MASTER'S DEGREE THESIS

MSc in Statistics and Operations Research

Title: Methods to impute links in an indirect sample. "Sans domicile" survey Sd2001.

Author: Carme Caum Julio

Advisor: Mónica Becue Bertaut

Department: EIO

University: UPC

Academic year: 2011-2012

Facultat de Matemàtiques i Estadística

UNIVERSITAT POLITÈCNICA DE CATALUNYA

À tout l'équipe de sans-domicile, pour sa motivation sur le projet,
À Maryse Marpsat, pour son encadrement et tout ce qui m'a appris,
À Lionel Viglino, pour tout son temps extra que m'a consacré,
À Monique Bécue pour toujours m'encourager à faire ce séjour,
À Bernadette, pour sa bonne humeur au bureau.

À ma famille: gràcies per respondre sempre al telèfon!

METHODS TO IMPUTE LINKS IN AN INDIRECT SAMPLE. "SANS DOMICILE" SURVEY SD2001.

CONTENTS

CONTEXT

From February to August 2011 I was contracted to work in the *Institut National de la statistique et des études économiques* (Insee) and the *Institut National d'Études Démographiques* (Ined), in Paris. My task was focused on the preparation of the next edition of the ***Sans-domicile (SD)* survey** (February 2012) and I worked within the *Unité des méthodes statistiques* (UMS) and an experts team from the departament of *Prix à la consommation, des ressources et des conditions de vie des ménages* and consisting of statisticians, sociologists and study engineers[1].

I deepened the knowledge of *sans-domicile* survey in the first months; a great deal of documentation from 2001's edition and related papers were available. I also had the chance to closely follow the preparation of the 2012's edition attending the regular experts meetings where important and difficult decisions were taken. All these experiences made me aware about the complexity of preparing a survey of such caracteristics and motivated me to enlarge my task focusing on a particular part of the questionnary called "*semainier"* which is essential for weighting the questionnaries and make inferences. To do so, I used the collected data from 2001's first edition (*Sd2001)*.

Combining all this work has led to this master's degree thesis.

[1] Experts team of S*ans-domicile* survey: Maryse Marpsat, Françoise Yaounancq, Michel Durée, Sylvain Quenum, Lionel Viglino, Daniel Verger, Pascal Ardilly, Bernadette Rocca and Cécile Brousse (Insee). Martine Quaglia and Stéphane Legleye (Ined).

THE OBJECTIVES OF THE MASTER'S DEGREE THESIS

The objectives of this master's degree thesis focus on how to impute missing data in a real survey. Particularly, we work on imputing a set of variables that determine the weigths assigned to each observation unit in an indirect sampling, such as the applied in *Sans-domicile* survey. *Semainier* is the questionnary part that refers to the subset of these variables.

The objectives are divided in the following points:

- To theoretically **justify** the importance of collecting *Semainier's* data in a survey based on an indirect sampling.

- To **analyze** and **validate** the imputation method for the *Semainier* applied in the first edition of the *Sans-domicile* survey (2001).

- To **modify** some of the parameters of the reference method and **contrast** the resulting method with the reference of 2001.

- **Estimate** a Bayesian model to explain and predict the marginal information of the *Semainier.* New improvements of the reference method might be introduced thanks to the model interpretation.

- To **develop** an action plan for imputing *Semainiers* of the new edition survey's edition in 2012.

INTRODUCTION

2.1 CONCEPTION OF *SANS-DOMICILE*'S SURVEY

In this section, we focus on contextualizing our work, deepeing the features of the *sans-domicile* survey (SD). First, survey's background is introduced as well as the basic nomenclature used along this master's degree thesis. Then, we explain the way of reaching *sans-domicile* population through the *Time Location* sampling method. Finally, related to the latter, the *Weight Share method* is introduced.

2.1.1 Sans-domicile survey's background

Sans-domicile[2] is the name of the first European national survey on users of accommodation services and hot meals, whose first edition was carried out in 2001 by the ***Institut National de la statistique et des études économiques (Insee)***.

Previously, two main different surveys had been performed in the metropolitan region of Washington by the *Research Triangle Institut* in 1991 and by the *Bureau of the Census* in 1996[3]. In France, a pilot work was performed by the ***Institut national d'études démographiques (Ined)*** on the Parisian *sans-domicile* people in 1995, under the supervision of the *Conseil national de l'information statistique* (*Cnis*). Nowadays, after the experience of 2001, an esperts team is working on the preparation of the new *sans-domicile* edition on January 2012.

The main *sans-domicile* survey aims are:

First, to count the number of people considered *sans-domicile* according to the definition of this survey.

Second, to describe the different life conditions of *sans-domicile* people as well as their difficulties for accommodation access. Moreover, to better know the entrance and maintenance process in the social exclusion state that they face.

[2] We will keep the French term «sans-domicile» as it is the real name of the survey. It can not be directly translated to the term «homeless».

[3] For further information see: Marpsat, Maryse: « L'enquête de l'Insee sur les *sans-domicile*: quelques éléments historiques». Courrier des statistiques 123, January-April 2008.

2.1.2 Nomenclature of Sans-domicile survey

Some essential concepts will be defined in this section to make the reader more familiar with the terms used in this field and, particularly, in this survey. Notice that we will try to be as faithful as possible to the original French nouns used in the survey[4].

A person is called **«*sans-domicile*»** (always in the survey context) if he/she has spent the night previous to the interview in:

- **A place not designed for habitation:** makeshift shelters (a tent, a parking, an attic) or public centers (the underground, the railway station, under a bridge, a public park, etc).
  They are also called: ***non-roof* people** (NR), considered the subpopulation with the most precarious situation.

- A freely accommodation cared by an **organism**. Also in an exceptionally opened center during the *Très Grand Froid* plan (gym, municipal centers, etc), in a hospital, in a prison, in some relative's house or in a squat.
  They are also called: ***no personal accommodation*** people (NPA).

Some other related terms have to be introduced briefly here:

- **Place:** physical place where the interviewed person has spent the night or has eaten. It can be a center or not.
- **Center:** physical place where a set of benefits are offered. It is cared by an organism. There are different typologies of centers depending on the type of individual service offered.
- **Users:** people that benefit from an individual service offered by a center, at least once during a month.
- **Benefit:** individual service offered. Two types are considered: accommodation benefit (a bed in a hotel room or in a grouped room) or meal benefit (a lunch or a dinner plate). A bed in a relative's house, a center under a bridge, etc **is not** considered to be a benefit.
- **Visit:** the intersection of a center and a day. It represents the appointment of the interviewers that has been decided with the head of the organism.

[4] From now on we will not use the term *homeless* any more because it is too general in the context of this thesis.

- **Target population**: the population that we want to study with the survey: *Sans-domicile* population.
- **Reached population**: the population that we actually interview: the users. It defines the survey's population couverture. Ideally, reached population should match with the target population.

- **Sampling unit:** the unit that is sampled according to the sampling plan. We will see why benefit will be the sampling unit.
- **Observation unit:** the unit that we are interested in and from which we want to make our inferences. In this case, the user is the observation unit.
- **Interviewed person:** the user asked for answering the questionnaire. Note that she/he is necessarily a user at the moment of the interview but she/he can be not always a frequent user.
- ***Semainier*:** a section of the questionnary where we count the number of benefits used in the 7 days previous to the interview is reported[5]. For example, concerning accommodation services, the interviewed person has to answer to the question: *Where have you spent the past seven nights?* Then, it is written down day by day by the interviewer.
- **Link**: each time that an interviewed person says that she/he has taken a benefit.
- **Weekly attendance**: the number of links counted for a user during the 7 days previous to the interview.
- **Non-responder**: a user for whom we do not know her/his weekly attendance because her/his *semainier* presents some missings (she/he does not remember or the data has been badly introduced).

2.1.3 How to reach the «sans-domicile»? The Time Location sampling method (TLS)

Remember that, **at applying a sampling plan, we aim to obtain a sample of *sans-domicile* that allows for inferring the results to the target population with the least baised estimations.**

[5] See the original *Semainier* in the Appendixes: *9.1 The Semainier's extract for the face to face questionnaire*

To start with, we should be aware that our **target population –the *sans-domicile* people-** is considered to be a **hard-to-reach population.** Several types of difficulty have led researchers to classify a population as "hard-to-reach":

- The target population is relatively small, which makes an investigation throughout the general population very expensive (e.g tourists, people with a very high income).

- Its members are hard to identify, partly because some of them might not wish to disclose that they are belonging to it for many reasons: because their behavior is illicit, because it is socially stigmatized, because they have no desire to revisit a painful past, etc (e.g *sans-domicile,* injecting drug users, people having unprotected sex with multiple partners, prostitutes, people who have been in foster care).

- No sampling frame is available or only an incomplete one, leading to biased estimations.

- The population's behavior is badly known, which leads to a poor choice where to reach them.

Due to this "hard-to-reach" characteristic, in the following we will expose and discuss the sampling method used in this survey. Other methods are available[6].

When studying a hard-to-reach population without very few expert knowledges (extremely bad ones for the *non-roof people*, the hiddest population), **we are not capable to fix adequate quotas in an empirical method**. In addition, with this approach, we would **not be able to associate a weight to each individual** (classically being the inverse of the probability of being sampled). Concerning the probabilistic methods, although all of them have their own limitations due to the complexity that the problem implies, we will apply the *time-location indirect sampling* that for *sans-domicile* survey.

The **Time-location sampling** (TLS, Marpsat and Firdion, 2000; Ardilly and Le Blanc, 2001; Brousse *et al.*, 2001; Brousse *et al. ,* 2006, Hauddret-Roustide *et al.,* 2008; 2009, Pollack *et al.,* 2005;

[6] See the Appendixes: *9.2 Sampling methods to study a hard-to-reach population*.

MacKellar *et al. ,* 2007) belongs to the family of **indirect sampling** methods (Lavallée, 1995; 2002; 2007; Lavallée and Rivest, 2009). This familly of sampling method follows the principle that the target people are present in a network. For instance: if we we are interested in drug users, we will probably have a network of drugs users who know each other or, in the *sans-domicile* case, a centers network.

Time-location sampling is based on selecting a sample from a population that is not the target one but which is **linked** to the target we are interested in. In our context, a set of **benefits** is the **sampled population** and **users of helping services, the reached population** (close to the target population). This principle is illustrated in figure 2. The interviewed users will be weighted correcting differences between individuals in terms of the frequency of their attendance of the services.
Its main weak points are that establishing and updating the list of benefits is often time and cost-consumming, being the collected information not always reliable. If a high proportion of the target population does not ask for benefits or only very rarely, this leads to a coverage bias. Finally, there might be data collection problems associated with the centers: refusal by managers or rapid departure of users, etc.

The Time-location sampling has been applied in some other studies with common features:

- Since 1996, in a study based on the mental health and access to health facilities of *non-roof people* from Paris (Kovess, Marngin-Lazarus, 1996).

- In different surveys in France (Coquelicot by the *Institut national de veille sanitaire,* 2002-2004) focused on hard-to-reach population like drugs users.

2.1.4 The Weight Share Method (WSM)

In a time-location sampling context, the **observation unit** (that is, user) has two particularities. First, he/she **differs from the sampling unit** (that is, benefit). Second, he/she can be interviewed through more than one sampling unit (otherwise said, he/she **is related to more than one row in the sampling frame).**

As examples, not related with *sans-domicile* survey, we can cite:

- We accept to interview a physical person, owner of a sampled housing, without taking into account its category (main or second). People owning more than one house will have more probability to be included in the sample than others with just one house. Thus, we will have to know, **for each interviewed person, how many houses this person owns (number of links**, in the survey's nomenclature).

- We are interested in studying the number of serious car accidents (with hospitalisation but without deaths) and we have a list of hospitalised people for this cause. The accidents with less implied people will have a lower chance to be studied. Hence, we have to ask in the questionaire **how many people where hospitalised due to this concret accident (number of links).**

In the *sans-domicile* context, we have to take into account that a user might ask for several benefits so that **unequal levels of attendance between one user and another lead to different probabilities of inclusion in the sample**. To integrate this information, Lavallée proposed in 1995 the Weight Share Method (*WSM*) which provides an unbiased estimator. Lavallée has also shown that this unbiased estimator keeps this property independently of the sampling method used to collect the data. Therefore, this method is an **essential element of the survey as we need to give to each user his proper weight.**

First, we consider *U* a population of benefits and *V* a population of users, where each one has at least one link with a benefit (r_{ik}=1 if the user *k* has asked for the benefit *i* and r_{ik}=0 otherwise). In our context, each benefit is linked to only one user from the *V* population.

Note that in figure 1 below, **there are some *sans-domicile* people for whom we do not have any link** because they are not users and others that do not belong to the target population but they are also interviewed (see *2.3.1 A limited target population's coverture*).

Here is the **caption text**[7] for figure 1:

: sampled benefit
: unsampled benefit
: interviewed person
: non-interviewed person
: link between a user and a sampled benefit
: link between a user and an unsampled benefit
: ***non-roof*** **interviewed person**

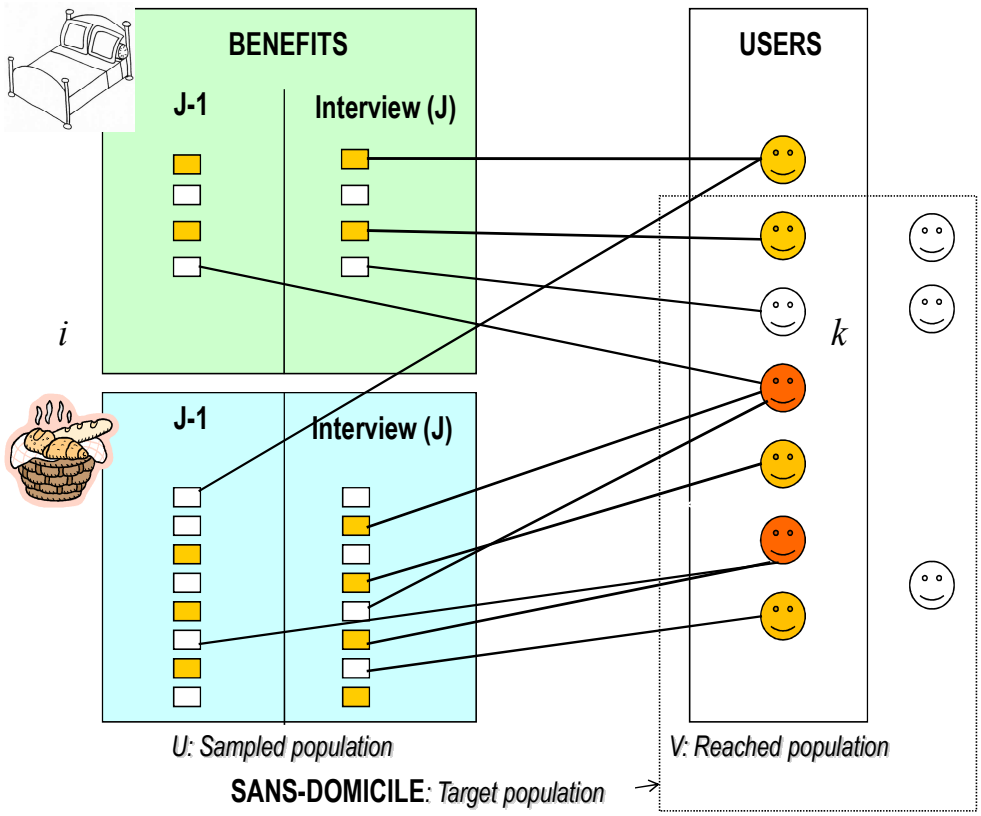

Figure 1. Outline of the **time-location indirect sampling** application in *SD*'s survey.

[7] Some remarks about this outline: In order to avoid overfilling the outline, we have omitted the people interviewed the previous day of the interviewer. Moreover, here the *semainier* appears reduced to just two days: the interview day and the previous day. In the real *semainier,* the links for the same day of the interview (day *j*) are not counted, as it is hard for the user to know were he/she is going to sleep that evening or where he/she is going to eat. That is why **the weekly attendance is based on the day *j-1* until the day *j-7***. To end with, note that in the outline we suppose that there are no users interviewed twice.

We are interested in the **total of a target variable** noted by *N*:

$$N = \sum_{k \in V} y_k$$

For instance, the total is the number of ***non-roof people*** (N_{NR}, where y_k=1 if the user is considered "non-roof" and 0 otherwise).

We define the number of benefits from *U* used by each *k* user as follows:

$$r_k = \sum_{i \in U} r_{i,k} \quad (1)$$

Using the next expression:

$$z_i = \sum_{k \in V} \frac{r_{i,k}}{r_k} y_k$$

We deduce the following equality:

$$N_{NR} = \sum_{k \in V} y_k = \sum_{i \in U} z_i$$

We suppose that we have a benefits sample noted by S_U from the population *U*. Implicitely, $S_V = \{k \in V_{NR} \mid \exists i \in S_U : r_{i,k} = 1\}$ and we have a set of weights for each benefit: $(w_i)_{i \in S_U}$. **All links between interviewed people and the *U* univers are supposed to be known.**

Finally:

$\hat{N}_{NR} = \sum_{k \in S_V} \tilde{w}_k y_k$, where $\tilde{w}_k = \frac{1}{r_k} \sum_{i \in S_U} w_i r_{i,k}$ is the weight of an interviewed user[8] (2)

A serious and systematic collection of the data concerning the levels of attendance of the interviewed users is required, being a crucial criterion for the quality of this survey. Without this data, it is not possible to weight the survey.

[8] The most suitable situation would be to have just one sampled benefit per user because, otherwise, we would have people interviewed twice. Nevertheless, concerning the user weights, nothing would change taking for granted that a person interviewed in two different centers and moments would give us the same answers to our questions (an unrealistic hypothesis in this context).

2.2 DESIGN OF *SANS-DOMICILE* SURVEY

In this third section, we detail the sampling plan, divided in three stages. Then, we discuss about the survey's reached population and, according to the survey's objectives, which is the survey's couverture.

2.2.1 Sampling plan

Notice that this survey, contrary to other surveys carried out by the Insee (for example, housing survey) has **a high weight dispersion which deteriorates the quality of the survey (a group of users have a strong impact in the estimations).** This fact generally[9] leads to estimators with high variance. Passing from benefits weights to user's weights through the *Weight Share Method* can have a strong impact in the weights dispersion. The larger the reference period is, the more the user's weights disperse. For weekly weights (*Semainier*), the impact of the weight shared can be at most 21 (a user can take at most 3 benefits per day: lunch, dinner, bed).
Thus, **the sampling plan has been designed with the aim that <u>the selection probabilites of benefits are approximately equal</u> in order to control the weight dispersion.**

A three-stage complex sampling plan is applied with the following stages:

1) **Primary Unit: Agglomerations sampling of more than 20.000 inhabitants**
   To create the sampling frame it is necessary to take a census of the centers. 80 agglomerations (of more than 20.000 inhabitants) have been sampled proportionally to the daily number of benefits offered in the agglomeration. Less of 20.000 inhabitants agglomeration are not considered.

   According to the following notation:

   - A, the set of agglomerations in France
   - T_a, the daily number of benefits offered in the agglomeration a.
   - q, the number of sampled agglomerations (q=80).

[9] Except if the target variable is not correlated with weights.

- π_a , the selection probability of the agglomeration *i.*

$$\text{Hence: } \pi_a = q\frac{T_a}{\sum_{a\in A} T_a} \quad (3)$$

2) Secondary Unit: *Visits* sampling

In this stage, a stratification of centers has been done crossing the variable type of benefit[10] (dispersed accommodation, indoor lunch, indoor dinner, etc) and the variable users demographic features (man, woman, man coming with a child, woman coming with a child, a couple, etc) in order to **include the maximum variety of users asking for different type of benefits.** Inside each stratum a ***visit,* defined as a couple (center, day),** is sampled applying the systematic method with a fixed jump.

Two conditions are imposed in this second stage (equations (4) and (5)):

First, the visit sampling probability must be proportional to the number of benefits daily offered in a center:

$$\pi_{v|s,a} = q_{s,a}\frac{T_{v|s,a}}{\sum_{v=1}^{S} T_{v|s,a}} \quad (4)$$

Where:

- $q_{s,a}$, the number of **visits** to sample (inside the stratum *s* in the agglomeration *a)*
- $T_{v|s,a}$, the number of daily benefits summed on all the centers of stratum *s* in agglomeration *a*
- *S,* the number of **visits** that the stratum *s* contains.

[10] Typologie of benefis conditioned to the nomenclature used in the telephone survey to take a census of the structures in order to create our frame sampling. See Appendage *Definition of the different types of services of the SD's survey* for further information.

Second, benefits have the same final probability of being selected (even if agglomerations have been sampled with unequal probabilities)[11]:

$$\pi_{v|s,a} = \alpha_s \frac{T_{a|s,a}}{\pi_a} \qquad (5)$$

Where: α_s is a coefficient previously determined for the stratum *s,* depending on its target population coverage.

With this double aim, we have determined the number of visits to sample through the following expression:

$$q_{s,a} = \alpha_s \frac{\sum_{v=1}^{S} T_{v|s,a}}{\pi_a} \qquad (6)$$

Note that, expression (6) is obtained from combining expression (4) and (5).

3) Tertiary Unit: Benefits sampling

In the majority of the cases[12], four benefits (n_v=4) are sampled in each visit. Hence, the probability of selecting the *i* benefit ($\pi_{i|v}$) is defined as:

$$\pi_{i|v} = \frac{n_v}{T_{v|j,a}}$$

Where n_v is the number of benefits sampled in each center *v*:

Depending on the typology of the center, two ways of selecting the benefits have been carried out:

[11] The deflation factor $\left(\frac{1}{\pi_a}\right)$ in the expression (4) increments the number of visits to do in the medium and little agglomerations (having a lower probability of sampling that agglomeration).

[12] An exception to the four fixed contacts happens in the hotel rooms (where we limite the number to two) having less people that could take advantage of this service.

a) Through a **list of users**: for instance, in a center offering grouped accommodation for a long period of time, a list of users was available.
b) **Counting** the number of people in the actual queue (or queues): for example, in centers offering meal benefits which is a fast benefit, very dinamic and this increases the difficulty when performing the visit. We have to mention that in these centers a pre-visit is done to have an estimation of the number of users asking for a benefit.

A sampling table is required in both cases. In the case of a refusal, the sampling table proposes the possible remplacers (maximum 3 by contact). The table has been created such as to distribute the sampled benefits along the users of a visit. In that way, **we have maximized the chance of interviewing people with different characteristics during the period of time considered for visits**.

Number of potential users, between 116 and 136																
Starting point to look for **Contacts**	**15**	…			**47**	…			**79**	…			**111**	…		
Potential individual **replacer**		23	31	39		55	63	71		87	95	103		119	127	135

Figure 2. Example of sampling table for a visit where is planned to fill four questionaires. A number of potential users from 116 to 136 is considered. The 15th user waiting in the queue will be chosen to be interviewed, then the 47th, 79th and 111th. If, for instance, the 15th refused to answer, we would ask the 23th as first replacer.

Moreover, in the case of multiple visits to the same service, in order to avoid a high correlation between samples, as many different tables as visits have been built.

2.2 .2 The survey's reached population

The *Sans-domicile* population is not studied by the ordinary French national housing survey (*Enquête du logement, EL*) carried out since 1995 and repeated every 4 years. The main reason

for not being included is clear: the individuals interviewed in the *EL* survey are selected through a particular sampling method starting from a housing frame. Consequently, according to their definition, *sans-domicile* people obligatory are not reached by this process.

To illustrate this fact, a comparison of the two surveys' field has been done:

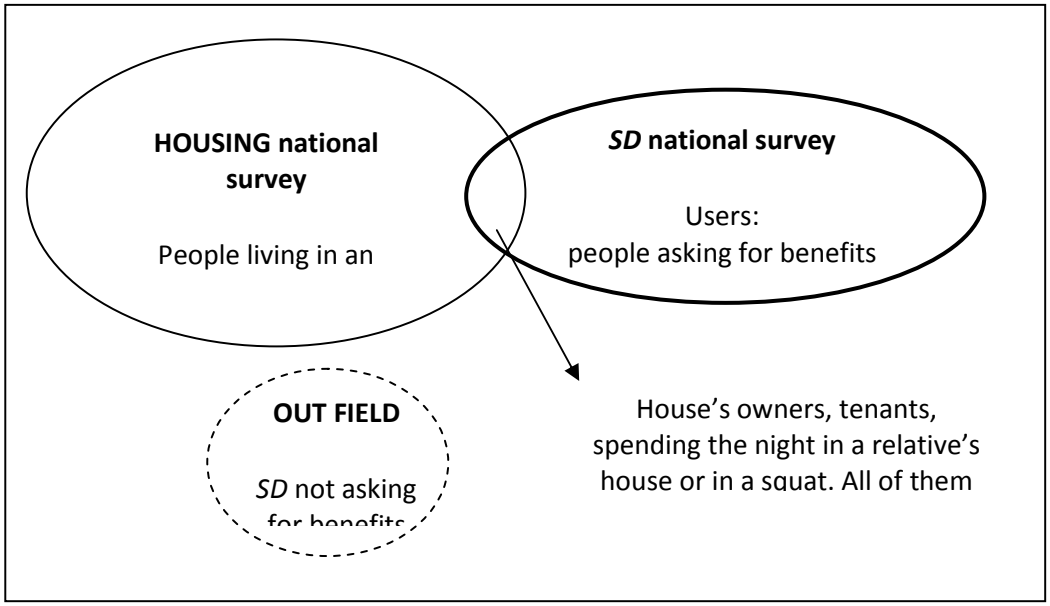

Figure 3. Different field coverture from Housing and *SD* national surveys

2.2.3 Reference period

A month has been fixed as the period of time for the collection of data. This period has been fixed looking to maximise the chances of covering an important proportion of *sans-domicile* and, at the same time, being possible to implement in the real practice. Thus, **it is assumed that users non-covered during a winter month are negligible.**

2.3 LIMITATIONS OF THE *TLS* AND THE *WSM* IN SD SURVEY

We have focused on two limitations that presents the applied methodology for reaching the *sans-domicile* population and that will be explained in this last section of the Introduction:
First, *Time-Location* sampling (TLS) method limitates the target population's coverture.
Second, as it has been noted in *2.2.4 The Weight Share Method (WSM),* the proposed unbiased estimator suppose that all links between interviewed people and the benefits population are known, which is a strong hypothesis far to be true. Basically for two raisons: first, it is technically impossible to know the exact number of links during the reference period (a month) and second, even if we count the number of links during a week, missing links appear quite frequently. Our master's degree thesis focuses on this second limitation.

2.3.1 A limited target population's coverture

Our baseline idea is that we are going to sample benefits offered by a recensed center (3rd stage of the sampling plan). Although sampling benefits is a good way to overcome our practical constraints, it induces **two main distortions** between the target population and the survey's reached population:

- The **reached population include people who actually do not belong to the initial target population** (for example, people who live in an ordinary house but in precary conditions and ask for meal benefits). Despite this, these people are also interviewed (people out of the *sans-domicile* square in figure 1).

- **The target population cannot be totally reached by the survey** (people out of the users square in figure 1): only the people that take advantage of benefits have a positive probability to be sampled. Thus, several *sans-domicile* profiles have not been studied in the real survey. For instance: people who sleep in the street for a short period of time without asking for help (probably because they ignore their existence or they refuse to use them). We can not either reach the *sans-domicile* living in urban agglomerations without help centers. Finally, we also miss people who arrive to the centers at moments not considered for visits.

2.3.2 Missing links in the Weight Share estimator

The *Weight Share* estimator takes for granted that we are capable to know the total number of services that each user has asked for during the survey's period (a month). Unfortunately, these quantities are not entirely known for two reasons:

a) The data collection is done along the period time in order to have the best coverage of the target population. So, for instance, a person interviewed at the beginning of the survey's period can not guess where he/she is going to sleep and eat in the future days.
b) The memory of the interviewed people is limited to some days.

Thus, it is impossible to estimate without any bias a total of interest on the survey's period without doing some hypotheses (Ardilly et Le Blanc, 2001).

The B part of the questionnaire (« *Fréquentation des services et situation vis-à-vis du logement »*) **collects the links[13] but only concerning the previous week of the interview** (called ***Semainier***).

The estimator given by the *WSM* is unbiased only if each interviewed person uses the same number of benefits from a week to another, a quite strong hyphotese. In this these, our estimations will be limited to an **«average week»** using the information of the *semainer.*

Nevertheless, a persistent problem is the **link non-response**, a phenomenon first studied by Lavallée and Xu[14] who investigated methods to **correct the overestimation** caused by the link non-response.

Explaining this fact in detail:

Being $\Delta = \Omega^U \setminus S_U$ a set of all the non-sampled benefits linked to an interviewed person and Δ_0 a subset of Δ of known links. Then:

[13] Each time that an interviewed person says that she/he has taken a benefit

[14] Xu Xiaojian, Lavallée Pierre : « Traitements de la non-réponse de liens dans l'échantillonnage indirect ». Techniques d'enquête. Canada, Décembre 2009.

$$r_k = \sum_{i \in S_U} r_{i,k} + \sum_{i \in \Delta_0} r_{i,k} + \sum_{i \in \Delta - \Delta_0} r_{i,k} \qquad (7)$$

Where $\sum_{i \in S_U} r_{i,k}$ is the set of sampled benefits used by the *k* interviewed person, $\sum_{i \in \Delta_0} r_{i,k}$ is the set of non-sampled benefits that the interviewed person state to have used during the previous week and $\sum_{i \in \Delta - \Delta_0} r_{i,k}$ **the set of non-sampled benefits missed**.

If we have a non-responder[15] and we ignore the third term of (7), we will apply the *WSM* with r_k^*, being $r_k^* \leq r_k$. Thus, according to the expression (2) we will get $\hat{N} \leq \hat{N}^*$, suffering from an **overestimation of our parameter** which is undesirable.

We see this effect through a hypothetical example:

Our aim –not far from reality- is to obtain an **estimation of the total of *non-roof* people** (noted by $\hat{Y}_{NR}$). In this case, two *non-roof* people have been interviewed. Notice that this figure is different from figure 2 because we have **introduced missing links** (dashed orange line).

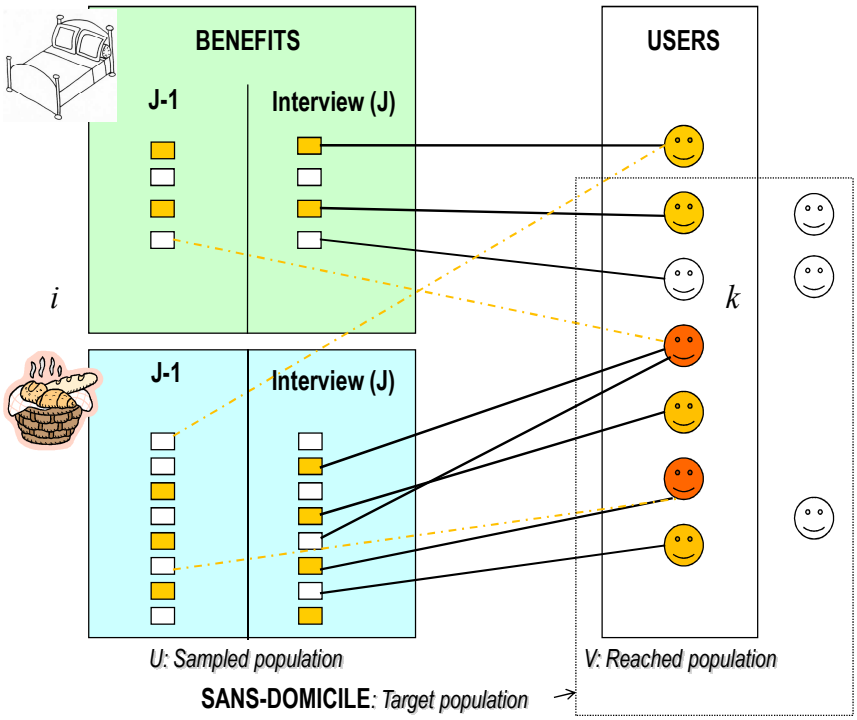

Figure 4. Example to estimate the total of *non-roof* people.

[15] A user for who we do not know her/his weekly attendance because her/his *semainier* is uncompleted.

Caption text:

: sampled benefit
: unsampled benefit
: interviewed person
: non-interviewed person
: link between a user and sampled benefit
: link between a user and an unsampled benefit
: **missing link between a person and an unsampled benefit.**
: ***non-roof* interviewed person**

Basing our calculations in the expression:

$$\hat{N}_{NR} = \sum_{k \in S_V} \tilde{w}_k y_k \text{ , where } \tilde{w}_k = \frac{1}{r_k} \sum_{i \in S_U} w_i r_{ik}$$

We figure that all benefits have the same weight or the same sampling probability (desirable but hard to manage it in a real situation) which means that: $\mathbf{w_i=w}$ for $\forall i \in S_U$. We fix this weight to 160 ($\pi_{i|v}$=0,006). Remember that r_{ik}=1 for $\forall k \in S_V$ for a benefit *i*.

We estimate our parameter under two different conditions:

a) We know **all** their links (figure 1, all lines are black and continuous). The first interviewed *non-roof* has 3 links and the second 2. Here we have the estimation:

$$\hat{N}_{NR} = \left(\frac{160}{3} + \frac{160}{2}\right) = 160\left(\frac{5}{6}\right) = \frac{400}{3} = 133,\hat{3} \text{ } \textit{non-roof} \text{ people.}$$

b) We know **some** of the links (figure 4, we have **link non-response**). In this case, the estimation would be :

$$\hat{N}^*_{NR} = \left(\frac{160}{2} + \frac{160}{1}\right) = 160\left(\frac{3}{2}\right) = 240 \text{ } \textit{non-roof} \text{ people.}$$

So, we have proved that $\hat{N}^*_{NR} \geq \hat{N}_{NR}$. Note that it will be equal only when the link non-response is, actually, that a user has not ask for a benefit at that moment (for instance, an interviewed person does not remember where she/he has slept two days before and, actually, she/he has slept in a gare which is not a benefit).

METHODOLOGY

The explained metodology will help us to achieve the goal **of proposing, validating and comparing different imputations methods.**

All the treatments that we are going to explore in the next pages are based on the data from the Sd2001 survey (n=4084 individuals).

The different steps of our methodology are summarized below:

Step 1: Study of the missingness mechanism

The data set is splitted in two subsets: responders and non-responders (weekly attendance known or unknown, respectively):

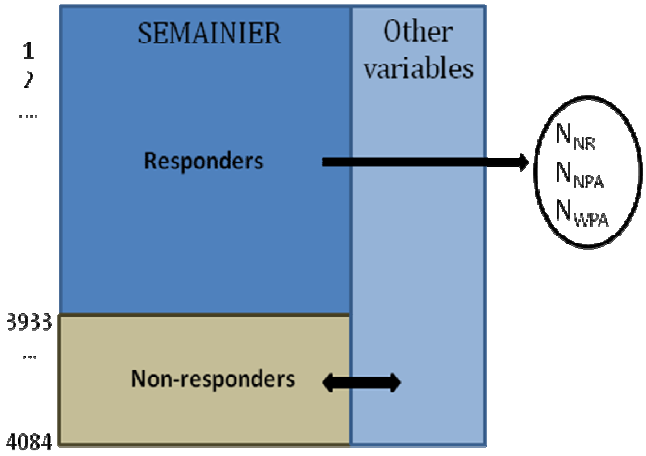

Figure 5. Outline of Step 1: Study of the missingness mechanism.

Three target totals are computed on the responders (n_1=3933): **N_{NR}, N_{NPA} and N_{WPA}** (the number of non-roof people, the number of people with no personal accommodation and the number of people with personal accommodation). These parameters, considered as **gold-standards,** are kept to be used in step 3 when computing the relatif error of estimation.

Then, we focus on the non-responders subset **(n_2=151 individuals)** in order to get closer to the **causes and reasons** of missings, which remains essential to avoid commiting the same mistakes in future survey's editions. A regression model for the probability of link response is estimated.

Step 2: Generation of missings

According to what we have learned in the previous step and other constrictions, a set of missings is generated (red interrogation symbols in figure 6) **using the responders subset (n_1=3933 individuals)**. Different missings rates scenarios have been considered. The missings generation algorithm will be explained in the section dedicated to this step.

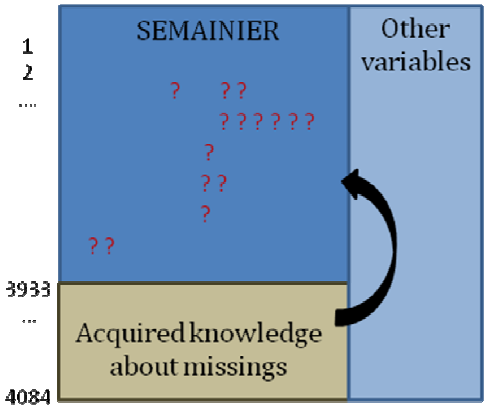

Figure 6. Outline of Step 2: Generation of missings.

Step 3: Application of three different missings imputation

Finally, three different imputation methods are applied on the missings generated in step 2. These three methods are explained in detail in the last part of this chapter.

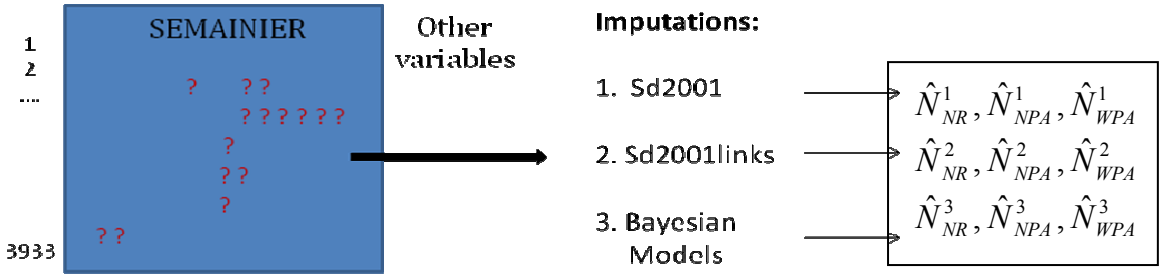

Figure 7. Outline of Step 3: Application of three different missings imputation.

3.1 Step 1: Study of the mechanism of the link non-response in SD2001

In this section we detail step 1:

First, we brefly recall a classic definition (Rubin, 1976) of the **three response mechanisms**:

- ***Uniform :*** the response probability is unrelated to any of the variables (auxiliary and target[16] variables). In this case, any piece of data is just as likely to be missing as any other piece of data and they are called ***Missings Completely At Random (MCAR).*** Without doing any kind of imputation (and thus, ignoring these data) estimated parameters remain unbiased, even if statistical power decreases.

  For instance: if the interview has been started but the individual is not feeling well and the interview has to be stopped, we might think that there is no implicit reason in that person to be a non-responder.

- ***Ignorable*** : missingness does not depend on the value of the target variable when controled by the auxiliary variables. They are classified as ***Missing At Random* (*MAR*).**

  For example: we want to report in the *semainier* at which center the interviewed person has eaten the previous days of the interview. As bad memory skills are usual in older people, more missings might be reported by this profile of users. But within each age group, missingness can be classified as *MCAR*.

- ***Non-ignorable:*** **the response probability is related to different variables (auxiliary and target variables)**. There is, inevitably, a bias caused by the missignesss. This hypoMaster's degree thesis is difficult to test. They are classified as missings ***Not at Random (MNAR).***

  For instance, women who have slept the previous nights in an urgency accommodation because they have been beaten do not want to report the type of benefit that they used (a particular benefit for beaten women) because they are afraid.

[16] Note that the target variable is the variable of interest that contains missings.

An **indicator *R* of link response** has been computed in order to explore the missingness mechanism: one dummy variable for each type of link of the *semainier* (accommodation, lunch and dinner from *j-1* to *j-7,* being *j* the interview's day).

Interviewed preson	A1	R_{A1}	…	A7	R_{A7}	L1	R_{L1}	…	L7	R_{L7}	D1	R_{D1}	…	D7	R_{D7}
1	21	1	…	11	1	1	1	…	1	1	5	1	…	99	0
2	99	0	…	99	0	10	1	…	1	1	99	0	…	7	1
3	42	1	…	42	1	8	1	…	99	0	1	1	…	8	1
4	11	1	…	99	0	8	1	…	99	0	7	1	…	99	0
5	11	1	…	99	0	99	0	…	8	1	1	1	…	1	1

Table 1. Extracted of 5 fictitious *semainiers.* « A » means Accommodation place. « L » means Lunch benefit used. « D » means Dinner benefit. The number added at the end of each variable refers to the day of the week, going from 1 to 7. Note that the cathegory *99* indicates that this link is a missing. Finally, the umbered columns are the indicators of link response, for each of the asked moment (meaning a total of 21 indicator variables for *SD2001*).

Here we are just interested in the indicators of link response (the set of umbered columns in table 1). With these 21 variables combined, we have obtained that: 61% of the individuals had **monotonous** non-response pattern. Monotonous non-response pattern is considered when, for instance, concerning accommodation links, R_{Ak}=0 from k=i, i-1, i-2, ...,i-7. The *i* moment is called the moment of the dropout.

Among the 4084 interviews in SD2001, 24 individuals (60 table cells) of the links with the accommodation places are unknown. Concerning the meal places, the link non-response is higher: 117 individuals (307 table cells) for the lunch benefits and 83 individuals (234 table cells) for the dinner benefits.

To **determine the non-response mechanism**, we had to fit a regression model for indicator *R.* Having a set of 21 binary values associated to each **non-responder**, a model for repeated measures has been chosen, taking into account the dependence within measures from the same person (autocorrelation, specified in a *working correlation matrix*). ***Generalized Estimation Equations*** (*GEE,* Liang and Zeger, 1986) for binomial response (*R*) have been used.

We had to convert the initial data set (table 1) into another format (table 2), where rows correspond to links. In table 2, three variables are associated to the weekly trajectory af a person:

- ✓ The **response indicator (R)**: R_{ijA} is equal to 1 if we know the place where the user *i,* at the moment *j* has slept (type of link A). Otherwise, R_{ijA} is equal to 0.
- ✓ *Semainier:* **the places**[17] **where they have slept and have eaten.** For accommodation, there are 17 different cathegories of places; for lunch an dinner, there are 10 modalities.
- ✓ **Link indicator:** it differentiates a place where benefits are offered (center) from a place where there are no benefits offered. For instance:
  - Sleeping in an accommodation centre (*Semainier=11)* and having lunch or dinner distribuited freely in a specific place or in a social restaurant (*Semainier=1)* **are considered benefits, and thus, a link.**
  - Sleeping in public places (*Semainier=42)*, not eating that evening (*Semainier=10)* or eating in the person's home or in a family or friends house (*Semainier=3, 4)* **are not considered benefits, and thus, there is no link**.

We decided to **make difference between the non-response of accommodation places and the meal places** (lunch and dinner together) following the decisions made in 2001 and found in the documentation of the survey[18]. Thus, a new variable called «**Type of link**» has been created to make this distinction (cathegory A, accommodation and M, meal) .

[17] Place: where the interviewed person has spent the night or has eaten. It can be a center or not (a link or a no link). See the details in the Appendixes: 9.1 The Semainier's extract for the face to face questionnaire

[18] From SAS programs used in 2001 for the links imputations (Sd2001).

Interviewed person	**Moment**	**Type of link**	***Semainier***	**Link indicator**	**R**	**Age**	**Understanding level**
1	1	A	11	1	**1**	<26	Good
1	2	A	11	1	**1**	<26	Good
i	*m*	A	**…**	**…**	**…**	<26	Good
1	7	A	99	99	**0**	<26	Good
1	1	M	1	1	**1**	<26	Good
1	2	M	99	99	**0**	<26	Good
1	…	M	**…**	**…**	**…**	<26	Good
1	7	M	99	99	**0**	<26	Good
2	1	A	99	99	**0**	>60	Bad
2	2	A	42	0	**1**	>60	Bad
2	…	A	**…**	**…**	**…**	>60	Bad
2	7	A	99	99	**0**	>60	Bad
2	1	M	1	1	**1**	>60	Bad
2	2	M	99	99	**0**	>60	Bad
2	…	M	**…**	**…**	**…**	>60	Bad
2	7	M	99	...	**0**	>60	Bad

Table 2. Extraction from the data set to explain the variable *D* (n=2 individuals in this concret case). The variables *individual (i), moment (j)* and *type of link (k)* are identifiers of the statistical unit for the model. The other variables (such as *age* and *understanding level*) are some of the explanatory variables of the model.

<u>Only the non-responders</u> (n_2=151 individuals) have been introduced. Two *GEE* models were estimated: one with the accommodation non-responders (*Type of link* being 'A', n_{21}=24) and the other with the meal non-responders (*Type of link* being 'M', n_{22}=142). Notice that interviewed people can be accommodation and meal non-responders at the same time.

In the following, we focus on the bases of this kind of models to justify its choice in this context:

Let be R_i a vector of 7 components for the individual *i* (*i* going from 1 to n_2). The model specification of a GEE model involves three elements:

- Systematic part: relates the expectation of an observation to the linear predictor via the link function *g*:

  $g(E[R_i]) = g(\mu_i) = x_i'\beta$, being x' the set of explanatory variables of the model

- Random part: specifies the variance function that has the following expression:

  $$V_i = \varphi A^{1/2} W_i(\rho) A^{1/2}$$ where,

  Φ is the dispersion parameter estimated using our data,

  A is a diagonal matrix of size *k x k* with: $v(\mu_{jj}) = \mu_{jj}(1 - \mu_{jj})$ in its diagonal,

  W_i is the autocorrelation matrix

- Correlation part: imposes the correlation structure for observations on the same interviewed person. In our case, the exchangeable (or compound symmetry) matrix has been used, as it does not require a high number of parameters to be estimated. Moreover, having observed in our data a mixture of two non-response patterns (monotonous and intermittent)[19], we have chosen a flexible structure, definied as follows:

  $$\text{Corr}(R_{ij}, R_{ik}) = \begin{cases} 1, j = k \\ \rho, j \neq k \end{cases}$$

In our case, the starting explanatory variables introduced in the model are socio-demographic variables (gender, age, housing situation and professional experience). In addition, the center of the interview and the understanding level[20] of the interviewed person (excellent-good, decent, bad) have been considered. All of them have been introduced according to the criteria of some experts from the working team. Likelihood ratio test has been used to compare

[19] At the beginning, we have tried to find a model for *D* for each of the non-response patterns: monotonous and intermittent separately. We have renounced to this approach due to the small size of interviewed people in each group leading to an undesired decrease of the statistical power.

[20] *Understanding level* was chosen from a set of variables belonging to a part of the questionary that the interviewer had to fill, according to his/her perception of the course of the interview. Variables such as the understanding level of the interviewed person, the facility to express himself/herself, his/her level of interest or his/her level of suspicion were introduced as active elements in a *MCA (Multiple Correspondance Analysis).* We finally summarized all this information with the understanding level (having a high contribution in the axes building and, at the same time, being well-represented in the factorial plan).

nested models and select our two final models. Notice that these variables were used to test de MCAR and MAR hypoMaster's degree thesis.

To summary, two variables have a statistical significance role in explaining the link response; *moment* (day of the week previous to the interview) and *understanding level* (of the interviewed person). We remark that reference levels have been chosen in order to have a high number of effectifs as recommended in the literature[21]; thus, for the *moment,* the previous day of the interview and, concerning the *understanding level,* the best understanding level.

The two temporary final models for the link response concerning the accommodation and meal places, showed that:

- ✓ For both kinds of benefits, **the further the referred moment is from the present the higher the link non-response probability is**[22]. This result could be expected as it is related to memory.

  In the model for accommodation links (p-value=0,02), the lack of memory starts affecting beyond the fourth day before the interview (95% CI OR not including the 1). For instance, passing to ask about the previous day of the interview to the fourth day before the interview, **multiplies the odds of response by 0,06.**

  In the model for meal links (p-value<0,001), the lack of memory has already started in the second previous day of the interview (95% CI OR not including the 1). For instance, passing to ask about the previous day of the interview to the fourth day before the interview, **multiplies the odds of response by 0,02.**

  According to these remarks, it seems that interviewed people find **harder to answer in a completed way all the questions concerning meal places** than accommodation places.

[21] Insee, métodologie statistique « L'econometrie et l'étude de comportements: Les modèles univariés à résidus logistiques ou normaux », page 44: « *La situation de référence: à quoi sert-elle? Comment la choisir?».*

[22] See Appendixes 9.3: Output GEE estimations. SAS Output 1 and 2.

- ✓ Only for meal places, **the more the interviewer perceives that the person has a bad understanding of the questions the higher the link non-response probability is**[23] (p-value=0,01)**.**

  Concretly, the effect of the understanding level is noticeable when the interviewed person has a bad understanding level: with a moment to ask fixed, passing from a person with an excellent-good level to a person with a bad level **the odds of response by 0,25.**

Taking into account these results, rejected the hypoMaster's degree thesis that the missing links are *MCAR.*

In order to test the ignorability assumption for the missing data (that is, missings are not *MNAR*), we had to impute their missing values using the method applied in 2001[24], assuming that this method is valid. After the imputation, we have introduced the imputed variable in our temporary final models as another covariate (*Semainier* and *Link Indicator* in table 2 without any missing or *99* cathegory). Notice that, among two possible covariates- *Semainier* and *Link Indicator*- for simplicity, we have chosen *Link Indicator* to be introduced in the existing models.

In both models, ***Link Indicator*** **has not a significant coefficient** (p-value=0,054 for the accommodation links and p-value=0,11 for the meal links)[25]. Thus, having asked for a benefit or not (*Link Indicator)* is not related with whether to answer or not to the question. For instance, interviewed people who have slept under a bridge have equal probability to answer that those who have spent the night in a center. Therefore, **we have found no empirical evidence to reject the ignorability hypoMaster's degree thesis of our missing data.**

[23] See Appendixes 9.3: Output GEE estimations. SAS Output 2.
[24] See section *5.3.1: Sd2001 method.*
[25] See in the Appendixes 9.3: SAS Output 3.

3.2 Step 2: Generation of missings

Once we have studied the missingness mechanism, we use the set of responders **(n_1=3933 individuals)** to generate new missings. We remind you that the reason why we simulate new missings: to test the three different imputation methods in terms of relative error, we need to know the real value of our totals of interest (N_{NR},N_{NPA},N_{WPA}). In this case, notice that we consider:

- **Parameters**: estimated totals in the sample of available cases (n_1=3933 individuals): N_{NR},N_{NPA},$N_{WPA.}$. They are considered as "parameters" because they are computed in the **real** sample.
- **Estimator:** estimated totals in the sample of available cases with generated missings and posterior imputation with method *j*: $\hat{N}_{NR}^{j}, \hat{N}_{NPA}^{j}, \hat{N}_{WPA}^{j}$

Concerning the generation of missings, **three different missing rates have been computed**, following the criterion employed by Allison (2010)[26]: 1, 5 and 20%.
Thus, two requirements have been imposed at the moment of designing the missings generation algorithm:

- To faithfully reproduce the real missigness mechanism studied in the previous section (8).

- To remain flexible in the generation of the missing rate to test the proposed methods under different interesting scenarios (9).

Below, the algorithm designed to obtain those missings and divided in two stages:

a) <u>Generation of the missing link « root »</u>

The basic idea is that we use the information of the GEE regression models of the previous step. Thus, we have two set of response probabilities:

[26] Allison, Paul: « Imputation of Categorical Variables with PROC MI ». SUGI 30. Focus Session. Paper 113-30.

Concerning the accommodation links, an estimated probability of response for every day of the week has been kept (*probA* in the SAS code[27]):

$$probA_1, probA_2, \ldots, probA_7 \quad (10)$$

Concerning the meal links, an estimated set of probabilities of response, crossing the variable *moment* and *understanding level*, are kept (*probM* in the SAS code):

$$probM_{1Good}, probM_{1Ok}, probM_{1Bad}, \ldots, probM_{7Good}, probM_{7Ok}, probM_{7Bad} \quad (11)$$

As a first approximation, we used these probabilites in a Bernoulli distribution to indicate whether a link is a missing or not (*R,* table 2*)*. In adition, we created a filter via a Bernoulli variable (*Alea,* table 3), according to requirement (9).

Hence, to generate a missing link « root » ($Root_{kij}$=1*)* fot the individual *k,* the moment *j* and the type of benefit *i* (accommodation or meal, were the person has been interviewed) is:

If (R_{kij}=1 and $Alea_{kij}$=1) then $Root_{kij}$=1 (12)

Below you have an example of accommodation benefits, this first stage of the algorithm would be completed as follows:

Individual	Moment	Understanding level	Probas A	R	Alea	Root
2	1	Good	0.96	0	0	0
2	2	Good	0.83	0	0	0
2	...	...	...	...	...	...
2	7	Good	0.46	**0**	**0**	**0**
...	...	...	...	...	...	...
501	1	Bad	0.47	**0**	**0**	***0***
501	2	...	...	...	...	...
501	...	Bad	0.96	0	0	0
501	6	Bad	0.45	**1**	**1**	***99***

Table 3. Table for the generation of the missing link ***root***

[27] For the detailed SAS code, see Appendixes: 9.4 SAS code for SD method.

Where:

- *Moment, understanding level*: covariates that determine the response probability (*Probas A/M,* expression (10) and (11), for accommodation or meal links, respectively). These probabilities are directly obtained from the regression model for accommodation or meal link response (step 1).
- *R* is a Bernoulli variable with probabilites *Probas A/M. Alea* is another Bernoulli variable with probability fixed by the user (depending on the missing rate; under 5%, 0,97 for accommodation links and 0,87 for meal links).
- ***Root* is an indicator variable of missing link**: if *R* and *Alea* are 1, then *Root* is '*99'* and thus, individual *k*, at the moment *j* and the benefit *i* has a missing.

According to the example in table 3, 1 links has been converted to missing. Thus, indivual 2 is considered to be a responder individual whereas individual 501 is a non-responder with a missing in the place where he/she slept six days previous to the interview.

b) <u>Configuration of the non-response pattern</u>

We face the fact that we had completly lost the dependence within each individual (while we had seen that 61% had monotonous patterns). Thus, according to one of the requirements for the algorithm (8), we have proposed to take into account: the percentage of monotonous patterns for the accommodation links (P*MonotA*=0,5) and meal links *(PMonotM*=0,75), as probabilites of two **indicators of the monotonous response pattern** associated to each individual.

Then, **for each missing link « root » (moment *j*):** if the individual had a monotonous response pattern, **since the moment *j-1* and until the furthest moment, all his/her links were also converted into missing links.**

Nota that all missing link « roots » are symboled with a « ***99***» and that all empty cells are different benefits cathegories which are not specified to simplify the example.

Individuals	Understanding level	*SEMAINIER*										Monot A	Monot M
		A1	***A2***	‥	***A6***	***A7***	***M1***	‥	***M5***	***M6***	***M7***		
1	Good					***99***					***99***	0	0
2	Good											*1*	*0*
3	Bad								***99***			*0*	*1*
4	Bad											*1*	*1*
k	...	...	...		...	*i*	...	...	...	...	...	...	...
501					***99***							1	*0*
3932	Good							***99***				*1*	*1*
3933	Bad		***99***									*1*	*0*

Table 4a. Data set after completing step 1: Generation of the missing link ***root.***

Individuals	Understanding level	*SEMAINIER*										Monot A	Monot M
		A1	***A2***	‥	***A6***	***A7***	***M1***	***..***	***M5***	***M6***	***M7***		
1	Good					***99***					***99***	0	0
2	Good											*1*	*0*
3	Bad								***99***	*99*	*99*	*0*	*1*
4	Bad											*1*	*1*
k	...	...	...	..	...	*i*	...	...	...	...	...	...	...
501	...				***99***	*99*						1	*0*
3932	Good							***99***	*99*	*99*	*99*	*1*	*1*
3933	Bad		***99***		*99*	*99*						*1*	*0*

Table 4b. Data set after completing step 2: Configuration of the **non-response pattern**.

To configure the non-response pattern, we have used:

- *MonotA* and *MonotM,* **two monotonous-pattern indicators,** with probabilities 0.5 and 0.75, respectively.

For example, individual 1 has no monotonous-pattern associated (neither accommondation nor meal places) so no more missing links will be created for her/him. Individual 501, having an

accommodation monotonous-pattern and a missing link ***root*** in acommodation the previous day before the interview (A1), he/she will have missing links in all the previous links to moment 2 from accommodation places.

3.3 Step 3: Application of three missings imputation methods for the « Semainier »

Once we have an algorithm to generate different missings links rates in the given data set of 3933 available cases, we focus on **explaining the three imputations methods** for the *Semainier* that have been applied on the 3933 interviewed people:

- ***Sd2001***: the reference method applied when exploiting data from 2001 survey. It is based on the nearest neighbourd principle.
- ***Sd2001links***: a version of *Sd2001* method. We slightly modify one of its parameters: the entered *Semainier*.
- ***Bayesian models****:* we performed predictions with an estimated Bayesian model of the number of weekly links.

We describe these methods in detail in the following subsections.

3.3.1 Sd2001 method

In 2001, after the survey's first edition between January and February 2001, an imputation method for the *Semainier* was written in SAS code. We had access to these files and that is why we found it. To take profit of this method, we had firstly to decode the SAS script[28].

The algorithm of the **reference method** of Sd2001 is a combination of two simple methods: **imputation by the nearest neighbourd and random *hot-deck*[29] imputation**. We explain each step through an example focused on the accommodation links imputation (the same process has to be considered for the meal links imputation).

[28] For the detailed SAS code, see Appendixes: 9.4 SAS code for SD method.
[29] The term *hot-deck* means that donors belong to the same data set of non-responders.

1. Selection of all the individuals (among the 3933) with at least one missing link (set of **non-responders**). Also, selection of those that have all the *semainier* filled up (set of **responders**).

2. Identification of all the responders individuals with the same gender and center of the interview than the non-responders. The set of individuals that meet these conditions are called ***candidate donors.***

Table 5 shows the **donnors strata** that are created after crossing *gender* and *type of benefit:*

	Dispersed A	**Dispersed A in rooms**	**Grouped A (>15 days)**	**Grouped A (<15 days)**	**Outdoor lunch**	**Outdoor dinner**	**Indoor lunch**	**Indoor dinner**
Man	473	116	780	326	16	132	472	198
Woman	559	76	498	131	4	18	95	39
	1032	**192**	**1278**	**457**	**20**	**150**	**567**	**237**

Table 5. Number of donors for each center of interview (from 2001) and gender.
Note that the fourth type of benefit refer to accommodation benefits (A).

3. Weekly attendance of helping benefits for the **non-responders** and their **candidate donors**. We count the number of links with each type of place where individuals have spent the night[30].

[30] We would do the same treatment with the meal links.

Non responder	Candidate donors	Sex	Type of Benefit	NUMBER OF DIFFERENT ATTENDED PLACES (BENEFITS OR NOT)							
				Non-responder				Candidate donors			
				Centre	*j*	Public place	99	Centre	j	Public place	99
501	**2**	M	I.Dinner	3	…	1	2	3	…	2	0
501	287	M	I.Dinner	3	…	1	2	7	…	0	0
501	299	M	I.Dinner	3	…	1	2	5	…	1	0
501	307	M	I.Dinner	3	…	1	2	7	…	0	0

Table 6. Number of attended places counted for each cathegory for the non-responders and their candidate donors.

501 non-responder has 4 candidate donnors: **2**, 287, 299 and 307, because they have the same gender (male) and they are interviewed in the same type of center (*indoor dinner*).

4. **Proximity indicators between non-responders and candidate donors.** For each row, we compute the *distance* between two individuals as the difference between the number of attended places of each cathegory (benefits or not) from the non-responser and from every candidate donor. Notice that missing link (*99*) is also considered a cathegory of the type of place; thus, the candidate donors have 0 counts of missing benefits as they are all responders.

   It is summarized with the following expression:

$$\textbf{Distance}_{kk'} = \sum_{q=1}^{17} \left| NbPlaces_{kq} - NbPlaces_{k'q} \right|,$$

   where *k* is the index for the non-responser individual, *k'* the index for one of his/her candidate donor and *q* the modality of place where the person has slept (from a total of 17 different cathegories).

   This distance will always be over or equal to 2 because, at least, one attended place is missing in the non-responders trajectory.

5. **Selection of couples « non-responser x donor » with the minimal distance for each non-responser individual.** In the case of having several candidates equally close to the non-responser (desirable), we will randomly select a couple.

6. **Missing links replacement** from each non-responser individual by the type of places that its donor has attended.

It is important to mention that, as it has been introduced in step 2 of this algorithm, candidate donors are searched within a strate of responders. There is an underlying hypoMaster's degree thesis: **the weekly attendance profiles are similar within people interviewed in the same type of center and from the same gender.** Moreover, note that it would be quite expensive (in terms of time and memory resources[31]) to ignore it and execute the algorithm without any kind of selection of the donors, meaning that all the responders could be donors for each non-responser.

3.3.2 Sd2001links method

A slight modification has been done compared to the reference method of 2001. The *semainier* information has been simplified in three possible cathegories, aggregating all the original ones. Table 7 shows the aggregations that have been carried out:

Original cathegories		**New cathegories**
Accommodation	**Meal (lunch and dinner)**	
11, 12, 21, 31	1	**1**
99	99	**99**
Others (12)	Others (8)	**0**

Table 7. New aggregation of the original link cathegories for accommodation and meal places

[31] For example, suppose we have 100 partial non-respondents (which is 3833 donors per individual). We will work with a table of 383.300 rows to calculate distances and choose the best donor for each non-respondent. By limiting donors with strata (gender and place of the interview), for the same size of non-respondents, the table has about 70.000 lines.

Thus: **the cathegory *1*** of the new set of cathegories means **a link** because the person has used an individual benefit reported in the recensed centers data set and, hence, her/his chance of being interviewed is higher. ***0* means a no-link** because, even if the person has slept or has eaten somewhere, this place is not in the recensed centers data set so there is probability 0 to sample it. Finally, ***99*** value keeps the idea of missing answer as originally.

Although **any of the explained steps of the 2001 methods has been modified**, **the *distance* notion is only based on three counts**: the number of links, the number of no-links and the number of unknown links. Therefore, for our imputations we will ignore the concrete cathegory of place which is logical having in mind our aim: **to complete the *semainier* with the correct counting of the number of weekly links for each individual.**

3.3.3 The Bayesian Model

There exist different reasons why we decided to explore this third methodology. They are:

- ✓ In the article of reference for this Master's degree thesis (Ardilly et LeBlanc, 2001), when they explained how to perform estimations over the period of time, weekly links needed to be extended to monthly links and, without any knowledge, strong hypotheses of regularity along the weeks needed to be done. Then, they proposed to exploit the information of the survey concerning the behavior of centers attendance.

- ✓ In the second article of reference (Lavallée, Xiaojian, 2009), they proposed to assume that the probability of a link between a unit in sampling population (benefits) and a unit in target population (users) depends on some auxiliary variables through a logistic regression model. They also cite Draper and Smith who state that the choice of which model should be employed is not always clear in practice.

- ✓ Strata used in Sd2001 need to be validated to decide if continue using them or replace them. Models can help us to do it and to propose other strata if the results show this.

Therefore, **our objectives with this Bayesian Model** are:

1. To get a first idea of what kind of models can we use to explain the weekly number of links.

2. To learn and interpretate the covariates that have a significant role.

3. To verify the reference stratum and, if possible, to propose other strata.

4. To explore and compare the imputation capacity of our model with imputations carried out with the two previous donors methods.

3.3.3.1 Formulation of the Bayesian model

The first idea that we tried was to formulate a Bayesian model like this, being **Y_i the number of <u>total</u> links per week for the individual *i.*** Note that we have dividided the available individuals (n=3933) in a model-subset and a test-subset. The model-subset (n=3117) are the set of individuals used to estimate the model. Hence:

$$y_i \sim \text{Binomial}(n_i=n=21, p_i), \text{ for each } i=1,..,3117$$

$$\log it(p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_q X_q$$

$$\beta_k \sim Normal(\mu = 0, \tau = 10^{-6}), \text{ for k=1,..,} q \text{ regression covariates}$$

The figure 8 below helps to understand that the formulation above is too simple for our data:

The first barplot corresponds to response variable for the first Bayesian model formulated above. As you can see, there is a mode in 7 links extremly separated from the other values. It seems that the distribution is showing a bi-population: those who have 7 links and the others. Studying the distribution of the accommodation links, the lunch links and the dinner links we can observ that the first has its mode in “7 links” and lunch and dinner have their mode in “0 links”.

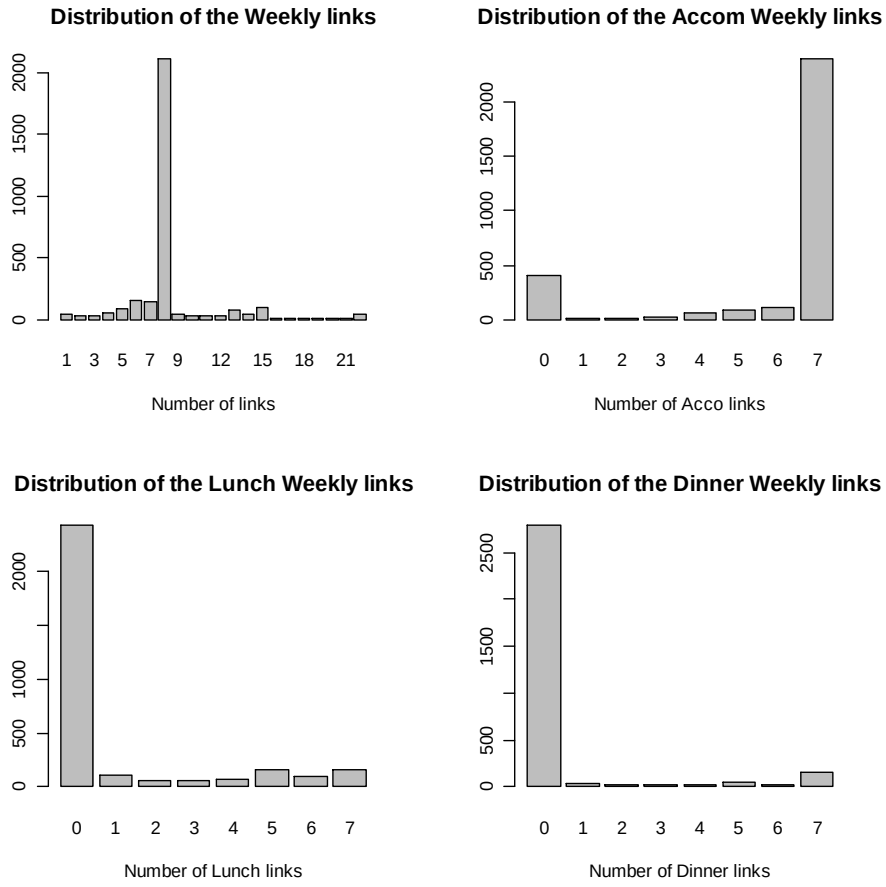

Figure 8. Distribution of the total number of links, the accommodation links, the lunch links and the dinner links per week.

Thus, we propose another formulation of the Bayesian model in order **to capture the behaviour of our data**. The model is based on **a 2-dependent-level model**. We have one for each type of link: accommodation, lunch and dinner. The formulation of the model is:

For the **<u>accommodation links</u>**:

Level 1: We define the indicator variable $Y_1 = I_{[Y=7]}$. Then :

$$Y_{1i} \sim Bernoulli(p_i^1), \text{ for each } i=1,..,3117.$$

$$\log it(p_i^1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{q_2} X_{q_2}$$

$\beta_k \sim Normal(\mu = 0, \tau = 10^{-6})$, *for each* k=*1,...,q₁* covariates of the regression model

N_j individuals are assigned to the cathegory "0..6 links" and N-N_j are assigned to "7 links", being j the number of the iteration. Note that N_j is a random variable that depends on the iteration.

Level 2: The response variable for level 2 is **Y_2= Number of accommodation links.** Then:

$$Y_{2i} \sim Binomial(n_i = n = 7, p_i^2)\text{, for each i=1,..,750}$$

$$\log it(p_i^2) = \lambda_0 + \lambda_1 X_1 + \lambda_2 X_2 + ... + \lambda_{q_2} X_{q_2}$$

$\lambda_k \sim Normal(\mu = 0, \tau = 10^{-6})$, for each k=1,...,$q_2$ covariates of the regression model

Note that **the prior distribution chosen for our q_1 and q_2 parameters are non-informative as this is a "pilot model".**

For **lunch or dinner links**, the formulation is the same except that the indicator variable in level 1 is **$Y_1=I_{\{Y=0\}}$.**

We have performed with WinBUGS software 1000 simulations of the model for level 1 and 1000 simulations for the model of level 2.
Below you have an outline (figure 9) to understand the final formulation of our Bayesian Model for two iterations:

According to these examples, in iteration *k,* 3 individuals are classified to group "0..6 links" (red) and, thus, n_k=3. In iteration *k+1,* 6 individuals are classified to group "0..6 links" and, thus, n_{k+1}=6.

Note that, at the end of the 1000 iterations, we obtain a vector of 1000 estimations of the q_1 parameters of the first level, estimated with the total number of individuals (n=3117) and a vector of 1000 estimations of the q_2 parameters of the second level, estimated with the number of individuals who where classified to "0..6 links", **a random number (n_k) that depends on the iteration *k*.** Remark that distributions of parameters for level 2 will be more dispersed.

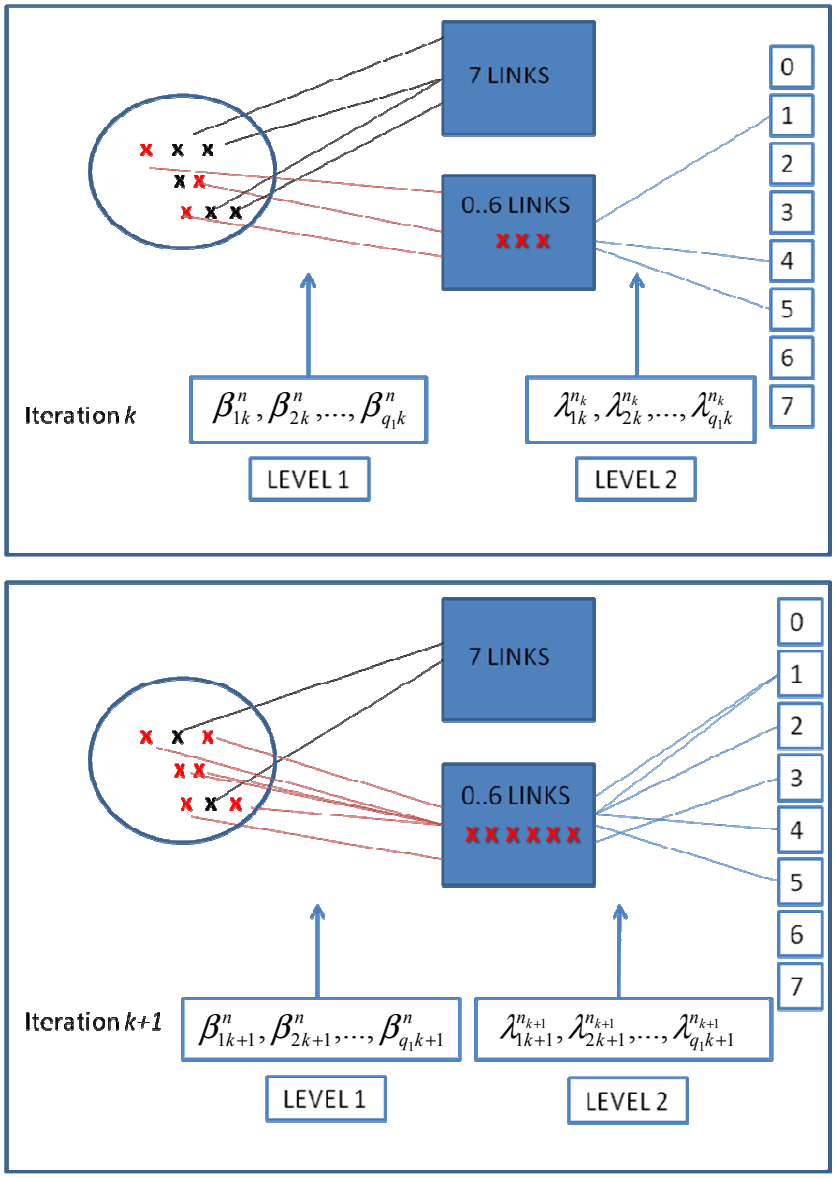

Figure 9. Outline of two iterations in the estimation of the accommodation links model.

3.3.3.2 The model covariates

A set of **25 questions** have been retained that we think can have a role in explaining the weekly number of links. We have performed a Chi-Square Independence test between each of them and the explanatory variable, fixing a significance level of 10% (to filter variables which are not very important, as we have a lot of statistical power). After this filter, we have finished **with 18 variables**. They can be classified as:

- **Demographical**: gender, place of birth (France, Out of France), couple, children, age (<37, between 37 and 56, >56), professional situation (never has worked or never <6 months, works at the moment, has lost the job).

- **Personal**: has already slept in the street, will sleep in the same place than yesterday, has eaten always in the previous week.

- **Help**: visiting a doctor in the previous 12 months, has received a food basket, has a social assistant, weekly attendance of the centers (general questions complementary to the *Semainier* where the information is detailed day per day: always sleeps in the center, sleeps >4night/week, sleeps 1-3 nights, sometimes), type of center of the interview.

- **Situation**: subpopulation (NSD, ASD, SA), solitude feeling, alcohol consumption.

3.3.3.3 Selection of the model

Each model from each level has been selected iteratively starting from the completed model (with the 18 covariates) and eliminating one at each step. Thus, **backward selection** has been performed. A probability of significance has been computed for each covariate (*k*). It is defined as:

$$p_k = min\ \{P(\beta_k<0), P(\beta_k>0)\}$$

The algorithm has stopped once all the covariates have, at minimum, one cathegory (dummy in the model) statistically significant. Then, DIC criterion has been used to finally choose the model for the corresponding level.

Concerning the validation of those 6 models and the interpretation of their covariates, it is all explained in the next chapter: Results and Discussion.

RESULTS AND DISCUSSION

In this chapter, we introduce our target parameters and the statistic to measure the validity of an imputation method. Then, we start comparing the two donors methods: Sd2001 and Sdlinks. Next, we focus on the Bayesian models. We start by explaining their contributions compared with the donors methods. Then, we validate the Bayesian models to understand in which part they are not working. Finally, we compare them with donors methods. Once all the imputations methods compared, we interpretate the Bayesian models and propose new strata for Sd2012. At the end of this chapter, we analyze future perspectives for Bayesian models, focused on imputing the non-francophones.

4.1 Target parameter

The first objective of Sd2001 survey is to determine the number of people considered as "sans-domicile" (Ardilly and Le Blanc,2001)[32]. According to this objective, we consider here 3 target parameters. we are interested in:

- The total number of non-roof people: $\mathbf{N_{NR}}$
- The total number of non-personal accommodation people: $\mathbf{N_{NPA}}$
- The total number of with personal accommodation people: $\mathbf{N_{WPA}}$

These totals have to be computed from the users of helping services in France (having used at least once from 15th January to 15th February 2001). The first two subpopulations are two kind of *sans-domicile* (see Introduction, 2.1.2 Nomenclature of *sans-domicile* survey).
We recall that the total of a population *i* is computed through the Weight Share estimator:

$$\hat{N}_i = \sum_{k \in S_V} \tilde{w}_k y_k \text{ , where } \tilde{w}_k = \frac{1}{r_k} \sum_{i \in S_U} w_i r_{i,k}$$

In this chapter, we want to compare and explore three methods of imputing the number of weekly links (r_k). Then, the **relative error,** also used in testing imputation methods (Lavallée and Xiaojian, 2009[33]) is computed as:

[32] Ardilly, Pascal et Le Blanc, David : « Echantillonnage et pondération d'une enquête auprès de personnes sans domicile : un exemple français ». Techniques d'enquête, volume 27, June 2001.
[33] Xu Xiaojian, Lavallée Pierre : « Traitements de la non-réponse de links dans l'échantillonnage indirect ». Techniques d'enquête. Canada, Decembrer 2009.

$$RE_{ikm} = \left(\frac{\hat{N}_{ikm} - N_i}{N_i} \right) x100 \qquad \text{where :}$$

- $\hat{N}_{ikm}$ is the estimation of the total of the **population *i*** (*no-roof, no personal accommodation, with personal accommodation*) under a **link non-response rate *k*** (1, 5, 20%) applying an **imputation method *m*** (*Sd2001, Sdlinks, Bayesian Models*).
- N_i the real total of the population *i.*

4.2 Donors methods: SD2001 vs SDlinks

First, we compare two donors methods: SD2001 and SDlinks. First, we explain our hypothesis about the results we expect to find. Then we show the obtained results.

These two methods are based on searching a donor for each non-respondent with the closest trajectory (places where they have gone during a week). These **two methods assume two hypotheses**:

a) Users who are close in terms of the known trajectory will be actually close for the unknown trajectory of the non-responders.
b) The candidate donors found inside a stratum should be the closest in terms of weekly trajectory. This could not be true if the strata are not well chosen.

Three parameters determine the donors method: the **variables to impute** (*Semainier*), the **strata** and the **criterion** to choose a donor.

Between Sd2001 and Sdlinks, the only difference is focused on the *Semainier*. With Sd2001, the number of cathegories of place is 17 for accommodation part and 10 for meal part; with Sdlinks the number of cathegories is simplified to 3: a benefit (a link), a no-benefit (no-link) and a missing. Remark that **with Sdlinks we try to assure the hypothesis (a) above. Because the information that we will keep from each individual is the number of benefits (links) that he/she has in the previous week, we choose donors who are close to NR in terms of the number of benefits.** For instance, **when Sd2001 separates two users** if they have not the same cathegory (the first slept in an accommodation centre (11) and the second in a hotel payed by an association (31), **Sdlinks bring them closer because both places are a benefit and, thus, a**

link. We will corroborate the assurance of hypothesis (a) above with the relative error (Figure 10, 11 and 12 below).

Another question concerns changing from Sd2001 to Sdlinks: if the criterion to choose a donor is based on data agrouped in different number of cathegories, does the number of candidate donors change?

We think that Sd2001 tends to **get a lot of "separated individuals" (useless donors) than to get "closer individuals" (useful donors that will be candidates)[34].** Thus, with **Sdlinks we increase the number of candidate donors**; that is, for a non-responder, we have more donors with a minimal distance.

According to table 8, under the assumption of 20% of missings, SD2001 associates only 1 donor as candidate donor to a non-responder (NR). Therefore, there is no variability which **makes the imputation poorer**. If, in this particular case, we do not choose the donor with the minimal distance and we choose donors among those with distance 6 the problem would be solved; however, this possibility is not considered by the SAS algorithm. Notice that this happens with a NR of a medium stratum (table 8) and also with small stratum.

Distribution of donors distance with a NR in a **MEDIUM STRATUM** (n=116 donors)					
Sd2001			**Sdlinks**		
Distance	Cases	%	Distance	Cases	%
4	**1**	**0,86**	4	**47**	**40,52**
6	19	16,38	6	3	2,59
8	1	0,86	8	5	4,31
10	2	1,72	10	4	3,45
12	2	1,72	12	5	4,31
14	91	78,45	14	52	44,83

Table 8. Distribution of donors distance with a NR in a medium stratum. Comparing distance from Sd2001 and distance from Sdlinks

[34] See the Appendixes: 9.5 Example of Sd2001 and Sdlinks.

Applying the Sdlinks method, we will have 47 candidate donors. We also have some useless donors (45%) but they are less than with Sd2001 (79%). It has to be mentioned that, of course, it is also important that we reject some donors from the stratum that are very different of the NR but our priority is to find a set of useful donors. Note that **if the stratum is well chosen, the number of useless donors should be small.**

We summarize this idea with some statistics from a sample of accommodation NR (n=788). **We have focused on the number of NR that has less than 10 candidate donors.** It has to be mentioned that, in the majority of those cases (over the 90%), this happened in medium-big strata. Thus, it seems that having only 10 candidate donors **is not a question of the stratum size** but a question of the *distance* distribution, which is right-skewed for these pathological cases and, consequently, we get a very few candidate donors. For 788 NR, **we found 12% of pathological cases with Sd2001 and only 7% with Sdlinks.**

Figure 10, 11 and 12 show the results issued from 40 simulations which consider generation of missings and imputation from both Sd2001 and Sdlinks methods (for the processus, see 3.2 and 3.3).
Remark that each row of plots has different scales. Rows of plots from figure 10 and 11 are comparable because they have the same scale. We combine two different scales because the sizes of the estimated populations are different. From the left: 3933 (Total), 3299 (NPA), WPA (522) and NR (112). We can expect that bigger populations will have less dispersed distributions.

Notice that, according to figure 10 below, assuming 1% of missingness (mean of 48 accommodation non-responders and 339 meal non-responders), we can not perceive differences between Sd2001 and Sdlinks. Both methods work correcly with the estimation of N_{Global} and with all the different subpopulations: N_{NPA}, N_{WPA} and N_{NR}.

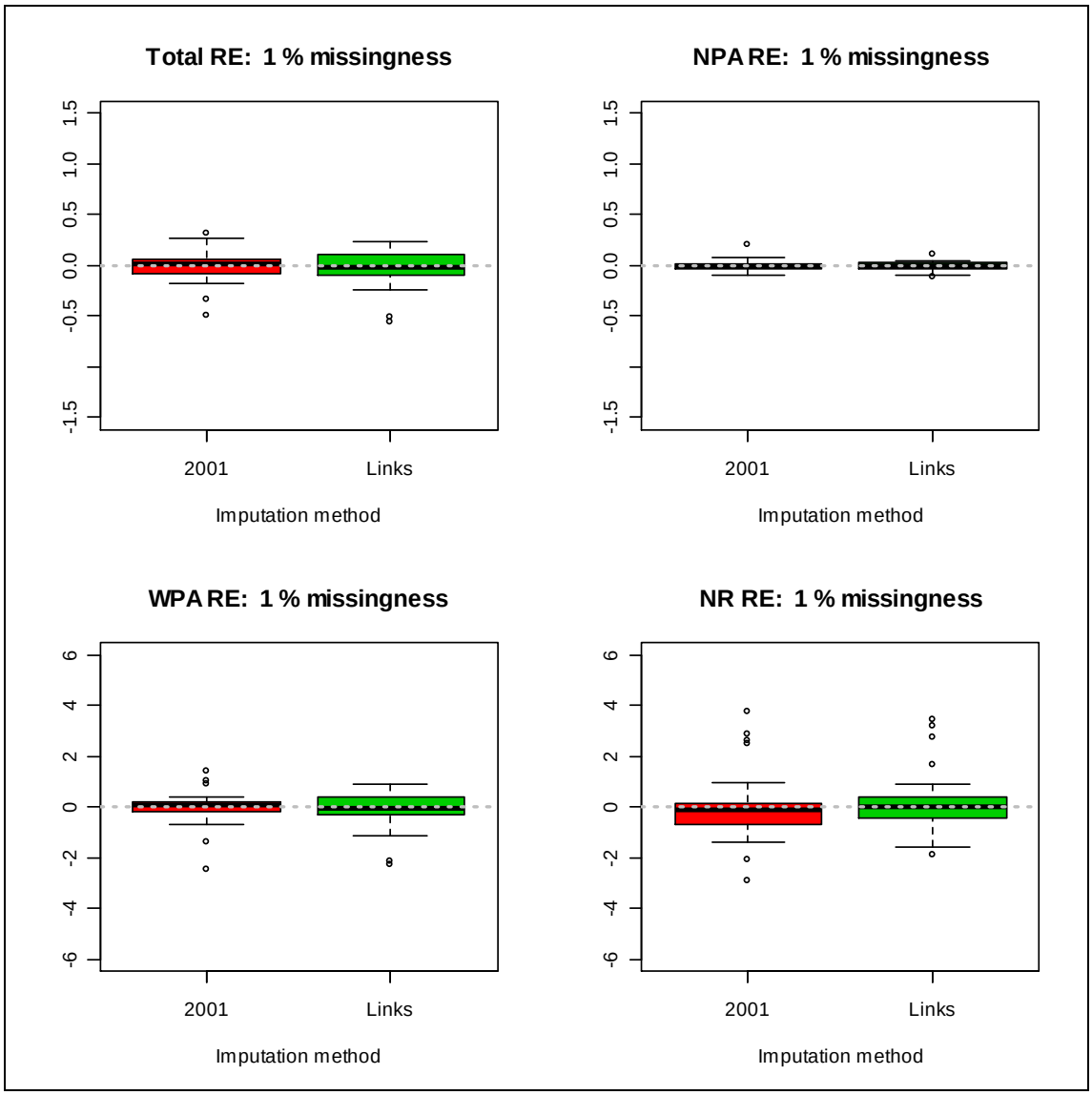

Figure 10. Relative error of 4 totals under 1% missingness: N_{Global}, N_{NPA}, N_{WPA} and N_{NR}. Sd2001 method is painted in red and Sdlinks method in green.

Notice that, according to figure 11 below, assuming 5% of missingness (mean of 195 accommodation non-responders and 1520 meal non-responders), the median of errors is not null and the variability increases. **Sdlinks remains less sensitive to the increase of the missing rate (unless for the NR population).**

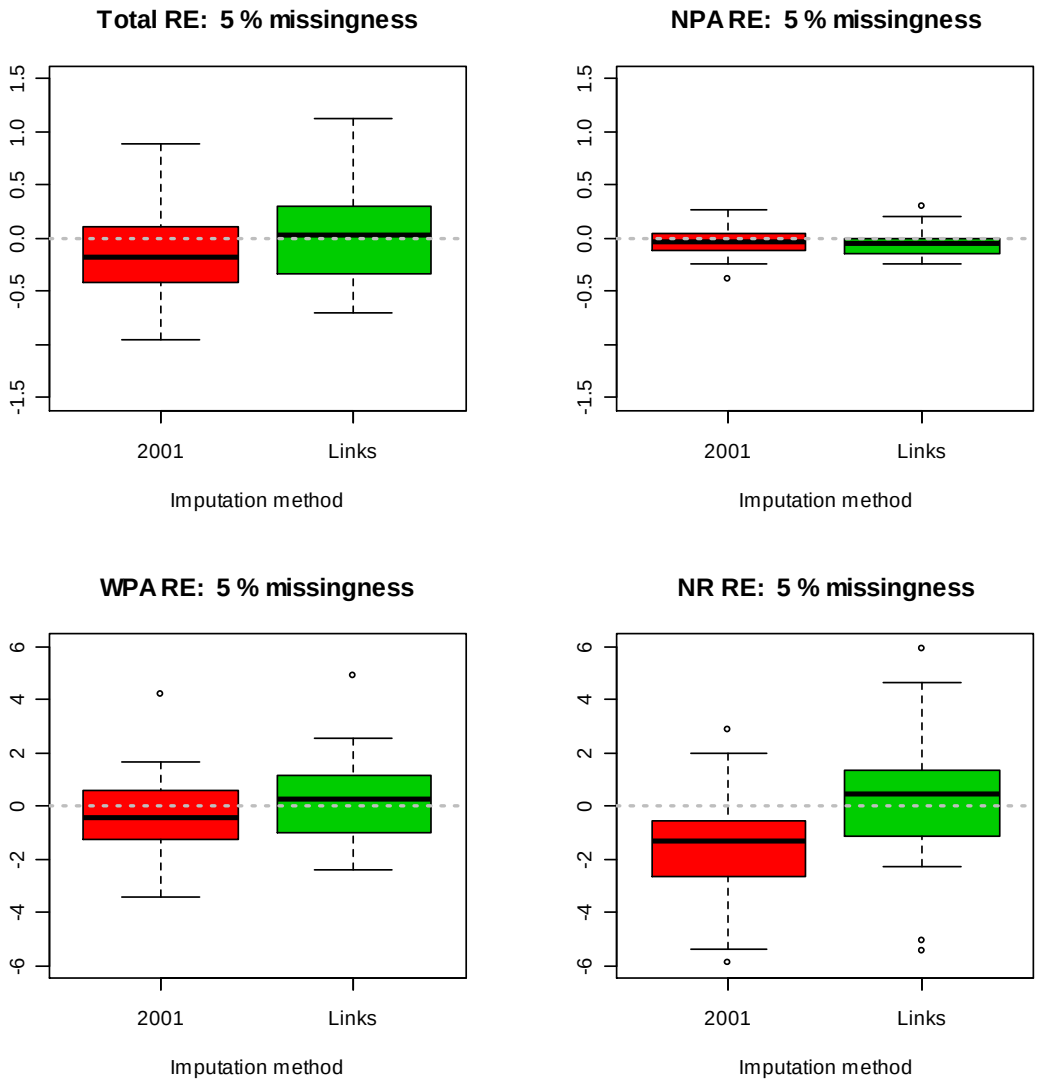

Figure 11. Relative error of 4 totals under 5% missingness: N_{Global}, N_{NPA}, N_{WPA} and N_{NR}. Sd2001 method is painted in red and Sdlinks method in green.

Notice that, according to figure 12 below, assuming 20% of missingness (mean of 740 accommodation non-responders and 3631 meal non-responders). Furthermore **22 non-responders, in average, are <u>not imputed</u> with none of the two methods.** This fact is a serious problem because if we do not find a donor for a NR the algorithm works without imputing the NR and his/her final number of links will be the same that he/she has at the beginning. That leads to underestimate the number of links related to this individual and, thus, to give him/her a higher weight in the posterior inferences.

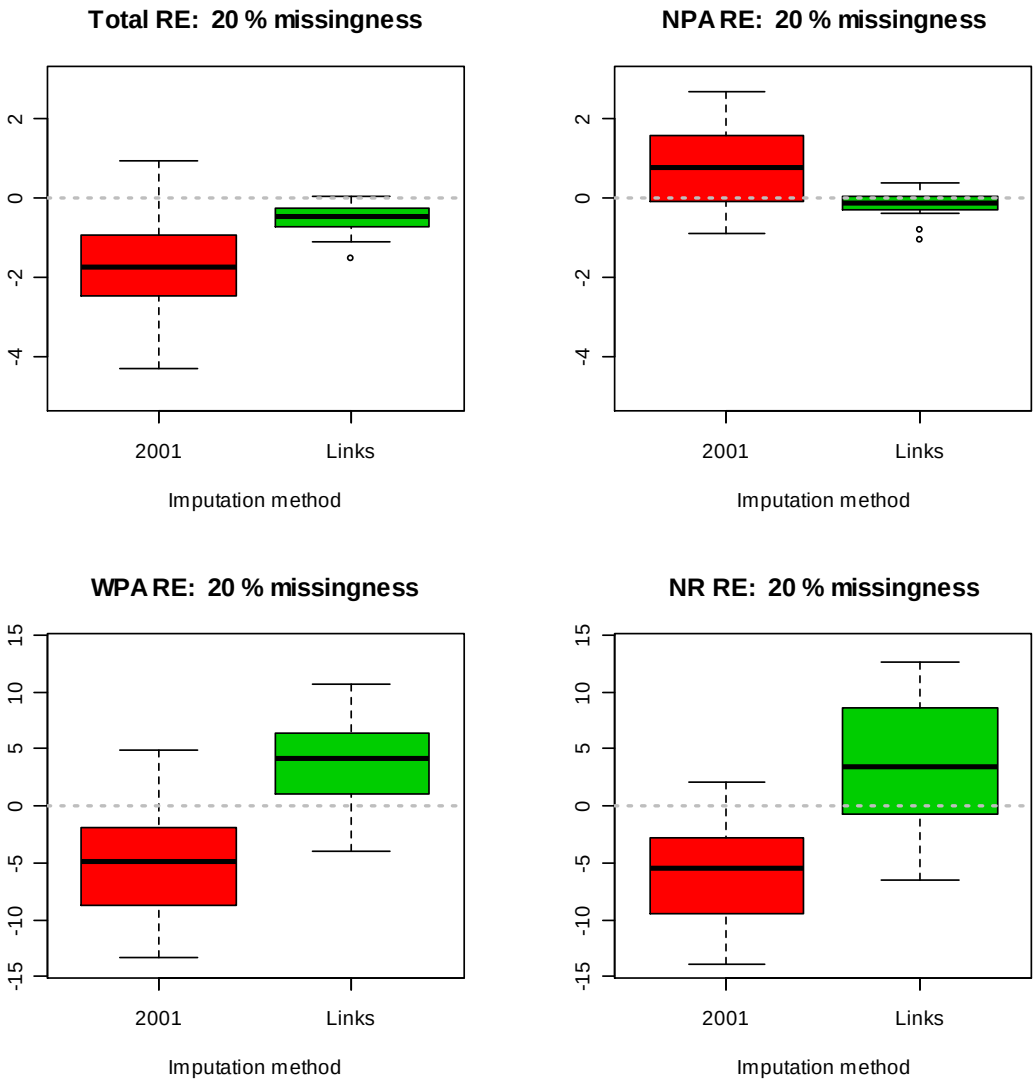

Figure 12. Relative error of 4 totals under 1% missingness: N_{Global}, N_{NPA}, N_{WPA} and N_{NR}. Sd2001 method is painted in red and Sdlinks method in green.

Comparing both methods**, the median of RE is always closer to 0 with Sdlinks**, especially for NPA subpopulation (being 0.75% for Sd2001 and -0.14% for Sdlinks).

We have applied U-Mann-Whitney test for all comparations and, with 20% of missings we find empirical evidences that **Sdlinks is statistically better than Sd2001, only in the case of NPA population that (p-value= 7.657e-05)[35].** We try to understand why Sdlinks seems to better work with populations such as NPA (n=3299):

As seen before, both methodologies differ when a non-reponser attends different cathegories of service that, in fact, are all links or all no-links. According to this, we differentiate two users:

- ***Regular user***: 11, 11, 11, 11 (4 links) or 42, 42, 42, 42 (0 links). *Sd2001* and *Sdlinks* choose the **same** set of donors.
- ***Irregular user:*** 11, 21, 21, 31 (4 links) or 42, 15, 24, 16 (0 links). *Sd2001* and *Sdlinks* choose a **different** set of donors.

According to the experts of our team, **more *«irregular»* a person is more precarious his/her situation is.** Then, it seems that the population of *sans-domicile* (NPA and NR) are the most irregular individuals. For instance:

- People with no personal accommodation (NPA) ask for all kind of services (accommodation and meal benefits) and use a great variety of benefits categories. Thus, they are very irregular in this sense and donor's methods give different results. As discussed before, in general, the reduced Semainier considered by Sdlinks works better for detecting candidate donors, specially because it has less "pathological cases" (see table 8).
- 83% of people with a personal accommodation (WPA, n=522) only combine two cathegories: eating in a meal service (cathegory 1 in the semainier, that is, a link) and eating in a friend's home (cathegory 4, that is, a no-link). Thus, they are very regular in this sense and, donors methods mostly agree one another.

To end this first discussion, you may have noticed that, assuming 20% of missingness in figure 12, both methods lead to more dispersed results. In this case, **when increasing the number of NR the number of candidate donors decreases (independently of the used method).** In

[35] Nevertheless, this conclusion is supported on a statistical test with low statistical power because of the small number of simulations (n=40).

addition, performed simulations show that **the number of missing links per person increases at the same time**:

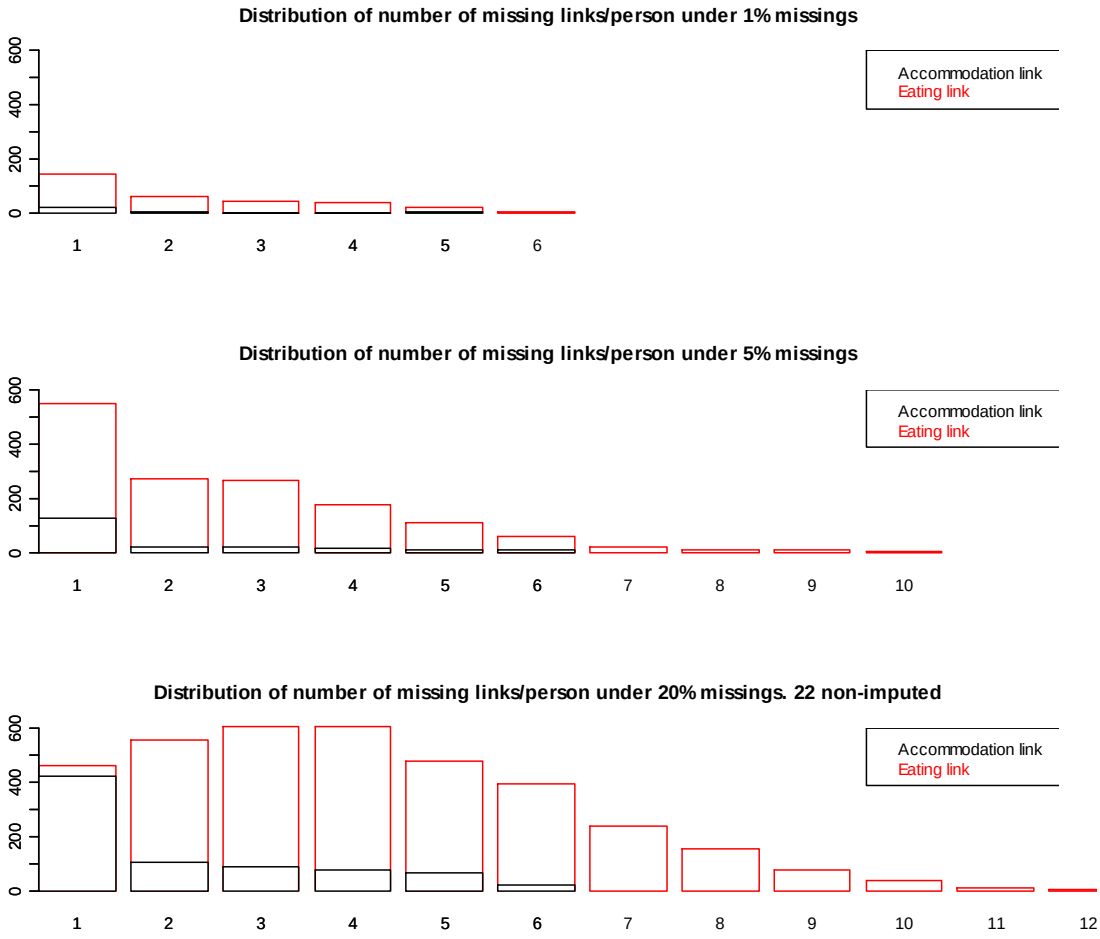

Figure 13. Distribution of the number of missing links per person under 1, 5 and 20% of missingness. Black bars refer to accommodation links (1..7) and red bars reffer to eating links (1..12).

Especially regarding the meal links, the tails of the distribution streches and the distribution becomes more symetrical.

This fact leads to another issue that has to be mentioned: **the more the number of missing links per person increases, the more the distance criterion deteriorates.** We have progressively less information about the trajectory of the non-responders that we use to find our candidate donors. **No method- neither Sd2001 nor Sdlinks- can afford this difficulty.**

4.3 Bayesian models

4.3.1 Contributions of Bayesian models

To start, table 9 contains the balance of donors methods compared to Bayesian model:

METHOD	DONORS METHODS Sd2001 and Sdlinks	BAYESIAN MODEL
OBJECTIVE	To fill the *Semainier*	To predict the weekly links
STRENGTHS	Easy to understand	Richer results
	Easy to compute	Used variables with lower rate (or 0) of missingness
	Use of direct information	Unsensitive to the incrementation of missing links/person
WEAKNESSES	Can not face problems with very few or 0 donors	Difficult to formulate
	Can not face problems with the incrementation of missing links/person	Difficult to interpretate
		Use of indirect information
CONSECUENCES OF WEAKNESSES	Use of poor donors. No imputation	More ambitious objective, more variable results
	Deformation of the distance criterion	Complexity in the statistical decisions

Table 9. Balance of the three methods: Sd2001, Sdlinks and Bayesian models.

A first and fundamental appreciation has to be done: donors methods and Bayesian models differ in their objective:

- Donors methods aim at filling the missing links in the *Semainier* of non-respondent interviewed users. They use direct information: the *Semainier* of a chosen donor. No other variables- except those from the stratification- are used.
- Bayesian model has the more ambitious objective of imputing the complete weekly links for the NR, without using any of her/his *Semainier* information. It starts from

other covariates that appear to be related to the weekly links (introduced after in the interpretation's part).

We have to notice that donors methods have two main weaknesses:

- We can deal with cases where we have less than 10 donors for a NR. This leads to **over-using donors** (not desirable) and lower probability to select a donor with a similar weekly trajectory (that happens if the strata are not very appropiate to find our donors and users are different inside a stratum).

  Remark that our strata have not been critized, because any other options have not been explored yet.

  In the worst case, we may not find any donor for a NR and the weekly count of links for that individual will be proceed from his/her known links.

  Bayesian models are able to predict the weekly links under any rate of missingness: covariates that are used have, in general, negligible percentage of missings.

- **The distance variable deteriorates with the incrementation of missing links/person** and neither the chosen strata nor the donors used will help to correctly fill the missing links of the NR.

  Bayesian models are unsensitive to the incrementation of missing links/person because this information is not used in the model prediction. For the Bayesian model, we have the same information of a NR with 1 missing link than another with 7 missing links: only their covariates have a contribution to the model.

Furthermore, Bayesian models **results are richer**. First, it is hard to learn about the weekly trajectory via donors methods. With Bayesian models, **we can interpretate which is the role of the significant covariates** (last section of the chapter), **justify and propose new strata** of people with similar weekly trajectories. Also, we have a great deal of tools to validate our models and detect in which circumstances they are not working well enough. Finally, a particularity of Bayesian models is that we are able to **use the knowledge issued form Sd2001 (kept as a posterior distribution of the parameters of our models)** to answer questions that

will arise about the population of Sd2012. On the contrary, frequentist models would ignore the information brought by Sd2001, as if it did not exist.

Despite all these positive points, Bayesian model is computationnally more sofisticated. We have also to notice that the variable "weekly links" is difficult to modelize for three reasons (see 3.3.3.1 Formulation of the Bayesian model):

a) The distribution of the response variable: it presents some picks that suggest us that a bimodal population is underlying.
b) The lack of prior knowledge about this topic: nobody has fitted yet a model for this survey to explain the weekly links.
c) At least, two different models have to be fitted: one for accommodation links (0..7) and another for meal links (0..14). Notice that this differenciation is also considered in the donors methods as they perform the imputation in two stages: first the accommodation links and second the meal links. We adopt here this approach and we go further also differentiating between lunch links (0..7) and dinner links (0..7).

4.3.2 Validation of Bayesian models

We present here the results concerning Bayesian models, according to the formulation introduced in the Methodology chapter:
First, we **justify the choice** of a 2-level Bayesian model. We show the expected marginal posterior predictive in each case, that is, the expected number of weekly links for a set of 362 NR estimated by the Bayesian model. For each NR we have the distribution of the parameter: **probability of having a link** ($\theta_{1i}, \dots, \theta_{500i}, NR = i$). We simulate **10 replicates of the response variable for each NR**, with a the parameter θ chosen randomly from its distribution.

In figure 14 and 15, in black, you have the real histogram of the total number of links (y) and in red the 10 histograms overlapped corresponding to the 10 repliques of the *y*, according to our Bayesian models. **The closer are the red histograms to the black histograms, the better is the Bayesian model.**

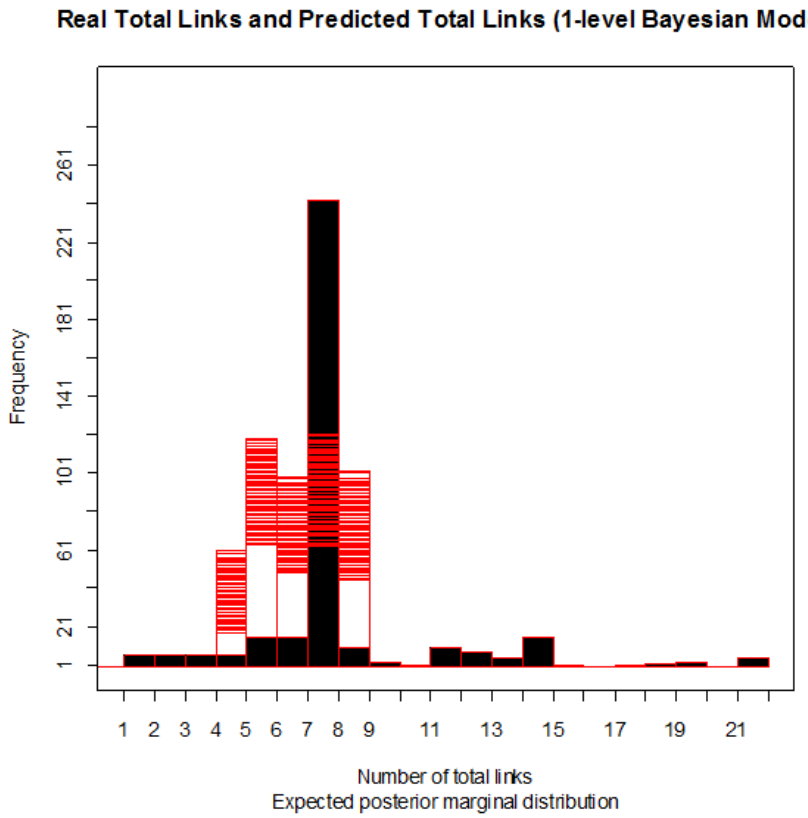

Figure 14. Real total links and predicted total links with the **1-level Bayesian Model.**

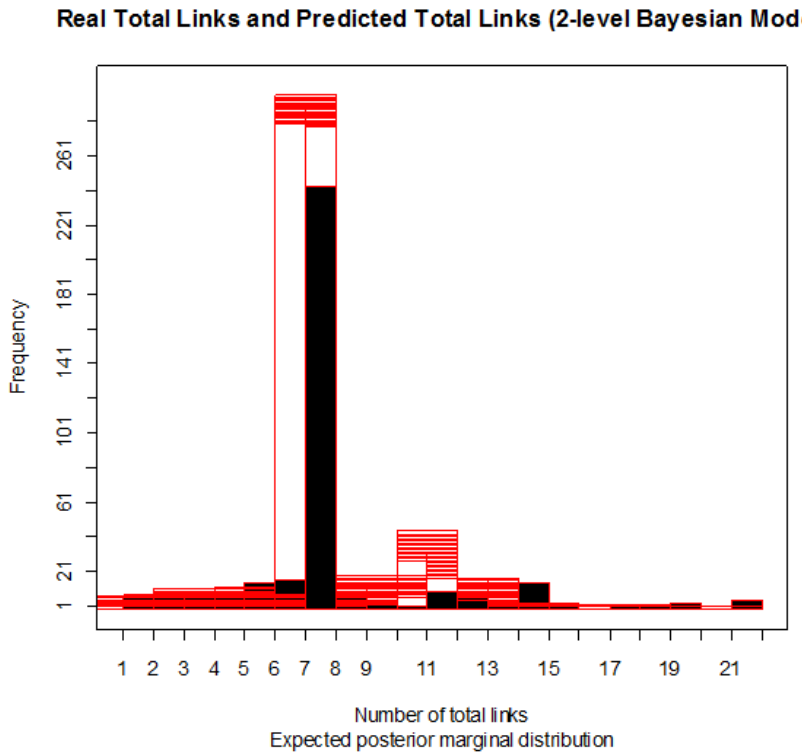

Figure 15. Real total links and predicted total links with the **2-level Bayesian Model.**

You can see that the **1-level Bayesian model is highly influenced by the mode of the real distribution** (7 links) and this **disrupts the model** which limits the range of predictions to be between 4 and 9 links, underestimating the dispersion of the variable. The 2-level Bayesian Model also takes into account the high pick corresponding to 7 links but, at the same time, **captures the variability of the real variable**, enlarging the range from 0 to 21, which is more realistic. However, it imputes more frequently than expected according to reality (black) 6, 10 and 11 links.

Furthermore, **362 individuals** have not been used in the estimation of the Bayesian models allowing for a **cross-validation** with an external set of individuals (**test-subset**). This kind of validation offers worst results than other kinds of validations of the model.

Therefore, from now until the end we focus on the **2-level Bayesian model which implies a total of 6 models:** a 2-level for sleeping links, a 2-level for lunch links and a 2-level for dinner links.

Now, we present another indicator of the **goodness of prediction** of our model. We only focus on **the percentatge of people underestimated and overestimated in the first level.** We show these indicators for people used in the model and for the 362 test-subset. It is evaluated only in the first level -a binary response- for lunch, dinner and accommodation links. For accommodation links, two candidate models are considered.

Variable: Number of <u>lunch</u> links		
Selected model : DIC=2161		
Model-subset		
Real\Model	0 links	1..7 links
0 links	544	**152**
7 links	**289 (9,3%)**	2132
Test-subset		
Real\Model	0 links	1..7 links
0 links	257	**14**
7 links	**41 (11%)**	50

Variable: Number of <u>dinner</u> links		
Selected model : DIC=1384		
Model-subset		
Real\Model	0 links	1..7 links
0 links	2739	**45**
7 links	**189 (6%)**	144
Test-subset		
Real\Model	0 links	1..7 links
0 links	303	**3**
7 links	**36 (9,9%)**	20

Variable: Number of accommodation links						
MODEL H1.1 (DIC=1740,73)				**MODEL H2.1 (DIC=1418,11)**		
Model-subset				**Model-subset**		
Real\Model	0..6 links	7 links		Real\Model	0..6 links	7 links
0 links	524	**194 (6,3%)**		0 links	153	**565 (18%)**
7 links	**118**	2281		7 links	**43**	2356
Test-subset				**Test-subset**		
Real\Model	0..6 links	7 links		Real\Model	0..6 links	7 links
0 links	57	**26 (7%)**		0 links	22	**61 (17%)**
7 links	**13**	266		7 links	**5**	274

Table 10-11. Concordance tables to measure the goodness of prediction of each lunch, dinner and accommodation models in level 1. They are issue of 1000 iterations of each model.

Concerning the two models of the first level for the accommodation links, it is quite confusing that model H2.1 has less DIC (it is better according to this criterion) but, at the same time, has worse predictions. The difference between them is that variable *Center of the interview* is introduced with 7 categories (H2.2) or with an agroupation of 3 categories (H2.1). **According to the predictions, the simplest model H2.1 is preferible[36].** Focusing on this model, notice that it tends **to overestimate the number of people with 7 links**, because "7 links" is the mode of the distribution for accommodation links. Despite this, results are not much worse when predicting in the model-subset or the test-subset; thus, **it seems that model for accommodation is quite robust.**

Lunch and dinner models underestimate the number of "1..7 links" because in their case, the opposite of accommodation links, the mode is "0 links". Lunch model has the higher underestimation proportion.

The consequence of using a **two-level model is that they are not independent estimations**:

- Unfortunately, if we overestimate (accommodation links) or we underestimate (lunch, dinner links), the corresponding individuals will be badly classified and will be imputed as 7 in the accommodation case and a 0 in the lunch and dinner case. They will have

[36] See in the Appendixes*: 9.6 Comparing candidate models for the first level of the accommodation links.*

no chances to be correctly predicted. **That is why we wanted to show these concordance tables for the first level,** as it is important that models do not commit this type of error as it will be irreparable.

- However, for individuals that were wrongly classified into "0..6" accommodation links or "1..7" in meal links, in the second level they will have another chance to be predicted correctly.

Now, regarding the second level of each model, it has to be mentioned that, as the range of values to predict increases to 8 (0,..,7) it is harder to predict correctly. For instance, we show the concordance table of the accommodation model. Remark that the estimation has been done with a subset of people who were predicted "0..6 accommodation links" in the first level (a mean of n=641).

Real\Model	0	1	2	3	4	5	6	7
0	**301**	47	5	5	0	1	0	0
1	0	**1**	1	0	0	0	3	0
2	0	0	**1**	0	0	0	2	0
3	0	2	3	**0**	0	1	11	0
4	1	3	0	0	**0**	0	35	0
5	0	3	3	1	0	**3**	42	0
6	0	3	2	0	0	5	**43**	0
7	1	6	4	4	19	24	56	**4**

Table 12. Concordance table accommodation model in **levels 2**. They are issue of 500 iterations of the model, each time with a random number of individuals classified in level 1.

You can see the pink diagonal of the matrice; that is the number of good predictions of the model. This model predicts the 55% of individuals that enter at each iteration. Remark that there are 4 people that were badly predicted in level 1 but correctly predicted in level 2.

We focus now on the **relative error computed from the 500 imputations of the 362 individuals of the test-subset**, to finally compare it with the results of donors methods. First, Figure 16 presents the underlying idea for comparing how each model –accommodation, lunch and dinner- works in terms of RE for each of our target parameters (our priorisation in this thesis). In general, it is hard to impute the NR population because we have only a small subset (n=12 in the test-subset) with a quite unstable behaviour. Notice that this population has always the highest variability.

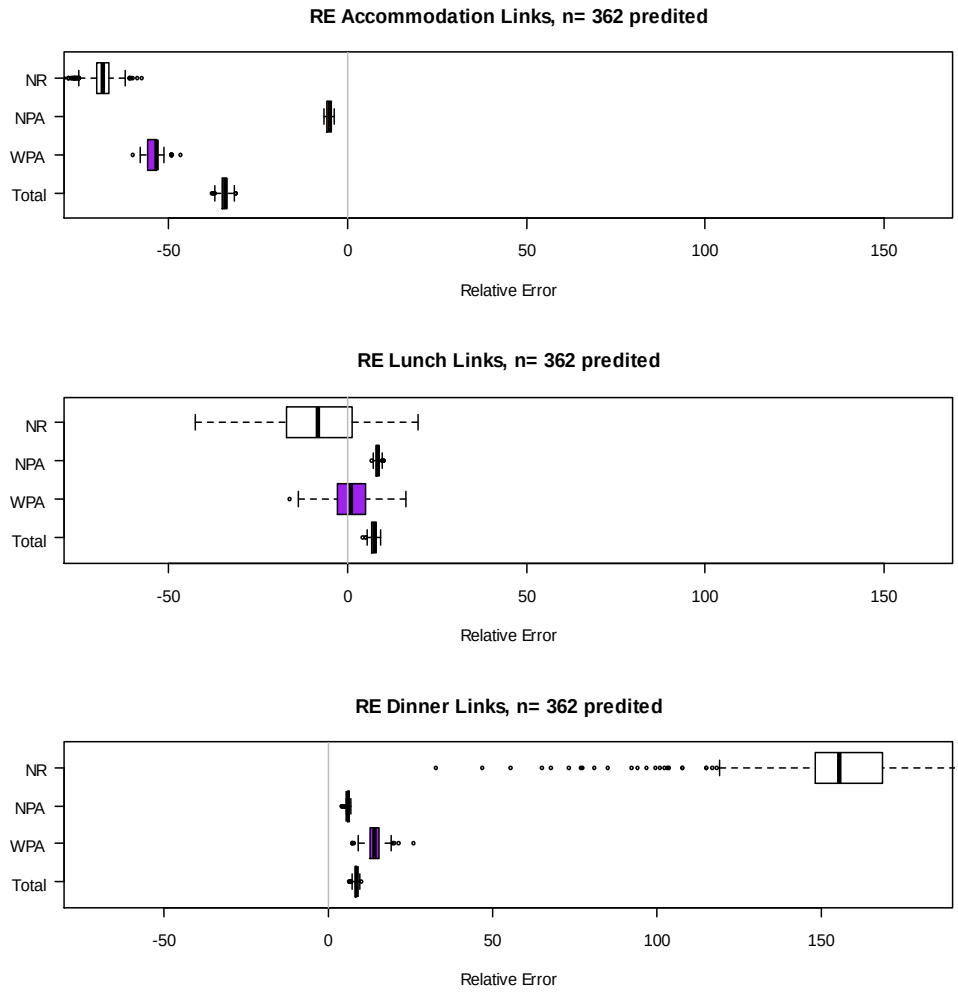

Figure 16. Distributions of the relative errors after 500 estimations of 4 totals : N_{NR}, $N_{NPA,}$ N_{WPA} and the global number (Total). In the expression of the *Weight share estimator,* we only take into account the number of weekly accommodation links (first plot), lunch links (second plot) and dinner links (third link).

Globally, **accommodation model has the further RE from 0 when estimating the N_{Global}** . We think that the weakness of this model is particullaly related with population of WPA (n=49 in the test-subset) which is also very badly estimated. Despite this, NPA (n=301) is better predicted.

Remark that distribution of RE from accommodation models are under 0 because we have said that the number of weekly links are overestimated and, according to the Weight Share estimator, they have more weight that in reality. The opposite happens with models for lunch and, specially, dinner links for the same reason.

Finally, we can see a clear difference between levels of RE: **population of NPA are usually better imputed than others**. The variance of the distribution of RE depends basically on the size of the population. That is clear. What it is not clear is **why we do not obtain similar medians of RE for the different populations,** specially in the case of accommodation models. It seems that, altough the variable *subpopulation* appears in all of our final models (see interpretation in 4.3.4), **there is implicit information related with *subpopulation* that affects the model prediction** and that we have not entered yet in our model.

4.3.3 Donors methods vs Bayesian model

Figure 17 shows the comparison of the two familly of methods: donors methods and Bayesian model, assuming 1, 5 and 20% of missing rate.

Distributions of RE from Sd2001 are in red, from Sdlinks in green and from Bayesian models in blue. **The first surprising aspect is that Bayesian models imply the highest variability under any missigness scenario, which is not at all desirable.** Two reasons can be mentioned:

a) **The 2-levels are not independent** : when we perform more precise predictions in level 2, they depend on individuals coming from level 1 at each iteration. This generates, inevitably, more variability in the imputation. 7% of the individuals of level 2 got different imputations over the 500 iterations (10 NR, 4 NPA and 12 WPA). This happens because this individuals are on the vergeof one or another imputation and

they oscillate all the time. This also implies another **source of variability** (see figure 9 in methodology for the formulation of the Bayesian model).

b) **The statistical object used with donors methods and Bayesian model is different** (see table 9). Once the first have a vector of 21 observations to fill when there is a missing (but they never have a vector of 21 missings), the Bayesian model starts with nothing and has to arrive to a number between 0 and 21. We see that results will be more dispersed in the Bayesian model because the initial state is more complicated.

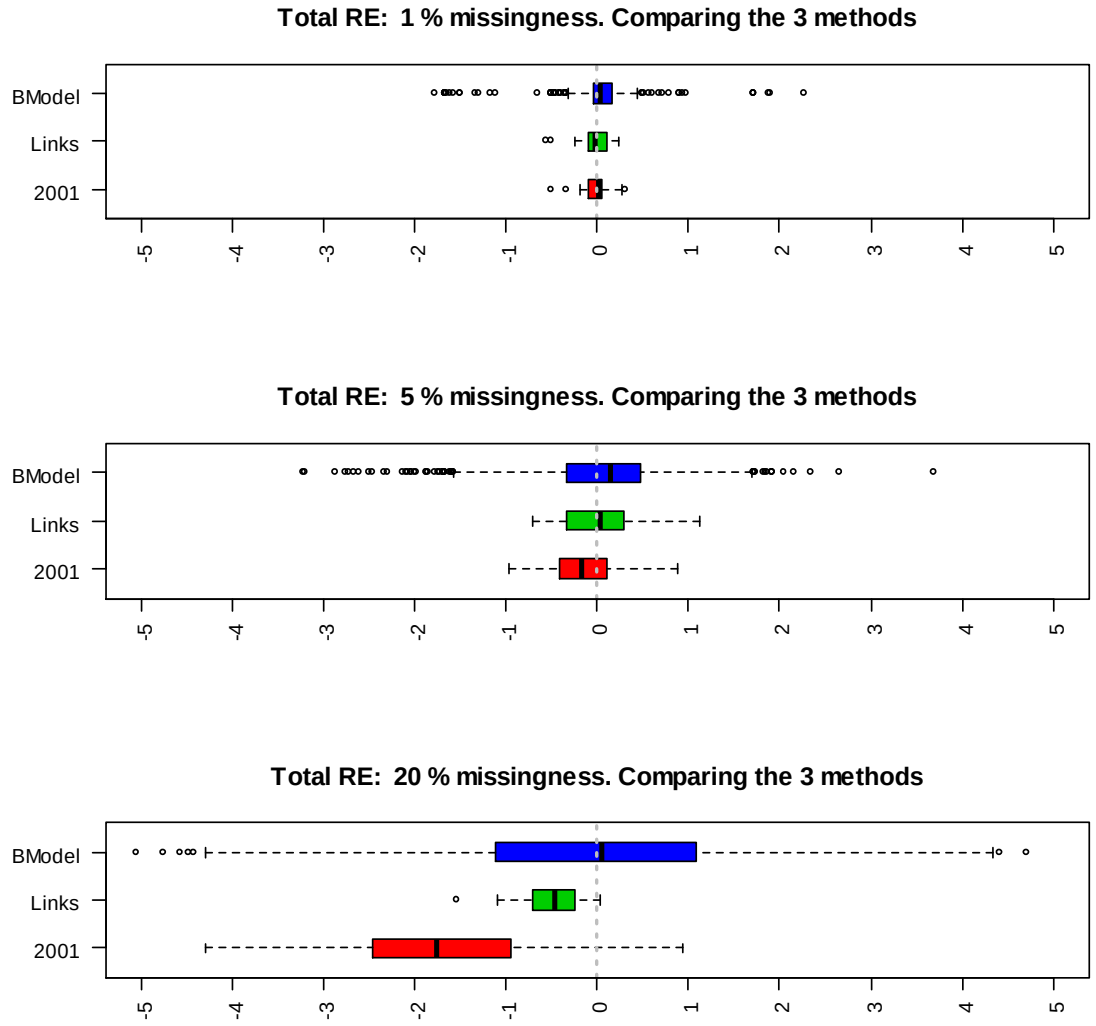

Figure 17. Distributions of the relative errors of 4 totals : N_{NR}, $N_{NPA,}$ N_{WPA} and the global number (Total). Comparation of the Sd2001, Sdlinks and Bayesian model methods, under 1, 5 and 20% of missing rates.

Apart from that, we notice that, **assuming 20% of missing, the median of RE of the Bayesian model is closer than donors methods (even Sdlinks, that we saw that, in general, worked better than Sd2001).** This fact corroborates our hypothesis introduced in table 9, that **this method remains, in median, unsensitive to an increase in the missings rate** and it has be taken into account.

Therefore, although we introduced non desirable sources of variability with Bayesian models, we are capable to get, in median, closer RE to 0 **assuming 20% of missing** than with donors methods.

4.3.4 Interpretation of Bayesian models. New strates for Sd2012

We interpretate the 6 regression models that have been estimated. Remark that the objective of this section is not to understand in detail the role of all the covariates (we do not have enough knowledge) but **to obtain a general idea of some variables that constantly appear in the different models**. Moreover, the tables presented below summarize the information about the most common variables. These variables will be proposed at the end of the section as new strates for Sd2012. Notice that a minimum of one variable will be proposed as a stratum for each type of link: accommodation, lunch or dinner (in 2001 we had a unique stratum for all kind of links).

Each model has a table indicating the level (1 or 2), the number of individuals used in the estimation and the number of covariates (p). Note that in level 2, as we have estimated the model with 500 different sets of individuals returned by level 1, we summarize this number with its mean.

Each table has 9 columns:

- ✓ The name of the variable (VARIABLE), its cathegory (CATEGORY) as all covariates are categorical.
- ✓ 3 percentils of the posterior distribution of the OR associated with the variable (P5, P50, P95). Also, a probability[37] to show the significance (Signif. Proba).

[37] See Methodology 5.3.3.3.2 to know how this probability is computed.

- ✓ The probability of the event (Proba “Links=0”, “a link” in lunch model level 1 and 2, respectively) is computed via considering that the individual has the same characteristics that the reference individual. But, he has changed of cathegory in the variable that we are interpreting. The probability of the first row is the probability associated to the reference individual. Thus, for instance, in the first model, the reference individual has a probability of 0,86 to have 0 lunch links. An individual with the same profile except that he belongs to the NPA subpopulation (reference is WPA) has a probability of 0,95 to have 0 lunch links.
- ✓ The expected cathegory (in level 1) or number of links (in level 2) according to its probability of event (EXPECTED). Finally, each variable has a position relative to the other variables of the model: the higher is its effect on the response, the higher is its ranking. Ranking varies from 1 to 3. Thus, only the most influenciable variables of each model have been included in these tables that aspire to simplify the interpretation of the models.

Concerning the rows:

- ✓ The first row, highlighted in grey, is always the reference profile. We are interested in the probability of the event and its expected value.
- ✓ Each variable has as many rows categories minus one (the reference category).
- ✓ Rows in yellow represent the first three variables with the highest contribution to increase the odds of the response (OR>1).
- ✓ Rows in blue represent the first three variables with the highest contribution to decrease the odds of the response (OR<1).
- ✓ Rows in white are categories from covariates that are not significant in the model and that are prone to be agrouped with other categories in future versions of these models.

Aiming at finding new strata or corroborrate the validity of ancient strata (*Center of the interview x Gender*), we fix some **guidelines to select these variables**. Remember that, to find strata, we look for variables such as the individuals presenting the same category have similar number of weekly links and the individuals presenting different categories have different number of weekly links.

Our guidelines are:

- All the selected variables must be in the ranking position (highlighted).
- The selected variables have to appear in both levels of a model.
- The selected variables must have , in one level, at least one categoriescategory with OR>1 (yellow) and another with OR<1 (blue). This means that its categories well discriminate the number of weekly links.
- If we have a set of binary variables in the ranking (yellow and blue) and we want to choose only one, we decide according to the probability of event for the reference individual. If it is high (>0.5), we look for a variable within the blue ranking to get more contrast. If it is low (<0.5), we look for a variable within the yelow ranking.
- If the chosen variable has some categories that are not significant categories, a reagroupation of categories has to be performed before proposing it as a new stratum for Sd2012.
- If we find more than one variable for a model and we want to cross them, we verify that there are no empty intersections and it is preferable that variables are very not associated. With them, we work with strata of aproximatetly the same size.
- Unluckily, there are some variables of the ranking that can not be selected to build strata because the real non-respondent of Sd2001[38] have some missings. They are: *eating, contact with a person, basket of food, working, solidary shop* and *week attendancy.*

Our decisions about strata are based, in general, on these premises.

Finally, the **reference individual** has been chosen having compatible features and, specially, with categories with a high count. He has the following profile:
Man, WPA (with personal accommodation), born in France, without children, single and interviewed in a center for long duration sleeping (they can stay more than 15 days), without contact with another person by phone, not working at the moment, who has never slept in the street, who has not gone to the doctor in the previous 12 months, less than 37, not receiving a basket of food, eating during all the week and declaring that the night of the interview, he will sleep in the same place than the previous night.

[38] See *Methodology. 5.1 Step 1: Study of the mechanism of the link non-response in SD2001.*

Here you have **the lunch models**:

LUNCH LINKS: Level 1 (n=3117), p=12 covariates								
VARIABLE	**CATHEGORY**	**P5**	**P50**	**P95**	**Signif. Proba**	**Proba Links=0**	**EXPECTED**	**RANKING**
	Reference	3,82	6,31	11,58	---	0,86	0	---
Subpopulation	NPA	2,19	3,05	4,18	0,00	0,95	0	1
Children	Children	1,59	2,15	2,94	0,00	0,93	0	2
Sex	Woman	1,46	1,85	2,38	0,00	0,92	0	3
Center of the Interview	Outdoor meal	0,03	0,05	0,07	0,00	0,22	0..7	1
	Indoor dinner	0,12	0,17	0,24	0,00	0,52	0	1
	Indoor lunch	0,14	0,18	0,25	0,00	0,54	0	1
	Urgence C.	0,28	0,46	0,72	0,00	0,74	0	1
Subpopulation	Non-roof	0,86	1,53	2,69	0,11	0,91	0	
Center of the interview	C. Rooms	0,58	1,00	1,76	0,50	0,86	0	
	Dispersed C	0,65	0,90	1,22	0,26	0,85	0	

Figure 18.Lunch links model. Level 1

In this level 1 for lunch model, it is interesting that categories of the *Center of the interview* have 90% credibilities intervals that are not overlapped. Thus, the type of center is important to differentiate the individuals having 0 lunch links from the others. Remark that all the categories have an OR<1 (blue); thus they decrease the probability of having "0 lunch links", specially when the individual is interviewed in a meal center (outdoor or indoor), which seems coherent. Note that *center with rooms* and *dispersed center* are not significant. They probably can be grouped with the reference category: *long duration centre.*

LUNCH LINKS: Level 2 (mean of n=559), p=15 covariates								
VARIABLE	**CATHEGORY**	**P5**	**P50**	**P95**	**proba**	**Proba a link**	**EXPECTED**	**RANKING**
	Reference	0,23	0,34	0,47	---	0,25	2	---
Eat	Not always	1,56	1,77	2,04	0,00	0,38	3	1
Working	Working	1,38	1,62	1,92	0,00	0,35	2	2
C. Interview	For lunch	1,29	1,58	1,90	0,00	0,35	2	3
Where Sleep tonight	Different place	0,17	0,24	0,36	0,00	0,08	0	1
C. Interview	For dinner	0,52	0,66	0,82	0,00	0,18	1	2
Subpopulation	NPA, NR	0,59	0,67	0,76	0,00	0,18	1	3

Figure 19.Lunch links model. Level 2

From level 1 to level 2, some variables are groupep because, in the second level, we work with less individuals (and the number is random) and we wanted to garantee that there were no empty intersection between covariates. It is the case of variable like subpopulation (non-roof people is a small group) and center of the interview (initially, there were 8 categories).

To summarize, in both levels, **the center of the interview appears in the ranking of variables**. In this second level, centers for lunch are separated from centers for dinner in the ranking. For lunch, OR>1; thus, increment of the odds of having a lunch link, which is also coherent. When the center is for dinner, the variation is opposite. Also, variable *Where to sleep tonight* indicates that people thinking that they have slept than the day before (thus, irregular people) decrease their probability of having lunch links, maybe because they move from one place to another and they find different ways of eating (without using a individual benefit or a link).

According to our previous guidelines for strata, we have chosen three different strata with variables: Where to sleep tonight (SLEEP), Subpopulation (2 categories: WPA, NPA+NR) and Center of the Interview (4 categories). Notice that the first two variables reffer to

accommodation information. The following tables indicate the number of cases of each stratum.

STRATUM 1: Subpopulation2 x SLEEP				**STRATUM 2: Subpopulation2 x Cinterview5**				
	Same place	Different Place	Unpredicted		Accom (3)	Indoor L	Indoor D	Outdoor
NPA+NR	3363	108	53	NPA+NR	2471	326	183	77
WPA	510	31	18	WPA	79	280	73	115

STRATUM 3: Subpopulation2 x SLEEP x Cinterview5				
Same Place	Accom (3)	Indoor L	Indoor D	Outdoor
NPA+NR	2421	284	163	62
WPA	71	266	66	106
Different Place	Accom (3)	Indoor L	Indoor D	Outdoor
NPA+NR	41	15	15	13
WPA	5	4	5	6
Unpredicted	Accom (3)	Indoor L	Indoor D	Outdoor
NPA+NR	9	27	5	2
WPA	3	10	2	3

Figure 20. Proposed strata for lunch links imputation.

We can already say that the third proposition for strata will not work if the missing rate increases in Sd2012. It must be taken carefully as we have intersections with a very few number of individuals.

The following tables present the **dinner models**:

DINNER LINKS : Level 1 (n=3117) , p= 7 covariates								
VARIABLE	**CATHEGORY**	**P5**	**P50**	**P95**	**Signif. Proba**	**Proba LinkS=0**	**EXPECTED**	**RANKING**
	Reference	4,12	6,04	9,10	---	0,86	0	---
Subpopulation	NPA	2,84	3,88	5,24	0,00	0,96	0	1
Sex	Woman	2,05	2,86	3,95	0,00	0,95	0	2
Couple	Couple	1,49	2,46	4,15	0,00	0,94	0	3
C. Interview	For lunch	0,04	0,05	0,07	0,00	0,24	0..7	1
Help	Soldary shop	0,32	0,42	0,56	0,00	0,72	0	2
Subpopulation	NR	0,29	0,51	0,92	0,03	0,76	0	3
C. Interview	For dinner	0,51	0,73	1,05	0,08	0,82	0	

Figure 21. Dinner links model. Level 1

Notice that women as well as people that have a couple have more probability to have 0 dinner links.

DINNER LINKS: Level 2 (mean of n=189), p=10 covariates								
VARIABLE	**CATHEGORY**	**P5**	**P50**	**P95**	**Signif. Proba**	**Proba a Link**	**EXPECTED**	**RANKING**
	Reference	0,95	1,28	1,75	---	0,56	4	---
Children	Children	1,77	3,98	9,87	0,01	0,84	6	1
Working	Working	1,37	1,81	2,23	0,00	0,70	5	2
Drink	Sometimes/rarely	1,33	1,65	2,03	0,00	0,68	5	3
Weekly Attendance	Sleep 1-3 nights	0,02	0,05	0,12	0,00	0,06	0	1
Couple	Couple	0,11	0,21	0,35	0,00	0,21	1	2
Doctor	Doctor	0,39	0,50	0,64	0,00	0,39	3	3

Figure 22. Dinner links model. Level 2

You can see that people who have children, who are currently working or who rarely drink have an OR>1 to have a dinner link. People who have a couple have not tendency to ask for dinner benefits.

Thus, the chosen strata according to our guidelines are:

STRATUM 1: Subpopulation x Couple			**STRATUM 2: Cinterview2 x Couple**			**STRATUM 3: Subpopulation x Cinterview2**		
	Couple	Single		Couple	Single		Accom+Dinner	Lunch
NPA	491	2896	Accom+Dinner	501	2957	NPA	3039	260
WPA	53	506	Lunch	46	480	WPA	242	280
NR	3	135				NR	65	47

Figure 23. Strata for dinner links models.

Finally, here you have the **accommodation links model**:

ACCOMMODATION LINKS : Level 1 (n=3117), p=11 covariates								
VARIABLE	**CATHEGORY**	**P5**	**P50**	**P95**	**Signif. Proba**	**Proba Links=7**	**EXPECTED**	**RANKING**
	Reference	0,47	0,70	1,10	---	0,41	0..7	---
Subpopulation	NPA, NR	21,67	29,27	39,41	0,00	0,95	All	**1**
Children	Children	1,56	2,18	2,97	0,00	0,61	All	**2**
Couple	Couple	1,26	1,86	2,88	0,00	0,57	All	**3**
Weekly attendance: sleeping	Sleep 1-3 nights	0,01	0,03	0,06	0,00	0,02	0..7	**1**
	Sleep >4 nights	0,02	0,03	0,04	0,00	0,02	0..7	**2**
	Occasionnaly	0,04	0,13	0,35	0,00	0,08	0..7	**3**

Figure 24. Accommodation links model. Level 1.

Remember that it is not possible to use the variable *Weekly attendance* because, actually, we have 13 missings from 24 non-responders of accommodation.

ACCOMMODATION LINKS : Level 2 (mean of n=641), p=10 covariates								
VARIABLE	**CATHEGORY**	**P5**	**P50**	**P95**	**Signif. Proba**	**Proba A Link**	**EXPECTED**	**RANKING**
	Reference	1,05	1,48	2,14	---	0,60	4	---
Subpopulation	NPA, NR	18,99	26,34	37,05	0,00	0,98	7	**1**
Couple	Couple	1,16	1,62	2,26	0,01	0,71	5	**2**
Contact	Have contact	1,24	1,48	1,76	0,00	0,69	5	**3**
Center of the interview	For dinner	0,02	0,03	0,03	0,00	0,04	0	**1**
	For lunch	0,04	0,05	0,06	0,00	0,06	0	**2**
Weekly attendance	Occasionnally	0,09	0,14	0,20	0,00	0,17	1	**3**

Figure 25. Accommodation links model. Level 2.

In this case, **to have a couple increases in both levels the odds of having "7 links" or a link one day.** Moreover, ***subpopulation* variable has an enormous OR**: passing from being WPA to NPA or NR increases the odds of having a link 26,34 times (in median).

According to our guidelines, we propose **the same strata than when we impute the dinner links.** Notice that **dinner attendance is related with accommodation attendance**, which is coherent because they are very close in time.

4.3.5 Future perspectives for Bayesian models. Imputation of non-francophones.

One of the new objectives of Sd2012 is to enlarge the population couverture to the non-francophons users of help services. A self-administered questionnaires has been designed in 11 different languages. It has to be mentionned that **distributing self-administered questionnaires involves significant risks as compared with a questionnaire face to face**, which can lead to several types of errors costly in terms of quality of the collected information: errors in understanding the issue, data entry errors, higher non-response,etc.

In May of 2011, the *sans-domicile* team carried out a test concerning this questionnary. Thanks to it, problems in filling the reduced *Semainier were detected*:

- The respondent does not understand that the period asked about his attendance is a week. Often, the counts exceeded the maximum services links (28 in SD2012). How can we treat this information?
- The respondent has selected an option without specifying the number of times he has used the service. Thus, we know that he has used at least once this service but there is no further information to clarify the exact attendance.
- The respondent has marked out and counted some services and, at the same time, he has left without filling (even if it is specifically requested to state 0 if it has never used this service during a week). Can we assume that this person has not used this type of benefit? Or, rather that he/she does not remember and this empty cell represents an unknown attendance?
- The respondent has not completed any of the questions related to this part. Thus, it is impossible to assign him/her any weight.

A total of 98 non-francophone-users received the questionnary and 30 of them did not fill completly the reduced *Semainier* (**30% of NR**). If the rate of NR would be as highas in this preliminar study, the Bayesian model could be applied. A new Bayesian model would be estimated starting from an informative prior considering the knowdlegde adquired from Sd2001. Prior information (posterior distribution of the parametres for the Bayesian model Sd2001) and information of the self-administrated questionnaries (Y=y) would be combined through the Bayes Theorem and we would have an updated Bayesian model richer in knowdledge and more accurate concernong the imputations.

The schema is presented hereafter:

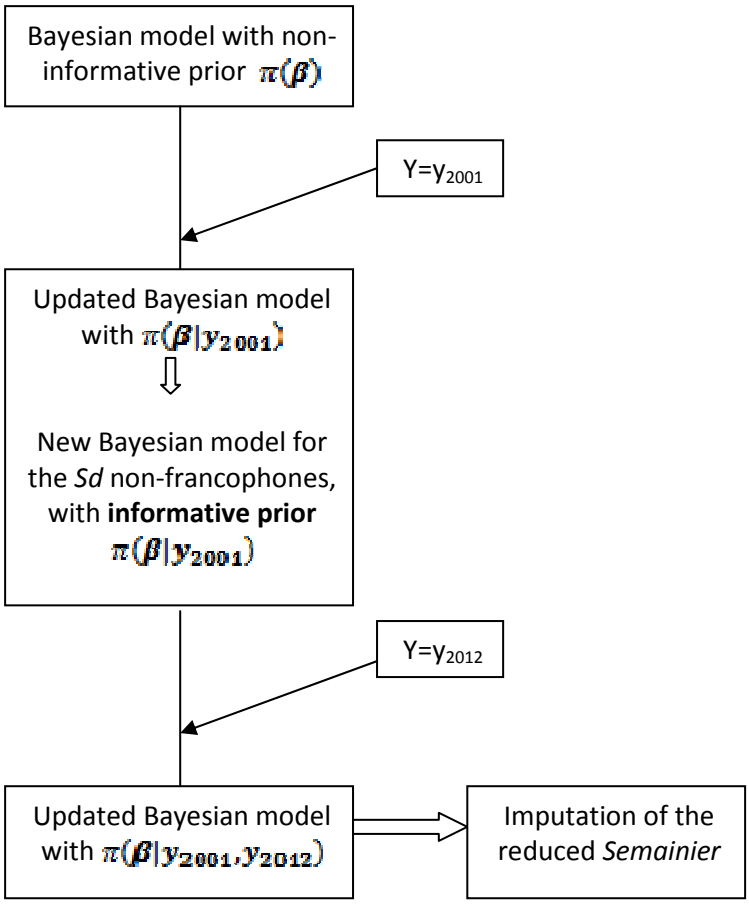

Figure 26. Future perspectives for the Bayesian model in the Sd non-francophones.

CONCLUSIONS AND PERSPECTIVES

- Concerning donors methods, the application of the distance function between a non-responder (NR) and a donor sometimes drastically reduces the number of candidate donors. This happens in the 12% of the cases for Sd2001 (reference method) and 9% for Sdlinks.

- Although Sdlinks works with a simplified Semainier (apparently loosing information), candidate donors are chosen according to their number of links which is the necessary information to weight each individual.

- According to 40 simulations, the median of relative error (RE) is always closer to 0 with Sdlinks. With 20% of missings, we find empirical evidences that Sdlinks is statistically better than Sd2001, only in the case of NPA population (p-value= 7.657e-05, being -0.14% for Sdlinks and 0.75% for Sd2001).

- When missingness rate increases, Donors methods are not able to face the limitations that the strata with very few donors present (with 0 donors a NR is not imputed). In addition, the number of missings per person increments and the criterion of the distance loses all meaning (with an empty Semainier, a NR is not imputed).

- These weaknesses do not exist with the Bayesian methods, because they do not use any information of the Semainier from NR and neither donors. Thus, supposing that covariates of models have no missings, Bayesian models can always impute a NR even if, in the worst of the cases, NR has an empty Semainier.

- Compared with the donors methods, Bayesian models remain, in median, more insensitive to an increase in the missings rate (median of RE of 0,14% for Bayesian models, -0.47 for Sdlinks and -1.76 for Sd2001).

- However, Bayesian models need to be improved. Different validation analyses show that:

  - Through correspondence tables computed with a test-subset (no included in the model estimation) at the first level, accommodation links are underestimated in a proportion of 17%, lunch links are overestimated in a proportion of 11% and dinner links are overestimated in a proportion of 9,9%. Improvements have to be studied to minimize those percentages.

  - The distribution of the relative error (RE) indicates that the size of subpopulations ($\tilde{N}$) that are the most difficult to estimate by the Bayesian model are those corresponding to *with personal accommodation* (WPA) and *non-roof* people (NR). In these latter cases, the variability of RE explodes. Different sources of variability are introduced:

    A first is issued from the 2-dependent-levels model, prediction errors from the first level are carried over the second level which, also, can produce new errors. In addition, model of level 2 is estimated at each iteration with different individuals.

    A second source, in our case, comes from an excess of statistical power which can lead to overparametrized models where covariates are actually introducing noise instead of information. For future analysis, we propose to apply a forward procedure starting with the null model in order to simplify the model.

- Reference stratum from 2001 (Center of the interview crossed with gender) does not contradict with the considered Bayesian models in this master's degree thesis. However, gender variable does not appear in all the models. In addition, other covariates could also be good strata depending on the type of link:
  - For **lunch links**: 3 strata. Subpopulation (WPA, NPA+NR) with Where to sleep tonight, Subpopulation with Center of the interview (5 cathegories) and a triple stratum Subpopulation, Where to sleep tonight and Center of interview (5). The last one only when missing rate is low.
  - For **dinner links**: 3 strata. Subpopulation (WPA, NPA, NR) with Couple, Center of interview (3 cathegories) with Couple and Subpopulation (3) with Center of Interview (3).

- For **accommodation links**: the same strata that for dinner links.

- Concerning the face to face interview, donors methods using Sdlinks can be applied in regular conditions (at least 1 and 5%), choosing a different stratification for each type of link.

- First information (May 2011) about filling the Semainier in the self-administrated questionnaires by non-francophone people is disappointing: 30% of the returned questionnaires have incompleted or empty Semainiers. If this is the final scenario in the part concerning the non-francophone people of the real survey in 2012, donors methods for imputation will prove impossible to implement. Then, a Bayesian model should be formulated according to our results. Distribution of prior parameters for Sd2012 will be the posterior parameters of the Sd2001 Bayesian model. Thus, predictions performed with the updated Bayesian model will be more accurate.

BIBLIOGRAPHY

6.1 Main papers

- Ardilly, Pascal et Le Blanc, David : « Echantillonnage et pondération d'une enquête auprès de personnes sans domicile : un exemple français ». Techniques d'enquête, volume 27, June 2001.

- Ardilly, Pascal : « Les techniques de Sondage ». Editions Technip. 2006.

- INSEE Méthodes (116, August 2001) : l'enquête *sans-domicile* 2001.

- Marpsat, Maryse: « L'enquête de l'Insee sur les *sans-domicile*: quelques éléments historiques». Courrier des statistiques 123, January-April 2008.

- Marpsat Maryse et Nicolas Razafindratsima : « Les méthodes d'enquêtes auprès des populations difficiles à joindre ». Methodological Innovations Online, 2010.

- Xu Xiaojian, Lavallée Pierre : « Traitements de la non-réponse de links dans l'échantillonnage indirect ». Techniques d'enquête. Canada, Decembrer 2009.

6.2 SAS documentation

- Johnston, Gordon: « Repeated Measures Analysis with Discrete Data Using the SAS System ».
- The GENMOD Procedure. Chapitre 29. SAS System help.

6.3 Complementary Articles

- Ake, Christopher: « Rounding after multiple imputation with non-binary categorical covariates ». SUGI 30. Focus Session. Paper 112-30.

- Allison, Paul: « Imputation of Categorical Variables with PROC MI ». SUGI 30. Focus Session. Paper 113-30.

- Nacache, Gueguen : « Analyse multidimensionnelle de données incomplètes ». Statistique Appliquée, 2005.

APPENDIXES

7.1 The Semainier's extract for the face to face questionnaire

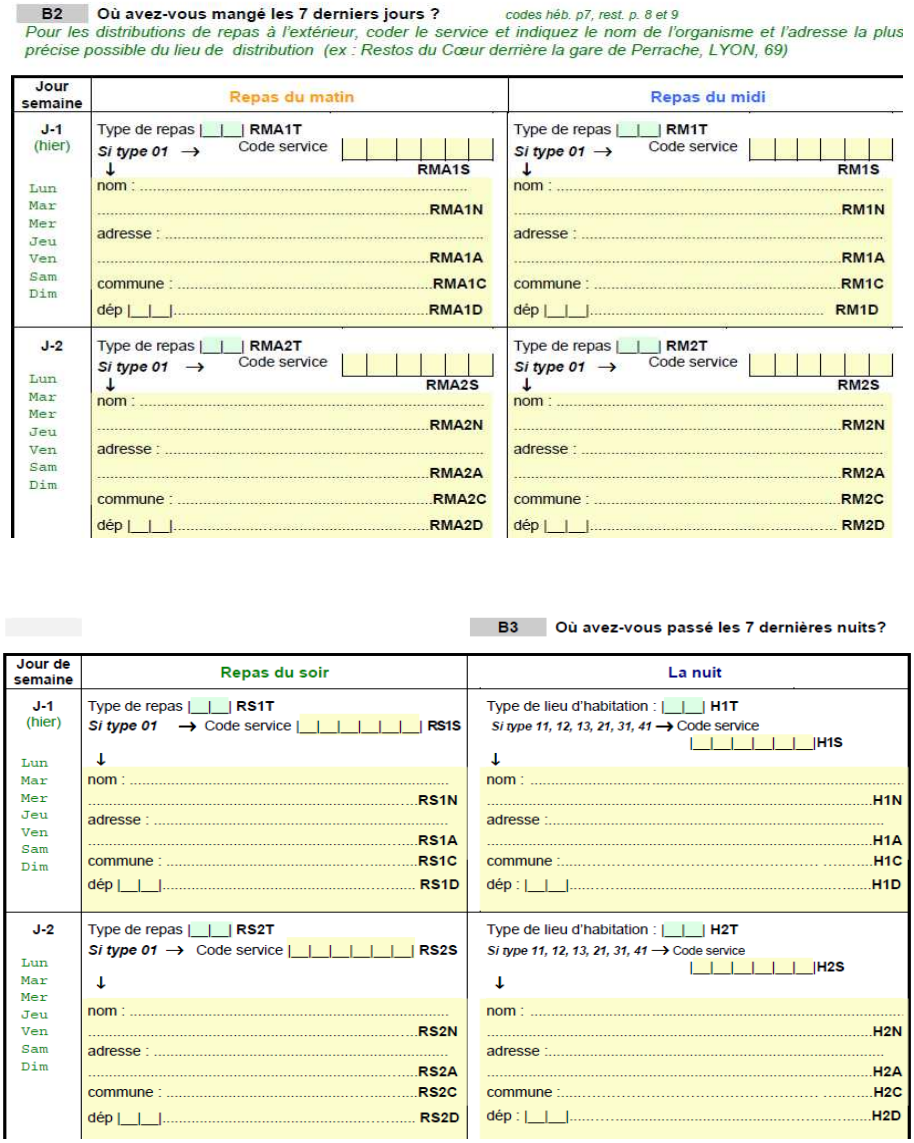

Figure A1. «*Semainer*» in the questionnaire of the test for *SD2012* from only two days.

We asked the person where she slept (***la nuit***), where she had breakfast (***repas du matin***)[39], lunch (***repas du midi***) and dinner (***repas du soir***) in the past two days before the interview.

The type of accomodation (or meal) place and the service's code are two essential informations of this section. The interviewers are also supposed to collect the structure's name, its postal address, the township and the departement that it belongs to.

Below we introduce a list of all the **different types of accommodation places** defined for this survey:

Sleeping accommodation or room in a collective housing

11. *Accommodation centre:*

Urgency accommodation in CHRS

Stabilization accommodation in CHRS

Insertion accommodation in CHRS

Urgency, stabilization or insertion accommodation out CHRS

Maternal centers

Social hotel

Working community

Stopover Beds for Healthcare

12. *Reserved place as a social accommodation in:*

A young-workers centre

A migrant-workers centre

A social residence

13. *Exceptionally opened places in the Plan Against the Great Cold (gyms, underground stations, municipal places, etc) with some installed beds*

14. *Other centers where the person is taking advantage as a resident*

15. *Hospital, clinic, nursing home or recovery home*

16. *Prison*

17. *Others.*

[39] In Sd2001, it was not asked where the person had breakfast.

Accommodation (individual home included, appartament) or mobile habitation (caravan, mobile-home)

- *Accommodation depending of an association or organism*

22. *Accommodation that the person owns, tenant, subtenant, resident.*

23. *Squated accommodation, occupation without any title deed.*

24. *Sleeping in a friend or family accommodation where they live in too.*

25. *Caravan, mobile-home, etc*

Hotel room

31. *Hotel room payed by an association, an accommodation centre or an organism*

32. *Hotel room payed for the person*

Places not intended for habitation

- *Emergency shelter (a tent, a building entrance, a separated building, a car, etc)*

42. *Public places (train station, underground, airport, commercial centre, bridge, parking, public park, etc)*

43. *Night shelters (without sleeping) including day shelters opened during the night for the Plan Against the Great Cold*

99. *Unknown*

The different **types of meal benefits** are:

01. *Breakfast, lunch or dinner distribuited freely in an especific place or in a social restaurant with really cheap prices*

02. *Meal had in the restaurant associated to the accommodation centre where the person has spent the night*

03. *Meal cooked in the person's home*

04. *Meal had in the family or friends' house*

05. *Meal had out the accommodation centre (a café, restaurant, fast-food place)*

06. *Taken-away food the previous day of the interview from a free distribution*

07. *Given food (out of free distributions)*

08. *Reclaimed food*

09. *Other*

10. *She/he has not eaten*

99.Unkown

7.2 Sampling methods to study a hard-to-reach population

We introduce and discuss different sampling methods available to study a hard-to-reach population, divided in two approaches:

a) Empirical sampling

The quota's method. It consists in studying a sample with exactly the same structure as the target population. Some criterions are fixed to limitate the **bias of selection** introduced by the interviewer. It is the **non-probability version of stratified sampling**. In practice, the quotas or "strata" are usually based on gender, age and socioprofessional categories. We should be able to guarantee that the target variable's values depend only on the criterions.

Compared to the probability approach (b), the quota's method is faster and less expensive, due to its simplicity. Moreover, we could easily achieve the *non-roof* that would be missed with the second approach.

However, this approach presents some **weak points**: the criterions must be conveniently chosen according to the population target that, forcely, should be well-known a priori. The worse are the fixed criterions, the more bias of selection we get and the worse estimation results we achieve. Furthermore, the auxiliary information from our target information must be updated.

b) Probabilist sampling

It is divided in two families:

1. **Two-phase sampling (G. Kalton, 2009).**

   At the first stage, a "filter" survey is undertaken on a large sample by providing a simplified questionnaire that allows members of the target population to be

identified. At a second stage, the sample is selected. The main disadvantage of these techniques is that they are expensive to implement and that they can only be used when the population is fairly stable and easily identifiable, which is not the case with the *sans-domicile*.

2. ***The indirect method (Lavallée, 1995; 2002; 2007; Lavallée and Rivest, 2009).***

In this three methods presented hereafter, we follow the principle of finding out where the target people are present (in specific places, in a social network).

I. Respondent driven sampling (Heckathorn, 1997):

Individuals initially interviewed receive a limited number of coupons that they use to recruit other people. Whenever a person is recruited, the recruiter is paid. The person recruited- who is also paid for filling out the questionnaire- receives the same number of coupons, etc. The survey is stopped when the size of the sample has been reached and the composition of the sampls is stable in terms of those characteristics that form the subject matter of the research. We have to mention that the recruited people can accept or refuse. It is essential to trace who recuited whom (to link the recruiter population and the recruited population) and to collect information concerning the size of each person's network to weight each interviewed person.

The limitations of this method are: people with a very poor social network have a lower probability of being reached; the recruited individual must be able to identify those that are members of the target population as well as we should be able to check that the recruited person actually belongs to that population, which could be intrusive. Finally, is not easy to find out the size of the recruiter's network.

II. Capture-recapture (Cowan, 1991):

The technique is based on at least two independent observations (or sources) of this population. In order to estimate the size *N* of the population, we need to know: *n*, the number of people belonging to the population observed at the first stage (or at the first source), *m*, the number observed the second

time (or in the second source) and *M* the number of people observed on both occasions. *N* is then estimated by *(nm)/M*. The people have to be identified in order to be included in *M* at the second occasion.

While this method's underlying concept is simple, the hypotheses that must be satisfied for being the model valid are fairly restrictive:

a) All individuals in the population must have the same chance of being selected at each stage.
b) The observations at the two different stages are independent.
c) The population remains fixed between the two stages.

An alternative would be to count the number of *sans-domicile* by making them filling in a brief questionnaire at each visit to one of the centres. Nevertheless, it is not realistic taking into account the reluctant attitude of the *sans-domicile* when they have to identify themselves. Moreover, we have to remind that the objective of SD's survey is double: **to count the size of this population and, at the same time, to improve our knowledge on it.**

III. Time-location sampling

The method selected for sans-domicile survey.

7.3. Outputs of the GEE estimations

Here are the essential SAS outputs that we have analyzed in order to get to the conclusions in section. The next table summarizes the statistical importance of the variable day of the week in the accommodation's links response:

24 accommodation non-responders. The variable *R* has 108 answers (*R =1*) and 60 missings (*R =0*). The statistical significance of the variable *day of the week* is **0,0158.** The estimated correlation parameter between two days (rho) is 0,08.

Explanatory variable: Day of the week	**Punctual OR**	**95% CI OR**
Day 1	Reference	Reference
Day 2	0,22	[0,04 ; 1,28]
Day 3	0,11	[0,01 ; 1,05]
Day **4**	0,06	**[0,01 ; 0,6]**
Day **5**	0,04	**[0 ; 0,37]**
Day **6**	0,04	**[0 ; 0,44]**
Day **7**	0,04	**[0 ; 0,27]**

SAS Ouput 1. Summary of the GEE model for the accommodation's links response.

Going on with the same process for the meal links, here we have:

Explanatory Variable	**Levels of the explanatory variables**	**Punctual OR**	**95% CI OR**
Day of the week	Day 1	Reference	Reference
	Day 2	0,06	[0,06 ; 0,23]
	Day 3	0,03	[0,01 ; 0,11]
	Day 4	**0,02**	[0 ; 0,06]
	Day 5	0,01	[0 ; 0,04]
	Day 6	0,01	[0 ; 0,03]
	Day 7	0,01	[0 ; 0,03]
Understanding level	Excellent, good understanding	Reference	Reference
	Decent	0,72	[0,41 ; 1,26]
	Bad	0,25	[0,11 ; 0,56]

SAS Ouput 2. Summary of the GEE model for the meal links response.

142 meals nonresponders. The variable *R* has 1447 answers (*R =1*) and 541 missings (*R =0*). The statistical significance of the variable *day of the week* is **lower than 0,001.** The variable *understanding level* has a p-value of **0,011.** The estimated correlation parameter between two days (rho) is 0,19.

In order to test the hypothesis of ignorability of the missing links, we have introduced to the model the variable *Semainier* after being imputed applying the SD2001 imputation method:

Explanatory variables: Day of the week and Imputated *Semainier*	**Punctual OR**	**95% CI OR**
Day 1	Reference	Reference
Day 2	0,19	[0,03 ; 1,14]
Day 3	0,09	[0,01 ; 0,93]
Day **4**	0,05	[0,01 ; 0,55]
Day **5**	0,03	[0 ; 0,33]
Day **6**	0,03	[0 ; 0,36]
Day **7**	0,03	[0 ; 0,22]
Link in the imputed *semainier*	Reference	Reference
No link in the imputed *semainier*	2,4	**[0,99 ; 6,72]**

SAS Output 3. Testing the ignorability of the accommodation's link nonresponse.

24 non-responders. The dependent variable *R* has 108 answers (*R =1*) and 60 missings (*D=0*). The statistical significance of the variable *day of the week* is **0,01.** The p-value of the variable *Imputated "Semainier"* is **0,054** (being **0,11** in the meal links model).

7.4 SAS code for the SD2001 method

Below is the SAS code that has been deciphed, used and adapted for the other methods based on the nearest neighbor:

```
%macro Imputations;

    proc sql;  /*Selection of the accommodation non-responders (NR)*/
        create table France.Imputations_Acco1 as
            select
                Ident, Qnlot, Qreg, Jourcl, Qresp, Qaccep,
                Sexe, Anais,
                H0T, H1T, H2T, H3T, H4T, H5T, H6T, H7T,
                RM1T, RM2T, RM3T, RM4T, RM5T, RM6T, RM7T,
                RS1T, RS2T, RS3T, RS4T, RS5T, RS6T, RS7T,
                Dormid, Endanc, Endsfreq, Endroicl
```

```
            from
                France.fini                 /*SD2001 survey data*/
            where
                Qvisite ne ''                           and
                (H1T='99' or H2T='99' or H3T='99' or H4T='99' or H5T='99'or H6T='99' or H7T='99');

    proc sql;   /*Selection of the accommodation responders (D): donors*/
      create table France.Imputations_Complets as
          select
                Ident as Identd, Jourcl as Jourcld, Sexe as Sexed, Anais as Anaisd,
                H0T as H0Td, H1T as H1Td, H2T as H2Td, H3T as H3Td, H4T as H4Td, H5T as H5Td, H6T as H6Td,
                H7T as H7Td, Dormid as Dormidd, Endanc as Endancd, Endsfreq as Endsfreqd, Endroicl as
                Endroicld
          from
                France.fini                 /* SD2001 survey data*/
          where
                (Qvisite ne '' and H1T<>'99' and H2T<>'99' and H3T<>'99' and H4T<>'99' and H5T<>'99' and
                H6T<>'99' and H7T<>'99');
    quit;

    proc sql;       /*Merge of all the accommodation non-responders and their available donors*/
        create table France.Imputations_Acco2 as
            select Imputations_Acco1.*, fini.*
            from
                France.Imputations_Acco1, (select Ident as Identd, Jourcl as Jourcld, Sexe as Sexed, Anais as
Anaisd,
                                                H0T as H0Td, H1T as H1Td, H2T as H2Td, H3T as H3Td, H4T as H4Td,
                                                H5T as H5Td, H6T as H6Td, H7T as H7Td,
                                                Dormid as Dormidd, Endanc as Endancd, Endsfreq as Endsfreqd,
                                                Endroicl as Endroicld
                                                From
                                                    France.fini
                                                 where
                                                        (Qvisite ne '' and H1T<>'99' and H2T<>'99' and H3T<>'99'
                                                        and H4T<>'99' and H5T<>'99' and H6T<>'99' and
                                                        H7T<>'99')) as fini
            where
                substr(Imputations_Acco1.Ident,5,1)=substr(fini.Identd,5,1)
                /*Donors from the same type of service than the non-responder (this information is contained
                in the personal identifier*/
                and
```

```
                Imputations_Acco1.Sexe=fini.Sexed;     /*Donors of the same sex that the non-responder*/
    quit;


    %macro Imputations_Acco;

        data France.Imputations_Acco3;          /*Updating  the semainier of the non-responder (NR) and their
donors (D)*/
        set France.Imputations_Acco2;
            if Jourcl='LUNDI' then Jourclnum=1;
            if Jourcld='LUNDI' then Jourcldnum=1;
            if Jourcl='MARDI' then Jourclnum=2;
            if Jourcld='MARDI' then Jourcldnum=2;
            if Jourcl='MERCREDI' then Jourclnum=3;
            if Jourcld='MERCREDI' then Jourcldnum=3;
            if Jourcl='JEUDI' then Jourclnum=4;
            if Jourcld='JEUDI' then Jourcldnum=4;
            if Jourcl='VENDREDI' then Jourclnum=5;
            if Jourcld='VENDREDI' then Jourcldnum=5;
            if Jourcl='SAMEDI' then Jourclnum=6;
            if Jourcld='SAMEDI' then Jourcldnum=6;
            if Jourcl='DIMANCHE' then Jourclnum=7;
            if Jourcld='DIMANCHE' then Jourcldnum=7;
            array HT(7) H1T--H7T ;
            array NT(7) NT1-NT7;
            array HTD(7) H1TD--H7TD;
            array NTD(7) NTD1-NTD7;
            do i=1 to 7;  /*Taking into account that the days of the inquiry can be different in the NR and D*/
                l=Jourclnum-i;
                m=Jourcldnum-i;
                if l<=0 then l=l+7;
                if m<=0 then m=m+7;
                NT(i)=HT(l);
                NTD(i)=HTD(m);
            end;
            %let MOD= 11 12 13 14 15 16 21 22 23 24 31 32 41 42 99; /*Modalities from the accommodation
service*/ ACCOSEM(15) ACCOSEM1-ACCOSEM15;
            %do i=1 %to 15;
                ACCOSEM(&i)=0;     /*Attendance's count of each modality per non-responder (NR)*/
                %do j=1 %to 5;
                  ACCOSEM(&i)=ACCOSEM(&i)+(NT(&j)=%scan(&MOD,&i));
```

```
            %end;
        %end;

        array ACCOSEMD(15) ACCOSEMD1-ACCOSEMD15;
        %do i=1 %to 15;
            ACCOSEMD(&i)=0;  /*Attendance's count of each modality per non-responder (NR)*/
            %do j=1 %to 5;
                ACCOSEMD(&i)=ACCOSEMD(&i)+(NTD(&j)=%scan(&MOD,&i));
            %end;
        %end;

        /*Differences among NR and their D regarding the attendance of each modality*/
        array diff(30) diff1-diff30;
        do i=1 to 15;
            diff(i)=abs(Accosemd(i)-Accosem(i));
        end;
        distance=sum(of diff1-diff30);      /*Distance function among NR and D. Each couple has an
associated distance*/
        alea=uniform(0);
    run;
%mend;

%Imputations_Acco;

proc sql;       /*Selection of the candidate donors (with the minimum distance)*/
    create table France.Imputations_Acco4 as
        select
            Ident as Identr, Identd, H1Td, H2Td, H3Td, H4Td, H5Td, H6Td, H7Td, distance, alea
        from
            France.Imputations_Acco3
        where
            distance=(select min(distance) from France.Imputations_Acco3 where ident=identr);

    create table France.Imputations_Acco5 as    /*Random sampling of a donor among all the candidates*/
        select
            Identr as Identrr, Identd, H1Td, H2Td, H3Td, H4Td, H5Td, H6Td, H7Td, distance, alea
        from
            France.Imputations_Acco4
        where
            alea=(select min(alea) from France.Imputations_Acco4 where identr=identrr);
quit;
```

```
data Impute1;
set France.fini(keep=Ident Qnlot--Endhtyp);
(…)

/*Same code for the meal missing links imputation*/

proc sql;  /*Merge with the global data set (Impute1), containing all the other variables*/
      create table Fusion_imputation1 as
            select Impute1.*, Imputations_Acco5.*
                  from
                        Impute1 left join France.Imputations_Acco5
      on
            Impute1.Ident=Imputations_Acco5.Identrr;
quit;

%macro Modification_imputation;

      data France.Imputation;
      set Fusion_imputation2;
            if Qvisite='' then delete;
            %do i=1 %to 7;
      /*If the modality for day i-th is unknown (we have a NR), it will be replaced by the modality of the day i-
      th from its chosen donor*/
      if H&i.T='99' then H&i.T=H&i.TD;
            %end;
      run;

%mend;

%Modification_imputation;

%mend;

%Imputations; run;
```

7.5 Example of Sd2001 and Sdlinks

Pedagogical example	***SD2001' Semainier***				***SD2001links' Semainier***			
Individual	Acco 1	Acco 2	Acco 3	Acco 4	Acco 1	Acco 2	Acco 3	Acco 4
NR	21	11	99	99	1	0	99	99
Donor 1	11	12	31	42	1	0	0	1
Donor 2	21	15	11	11	1	1	1	1

Individual	**21**	**11**	**12**	**31**	**15**	**42**	**99**		**1**	**0**	**99**	
NR	1	1	0	0	0	0	2	**Distance (NR, donor)**	2	0	2	**Distance (NR, donor)**
Donor 1	0	1	1	1	0	1	0	6	**3**	**1**	**0**	**4**
Donor 2	**3**	**1**	**0**	**0**	**0**	**0**	**0**	**5**	4	0	0	4
SD2001									***SD2001links***			

Note that, with Sd2001 we select only one candidate donor whereas, with Sdlinks, we have 2 candidate donors.

7.6 Comparing candidate models for the first level of the accommodation links.

Remark that we have validated it with the corresponce tables. Here, you can clearly see that H1.1 is better than H1.2 as the RE are closer to 0 in the first case.

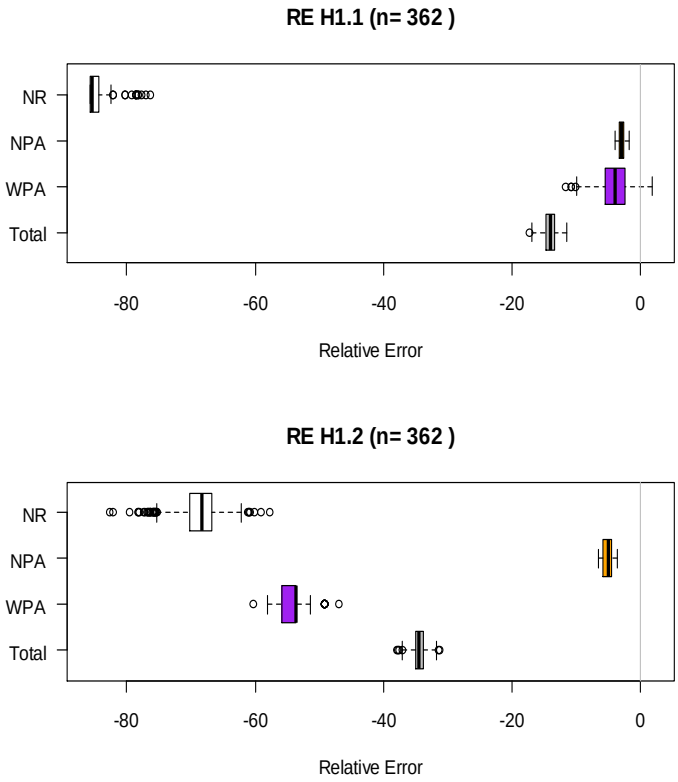

Figure A2. Distributions of the relative errors after 500 estimations of 4 totals : N_{NR}, $N_{NPA,}$ N_{WPA} and the global number (Total), with H1.1 model and H1.2 models.