Discrete Choice Models for Bike-Sharing Transportation Systems

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

J. Raimbault[1,2]

[1]Graduate School, Ecole Polytechnique
[2]LVMT, Ecole Nationale des Ponts et Chaussées

PIL Presentation - Dpt VET, ENPC
under the direction of Z. Christoforou, LVMT, ENPC
November 6, 2014

Outline

Research Question

User surveys in discrete choice are very expensive, and one often has bad quality data. However possible to cross different data source and methods to improve results robustness, as recent work show [Crabtree and Kohler, 2012].

Research Question : *To what extent can we improve the estimation of discrete choice parameters by using user questionnaire data with system dynamics raw data, and coupling statistical analysis and Agent-Based Modeling ?*

Why study bike-sharing systems ?

Quick development across the world since 2000, starting from Europe ([DeMaio, 2009]).

Around 200 systems in the world. Ecological and compatible ("sustainable") transport mode ([O'Brien et al., 2013]).

Extensions to unexpected places ? USA ([Gifford and Campus, 2004]) where car is dominant, or China ([Liu et al., 2012]) where relation to bikes has strongly changed these last years.

Already well studied : statistical models ([Borgnat et al., 2009b, Borgnat et al., 2009a],[Michau et al., 2011]) or data-mining analysis ([O'Brien et al., 2013],[Vogel et al., 2011, Kaltenbrunner et al., 2010])

But Intrinsically non-performant systems...

Figure: Full or empty docking stations in Paris: decrease in the level of service (source www.velib.paris.fr)

Discrete Choices

- Discrete Choice Modeling : theoretical and practical framework to formalize user choice (used in transportation, marketing, politics) [Ben-Akiva and Bierlaire, 1999], in fact supervised learning with particular loss function)

- Ergonomic tools to estimate models [Bierlaire, 2006]

- Bike-sharing studied from this point of view only for modal choice ; should be a good tool to improve knowledge on system and better design or manage it.

Project Description

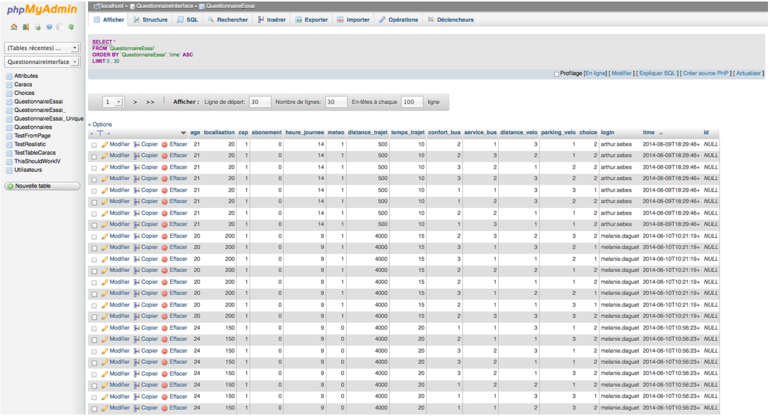
Sequence of Problematics :

1. Conceive and realize a precise survey to estimate discrete choice models.
2. Problem with Questionnaire administration : how to use this poor quality data ?
3. On the other hand, data available on raw dynamics but also incomplete.
4. Proposition of indirect inference of DC Parameters by coupling approaches. Core results more methodological than practical.

Discrete Choice Questionnaire

- Generic web-application for questionnaire administration
- Php server-side application, with standard SQL database
- Direct Biogeme export (specification file with format [BIOGEME_VARIABLE_NAME ; BASE_VARIABLE_NAME ; BASE_VALUE])
- Demo at http://37.187.242.99/Questionnaire

Generic Database



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

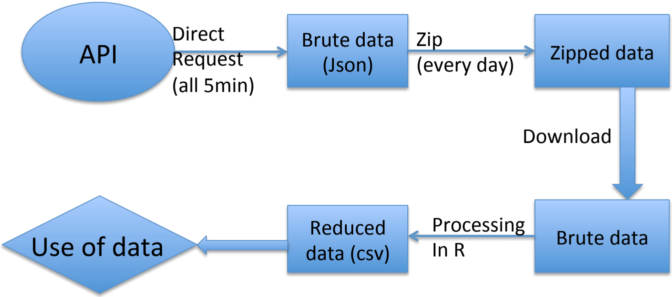
Raw Data : why Open Data ?

- Public data provided by the operator in real time. Problem: need a constantly running collection data process, and only docking station status (incomplete data).

- Why not ask full travel data to operator ? Independent and open research ([Banos, 2013]), reporting bias (in [Nair et al., 2013] results are not presented complete because company did not want for commercial reasons). We do a compromise, and see if we can however have good results.

- Also risk of unconscious spin in the description of results [Boutron et al., 2010].

Raw Data collection Process



Precise Discrete Choice Questionnaire

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

Simplified Discrete Choice Model

- Stated preference model, Conception and estimation in [Bourcet et al., 2014]
- Variables :
  - age
  - professional categorization
  - regular user
  - average distance to public transport
- Choice of mode between bus and bike-sharing, with attributes : 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 (3 discretization levels).
- Utilities : $U_{bus} = \sum \beta_{X_{bus}} X_{bus} + \varepsilon_1$ and $U_{bike} = \sum \beta_{X_{bike}} X_{bike} + \varepsilon_2$.
- Results : $\beta_D \in [-0.06, 0]$ and $\beta_{t_B} \in [-0.25, -0.15]$

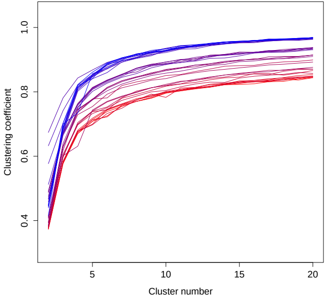
A case study on missing data

Many methods to fill incomplete data [Rubin, 2009]. Case study of comparison between two proposed in [Crowe et al., 2010] and [Mitra and Reiter, 2010].

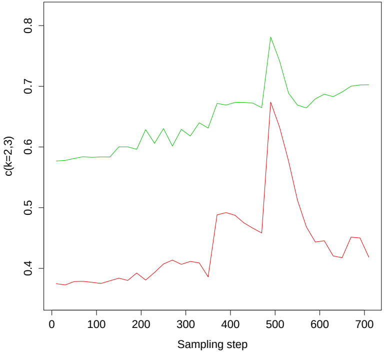
Main conclusions :

- Deleting rows with missing variables leads to less bias but more variance
- Use heuristic to know if complete before or after computing outputs (parameters of generalized estimator).

Data-mining for dimensionality reduction



(a) Clustering coefficient as a function of cluster number for different values of sampling step.



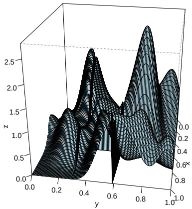
(b) Plot of the value of the clustering coefficient for k=2 (red) and k=3 (green), as a function of sampling step.

Inference of OD Fields

Core of the parametrization : estimation of O/D fields with gaussian kernels non-parametric estimation ([Tsybakov, 2004]) with package kernlab ([Karatzoglou et al., 2004]). With $(d_i(t))$ real arrivals at $(\vec{x}_i(t))$, $D(t)$ spatial field is given by

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

Similar to Geographically Weighted Regression Methods [Brunsdon et al., 1998],[Brunsdon et al., 2002]



Settings and agents

ABM proposed in [Raimbault, 2014].

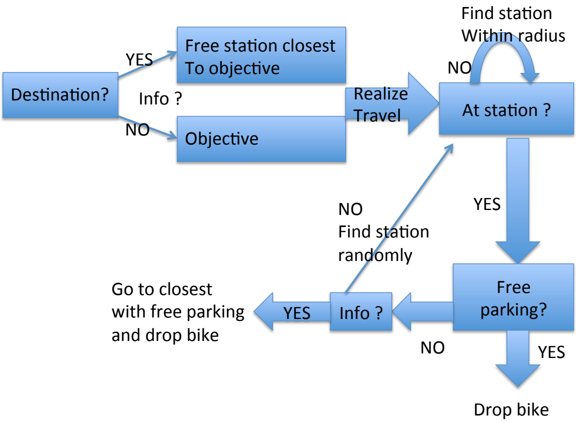
- Agents: bikers with information $i(b)$ (boolean), tolerated walking radius $r(b)$ and mean speed $\bar{v}(b)$; docking stations located in space with current standing bikes $p_b(s,t)$ and capacity $c(s)$
- Euclidian network $N=(V,E)$, representing the road network. Stations are nodes of the network and movement of bikers is embedded in the trace of N in $\mathbb{R}^2$
- Scale of the district; we suppose known temporal fields of origin $O(t)$ and destination $D(t)$ (probabilities of O/D given a trip), boundaries conditions $N(t)$ as flows (in- and outflows) at fixed boundaries points

Temporal Evolution

At each time step:

- Start new travels randomly using O, D, N
- Make bikers in travel advance of the corresponding distance
- Finish travels and redirect bikers when needed (see flowchart of bikers behavior)

Bikers behavior



Discrete Choice in Bikers behavior

Utilities functions when needs to drop a bike at a full station or take one at an empty one, with t_w average waiting time and $\tilde{d}$ distance to closest station

With information :

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

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

Without information :

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

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

Calibration Procedure

Calibration of model on mean MSE on load factors time-series, $E =< E(k) >_k$ with

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

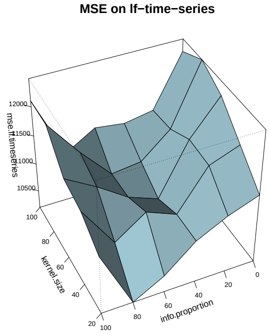
Parameters :

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

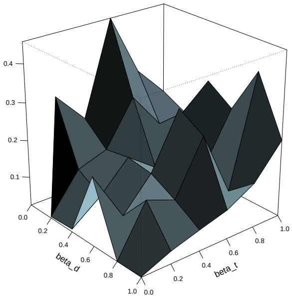
Parameter Space : Hypercube $\beta_d \in [-0.06, 0]$, $\beta_t \in [-0.25, -0.15]$, $\bar{r} \in [0, 1000], \sigma \in [50, 500]$ and $p_i \in [0.3; 0.7]$.

Calibration : Results

On convergent runs (76%) : $(\bar{r}, p_i, \sigma, \beta_t, \beta_d) = (238 \pm 51, 0.67 \pm 0.08, 321 \pm 69, -0.05 \pm 0.01, -0.16 \pm 0.02)$.



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



(b) Along (β_d, β_t).

Large Deviations of Gradient Algorithm

Markov Formalism : Master equation for system State

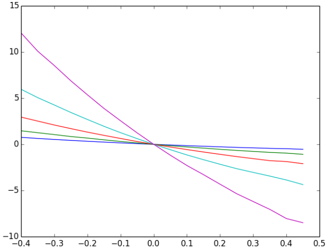
$$\partial_t P(\mathscr{C},t) = \sum_{\mathscr{C}' \neq \mathscr{C}} W(\mathscr{C}' \to \mathscr{C}) P(\mathscr{C}',t) - r(\mathscr{C}) P(\mathscr{C},t)$$

Large Deviation function with s conjugated with activity K:
$< e^{-sK} > \sim e^{t\psi(s)}$ s-modified dynamic :

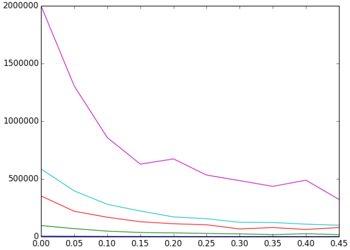
$$\partial_t P(\mathscr{C},s,t) = \sum_{\mathscr{C}' \neq \mathscr{C}} W_s(\mathscr{C}' \to \mathscr{C}) P(\mathscr{C}',t) - r(\mathscr{C}) P(\mathscr{C},t)$$

that drives the cloning algorithm, equivalent to a Markov dynamic with escape rate $r_s(\mathscr{C}) = \sum_{\mathscr{C}' \neq \mathscr{C}} W_s(\mathscr{C} \to \mathscr{C}') - r(\mathscr{C})$.

Large Deviations of Gradient Algorithm



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



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

Limitations of the approach

- Lack of external validation ; more methodological proposition than consistent results
- Limited DC Model and still no exploration of Parameter Space (became too huge).
- Many assumptions that would need to be relaxed ; however good thematic model.

Possible Developments

- More precise sensitivity Analysis to DC parameters
- Obtain good data and compare results (external validation)
- Internally valid DC extension and calibration procedure
- Explore strategy on user choice behavior
- Role of docking stations ?

Conclusion

- Broad approach, many point of views combined.
- Methodologically interesting, to be compared with existing work in quantitative social science (archeology, geography)
- Novel approach proposed (ex Large Dev for calibration algorithm)
- Promising as the basis of a further work.

References I

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.

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.

References II

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*.

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.

References III

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.

References IV

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.

DeMaio, P. (2009).
Bike-sharing: History, impacts, models of provision, and future.
Journal of Public Transportation, 12(4):41–56.

References V

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.

Kaltenbrunner, A., Meza, R., Grivolla, J., Codina, J., and Banchs, R. (2010).
Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system.
Pervasive and Mobile Computing, 6(4):455–466.

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

References VI

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.

Michau, G., Robardet, C., Merchez, L., Jensen, P., Abry, P., Flandrin, P., and Borgnat, P. (2011).
Peut-on attraper les utilisateurs de vélo'v au lasso.
In *Proceedings of the 23e Colloque sur le Traitement du Signal et des Images. GRETSI-201*, pages 46–50.

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

References VII

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.

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*.

References VIII

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

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.