An efficient hierarchical model for multi-source information fusion

Ismaïl Saadi[1*], Bilal Farooq[2], Ahmed Mustafa[1], Jacques Teller[1], Mario Cools[1,3,4]

[1]Local Environment & Management Analysis (LEMA), Urban and Environmental Engineering (UEE), University of Liège, Allée de la Découverte, Quartier Polytech, 4000 Liège, Belgium

[2]LITrans, Ryerson University, Department of Civil Engineering, 350 Victoria St, Toronto, Ontario M5B 2K3, Canada

[3]KULeuven Campus Brussels, Department of Informatics, Simulation and Modeling, Warmoesberg 26, 1000 Brussels, Belgium

[4]Hasselt University, Faculty of Business Economics, Agoralaan Gebouw D, 3590 Diepenbeek, Belgium

Email addresses: ismail.saadi@uliege.be (I. Saadi), bilal.farooq@ryerson.ca (B. Farooq), a.mustafa@uliege.be (A. Mustafa), jacques.teller@uliege.be (J. Teller), mario.cools@{uliege.be,kuleuven.be,uhasselt.be} (M. Cools)

*Corresponding author: Tel. : +32 4 366 96 44, Fax : +32 4 366 29 09

Abstract

In urban and transportation research, important information is often scattered over a wide variety of independent datasets which vary in terms of described variables and sampling rates. As activity-travel behavior of people depends particularly on socio-demographics and transport/urban-related variables, there is an increasing need for advanced methods to merge information provided by multiple urban/transport household surveys. In this paper, we propose a hierarchical algorithm based on a Hidden Markov Model (HMM) and an Iterative Proportional Fitting (IPF) procedure to obtain quasi-perfect marginal distributions and accurate multi-variate joint distributions. The model allows for the combination of an unlimited number of datasets. The model is validated on the basis of a synthetic dataset with 1,000,000 observations and 8 categorical variables. The results reveal that the hierarchical model is particularly robust as the deviation between the simulated and observed multivariate joint distributions is extremely small and constant, regardless of the sampling rates and the composition of the datasets in terms of variables included in those datasets. Besides, the presented methodological framework allows for an intelligent merging of multiple data sources. Furthermore, heterogeneity is smoothly incorporated into micro-samples with small sampling rates subjected to potential sampling bias. These aspects are handled simultaneously to build a generalized probabilistic structure from which new observations can be inferred. A major impact in term of expert systems is that the outputs of the hierarchical model (HM) model serve as a basis for a qualitative and quantitative analyses of integrated datasets.

Keywords. Iterative Proportional Fitting (IPF); Hidden Markov Model (HMM); Hierarchical Model (HM); Multi-source information fusion

1. Introduction

Forecasting activity-travel patterns is relevant to many applications and research domains, e.g. urban/transportation research, and social sciences (Liu et al., 2013, 2015; Saadi et al., 2017). The behavioral realism associated to the simulation of complex urban and transportation systems requires highly disaggregated and reliable datasets (Batty, 2007; Axhausen & Gärling, 1992). A major problem is that such disaggregated data are not always available (Barthelemy & Toint, 2013). Moreover, sampling rates are generally low, i.e. in the best case at most 10% of the total population, as data collection for travel surveys/micro-samples is costly, and large-scale surveys, i.e. censuses, are systematically subjected to privacy and confidentiality issues (Saadi et al., 2016b). Therefore, in urban and transportation research, efficient and flexible methods are required to fuse information stemming from multiple micro-samples and aggregate statistics, e.g. socio-demographic marginal distributions (Saadi et al., 2016b; El Faouzi et al., 2011; Saadi et al., 2016a; Wu, 2009).

In this paper, a methodological framework is presented that allows for an intelligent merging of multiple data sources. Furthermore, heterogeneity is smoothly incorporated into micro-samples with small sampling rates subjected to sampling bias. These aspects are handled simultaneously to build a generalized probabilistic structure from which new observations can be inferred. A major impact in term of expert systems is that the outputs of the hierarchical model (HM) model serve as a basis for a qualitative and quantitative analysis of integrated datasets. In this context, the decision-making process can be significantly simplified. Advanced knowledge for extracting relevant information from multiple datasets could be replaced by a simpler analysis of a unified dataset that incorporates all the information and variable interactions.

Section 1.1 presents a general overview of the existing methods. Section 1.2 lists the contributions of the current study with respect to the existing work.

1.1. Related work

In the literature, four types of methods - synthetic reconstruction, combinatory optimization (CO), sample free fitting, Monte Carlo Markov Chain (MCMC) simulation-based method - have been distinguished (Ye et al., 2017) to merge data from multiple data sources.

IPF sythetic recontruction-based approaches are commonly used for modeling populations for transport and urban systems (Arentze et al., 2007; Beckman et al., 1996; Zhu & Ferreira, 2014; Barthelemy & Toint, 2013). IPF procedures consist of fitting a multi-dimensional contingency table given a set of target marginal distributions and a single micro-sample derived, for instance, from a travel survey. Observed marginal distributions are used as targets for fitting the micro-sample via an iterative reweighting procedure. In practice, the contingency tables are initiated with micro-samples with low sampling rates, i.e. at most 5 to 10%. This dependency on micro-samples is particularly problematic as IPF procedures systematically preserve the error of the related multi-variate joint distribution despite the fact that the marginals are fitted quasi-perfectly. Furthermore, applying an IPF may be problematic in the case of unavailable micro-samples for disaggregate inputs. In addition, the quality of the sample influences the final IPF output. In some situations, when a combination of attributes with low probability occurrence is missing within the sample, the synthetic population will not include the corresponding set of combined attributes. Setting up the zero element cells with very small values has been proposed to tackle this issue; however this would add an arbitrary bias. In contrast, IPF procedures are particularly powerful in providing highly accurate synthetic populations, when the correspondence between the synthetic and observed populations is evaluated on the basis of the marginal distributions.

Besides, CO can be defined as a micro-data reconstruction approach which performs a random selection of households from micro-samples in order to reproduce the characteristics of a specific geographical unit. Different statistical metrics have been proposed to assess the goodness-of-fit of the model (Voas & Williamson, 2000). Similar to IPF, CO is a sample-based approach that also suffers from the zero-cell problem in the image of IPF.

Given the fact that disaggregated samples are difficult to obtain in some countries, sample-free methods emerged as interesting alternatives. Marginal and/or conditional distributions of partial attributes are adopted as input data in order to enable more flexibility. However, when the distributions are not consistent across the data sources, a problem occurring especially in the case of discrete variables, further adjustments are operated by performing individual shifts. Furthermore, sample-free methods are extremely time-consuming and generally require a heavy methodological procedure with multiple connected sub-models for generating an individual pool.

With respect to the Markov Process-based methods, Farooq et al. (2013) used, for instance, an MCMC method for population synthesis. Both the full and partial conditional distributions used by MCMC method can be calibrated on multiple micro-samples. Despite the relative flexibility in terms of data integration, the MCMC-based approach is insufficiently adapted for dealing with datasets that have variables with a high number of categories. This is due to the fact that the Multinomial Logit Models (MNL), that are used within the simulation procedure, are too sensitive to this type of variables. In addition, the method may over-fit the micro-samples if full conditional probability distributions are used and substantial information may be lost if partial conditionals are adopted. Besides, MCMC simulation-based method can be considered as a sample-free approach as it relies on conditional distributions which are calibrated on the basis of different data sources. Both discrete and continuous variables can be handled. However, inconsistencies in conditional distributions, may keep MCMC from converging towards a stationary state; which would never result in a correct population.

Saadi et al. (2016b) used an HMM-based approach for synthesizing the population of Belgium. The method is highly flexible for fusing multiple micro-samples and shows competitive prediction capabilities. Nonetheless, the full dependency on micro-samples often leads to less accurate simulated marginal distributions despite accurate simulated joint distributions. In this paper, we propose an extension of the HMM by integrating IPF, allowing an efficient multi-source information fusion.

1.2. Contributions

The contributions of the current study are defined as follows:

1. We develop a new hierarchical model for fusing an unlimited number of information sources irrespective of the level of aggregation.
2. The hierarchical model generalizes the HMM by incorporating IPF. In doing so, the quality of the simulated multivariate joint distributions is preserved in addition to quasi-perfect marginal joint distributions.
3. Efficient algorithms are designed for smartly calibrating the hierarchical model (HM).

The remainder of the paper is structured as follows. First, we describe the new modeling framework. In Section 3, the results are discussed and conclusions are formulated in Section 4.

2. The Hierarchical Model (HM)

2.1. Data

The methodology developed under the present study essentially handles (a) travel surveys which include socio-demographics or transport/urban-related variables and (b) corresponding aggregate marginal distributions. The variables can be either discrete or continuous but discretized to be handled within the model. Typically, gender (male-female), car ownership (yes-no), socio-professional status (student, worker, employee, etc.), residential location (ID of the commune) are, among others, considered as discrete variables. The surveys may also include continuous variables such as age - between 1 and 100 or travel time. In most cases, continuous variables are discretized into categories in order to enable data fusion. In practice, researchers mainly deal with discrete or discretized continuous variables. Data can be collected by means of face to face interviews or on-line questionnaires.

Besides, two types of input must be clearly distinguished in the current modeling framework. On the one hand, we have the micro-samples, e.g. travel surveys, which are relatively detailed but with small sampling rates, i.e. less than 10%. Also, the links in-between the variables are preserved as for each observation, one has information about, e.g. gender, age, socio-professional status and many other variables, of a specific anonymized person. On the other hand, we have aggregate data which can be derived from national organisms or bureau of statistics independently of each other, e.g. pyramid of ages, gender distribution, etc.

2.2. Model structure

The structure of the hierarchical model (HM), which enables multi-source information fusion, is illustrated in Figure 1. HM includes two important components, i.e. HMM and IPF. The N micro-samples and the M aggregate marginal distributions can be used simultaneously as inputs within the HM framework. The scaled-up and fused micro-sample enables the connection between HMM and IPF. As the multi-source fusion process already takes place within the HMM component, the scaled-up and fused micro-sample systematically includes all the variables of interest. IPF enables a direct fitting of the marginal distributions based on the observed targets, i.e. second set of inputs. Of course, the use of all the aggregate marginal distributions is not mandatory. It depends on data availability. Thus, HM is designed to allow enough flexibility towards unavailable marginal distributions. It is indeed possible to fit data against a number of marginal distributions which is lower than the total number of variables of interest, i.e. M. Finally, HM results in a fused and more accurate dataset that can be used in multiple applications, e.g. agent-based modeling of complex urban and transportation systems (Batty, 2007; Horni et al., 2016).

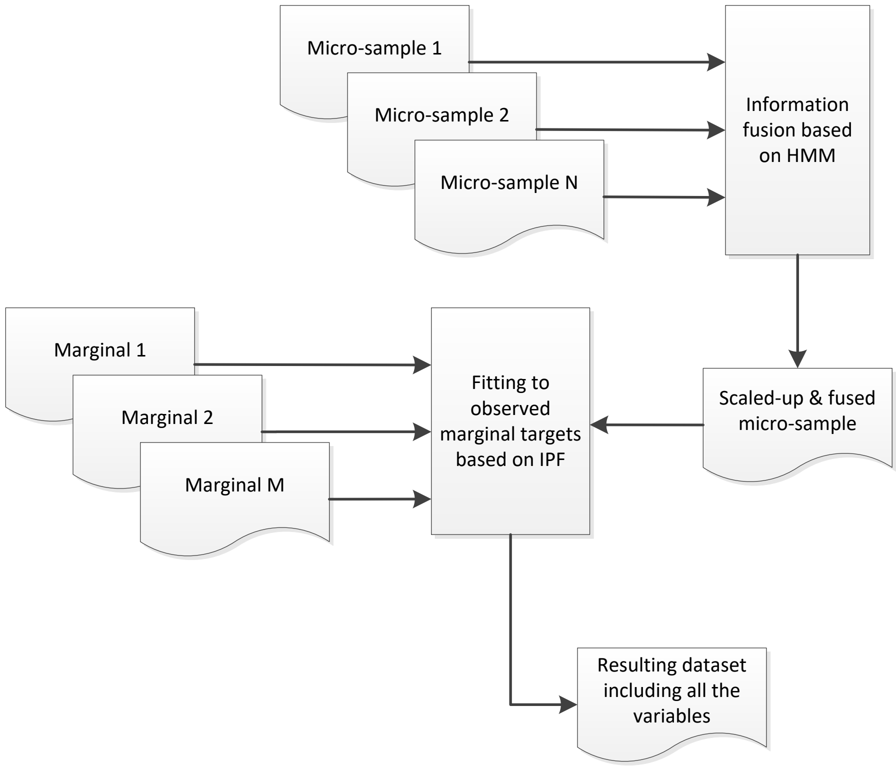


Figure 1: Structure of the hierarchical model

Regarding the fusion process, the N micro-samples are merged based on HMM using Algorithms 1 and 2. In doing so, the HMM sequentially learns the configuration structure of the pseudo multivariate joint distribution of the true population. Here, the word "pseudo" has been used because a sample with a very small sampling rate will never statistically replicate an accurate representation of the true population.

2.3. Learning

A new generalized algorithm is proposed in the context of this modeling framework to merge multiple data sources and handle missing values, i.e. not attributed (NA) and/or not-a-number (NAN). Indeed, (a) standard methods for estimating HMM are not adapted for handling data stemming from multiple sources. Instead of estimating the transition probabilities from a single micro-sample, the algorithm is designed such that the information about the transition probabilities from a variable to another are extracted from their corresponding data source.

In addition, (b) the way of handling missing data vary from a method to another. A naive way is to clear the row with partial information. For example, a full observation, e.g. row in a dataset, containing a single NA value can be cleared. This may be problematic if missing values are important within the dataset. The overall distributions of the variables contained within the "cleaned sample" might be subjected to major changes compared to the original one. Thus, even if the dataset includes observations with partial information, then HMM ignores NA or NAN values and uses the complementary available information for updating the model. This feature is enabled by Algorithm 2.

Two hypotheses have been formulated. (A1) In the case of a multi-source information fusion operation, we assume that the different micro-samples share at least a common variable in order to enable the shift from a sample to another, and for guarantying the fusion process. (A2) The categories within the variables are defined as integers starting from 1.

In order to understand the fusion process, Figure 2 presents an HMM with n variables. The variables are symbolized with states and the transition patterns with either continuous or dashed arrows. For example, setting up a synthetic dataset of 3 variables, e.g. age, gender, car ownership, would require an HMM of length 3, i.e. $n = 3$. The transition probabilities, T_1, T_2, ..., T_i, ..., T_n, which can also be defined as 2 way tables are estimated from a single data source if all the variables are included within the same dataset, from multiple datasets otherwise. For example, the link between age and gender would come from sample 1 and the link between car ownership and gender or age and car ownership from sample 2. In both cases, assumption A1 is respected as both samples share at least a common variable. Detailed descriptive aspects have been included within the Algorithms 1, 2 and 3 to understand how the algorithms are applied.

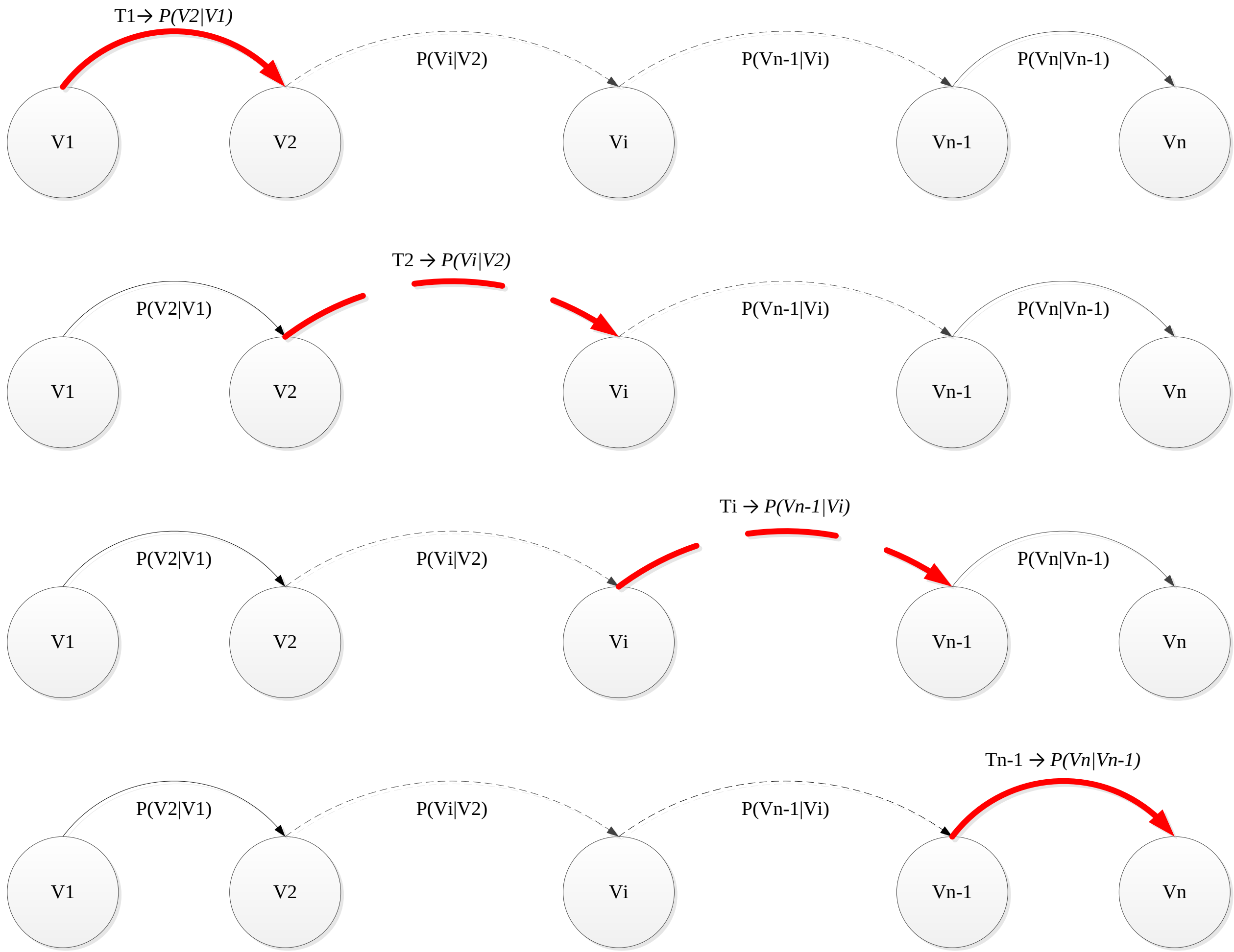


Figure 2: Representation of the transition patterns - V_i represents a variable with a specific number of categories. The objective is to systematically determine the relation between two adjacent variables by estimating a 2 way table. The red continuous-dashed arrows symbolize the transition patterns. They can either be estimated from a single or a combination of datasets. T_i represents a matrix which dimensions depend on the number of categories of the involved variables V_i and V_{i+1}.

Before running Algorithm 1, a pre-processing of the variables of interest should be realized. After selecting the variables, the micro-samples in which the variables are contained should be collected, e.g. from national travel surveys. The link between the transition patterns and their corresponding micro-samples needs to be clearly identified. Also, it must be ensured that common variables exist across the samples (Assumption 1) and that the categories are represented in terms of integers starting from 1 (Assumption 2). Finally, the location of the partial transition matrix T_k needs to be pre-defined to enable the sequential updating of T.

Algorithm 1 Updating of the transition probability matrix T

```
// Initialize K number of transition patterns
// Initialize N sum over all the variable categories
Set K and N

// Returns an N×N matrix
T ← CreateTable(N,N)

// Loop over the K transition patterns
for k=1 to K-1 do
  // Returns a two-columns sample with variables $V_k$ and $V_{k+1}$
  [$V_k$,$V_{k+1}$] ← GetMicroSample(k)

  // Returns the corresponding two-way table $P(V_{k+1}|V_k)$
  Tk ← Get2DCrossTab($V_k$,$V_{k+1}$)

  // Returns X-Y initial and final locations of $T_k$ with respect to T
  [xi,xf,yi,yf] ← GetPositions(k)

  // Assign Tk to T within the corresponding location
  T[xi→xf,yi→yf]← Tk
end for
```

Algorithm 2 Get2DCrossTab

```
function:  Get2DCrossTab($V_k$,$V_{k+1}$)

// Returns the number of levels within the input variable
nk1 ← getNumberOfCategories($V_k$)
nk2 ← getNumberOfCategories($V_{k+1}$)

Tk ← CreateTable(nk1,nk2)

for i=1 to length($V_k$) do
  if $V_k$[i]="NAN" or $V_{k+1}$[i]="NAN" or $V_k$[i]="NA" or $V_{k+1}$[i]="NA" then
    // Do not update
  else
    Tk[$V_k$[i],$V_{k+1}$[i]] ← Tk[$V_k$[i],$V_{k+1}$[i]]+1
  end if
end for
return Tk
```

2.4. Sampling

After the learning step, a desired number of observations is inferred from the estimated HMM structure using Algorithm 3. Theoretically, an infinite number of sequences can be generated based on the estimated HMM while preserving all the properties of the population/original dataset. In practice, it will depend on the application needs. In urban and transportation research, the number of sequences depends on the size of the populations that we need to synthesize. A sequence is defined as a combination of attributes or variables.

Algorithm 3 describes the adopted procedure for generating combination of attributes from the HMM component of HM. Based on the function `getDistribution()`, the distribution of V_1 is obtained and stored in $\mathbf{p}$. Q stands for the size of the population or the number of observations needed. After initializing the variables, we double loop along the columns and rows of A to generate sequentially the combination of attributes of the synthetic dataset.

Algorithm 3 Data sampling

```
Set Q // Number of observations - size of the dataset

// Returns the density distribution of variable V_1
p ← GetDistribution(V_1)

// Returns null table of dimensions Q×K+1 to store the set of observations
A ← CreatTable(Q,M)

for j=1 to Q do
  γ ← Sample from p
  A[j,1] ← γ
  for k=1 to K do
    // Returns the kth transition table T_k of T
    T_k ← GetTransitionTable(k,T)
    Sample V_{k+1} from T_k = P(V_{k+1}|V_k) based on V_k (or A[j,k]) and store in A[j,k+1]
  end for
end for
```

2.5. Fitting

After the sampling, the scaled-up and fused micro-sample is fitted to the target marginal distributions to operate the final step of the HM modeling framework. In doing so, an adjusted population/-dataset is obtained. Although the cells are updated until the target aggregate marginal distributions are fitted, there is no risk of losing the configuration structure of the multi-dimensional table. In this regard, Barthelemy & Toint (2013) highlighted that IPF preserves the correlation structure of populations based on the odd ratios technique. The preservation of the weights within contingency tables is demonstrated in details in Mosteller (1968).

3. Numerical experiments

The hierarchical model is tested based on a synthetic dataset of 1,000,000 observations and 8 random variables with 128, 16, 8, 8, 4, 4, 3 and 2 categories respectively. Data are deliberately heterogeneous and designed in the image of real world situations. In urban and transportation research, variables contain multiple categories for representing socio-demographics/transport-related variables. The number of categories is even more important if spatial information is included. Therefore, we also chose a complex categorical variable with 128 levels. Table 1 presents a detailed statistical description of the synthetic dataset.

Surveys might be subjected to missing information, e.g. encoding errors during data collection or presence of NA/NAN values. This issue is particularly important as the systematic removal of a combination of variables because of a missing one may lead to overall changes in terms of variable distributions. This aspect has been deeply discussed in Saadi et al. (2016b) by utilizing the survey on workforce. Indeed, data synthesis of a higher number of variables would increase the probability of finding a higher number of missing values. Saadi et al. (2016b) outlined that for the synthesis of three variables, the gender distribution was 49.55% and 50.45% for male and female respectively after data cleaning. Regarding the synthesis of 6 variables, the distribution shifted towards 53.97% and 46.03% after data cleaning. Furthermore, the synthesis of 6 variables has led to a huge decrease in the sample size compared to the original size, i.e. $\Delta = -68\%$. Thus, in the current study, a better algorithm has been defined to synthesize any number of attributes based on the original datasets. In this regard, performing data cleaning is no longer necessary. Valuable amount of information is preserved then.

Table 1: Statistical description of the synthetic dataset

Variable ID	Number of levels	Statisttical description
1	128	Truncated normal distribution
2	16	Normal distribution with the following proportions: 1:2% -2:3% -3:4% -4:6% 5:7% -6:8% -7:9% -8:10% 9:10% -10:9% -11:8% -12:7% 13:6% -14:4% -15:3% -16:2%
3	8	Poisson distribution with the following proportions: 1:5% -2:12% -3:18% -4:20% 5:18% -6:14% -7:9% -8:5%
4	8	Poisson distribution with the following proportions: 1:11% -2:19% -3:22% -4:20% 5:14% -6:8% -7:4% -8:2%
5	4	Poisson distribution with the following proportions: 1:15% -2:27% -3:31% -4:27%
6	4	Poisson distribution with the following proportions: 1:10% -2:22% -3:32% -4:36%
7	3	Poisson distribution with the following proportions: 1:8% -2:35% -3:57%
8	2	1:45% -2:55%

In order to underline the influence of the sampling rate on model outputs, five bootstrap samples are derived from the original dataset in the following order 10%, 5%, 1%, 0.1% and 0.06%. There is no point in considering sampling rates higher than 10%, since such data are typically not available. In Section 3.1, we present the practical procedure for model estimation using a single micro-sample and all the marginals. The results are compared on the basis of the joint and marginal distributions to highlight the performances of HM. In Section 3.2, we illustrate how to fuse multi-source information based on another case study considering multiple micro-samples and all the marginal distributions.

3.1. Model estimation

To run Algorithms 1 and 2, we identify the positions of the partial matrices T_k based on the number of levels (see Table 1). The full transition probability matrix T is of dimension $n \times n$ where

$n = 128 + 16 + 8 + 8 + 4 + 4 + 3 + 2 = 173$. The eight variables of the micro-sample are arranged in descending order of the number of categories an stored as a matrix of dimension $(\delta * 1,000,000) \times 8$ where δ is the sampling rate. We need to compute seven 2-way tables - transition patterns - as 8 variables are synthesized. T matrix is updated through a sequential read of the transition patterns. The values of each variable of V_k and V_{k+1} are used as subscripts by T_k for localizing the corresponding cell. If NA or NAN values are detected, then the algorithm does not update T_k. Thus, incomplete datasets can be used without cleaning procedure as they are implicitly handled by the HM model.

After estimating T, we run Algorithm 3 to generate a certain number of combination of attributes. In this case study, the generated dataset includes 1,000,000 observations to enable a direct comparison with the original one, see Table 1. It must be kept in mind that a single micro-sample and all the aggregate marginals are available in this case study. V_1 is the first random variable which contains 128 categories. A value between 1 and 128 is sampled based on the weights vector $\mathbf{p}$. Then, we loop over the transition patterns to systematically sample the next value based on the corresponding two-way table T_k. T includes all the two-way transition tables to sample the next variable from the current one, see Algorithm 3.

In this case study, T is defined by means of 7 two-way tables as 8 variables are handled, i.e. $T_{1\to128|129\to144}$, $T_{129\to144|145\to152}$, $T_{145\to152|153\to160}$, $T_{153\to160|161\to164}$, $T_{161\to164|164\to168}$, $T_{164\to168|169\to171}$ and $T_{169\to171|172\to173}$. Note that $T_{1\to128|129\to144}$ is not reported because of its dimensionality 128×16. The dimensions of each single table are associated to the number of categories of two adjacent variables. For example, variables 7 and 8 contain 3 and 2 categories, respectively. Thus, T is updated from rows 169 to 171 and from columns 172 to 173 using $T_{169\to171|172\to173}$ of dimensions 3×2. The same updating procedure is applied for the rest of the tables using Algorithms 1 and 2. Figure 3 shows how the interactions are occurring in-between multiple adjacent variables. As highlighted earlier in the paper, the transition patterns are defined as 2-way tables or bi-variate joint distributions. Each cell of a table represents the frequency of a combination of two categorical variables within the overall number of transitions. For instance, if we consider V_5 and V_6, then the dimension of the corresponding 2D table is 4-by-4 and it contains 16 cells.

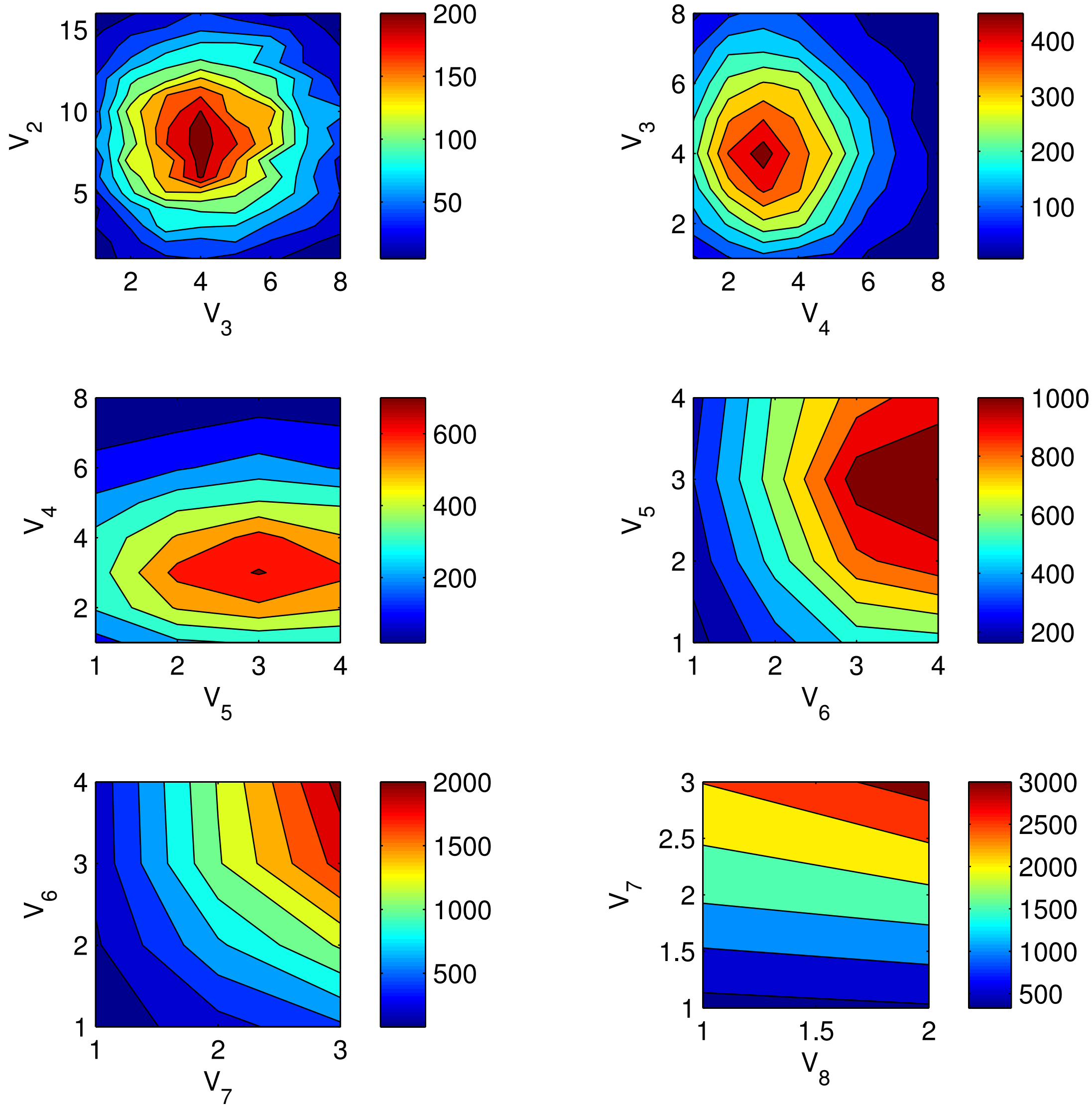


Figure 3: Variable interactions characterized by the probability matrix T, $T_{1\to128|129\to144}$, $T_{129\to144|145\to152}$, $T_{145\to152|153\to160}$, $T_{153\to160|161\to164}$, $T_{161\to164|164\to168}$, $T_{164\to168|169\to171}$ are respectively associated to the interaction maps $V_2 - V_3$, $V_3 - V_4$, $V_4 - V_5$, $V_5 - V_6$, $V_6 - V_7$, $V_7 - V_8$

The performance of the HM model that has been presented in this paper is compared with conventional methods reported in literature. In particular, the HM model is compared to the Direct Inflating (DI) approach, in which the sample is replicated multiple times to obtain the final dataset. In essence, the DI approach is a basic scaling-up process. A second comparison is made with Iterative Proportional Fitting (IPF,) as presented in Beckman et al. (1996). The comparison is made with Hidden Markov Models (HMM) (Saadi et al., 2016b).

Tables 2 and 6 present the marginal errors according to the benchmark methods (DI, IPF, HMM) and the new HM approach presented in this paper. One could depict that HM achieves comparable results to that of IPF with quasi-perfect marginals. In contrast, DI and HMM show important deviations. Moreover, the evolution of the marginal errors demonstrates that there is a relationship between variable dimensionality and importance of the RMSE. Also the RMSE increases if sampling rate decreases.

Table 2: RMSE according to the marginals based on DI, IPF, HMM and HM for a sampling rate of 10%

	DI	IPF	HMM	HM
M1	274.36	1.67E-12	281.64	2.48E-11
M2	651.12	4.46E-12	614.60	9.70E-11
M3	937.42	1.03E-11	777.41	1.48E-11
M4	1080.50	1.03E-11	1061.53	0
M5	1419.66	0	1301.56	0
M6	762.06	0	826.85	7.28E-12
M7	651.80	0	283.25	8.42E-12
M8	1954.00	0	2165.00	4.12E-11

Table 3: RMSE according to the marginals based on DI, IPF, HMM and HM for a sampling rate of 5%

	DI	IPF	HMM	HM
M1	336.18	2.06E-11	360.05	3.19E-10
M2	785.28	2.50E-11	768.66	2.06E-10
M3	772.67	1.65E-11	799.21	2.53E-11
M4	830.02	1.56E-11	1009.45	1.80E-11
M5	2182.65	3.25E-11	2158.18	3.25E-11
M6	1177.64	3.00E-11	1115.92	0
M7	186.40	0	1037.01	0
M8	464.00	0	242.00	0

Table 4: RMSE according to the marginals based on DI, IPF, HMM and HM for a sampling rate of 1%

	DI	IPF	HMM	HM
M1	876.90	1.91E-11	882.35	6.31E-11
M2	3804.71	2.67E-11	3901.08	8.24E-12
M3	3193.39	1.31E-11	3111.55	1.50E-11
M4	2941.89	1.82E-11	2901.29	1.07E-11
M5	2757.62	3.25E-11	3065.11	0
M6	4400.60	2.91E-11	4254.56	3.25E-11
M7	7349.19	3.36E-11	7234.25	3.46E-11
M8	8164.00	8.23E-11	7856.00	0

Table 5: RMSE according to the marginals based on DI, IPF, HMM and HM for a sampling rate of 0.1%

	DI	IPF	HMM	HM
M1	3546.39	1.70E-11	3555.91	2.14E-09
M2	13461.83	1.59E-11	13448.65	1.10E-11
M3	7254.36	2.37E-11	7242.95	1.80E-11
M4	11203.24	1.90E-11	11122.91	1.06E-11
M5	12686.65	4.37E-11	12700.80	1.46E-11
M6	12677.68	2.91E-11	12542.79	0
M7	18078.30	3.36E-11	18499.72	8.40E-12
M8	29264.00	4.12E-11	30059.00	0

Table 6: RMSE according to the marginals based on DI, IPF, HMM and HM for a sampling rate of 0.06%

	DI	IPF	HMM	HM
M1	3546.39	1.70E-11	3555.91	2.14E-09
M2	13461.83	1.59E-11	13448.65	1.10E-11
M3	7254.36	2.37E-11	7242.95	1.80E-11
M4	11203.24	1.89E-11	11122.91	1.06E-11
M5	12686.65	4.37E-11	12700.80	1.46E-11
M6	12677.68	2.91E-11	12542.79	0
M7	18078.30	3.36E-11	18499.72	8.40E-12
M8	29264.00	4.11E-11	30059.00	0

In order to investigate the propagation of the error through the HM, Table 7 presents the RMSE for different sampling rates based on DI, IPF, HMM and HM. DI means that the bootstrap sample has been directly scaled-up and compared to the observed dataset. RMSE of DI and IPF are almost equivalent because IPF re-weights the contingency tables with respect to targets while preserving the proportions. Thus, even the related errors are preserved. Also, HM and HMM show equivalent RMSE's for the three highest sampling rates. In the case of the extremely small sampling rate, i.e. 0.06%, a slight deviation can be observed because of the reweighting procedure enabled by IPF. Theoretically the errors of HMM and HM should be exactly the same as highlighted in Section 2, but small differences are observed. This can be explained by the fact that at the end of the reweighting of the multi-dimensional contingency table, the cell values are rounded. As the later contingency table contains a huge number of cells, the cumulation of rounding error leads to a small decrease of the errors especially for small sampling rates.

Table 7: Evolution of the RMSE according to multiple sampling rates and methods

	DI	IPF	HMM	HM
10%	0.85	0.85	0.40	0.40
5%	1.23	1.23	0.40	0.40
1%	2.81	2.83	0.40	0.41
0.1%	8.91	10.00	0.45	0.49
0.06%	11.5	13.65	0.49	0.54

Based on the results of Tables 2-6 and 7, we conclude that HM allows the best trade-off as multivariate joint distribution errors are almost preserved as well as those of the marginals. Also, HM is less sensitive to sampling rate variability, i.e. from 10% to 0.06%, as the RMSE increases by +35%. When IPF is considered independently, the RMSE increases by +1505.88%. The results reveal that the IPF component of HM affects only the marginals but HMM influences the multi-variate joint distribution. This can be explained by the fact that the HMM component of HM incorporates more heterogeneity into the micro-sample. Indeed, for small sampling rates, some combination of attributes are not necessarily covered. This problem is implicitly avoided by HM.

3.2. Multi-source information fusion

In this second case study, we suppose that the dataset that we want to synthesize contains the same number of variables and variable categories. The only difference is that the variables are included within 3 independent datasets in order to illustrate how to perform a multi-source information fusion. Table 8 presents the distribution of the variables through the 3 micro-samples (MS) with different sampling rates. The sampling rates are deliberately low in order to highlight how efficient is the HM. Each single micro-sample contains four variables.

Table 8: Description of the micro-samples (MS)

	$MS1$	$MS2$	$MS3$
$M1$			×
$M2$	×		×
$M3$	×		
$M4$	×		×
$M5$		×	×
$M6$		×	
$M7$		×	
$M8$	×	×	
Sampling rate	0.1%	1.0%	2.0%

Based on Table 8, we notice that $T_{1\to128|129\to144}$, $T_{129\to144|145\to152}$, $T_{145\to152|153\to160}$, $T_{153\to160|161\to164}$, $T_{161\to164|164\to168}$, $T_{164\to168|169\to171}$ and $T_{169\to171|172\to173}$, can be estimated with $MS3$ (micro-sample 3), $MS1$, $MS1$, $MS3$, $MS2$, $MS2$, $MS2$ respectively using Algorithms 1 and 2. In doing so, T is fully implemented based on partial micro-samples. Also, multi-source information fusion is made effective. The rest of the procedure is similar to what has been described in Section 3.2. Figure 4 presents the comparison between the simulated and observed datasets on the basis of the marginals. One could depict that HM leads to quasi-perfect marginals regardless of the variable complexity.

In addition, Figure 5 shows the comparison between the simulated and observed multi-variate joint distributions for different combination of variable patterns. There is no risk of under/over-estimation as the data points present a good symmetry on both sides of the straight line. Moreover, linear fits (in red) and straight lines (in green) are almost systematically overlapped. Slopes are ranging from 0.97 to 1.00 with extremely small intercepts. Important spread can be observed with respect to patterns $V_1 - V_2 - V_3$, $V_2 - V_3 - V_4$ and $V_3 - V_4 - V_5$ because of variable dimensionality. V_i are arranged in descending order of number of categories. Thus the combination $V_1 - V_2 - V_3$ has the highest number of cells. As a result, the density of data points is significant (Figure 5a).

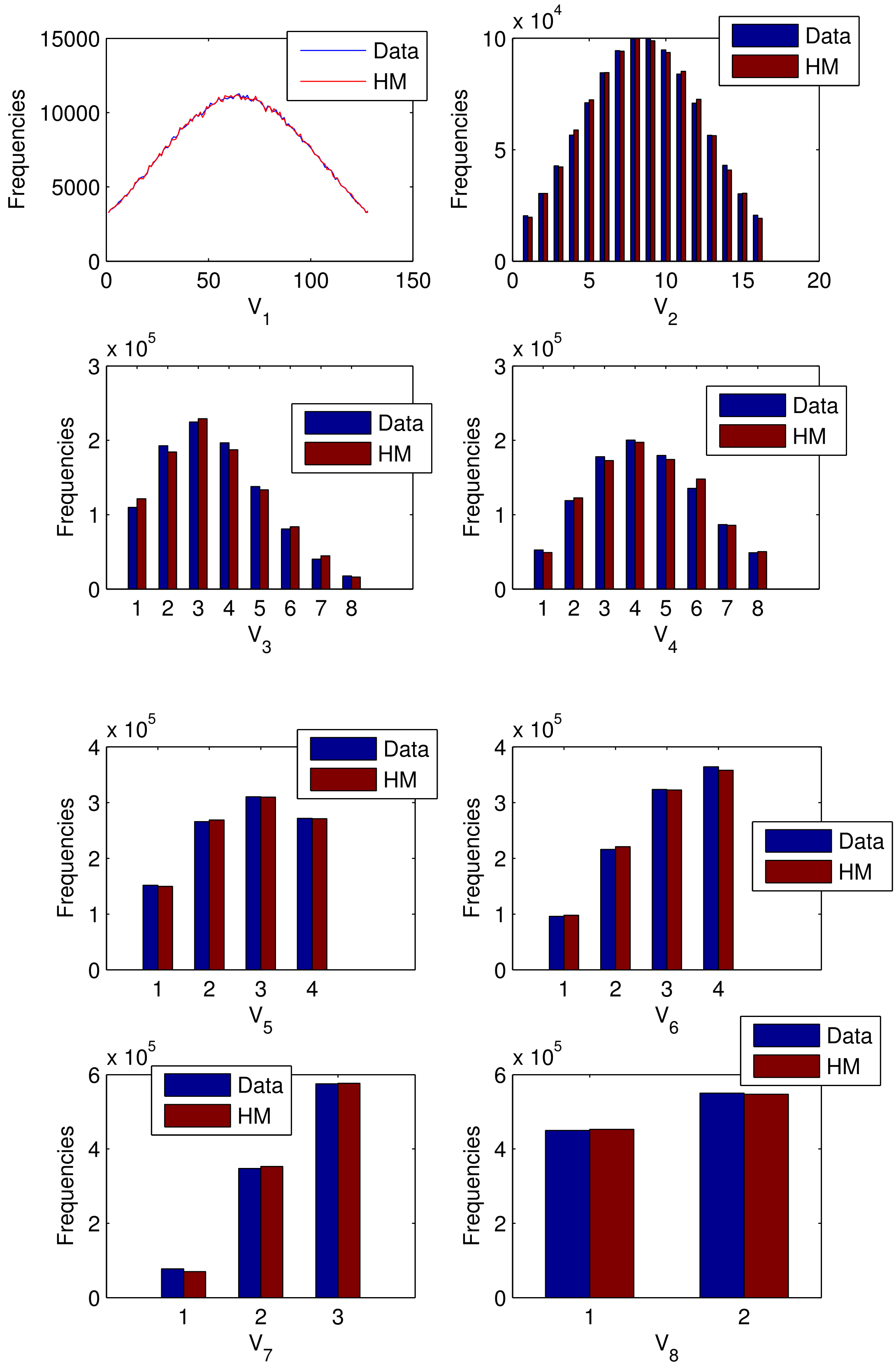
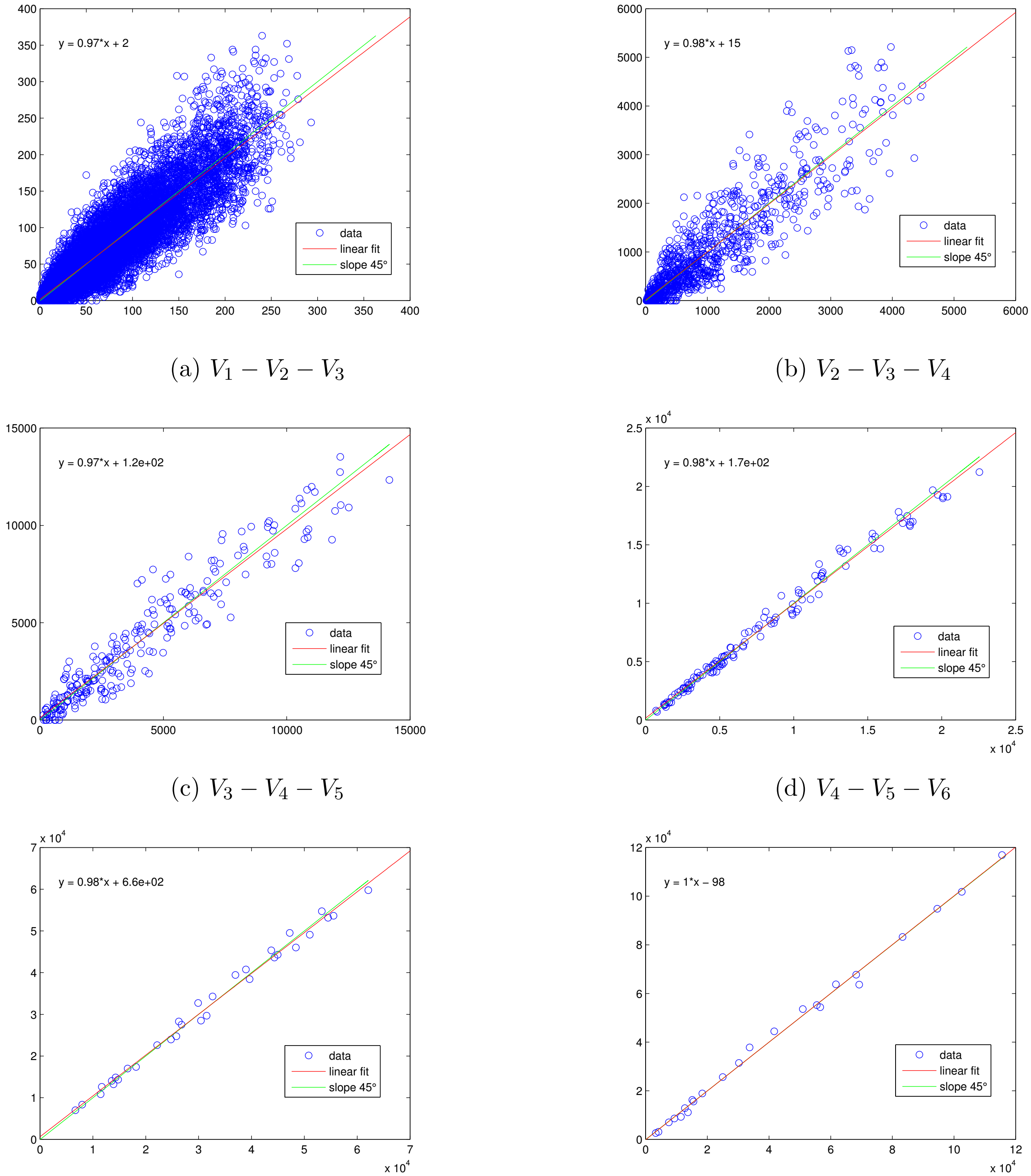


Figure 4: Comparison between the simulated and observed marginals



(a) $V_1 - V_2 - V_3$

(b) $V_2 - V_3 - V_4$

(c) $V_3 - V_4 - V_5$

(d) $V_4 - V_5 - V_6$

(e) $V_5 - V_6 - V_7$

(f) $V_6 - V_7 - V_8$

Figure 5: Comparison between the simulated and observed multi-variate joint distributions

3.3. Implications of the experimental outcomes

The experimental outcomes presented in the current study may have important implications in terms of modeling options. It has been now clearly demonstrated that (a) one should rather use a hierarchical procedure to ensure that the dataset is sufficiently accurate regardless the statistical indicators used. (b) Micro-samples may suffer from a lack of representativeness as combination of attributes with low probability of occurrence may not be captured during data collection. Thus, the HMM component of the HM simultaneously merges multiple datasets in addition to incorporating enough heterogeneity to avoid problems related to representativeness or sampling bias. (c) The presented framework would make the fusion process more straightforward for researchers and practitioners. (d) A major impact in term of expert systems is that the outputs of the HM model serve as a basis for a qualitative and quantitative analyses of integrated datasets. In this context, the decision-making process can be significantly simplified. Advanced knowledge for extracting important information from multiple datasets could be shifted towards a simpler analysis of a unified dataset that incorporates all the information and variable interactions.

3.4. Theoretical comparison

Table 9 compares HM with HMM and IPF in terms of the strengths and weaknesses based on several criteria. Aggregate data, e.g. populations age and gender distributions, are reliable and extremely stable. Disaggregate data, e.g. household travel surveys, provide detailed information about people, but are generally subjected to small sampling rates leading to a serious lack of representativity. HM clearly provides the best trade-off compared to the conventional IPF and the recent HMM-based approach.

	IPF	HMM	HM
Use of aggregate data	Yes	Partial	Yes
Use of disaggregate data	Partial	Yes	Yes
Quasi-perfect marginal distributions	Yes	No	Yes
Accurate multivariate joint distribution	No	Yes	Yes
Information fusion	Partial	Partial	Full

Table 9: Comparison between IPF, HMM and HM

4. Conclusions

In urban and transportation research, key information about agents, i.e. households or individuals, is often included within a wide range of small and independent datasets. To combine the information from these independent datasets, we presented a hierarchical model (HM) for (i) allowing multi-source information fusion and (ii) achieving higher prediction accuracies.

Based on the results highlighted in Section 3, the strengths of the proposed research can be formulated as follows:

- HM provides the best trade-off in terms of RMSE minimization, when marginals and joint distributions are simultaneously compared. This can be explained by the fact that the principal key features of IPF and HMM are combined within a single unified framework.

- Multiple micro-samples and aggregate marginals can be integrated within HM for allowing multi-source information fusion. Also HM shows a lot of flexibility in terms of data availability. We mentioned that a partial set of marginals can be used if there is absolutely no data.

- HM is extremely competitive and relatively robust with respect to sampling rate variability. This means that with a sampling rate of only 1%, it is possible to achieve results which are almost comparable to a HM calibrated with a micro-sample of 10%. Several applications within the field of urban and transportation research assume sampling rates which are around 1% using standard methods, i.e. IPF. But the results presented in Table 7 show that with IPF, a still commonly used method, the RMSE is equal to 13.65. In this context, HM emerges as a far better alternative for mitigating the error in micro-simulation.

Besides, further research is needed to overcome weaknesses of the proposed research method:

- Generalizing the method for handling a wide range of input data format is an important issue. A systematic expert system procedure could be more efficient to enable intelligent data fusion strategies. Indeed, although the developed fusion method provides interesting results, further methodological improvements can be integrated within the modeling framework to make it more universal. At this point, surveys and aggregate-based data are handled by the HM. However, fusing the current data format with other types of data, e.g. panel data, GPS traces of individuals, trip data is still a key challenge.

- The integration of a feature that allows for multi-level data fusion should be investigated. For example, in transportation research, decision-making process can be explained at both household and individual levels. Household data is more aggregated than individual level data.

- To extend the use of the current method within other research fields, additional efforts are needed to ensure that HM is relatively robust to scalability, referred to as the number of variables that should be synthesized. In this regard, an important issue raises up regarding the interaction between scalability and the increase of heterogeneity. Is there a risk of getting a reverse effect?

5. Acknowledgements

The research was funded by the ARC grant for Concerted Research Actions for project no. 13/17-01 entitled "Land-use change and future flood risk: influence of micro-scale spatial patterns (FloodLand)" and by the Special Fund for Research for project no. 5128 entitled "Assessment of sampling variability and aggregation error in transport models", both financed by the French Community of Belgium (Wallonia-Brussels Federation).

6. References

Arentze, T., Timmermans, H., & Hofman, F. (2007). Creating synthetic household populations: problems and approach. *Transportation Research Record: Journal of the Transportation Research Board*, *2014*, 85–91. doi:https://doi.org/10.3141/2014-11.

Axhausen, K. W., & Gärling, T. (1992). Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. *Transport reviews*, *12*, 323–341. doi:http://dx.doi.org/10.1080/01441649208716826.

Barthelemy, J., & Toint, P. L. (2013). Synthetic population generation without a sample. *Transportation Science*, *47*, 266–279. doi:https://doi.org/10.1287/trsc.1120.0408.

Batty, M. (2007). *Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals*. The MIT press.

Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, *30*, 415–429. doi:https://doi.org/10.1016/0965-8564(96)00004-3.

El Faouzi, N.-E., Leung, H., & Kurian, A. (2011). Data fusion in intelligent transportation systems: Progress and challenges–a survey. *Information Fusion*, *12*, 4–10. doi:https://doi.org/10.1016/j.inffus.2010.06.001.

Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, *58*, 243–263. doi:http://dx.doi.org/10.1016/j.trb.2013.09.012.

Horni, A., Nagel, K., & Axhausen, K. W. (2016). *The multi-agent transport simulation MATSim*. doi:https://doi.org/10.5334/baw.

Liu, F., Janssens, D., Cui, J., Wets, G., & Cools, M. (2015). Characterizing activity sequences using profile hidden markov models. *Expert Systems with Applications*, *42*, 5705–5722. doi:https://doi.org/10.1016/j.eswa.2015.02.057.

Liu, F., Janssens, D., Wets, G., & Cools, M. (2013). Annotating mobile phone location data with activity purposes using machine learning algorithms. *Expert Systems with Applications*, *40*, 3299–3311. doi:https://doi.org/10.1016/j.eswa.2012.12.100.

Mosteller, F. (1968). Association and estimation in contingency tables. *Journal of the American Statistical Association*, *63*, 1–28. doi:https://doi.org/10.1080/01621459.1968.11009219.

Saadi, I., Mustafa, A., Teller, J., & Cools, M. (2016a). Forecasting travel behavior using markov chains-based approaches. *Transportation Research Part C: Emerging Technologies*, *69*, 402–417. doi:https://doi.org/10.1016/j.trc.2016.06.020.

Saadi, I., Mustafa, A., Teller, J., & Cools, M. (2017). Investigating the impact of river floods on travel demand based on an agent-based modeling approach: The case of liège, belgium. *Transport Policy*, . doi:https://doi.org/10.1016/j.tranpol.2017.09.009.

Saadi, I., Mustafa, A., Teller, J., Farooq, B., & Cools, M. (2016b). Hidden markov model-based population synthesis. *Transportation Research Part B: Methodological*, *90*, 1–21. doi:https://doi.org/10.1016/j.trb.2016.04.007.

Voas, D., & Williamson, P. (2000). An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *Population, Space and Place*, *6*, 349–366. doi:https://doi.org/10.1002/1099-1220(200009/10)6:5<349::AID-IJPG196>3.0.CO;2-5.

Wu, S. (2009). Applying statistical principles to data fusion in information retrieval. *Expert Systems with Applications*, *36*, 2997–3006. doi:https://doi.org/10.1016/j.eswa.2008.01.019.

Ye, P., Hu, X., Yuan, Y., Wang, F.-Y. et al. (2017). Population synthesis based on joint distribution inference without disaggregate samples. *Journal of Artificial Societies and Social Simulation*, *20*, 1–16.

Zhu, Y., & Ferreira, J. (2014). Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. *Transportation Research Record: Journal of the Transportation Research Board*, *2429*, 168–177. doi:https://doi.org/10.3141/2429-18.