization is not our purpose here. However, as the applied method is quite generic, that aspect of our work could be transposed to other systems or fields in further works.

1.2 Project Description

As the final result of this scientific project is the result of many contradictory steps and concrete difficulties to follow the initial sketch of plan, in particular the impossibility to realize the expected questionnaire and gather the data that would have allowed a fine Discrete Choice modeling, the general plan may not seem natural at first sight, thus we propose to explicit a rough temporal timeline of research questions that drove the project, and its link to that final form. Sequentially, we explored the following problematics :

1. Conceive and realize a precise survey, in order to estimate discrete choice models, particularly on user choices done when taking or dropping a bike.
2. Insert the discrete choice model in an existing (but basic) agent-based model for Paris bike-sharing system described and explored in [Raimbault, 2014].
3. Because of practical issues, initial survey that was expected to be revealed preferences and with some characteristics, was replaced by a stated preferences survey with more simple characteristics and attributes.
4. The problem became : how to use this poor quality data ? Can we however answer the initial question ?
5. We elaborated the calibration method to infer DC parameters and partially tackle it. Core results became then more of methodological than practical type.

We can mention an explored direction that is not reported here because of issues in implementation and low priority regarding research question. We tried to proceed to a crossed systematic review on all bike-sharing and discrete choice literature, using a novel method which provide a precise dynamic scientific taxonomy of the fields (general method introduced and applied in [Chavalarias and Cointet, 2013]). Its advantages are that it provides information on the evolution of the field and therefore allows in general to make unexpected connections between concepts or fields, suggesting potential research directions. A sketch of Java implementation using Graphstream library for dynamic graph handling [1]. Text-mining of the corpus of citations extracted from Web of Science was done thanks to the online tool provided by authors [2]. Un-

[1] http://graphstream-project.org/
[2] http://manager.cortext.net/

fortunately the method was not successful so we do not develop it here as it has in that state of things few relation to our purpose.

Note that all source code and data used in the project are openly available on the Github repository of the project [3]. It is a crucial point for practice of Open Science and reproducibility [Ram, 2013].

We describe in a first section the different data collection processes which were a considerable part of our project. We then present Discrete Choice Models proposed and estimated. Sec. 4 describes statistical analysis of raw data, including a case study on missing data inference. Sec. 5 describes the agent-based model and its discrete choice extension, used in the last part to infer discrete choice parameters through model calibration. We conclude by proposing a method to study stability of multidimensional calibration.

2 Data Collection

2.1 Discrete Choice Questionnaire

In order to collect data for Discrete Choice modeling through surveys, we conceived and implemented a generic web-application for questionnaire administration, which could be used either for direct web survey or during terrain survey. Because of the uncertainty on the form of the questionnaire and on the feasibility of the survey, we proposed a totally adaptable application allowing variable questionnaire structure. Thanks to a restricted access interface, an administrator can create a questionnaire, specifying general description characteristics, attributes, choices, number of output scenarii , type of experimental design. Structure is stored in static tables in database, whereas a dedicated dynamic table is created to store user answers. Questionnaire administration can be done by an authorized user (personal survey) or by anonymous users on the web, for which security requirements (in particular an anti-bot captcha) were added. Furthermore, a data export module allows to export data directly to the specific Biogeme format (csv-like), given a filter specification file (for dummies formatting e.g.).

The application was implemented as a server-side php-application, including a large use of jQuery and Ajax request that allow lightweight navigation and a simplified architecture of contents and utilities pages. Database is efficiently managed by a classic Mysql base (since expected data size stays small, at least always smaller than 1Mo), exploited through current php drivers (PDO). It is installed on a web server, implying security management because of public access of a part of the questionnaire. For reproducibility sake, platform code is Open Source [4], as for data available at the public export URL given in the following.

Concerning data export [5], a particular treatment was needed as there is no necessarily direct correspondence between raw variables and Discrete Choice Models variables, for example in the case of dummy variables. Therefore, a text file declaring filter format is needed under the following format : each row is an output variable (column in output .dat file) and follows the template [BIOGEME_VARIABLE_NAME ; BASE_VARIABLE_NAME ; BASE_VALUE]. If BASE_VALUE = NULL, value is unchanged (quantified variables). In an other case, output is a dummy variable which is 1 if and only if value of the record correspond to the provided value. For more flexibility, we also implemented a mapping which allow to apply a function to quantitative variables (formatted by BASE_VALUE=NULL :v1-f(v1) :v2-f(v2) :... :vn-f(vn)).

[3] at https://github.com/JusteRaimbault/DiscreteChoicesBikeSharing

[4] available at https://github.com/JusteRaimbault/Questionnaire

[5] Data export public URL : http://37.187.242.99/Questionnaire/php/utils/export.php

This platform allowed to test experimental questionnaires, to realize the simplified survey and to quickly generate data used in the second discrete choice model described before.

2.2 Raw Data on System Dynamics

Type of data We also collected raw data data for the statistical analysis and parametrization of the final model. They are public available data (open data) from the bike-sharing system of Paris ("V'Lib"), provided by the operating company in direct time on a dedicated website (url `http://api.jcdecaux.com`, for which the format of request is specified on `developer.jcdecaux.com`). It provides only the status of docking stations at the request time so we had to automatize the data collection process on a large time period in order to have significant time-series. Process is detailed in the following.

We chose to gather such kind of data first because the obtention of more precise data from the company can raise several problems such as confidentiality issues or more constraining for our research, lead to an lack of independence in the design of the modeling process, since most of the time delivering of data had its price that is at least answering to some question asked by the company. Secondly, we argue that our experience will be one way of testing the possibilities and limits of open data: if the public provided data can lead to relatively good results compared to what can be obtained with a larger set. However, if our research process becomes quickly limited by the lack of precision or diversity in the data, that will bring one essential question on front, that is that open data does not necessarily means freedom not exhaustivity, and that the control of the provided data can implicitly be highly dangerous for the global opening process. On that point, we follow Banos in [Banos, 2013] when he argues that a necessary condition of an open scientific cumulative process is a total transparency in the methods and an exhausting sharing of implementations of models of simulation and of data. Furthermore we wanted to avoid any risk of implicit reporting spin since it stays a major issue today for the quality of research as it is claimed by Ravaud & *al.* in [Boutron et al., 2010]. The purpose of data sharing by the company in our case was surely, because of the nature of the available data, i. e. only current time stations status, nothing more than current time information and mapping. However, we will see that we can use them for statistical analysis and obtain quite good results.

Data collection process A script requesting current data to the API and saving it into a file have been written and scheduled each 5 minutes on a remote server (we did not choose finer temporal granularity for a material reason, because the size of data becomes quickly huge and storage becomes then an issue). Data on remote server is then zipped everyday for storage purpose. When needed we download the files and process them with R using [Couture-Beil, 2013] in order to store them locally on a reduced form (csv) that can be called directly by our data processing algorithms. Note that it would have been more logical to process the data remotely and store them under the reduced form but technical reasons were an obstacle (in particular the installation of R on the remote machine). We also extracted from extensive files static information such as numbering and coordinates of docking stations, what have been useful after for example to create a geographical file for map drawing with [Keitt et al., 2011]. Fig. 1 shows a flowchart of the data collection and primary processing process. We collected data for all Paris during around 3 month, following statistical analysis are done on these data.

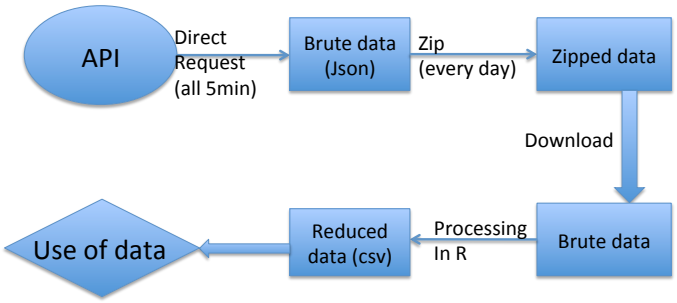


Fig. 1: Flowchart of raw data collection process

3 Discrete Choice Modeling

3.1 Towards a precise discrete choice behavioral model

The initial aim of our project was to collect survey data to estimate precise behavioral discrete choice models for bike-sharing users. Many factors of choice had to be considered, implying nested models. Among others, they included modal choice, and in the bike-sharing nest, route choice, destination choice, docking station choice and redirection procedure in the case of an adverse event (full station). A precise modeling of user behavior would have been essential for a precise parametrization of an agent-based model of bike-sharing. Indeed, rules where decisions are involved and basic probabilities are used in a root model, discrete choices are the key for a step of desegregation further.

Table 1 give a first sketch of revealed preference questionnaire, designed to answer the above questions. It could not be tested nor implemented, thus its use was nothing more than the inspiration of the simplified stated preferences survey described below.

3.2 Simplified incomplete model

As described in introduction, practical reasons made the initial survey impossible to realize, and a simpler stated preference survey with relatively few answers was imposed by external constraints. The conception and realization was done in a joint student project [Bourcet et al., 2014] which beneficed from the developed online interface described above. Following the general context of our study, interesting results from that modeling that we reuse are potential intervals for some discrete choice parameters, allowing an efficient bootstrap of calibration procedure.

Let review briefly the structure of the model and main results.

Accounted socio-economic variables of users are

- age
- professional categorization

Table 1: Proposed detailed survey for behavioral Discrete Choice Models.

Category	Variable	Observations
Survey context	Localization	Determined by geolocalization
	Day, time	
	Weather	Collected automatically
	Remarks	
Socio-economic profile	Age	Determine relevant age classes
	Occupation	idem
	Income	Delicate question
	Household infos	Size, residential localization
Transportation profile	Motorization	Personal and household
	Public transport subscription	Range
	Frequent O/D	By mode and motive
	Frequency	idem
	Mean distance	idem
	Bike-sharing subscription	
	Specific bike-sharing O/D	
	Car-sharing subscription	
	Car-sharing O/D	
Bike-sharing profile	Typical schedule of use	
	Frequencies	Weekly, monthly
	Typical week	
	Mean traveled distance	
	Factor of choice	Subjective
	Route choice procedure	idem
	Concurrent mode	
	Complementary mode	
	Walking distances to bike	Origin and Destination
	Use of electronic device	Type, moment, purpose
	Subjective impression	System, level of service, typical behavior
	Docking stations choice	Experience, expected charge, proximity, random

– regular user
– average distance to public transport

We also consider weather as a characteristic as it will be taken as exogenous in stated scenarii.

Choices are simplified as they are only between bus and bike-sharing (comparable ranges and speeds). Attributes of choices are, with respective notations : travel distance D, expected travel time t, time to find a bike t_B, time to drop a bike t_D, bus delay D and bus comfort C. Attributes levels are chosen with 3 level for each discretized attribute (from t_B to C).

Scenarii are given following a Random Experience Plan (with deletion of dominated scenarii), which thanks to the 150 data rows collected gave a reasonable random sampling of the space.

Discrete choices utilities are linear in the simpler model estimated : denoting X_{bus} or X_{bike} attributes and characteristics of each choice, we write $U_{bus} = \sum \beta_{X_{bus}} X_{bus} + \varepsilon_1$ and $U_{bike} = \sum \beta_{X_{bike}} X_{bike} + \varepsilon_2$.

Models were estimated with Biogeme and robust value were found for most of parameters, with expected signs. For the following, important results are the confidence intervals : $\beta_D \in [-0.06, 0]$ and $\beta_{t_B} \in [-0.25, -0.15]$.

4 Statistical Analysis

As explained before, the central question of our project has quickly become how to deal with bad quality data, in two sense : not necessarily expected variables observed (realization of an other questionnaire) and poor statistical robustness because of a small number of answers (around 150 rows in database). Therefore, a statistical analysis appeared to be the best way to

- Test for methods to fill missing rows (handling of missing data)
- Understand underlying data structure and therein system behavior
- Propose missing variables inference

the third point being crucial for the following as it allows to parametrize the agent-based model through Origin/Destination inference. This paragraph follows these three aspects as a plan.

4.1 Handling Missing Data : a case study

Facing the issue of poor quality data, we explored ways to tackle missing data. We studied more precisely some techniques as a case study. We comment on recent progresses done to compare different methods of handling missing data, that are called imputation methods for missing baseline data. Such methods are particularly scrutinized by statisticians working in Therapeutic Evaluation but methods and results can be easily generalized to any domain involving statistical models on potentially incomplete data. We focus on [Crowe et al., 2010] and [Mitra and Reiter, 2010] that try to compare methods for imputation, and the consequences of choices on final results of evaluation of treatment effect.

Missing data in dataset are an essential problem to tackle in statistical analysis of treatment effect. One does not simply remove patients with missing data since it brings more bias and leads to unbalanced data at it is explained in the discussion of [Crowe et al., 2010]. First methods for imputation of baseline data were proposed by Rubin[Rubin, 2009]. Multiple imputation methods are compared between themselves (and with mean imputation and with results with no missing data in the binary outcome case). The missing values in the set of values $\mathbf{X}$ are filled through statistical model (that differs between the methods), and it is done m times to obtain $(\mathbf{X}_i)_{1\leq i\leq m}$ completed datasets from which propensity scores are calculated.

Filling missing data The objective of the first work with binary outcome (death or not) by Crowe & *al.* in [Crowe et al., 2010] is to compare through numerical simulations the influence of the model chosen to proceed to multiple imputation on the bias and variance (confidence interval in fact) of the results. The method for numerical simulation is:

- Random normal generation of 6 explanatory variables and assignment to $\mathbf{X}$
- Assignment of treatment following a logit model controlled only by the first variable (with a variable parameter)
- Assignment of treatment effect also following a logit, depending on both treatment and the first variable (also variable parameters)
- Suppression of data (or creation of missing data in other terms), for some "completely at random", i. e. setting a given proportion of values to random uniform distributed values. If one rows is set to missing data, the two following variables will also have missing data on that row. Then some data are suppressed "at random"; this time the data supposed to be suppressed will be if it is positive (so it depends on the value of the variable).

- Then all possible methods for imputation are tested:
  - with complete data (as a comparison)
  - with treatment mean imputation (no multiple imputation, that mean replacing missing data by the mean)
  - multiple imputation using as controls in the regression model for imputation only explanatory variables
  - multiple imputation using also treatment
  - multiple imputation using explanatory variables, treatment and outcome
- bias and confidence intervals are estimated in each case

The main conclusion is that the best multiple imputation method was the one controlled by explanatory variables, treatment and outcome. It is also obvious to note that "complete data" method (i.e. deleting the row with missing data) leads to less bias but to strong variance, so quite less significance of the results on treatment effect, as we stated in introduction. Although the multiple imputation brings bias compared to a scenario with all data (called the "gold scenario" in the paper), it is sure that if baseline data are missing, best choice is to use multiple imputation with controls by variables, treatment and outcome.

Sensitivity of estimators to missing data An other way to tackle the question is to directly look at the sensitivity of estimators for mean treatment effect, noted τ, on missing data and methods used to reconstruct it. Indeed, two approaches are possible (called the Within and the Across approach). The Within is called like that because estimates first treatment effect in each completed dataset by matching and then estimates the treatment effect by taking the mean of all estimated treatment effect, so the matching is done "within" each completed dataset. In the Across method, a mean propensity score is first calculated from all propensity scores of completed datasets and then the matching is done on that mean.

Formally, if we note $\mathbf{e}(\mathbf{X})$ the vector of propensity scores for data $\mathbf{X}$ and $\tau(\mathbf{e})$ the corresponding estimated treatment effect through matching, we have the estimator in the Across approach given by

$$\hat{\tau}_{Ac}(m) = \tau\left(< \mathbf{e}(\mathbf{X}_i) >_{1\leq i\leq m}\right)$$

whereas the estimator in the Within approach is

$$\hat{\tau}_{Wi}(m) =< \tau(\mathbf{e}(\mathbf{X}_i) >_{1\leq i\leq m}$$

One can have a generalized point of view to gather both approach, by considering the extended Across approach: if $(\hat{\tau}_{Ac}^{(j)}(m))_{1\leq j\leq r}$ are r realizations of an estimator in the Across approach, we can take the mean and define the generalized estimator

$$\hat{\tau}(m,r) =< \hat{\tau}_{Ac}^{(j)}(m) >_{1\leq j\leq r}$$

that is quite practical since we obviously have $\hat{\tau}_{Wi}(m) = \hat{\tau}(1,m)$ and $\hat{\tau}_{Ac}(m) = \hat{\tau}(m,1)$.

- First, artificial data with two covariates are simulated with Logit models both for treatment distribution and indexes of missing data, and different parameters for the logit for treatment assignment are tested (i. e. treatment assignment depending on only one covariate, and one both with equal weights). Bias and variance are estimated in each case.

- The comparison is also done on real data for different values of (m, r) in the generalized estimator. An heuristic for the selection of the best values for the couple is proposed, what can be considered as a compromise between the two initial approaches.
- Theoretical formulation of limit variance estimators are proposed for both cases but they are not used. The development and study of such estimators is pointed out as a key-point in further research.

The main results to this second question are that, for artificial data, across method gives smaller bias than within method in the case where the covariate with most of missing data was most influent in treatment assignment, although variances were smaller for within method (with quite same mean square errors). In other parametrization of logit models, the two estimators gave quite the same results. Following the work done on genuine data, they finally advise to use generalized estimator, and use the heuristic for finding best values of (m, r) that can lead to the best compromise between bias and variance values.

4.2 Data-mining of raw data

4.2.1 Data visualisation

Many basic means for a global visualization of data behavior are available such as the ones proposed in [O'Brien et al., 2013], so we won't go too much into detailed representation since it is not the first purpose of our study. Note that this step is however essential, especially during the elaboration of algorithm and the choice of methods for statistical treatment.

To have an idea of the cyclic character of daily mobility patterns, we can plot the total number of available bikes at docking stations against time. If we suppose the total number of bikes constant over the time duration of the plot, what seems reasonable even on the all time period our data cover (even if there are surely variations because for example of bike reparations, they are surely negligible regarding the total number of bikes, which is around 15000), this plot is exactly the complementary of the quantity of current travel as a function of time, what allows to visualize mobility trends. Fig. 2 shows the obtained curve that fits the expected results, showing in particular the distinction between week days and weekends.

We can also for example draw maps for the understanding of spatial patterns in system use. One can expect for example to see distinction in time between residential and activity areas for the quantity of available bikes in stations. This allows to visualize global and local heterogeneity patterns. Fig. **??** shows an example of such maps on a particular district.

4.2.2 Extraction of patterns

A first step in the treatment of data is to extract typical patterns in use of the system. In [Vogel et al., 2011], data-mining techniques, and especially clustering of activity profiles, are used to extract typical patterns in station use. We propose to use similar methodology in order to identify typical overall day profiles and classify them. We expect to be able to differentiate weekdays from week-ends for example, but also see the influence of climate on use patterns. The clustering of time-series offer an alternative for a predictive model, as the cyclic model proposed in [Borgnat et al., 2009a,Borgnat et al., 2009b].

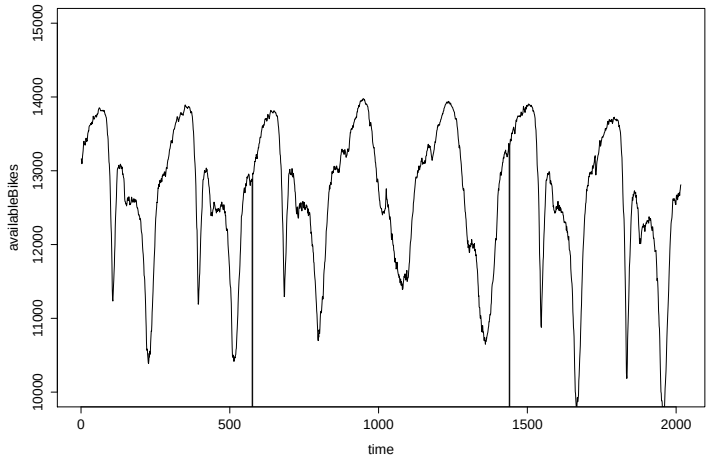
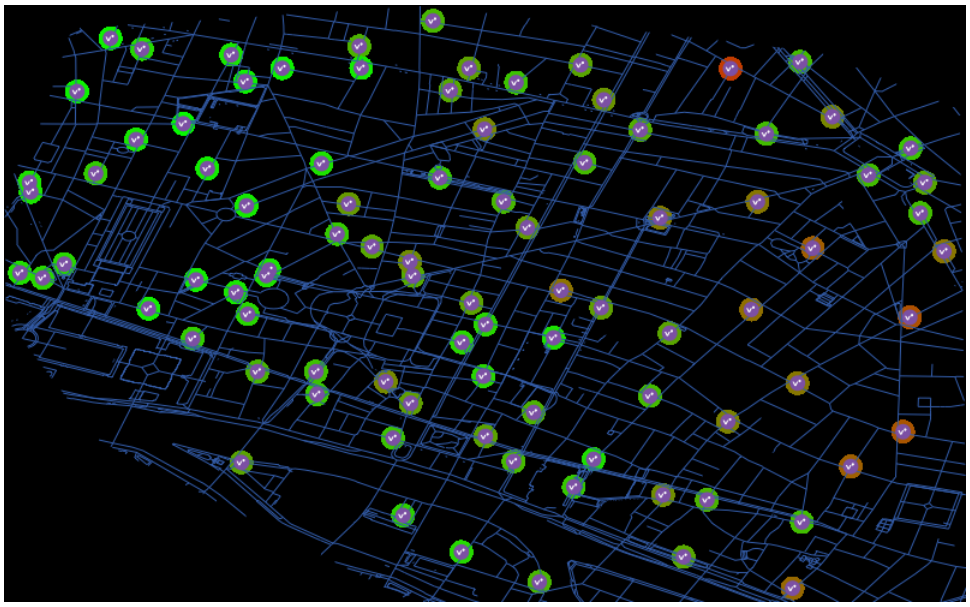


Fig. 2: Quantity of total available bikes over a week. We observe the typical patterns of the daily mobility, with two minima corresponding to morning and evening affluence. The two day in the middle correspond to saturday and sunday since the time-serie begins on a wednesday. These weekend days present only one minimum, what is logical (no affluence in the morning) and confirms the results of other studies.

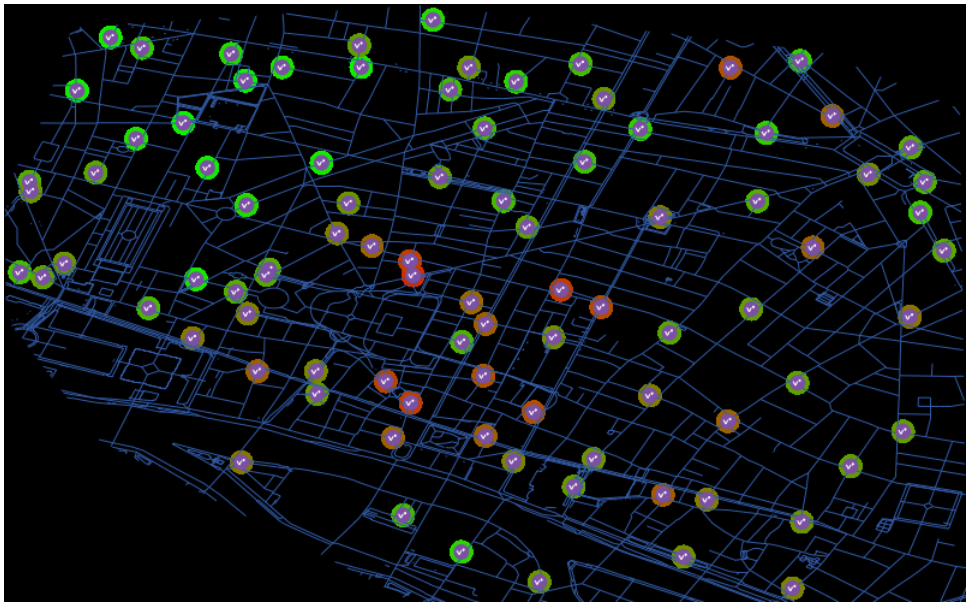
A day is exhaustively represented by the time-series, defined on all the stations of the system $s \in S$, and on a discrete time sample $T = \{0, \tau, ..., N\tau\}$ (with τ time step of the data, 5min in our case), $(b(s,t))_{s\in S, t\in T}$ of available bikes at each stations. Each station has a maximal capacity $c(s)$ that allow to define the number of free parking places $p(s) = c(s) - b(s)$ and the load factor which can be more convenient to work with since it is normalized $lf(s) = \frac{b(s)}{c(s)}$. The overall clustering process first aims to reduce the dimension of the representation of a day without loosing majority of information, and then to be able then to classify days and make predictions on the day characteristics from its data.

First the dimension is reduced through a sampling process that can be seen as a projection from the space of complete time-series to a space of smaller dimension. If $\varphi \in \mathbb{N}^{\mathbb{N}}$ is a extraction then the sampling is defined as the canonic projection $\mathcal{S} : \mathbb{R}^{|T||S|} \to \mathbb{R}^{|\varphi(T)||S|}$. The question of the value of the time step for sampling is important. We tried for many values and looked at the possible loss of information through the evolution of clustering coefficient regarding number of clusters. It appeared that we had still good precision for large time steps such as one hour. See fig. **??** for more precision on the influence of sampling step.

We proceed then for each day to a k-means algorithm on the sampled time-series (as described in [Warren Liao, 2005]), in order to reduce more the dimension needed to represent a day. Intuitively, that corresponds to a classification of stations according to their "profile". We take in practice 20 clusters, what allows to divide by 100 the dimension. The final step is to cluster the representations of the days for establishing a classification of days. With two clusters, one expect to isolate weekdays from weekends, although k-means can lead to bad results if cardinal of clusters appear to be imbalanced. In our case it worked quite well and we were able to reproduce that distinction. However, a finer distinction (e. g. between rainy and shiny days) was not possible and some work on a more specialized clustering algorithm (k-means is very general) would be needed to obtain more precise results. Fig. **??** shows the compar-

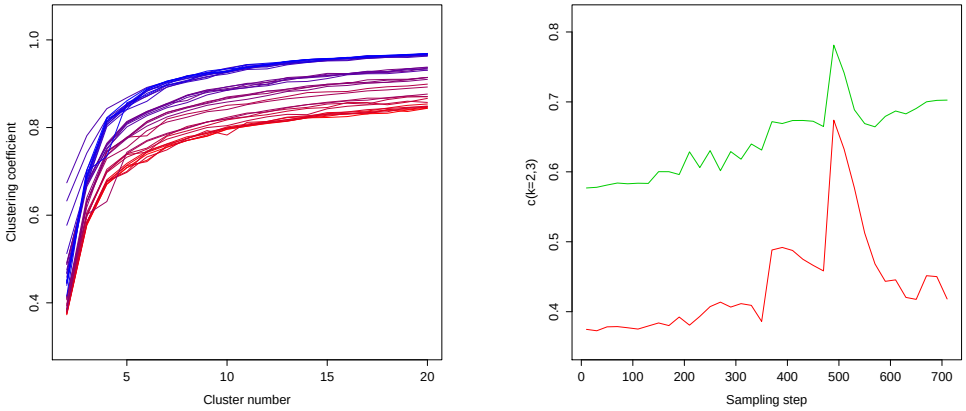


(a) Midnight



(b) Midday

Fig. 3: Examples of heatmaps at two different moments of the day for the district of Chatelet. The color indicates, from green that corresponds to an empty station, to red to a full one, number of available bikes. Since it is a working district and not residential, stations in the center are overloaded during the day but empty during the night as expected.



(a) Clustering coefficient as a function of cluster number for different values of sampling step. The more blue the curve is, the more sampling step is large. If the curve goes faster to 1, that means that points are less distinct and that statistical distribution contains less information. We observe a jump that is quantified in (b).

(b) Plot of the value of the clustering coefficient for k=2 (red) and k=3 (green), as a function of sampling step. We see the significative loss of information around a step of 400 minutes, which should correspond to the disappearance of pics in the curve, since they contribute significantly to the quantity of information.

Fig. 4: Influence of sampling interval on quantity of conserved information in the clustering process.

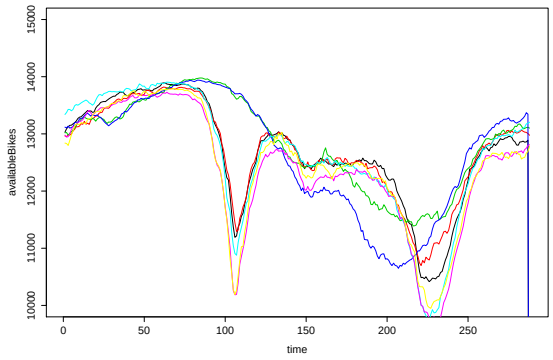
ison between real curves of available bikes and predicted curves by the clustering algorithm.

4.3 Inference of O/D fields

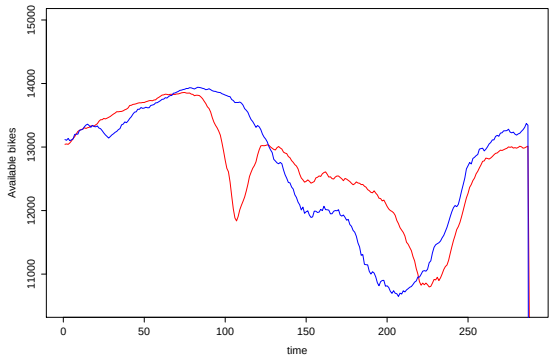
The core part of statistical analysis within the all project is indeed the inference of unknown variables, that are probability distribution of Origin and Destination fields of users, needed for model parametrization. Note that this question has its own scientific interest and has already been tackled in economic geography and transportation geography (see [Leurent, 2006] e. g.).

4.3.1 Geographically Weighted Regression

The method used for inference is indirectly linked to a powerful approach of geographical statistics, called Geographically Weighted Regression (GWR). It was introduced relatively recently by Bronson in order to integrate neighborhood effect in statistical analysis, i. e. integrate the spatial dimension in data analysis [Brunsdon et al., 1998]. The core feature is to build from data an hypothetical continuous spatial field by non-parametric estimation (kernel mixture) and proceed to statistical analysis such as regressions on sub samplings of this field. Kernel choice can be function of euclidian distance (gaussian, quadratic) but not necessarily (e.g. cultural or income distance) [Lu et al., 2011]. Kernel size plays a central role in final results, and calibration procedures by cross-validation were also developed [Brunsdon et al., 2002]. All current



(a) Curves of available bikes for all day of the week. Week days are superposed and correspond to the curves with two pics. the green and the blue curve are respectively saturday and sunday.



(b) Theoretical predicted curves for two clusters. As expected, we distinguish week days (red curve) from weekend (blue curve), according to the real curves.

Fig. 5: Results of clustering process for classification of days: distinction between weekends and week days.

methods linked to GWR are implemented in a R package [Lu et al., 2013], so it could be a powerful alternative to the empirical method described after (test should be object of further work).

4.3.2 Empirical inference

Our statistical model for the inference of field is a non-parametric estimation with Gaussian kernels (described in [Tsybakov, 2004]). Considering the real departures and arrivals in bike stations (that are easily calculated by discrete differentiation of data), we count each as a contribution to the global field at the current time step, smoothed with Gaussian kernel (that appeared to be enough in practice). At time t ,with a parameter σ fixing kernel sizes (each kernel has the same size, further work could be done to test the influence of multiple sizes, weighted by the maximum

of the kernel distribution for example) and a set of effective arrivals ($d_i(t)$) at the corresponding coordinates ($\mathbf{x}_i(t)$), the spatial field of destinations is estimated as, with K normalization factor,

$$[D(t)](\mathbf{x}) = \frac{1}{K} \sum_i d_i(t) \cdot exp\left(\frac{\|\mathbf{x} - \mathbf{x}_i\|}{2\sigma^2}\right)$$

We do the same for the origin field. Kernel estimations are done with the ergonomic package `kernlab` ([Karatzoglou et al., 2004]). These extrapolated fields are then discretized and used as parametrization for the agent-based models.

5 Agent-based Modeling

A core component of our approach is the use of an agent-based model for dynamics an overall bike-sharing transportation system, i.e. the coupling of social components (users) with the infrastructure (docking stations, information system) and the superstructure (bikes). Indeed, many studies of such systems focused on top-down optimization of the conception and management. It can be for example an optimal design of the implantation of docking stations, knowing estimators of temporal demand functions [Lin et al., 2011,Lin and Yang, 2011]. Many studies have also focused on the redistribution problem (or balancing issue) from an Operational Research point of view, consisting in proposing the optimal trajectory and number of bikes to move for the operator [Chemla et al., 2011,Contardo et al., 2012,Raviv et al., 2013]. These approaches always keep a top-down point of view, i.e. the idea to have all informations and possibilities of intervention on the system, whereas the reality is much closer from a bottom-up paradigm, at least for interventions, as concrete limitations force to partial and local solutions. Some suggestions towards more local approaches were given as in [Rainer-Harbach et al., 2013], but the problem was never totally considered from a bottom-up point of view. A first agent-based modeling of a bike-sharing system was recently done in [Raimbault, 2014]. We use this model and its implementation here, so we recall the model description in the following.

5.1 Description of basic model

The granularity of the model is the scale of the individual biker and of the stations where bikes are parked. A more integrated view such as flows would not be useful to our purpose since we want to study the impact of the behavior of individuals on the overall performance of the system. The global working scheme consists in agents embedded in the street infrastructure, interacting with particular elements, what is inspired from the core structure of the Miro model ([Banos et al., 2011]). Spatial scale is roughly the scale of the district; we don't consider the whole system for calculation power purposes (around 1300 stations on all the system of Paris, whereas an interesting district have around 100 stations), what should not be a problem as soon as in- and outflows allow to reconstruct travels entering and getting out of the area. Tests on larger spatial zones showed that generated travel were quite the same, justifying this choice of scale. Focusing on some particular districts is important since issues with level of service occur only in narrow areas. Time scale of a run is logically one full day because of the cyclic nature of the process ([Vogel et al., 2011]).

Formalisation The street network of the area is an euclidian network $(V \subset \mathbb{R}^2, E \subset V \times V)$ in a closed bounded part of $\mathbb{R}^2$. The time is discretized on a day, so all temporal evolution are defined on $T = [0, 24] \cap \tau\mathbb{N}$ with τ time step (in hours). Docking stations S are particular vertices of the network for which constant capacities $c(s \in S)$ are defined, and that can contain a variable number of bikes $p_b(s) \in \{0, \dots, c\}^T$. We suppose that temporal fields $O(x, y, t)$ and $D(x, y, t)$ are defined, corresponding respectively to probabilities that a given point at a given time becomes the expected departure (resp. the expected arrival) of a new bike trip, knowing that a trip starting (resp. arriving) at that time exists. Boundaries conditions are represented as a set of random variables $(N_I(i, t))$. For each possible entry point $i \in I$ ($I \subset V$ is a given set of boundaries points) and each time, $N_I(i, t)$ gives the number of bikes trips entering the zone at point i and time t. For departures, a random time-serie $N_D(t)$ represents the number of departures in the zone at time t. Note that these random variables and probabilities fields are sufficient to built the complete process of travel initiation at each time step. Parametrization of the model will consist in proposing a consistent way to construct them from real data.

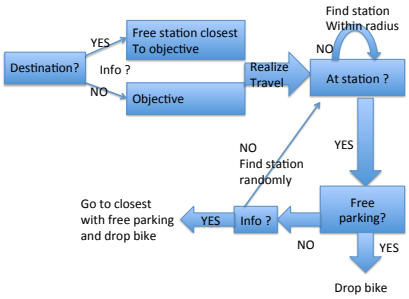


Fig. 6: Flowchart of the decision process of bikers, from the start of their travel to the drop of the bike.

Docking stations are fixed agents, only their functions p_b will vary through time. The other core agents are the bikers, for which the set $B(t)$ is variable. A biker $b \in B(t)$ is represented by its mean speed $\bar{v}(b)$, a distance $r(b)$ corresponding to its "propensity to walk" and a boolean $i(b)$ expressing the capacity of having access to information on the whole system at any time (through a mobile device and the dedicated application for example). The initial set of bikers $B(0)$ is taken empty, as $t = 0$ corresponds to 3a.m. when there is approximately no travels on standard days.

We define then the workflow of the model for one time step. The following scheme is sequentially executed for each $t \in T$, representing the evolution of the system on a day.

For each time step the evolution of the system follows this process :

- Starting new travels. For a travel within the area, if biker has information, he will adapt his destination to the closest station of its destination with free parking places, if not his destination is not changed.
  - For each entry point, draw number of new traveler, associate to each a destination according to D and characteristics (information drawn uniformly from proportion of information, speed according to fixed mean speed, radius also).
  - Draw new departures within the area according to O, associate either destination within (in proportion to a fixed parameter p_{it}, proportion of internal travels) the area, or a boundary point (travel out of the area). If the departure is empty, biker walks to an other station (with bikes if has information, a random one if not) and will start his travel after a time determined by mean walking speed and distance of the station.
  - Make bikers waiting for start for which it is time begin their journey (correspond to walkers for which a departure station was empty at a given time step before)

- Make bikers advance of the distance corresponding to their speed. Travel path is taken as the shortest path between origin and destination, as effective paths are expected to have small deviation from the shortest one in urban bike travels [Borgnat et al., 2009c].
- Finish travels or redirect bikers
  - if the biker was doing an out travel and is on a boundary point, travel is finished (gets out of the area)
  - if has no information, has reached destination and is not on a station, go to a random station within $r(b)$
  - if is on a station with free places, drop the bike
  - if is on a station with no places, choose as new destination either the closest station with free places if he has information, or a random one within $r(b)$ (excluding already visited ones, implying the memory of agents).

Fig. 6 shows the decision process for starting and arriving bikers. Note that walking radius $r(b)$ and information $i(b)$ have implicitly great influence on the output of the model, since dropping station is totally determined (through a random process) by these two parameters when the destination is given.

5.2 Discrete Choices Extension

The extension to this agent-based model is the insertion of a Discrete Choice Model at some decision steps by bikers. The more natural was the choice of waiting at a docking station to get a bike when it is empty, or similarly to wait for a free parking place when station is full, instead of moving to the closest station or taking an other mode. We do not consider the last hypothesis, assuming that users necessarily have already made an immutable mode choice.

Formally, choice are waiting w or moving m. We denote by t_w waiting time, d' distance to go to next station, and $\tilde{d}$ the distance difference to destination.

For bikers with $i(b) = 1$, they know exactly distance to closest station, and choice can be made accordingly, following the utilities

$$U_w(i=1) = \beta_t t_w + \beta_d \tilde{d} + \varepsilon_w$$

$$U_m(i=1) = \beta_t \frac{d'}{\bar{v}} + \beta_d \tilde{d} + \varepsilon_m$$

whereas for a biker with no information, terms in $\tilde{d}$ will vanish as he has no information how far he will go, and traveling distance is different (in that case, user only check the average lost time) :

$$U_w(i=0) = \beta_t t_w + \varepsilon_w$$

$$U_m(i=0) = \beta_t \frac{\bar{d}}{\bar{v}} + \varepsilon_m$$

It could be the object of further work to propose the addition of an "exhaustion"term, expressing the reluctance of the user to bike more. Also, risk aversion is not taken into account for uninformed users.

The integration of a nested logit including the choice of use of information if it is available might be interesting, but not necessarily efficient in our method as it would add too many parameters, weakening the calibration procedure.

The extension was implemented and code adapted in order to be able to estimate the two DC parameters we are interested in, i.e. β_t and β_d.

6 Inference of Unknown Parameters by Model Calibration

6.1 Extended Calibration Procedure

The calibration is done with objective to minimize the error on real data for load factor : if we note $(lf(s,t))_{s\in S, t\in T}$ real time-series, it is defined for the k-th realization of the model as

$$E(k) = \frac{1}{|S|\,|T|}\sum_{t\in T}\sum_{s\in S}\left(\frac{p_b(s,t)}{c(s)} - lf(s,t)\right)^2$$

and we want to minimize the empirical mean on many realization $E =< E(k) >_k$.

Parameter space contains

- $\bar{r}$ mean walking radius of bikers
- p_i probability to have information
- σ kernel size for fields inference
- DC parameters β_t, β_d

Dimension is too huge for a systematic grid exploration. We use first results of section 3 to have bounds on DC parameters and significantly reduce search size. We assume $\beta_t \sim \beta_{t_B}$ and $\beta_d \sim \beta_D$ so we get $\beta_d \in [-0.06, 0]$ and $\beta_t \in [-0.25, -0.15]$. Other parameters can be handily bounded : $\bar{r} \in [0, 1000], \sigma \in [50, 500]$ and $p_i \in [0.3; 0.7]$.

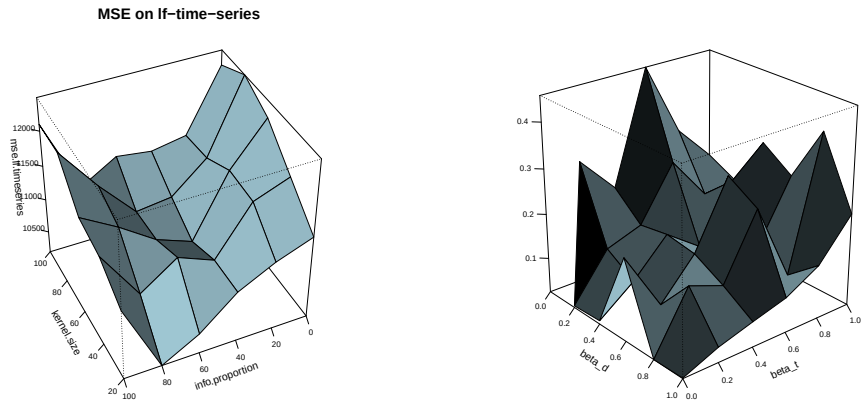
We build then a grid of size 10^5 within this bound and we follow a standard gradient descent for calibration. Fig.**??** shows examples of projected surfaces on subspaces. Among some dimension, behavior is chaotic and leaves less chances for the classic descent to work, so we add a touch of simulated annealing to help convergence towards an hypothetical global minimum. Empirically, we obtain roughly 76% of convergence to a small region, that is with 95% confidence intervals (normal hypothesis) $(\bar{r}, p_i, \sigma, \beta_t, \beta_d) = (238 \pm 51, 0.67 \pm 0.08, 321 \pm 69, -0.05 \pm 0.01, -0.16 \pm 0.02)$. We have this way inferred unknown DC parameters, learning that people are less reluctant to move than expected.

6.2 Towards calibration stability assessment through Large-Deviations functions simulation

A crucial issue in the proposed calibration procedure is the non-precise knowledge of the behavior of the response surface around the area where parameters seems to give the best results. However, the algorithm did not necessarily converge, and it is difficult to extract statistics because of the rather chaotic geometry along some dimensions (mean behavior on a precise point has no meaning if trajectories are sensitive to infinitesimal disturbances). To tackle the question of the accuracy of the calibration, we propose a novel method inspired from recent works in statistical physics that is the determination of large deviation functions.

Considering a stochastic system where "extreme"event (regarding a given observable, that we can isomorphically consider as an activity K) occur in exponentially decreasing probabilities, one can compute its large-deviations function that is linked to the rate of divergence to extreme configurations [Touchette, 2009].

Our system will be the model within its parameter space, a configuration given by parameter values, and dynamics following MSE gradient and simulated annealing.



(a) Response surface along (σ, p_i) dimensions.

(b) More chaotic response surface along (β_d, β_t) (rescaled).

Fig. 7: Example of Mean-square error response surfaces, obtained with $K = 100$ repetitions for each combination of parameters. As some parameters allow a regular, almost convex behavior, ideal for simplex calibration, other give chaotic landscapes.

Markov formalism and cloning algorithm We place ourselves in the continuous-time Markov chains formalism for describing the system. We consider the set of configurations $\{\mathcal{C}\}$ and the corresponding transition rates between states $W(\mathcal{C} \to \mathcal{C}')$. We have the master equation for probabilities in time, with the escape rate of a state $r(\mathcal{C}) = \sum_{\mathcal{C}' \neq \mathcal{C}} W(\mathcal{C} \to \mathcal{C}')$

$$\partial_t P(\mathcal{C}, t) = \sum_{\mathcal{C}' \neq \mathcal{C}} W(\mathcal{C}' \to \mathcal{C}) P(\mathcal{C}', t) - r(\mathcal{C}) P(\mathcal{C}, t)$$

In the case of our system, even gradient of MSE is stochastic as we estimate an empirical gradient on K realizations of the model, therefore escape rate will be empirically calculated at each step, as transition rates. One stays if $||\mathbf{grad} E|| < \varepsilon$ a given threshold (we take $\varepsilon = 1$). One jumps to neighbor states with probabilities linear in component contribution. Finally, one jumps further with a probability $\alpha \cdot Var_R(||\mathbf{grad} E||)$ where the variance is taken on a fixed radius R (we take $R = 100$ and $\alpha = 0.05$) (probabilities are then normalized). Activity is taken as the mean distance to the expected convergence point.

With the calculation of rates, we can implement the cloning algorithm proposed in [Lecomte and Tailleur, 2007,Tailleur et al., 2009] (discrete space, discrete time). The idea exploited in the cloning algorithm is to use population dynamics to obtain biased selection of large deviations that could not have been observed in standard runs. The large deviation function ψ is defined, for s conjugated with the activity K, by $< e^{-sK} > \sim e^{t\psi(s)}$ where the mean is on all histories.

One has then the s-modified master equation for the transformed probabilities $\hat{P}(\mathcal{C}, s, t)$, with only the modified rates:

$$\partial_t P(\mathcal{C}, s, t) = \sum_{\mathcal{C}' \neq \mathcal{C}} W_s(\mathcal{C}' \to \mathcal{C}) P(\mathcal{C}', t) - r(\mathcal{C}) P(\mathcal{C}, t)$$

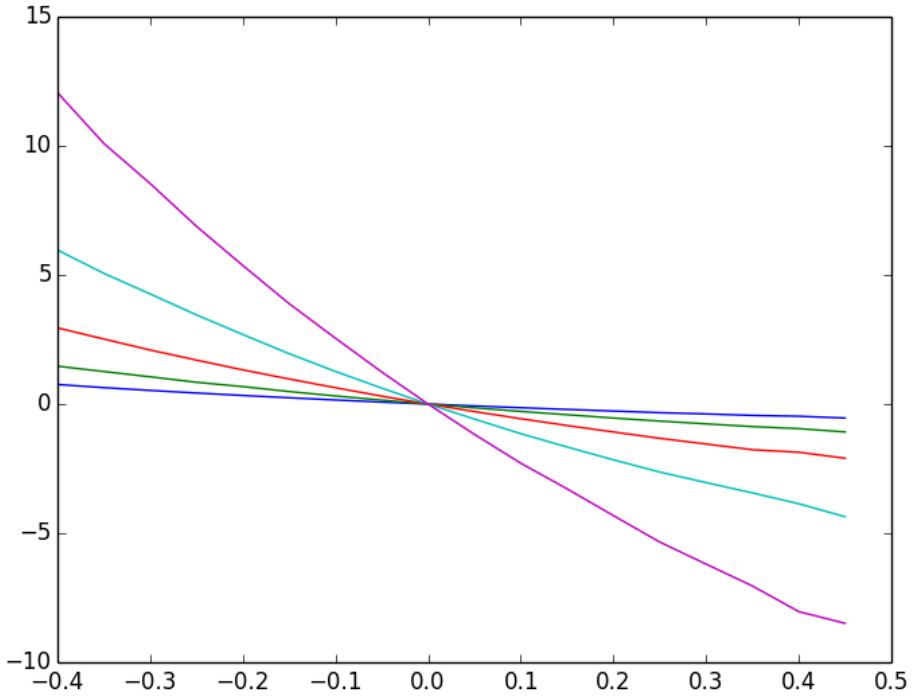
Introducing the s-modified escape rate $r_s(\mathcal{C}) = \sum_{\mathcal{C}' \neq \mathcal{C}} W_s(\mathcal{C} \to \mathcal{C}')$, one can rewrite the above equation as a Markov dynamic with a new escape rate $\delta r_s = r_s - r$, that we will consider as a cloning rate: if copies of the system evolve in parallel following the new Markov dynamic, and are cloned at rate δr_s, one can show that the size of the population follows the first master equation so one can evaluate ψ by evaluating the linear growth coefficient of the logarithm of population size at large time (detailed in [Lecomte and Tailleur, 2007]). One can also calculate the average of activities on histories, which is of interest for our question.

The algorithm was modified from the open python implementation provided by authors of [Tailleur et al., 2009]. The calculation of E and its gradient was done through a system call to headless Netlogo running the ABM. First result of curves for ψ and the activity are shown in figure 8 but results are too fuzzy yet to have a clear conclusion on the stability of the calibration. We have however proposed an interesting method that would deserve further exploration and developments.

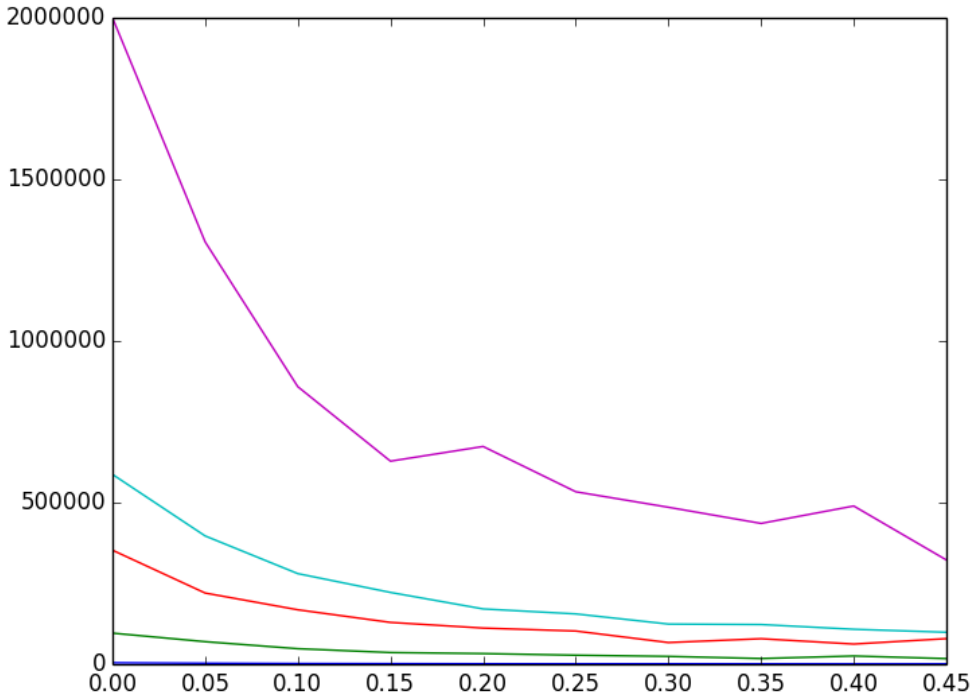
7 Conclusion

Despite the fail on precise survey conception and administration, we turned over the research question towards a less practical but rather interesting methodological and theoretical question : how can we infer missing data or parameters from poor and incomplete data. To answer that, we elaborated a complicated process allowing to reconstruct dynamics of the system by coupling statistical analysis (field inference) with agent-based modeling. This hybrid model can be extended and use to indirectly determine discrete choice parameters by model calibration, allowed by a first rough DC modeling used to obtain a bound on parameters. Finally, we propose a novel method to study the stability of the calibration, using ideas from statistical physics especially numerical calculation of large-deviation functions through a cloning algorithm, opening research paths although it was not directly conclusive.

Further work should first focus on a better junction between all parts of the project, and robustness reinforcement for each single part. Other point mentioned in text can of course be taken as independent developments.



(a) $\psi(s)$ for $K = 20...100$



(b) Mean activity for $K = 20...100$

Fig. 8: Large-deviation function and mean deviation from minimum, drawn for different values of K number of repetitions. As expected, growing K diminishes the number of deviating events, i.e. stabilize the surfaces. We can conclude than over $K = 60$ we have a quite satisfying convergence as ψ vanishes quickly.

References

Banos, A. (Décembre 2013). Pour des pratiques de modélisation et de simulation libérées en géographie et shs. *Thèse d'Habilitation à Diriger des Recherches, UMR CNRS 8504 Géographie-Cités, ISCPIF*.

Banos, A., Boffet-Mas, A., Chardonnel, S., Lang, C., Marilleau, N., Thévenin, T., et al. (2011). Simuler la mobilité urbaine quotidienne: le projet miro. *Mobilités urbaines et risques des transports*.

Ben-Akiva, M. and Bierlaire, M. (1999). Discrete choice methods and their applications to short term travel decisions. In *Handbook of transportation science*, pages 5–33. Springer.

Bierlaire, M. (2006). Biogeme: a free package for the estimation of discrete choice models. In *Swiss Transport Research Conference*, number TRANSP-OR-CONF-2006-048.

Borgnat, P., Abry, P., and Flandrin, P. (2009a). Modélisation statistique cyclique des locations de vélo'v à lyon. In *XXIIe colloque GRETSI (traitement du signal et des images), Dijon (FRA), 8-11 septembre 2009*. GRETSI, Groupe d'Etudes du Traitement du Signal et des Images.

Borgnat, P., Abry, P., Flandrin, P., Rouquier, J.-B., et al. (2009b). Studying lyon's vélo'v: a statistical cyclic model. In *European Conference on Complex Systems 2009*.

Borgnat, P., Fleury, E., Robardet, C., Scherrer, A., et al. (2009c). Spatial analysis of dynamic movements of vélo'v, lyon's shared bicycle program. In *European Conference on Complex Systems 2009*.

Bourcet, M., Lesturgie, J., Allain, T., Etienne, C., Sebes, A., and Raimbault, J. (June 2014). Le vélib comme choix de mode. Technical report, ENPC, D'epartement VET.

Boutron, I., Dutton, S., Ravaud, P., and Altman, D. G. (2010). Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA: the journal of the American Medical Association*, 303(20):2058–2064.

Brunsdon, C., Fotheringham, A., and Charlton, M. (2002). Geographically weighted summary statistics—a framework for localised exploratory data analysis. *Computers, Environment and Urban Systems*, 26(6):501–524.

Brunsdon, C., Fotheringham, S., and Charlton, M. (1998). Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(3):431–443.

Chavalarias, D. and Cointet, J.-P. (2013). Phylomemetic patterns in science evolution—the rise and fall of scientific fields. *PloS one*, 8(2):e54847.

Chemla, D., Meunier, F., and Calvo, R. W. (2011). Balancing a bike-sharing system with multiple vehicles. In *In proceedings of Congres annuel de la société Française de recherche opérationelle et d'aidea la décision, ROADEF2011, Saint-Etienne, France*.

Contardo, C., Morency, C., and Rousseau, L.-M. (2012). *Balancing a dynamic public bike-sharing system*, volume 4. CIRRELT.

Couture-Beil, A. (2013). rjson: Json for r. *R package version 0.2*, 13.

Crabtree, S. A. and Kohler, T. A. (2012). Modelling across millennia: Interdisciplinary paths to ancient socio-ecological systems. *Ecological Modelling*, 241:2–4.

Crowe, B. J., Lipkovich, I. A., and Wang, O. (2010). Comparison of several imputation methods for missing baseline data in propensity scores analysis of binary outcome. *Pharmaceutical statistics*, 9(4):269–279.

Fishman, E., Washington, S., and Haworth, N. (2013). Bike share: A synthesis of the literature. *Transport Reviews*, 33(2):148–165.

Geng, X., TIAN, K., ZHANG, Y., and LI, Q. (2009). Bike rental station planning and design in paris [j]. *Urban Transport of China*, 4:008.

Gifford, J. and Campus, A. (2004). Will smart bikes succeed as public transportation in the united states? *Center for Urban Transportation Research*, 7(2):1.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab-an s4 package for kernel methods in r.

Keitt, T. H., Bivand, R., Pebesma, E., and Rowlingson, B. (2011). rgdal: bindings for the geospatial data abstraction library. *R package version 0.7-1, URL http://CRAN. R-project. org/package= rgdal*.

Lecomte, V. and Tailleur, J. (2007). A numerical approach to large deviations in continuous time. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(03):P03004.

Leurent, F. (2006). *Modélisation du trafic, des déplacements sur un réseau et de l'accessibilité aux activités grâce au transport*. PhD thesis, Université Paris Dauphine-Paris IX.

Lin, J.-R. and Yang, T.-H. (2011). Strategic design of public bicycle sharing systems with service level constraints. *Transportation research part E: logistics and transportation review*, 47(2):284–294.

Lin, J.-R., Yang, T.-H., and Chang, Y.-C. (2011). A hub location inventory model for bicycle sharing system design: Formulation and solution. *Computers & Industrial Engineering*.

Liu, Z., Jia, X., and Cheng, W. (2012). Solving the last mile problem: Ensure the success of public bicycle system in beijing. *Procedia-Social and Behavioral Sciences*, 43:73–78.

Lu, B., Charlton, M., and Fotheringhama, A. S. (2011). Geographically weighted regression using a non-euclidean distance metric with a study on london house price data. *Procedia Environmental Sciences*, 7:92–97.

Lu, B., Harris, P., Gollini, I., Charlton, M., Brunsdon, C., and Lu, M. B. (2013). Package 'gwmodel'.

Midgley, P. (2009). The role of smart bike-sharing systems in urban mobility. *JOURNEYS*, 2:23–31.

Mitra, R. and Reiter, J. P. (2010). A comparison of two methods of estimating propensity scores after multiple imputation.

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

Raimbault, J. (2014). User-based solutions for increasing level of service in bike-sharing user-based solutions for increasing level of service in bike-sharing transportation systems. In *Proceedings of the Conference on Complex Systems Design and Management. CSDM, Paris 12-14 nov. 2014*.

Rainer-Harbach, M., Papazek, P., Hu, B., and Raidl, G. R. (2013). Balancing bicycle sharing systems: A variable neighborhood search approach. In *Evolutionary Computation in Combinatorial Optimization*, pages 121–132. Springer.

Ram, K. (2013). Git can facilitate greater reproducibility and increased transparency in science. *Source code for biology and medicine*, 8(1):7.

Raviv, T., Tzur, M., and Forma, I. A. (2013). Static repositioning in a bike-sharing system: models and solution approaches. *EURO Journal on Transportation and Logistics*, 2(3):187–229.

Rubin, D. B. (2009). *Multiple imputation for nonresponse in surveys*, volume 307. Wiley. com.

Tailleur, J., Lecomte, V., Marro, J., Garrido, P. L., and Hurtado, P. I. (2009). Simulation of large deviation functions using population dynamics. In *Aip Conference Proceedings*, volume 1091, page 212.

Touchette, H. (2009). The large deviation approach to statistical mechanics. *Physics Reports*, 478(1):1–69.

Tsybakov, A. B. (2004). Introduction to nonparametric estimation. (introduction à l'estimation non-paramétrique.).

Vogel, P., Greiser, T., and Mattfeld, D. C. (2011). Understanding bike-sharing systems using data mining: Exploring activity patterns. *Procedia-Social and Behavioral Sciences*, 20:514–523.

Warren Liao, T. (2005). Clustering of time series data—a survey. *Pattern Recognition*, 38(11):1857–1874.