doi.org/10.26398/IJAS.0029-009

SIMULTANEOUS ANALYSIS FOR AUDIENCE PROFILING AND AUDIENCE PREDICTION

Enzo Fenoglio[1]
Cisco Systems France - Chief Technology and Architecture Office, Paris Innovation & Research Lab, Issy-les-Moulineaux, France

Abstract. *TV platform operators have demographic insight of the primary account holder of a household but know little about the other residents. Operators and advertisers have to access the complete set of demographic profiles, but also estimate the demographic composition for each subscriber household to provide better services. This paper describes a methodology for audience profiling and audience predictions. Simultaneous Analysis (SA) of contingency tables is applied to the aggregated contingency tables of users viewing event counts for multiple guidebooks; logistic regression is used on the SA factor space to estimate the household compositions for single-membership households; multi-membership household demographic composition is estimated by a generative model of multiple distributions of single-membership household profiles; audience prediction is estimated as a weighted sum of class percentage of single and multi-users groups.*

Keywords: *Simultaneous analysis, Logistic regression, Audience profiling, Demographic prediction.*

1. INTRODUCTION

TV platform operators tend to have full demographic insight of the primary account holder of a household but know little about the other residents of the household which consume their service. An operator needs to have complete set of demographic profiles within each subscriber household to improve audience measurement reach across gender and age breakdowns and ensure a match between target groups and audience composition selecting broadcast services. Demographic user profiles aid marketing managers in their communication channel choices and allow for a closer match between demographic target groups and message receiving audiences, resulting in higher advertising effectiveness for advertisement campaigns and audience measurement. For example, information of

[1] Corresponding author: Enzo Fenoglio, email: efenogli@cisco.com

interest may be the percentage of male and female viewers for a given program or the age distributions of male and female viewers, but also, education and profession distribution or seasonal audience variations.

Classification of subscribers into specific audience groups may help in the estimation of where service is happening to be consumed, reducing the uncertainties of aggregate projections. One major factor affecting service consumption is the typology of the population which generates it, generally described by a set of categorical variables such as educational level, income, household structure, etc. By taking into account differences in population characteristics, one can better analyse expected differences in service consumption. Some studies about ads forecast strategies, which demonstrate the usefulness and flexibility of such approaches have been undertaken by service providers. However, such studies which account for the characteristics of the consumers and their households in the determination of consumer patterns and in the classification of populations into service consumption groups are not common practice yet. This is probably due to the scarcity of adequate dataset or the difficulties involved in merging data coming from different households at different aggregation levels. One of the issues with audience prediction is that the input dataset is loosely structured with a very high dimensionality and somehow the data are missed or reported incorrectly and it is important to perform some data reduction of the original dataset before applying classification methods. We will see in Section 5.2 that better classification results are obtained if the researcher uses factor scores as covariate in regression analysis instead of the original categorical covariates (Saporta and Niang, 2006). Another issue with data surveying is that operators are the more and more confronted to the problem of providing accurate report of users consumption while preserving the privacy of each user. We will see how aggregating data to pre-defined user groups is possible to protect subscriber privacy and at the same time, estimate the socio-economic profile of view ers and the characteristics of households that generate the service consumption. The methodology described in this paper is largely inspired by multiblock discriminant correspondence analysis (Williams et al., 2010) as we use pre-defined groups of subscribers and a set of contingency tables representing aggregated viewing event counts. However, we use simultaneous analysis of contingency tables (SA) (Zarraga and Goitisolo, 2009) to nicely describe the partial and the global structure of the tables and logistic regression with the factors obtained as covariates for prediction of unseen subscriber profiles.

The paper is organized as follows. Section 3, briefly describes the problem to solve with Audience profiling. Section 4, describes the dataset structure and,

in particular, the household and the audience datasets used to create the ground truth. Section 5, describes the four steps methodology applied for audience profiling and audience prediction. We illustrate in Section 6 the methodology on a real case example, and in Section 7 we provide some concluding remarks. Finally, an appendix provides some theoretical background of barycentric analysis and of the Principle of Distributional Equivalence that have been extensively used in this work.

2. RELATED WORK

Audience profiling and quantitative audience research is a well-established practice to decide which channels and timeslots the service providers will pay to run ads based on the demographic of the audience (Ballantine et al., 2006). An interesting research that shows the methodologies and the techniques employed in measuring audiences can be found in Bornman (2009). As for audience data exploration there are few studies that use correspondence analysis (CA) for TV audience profiling as in Redfern (2012) but many that use CA for demographics. As for prediction, it is common practice to use the original data set to build feature vectors as input for a regression method, as in De Bock and Van den Poel (2010) which we follow but only for the multimember household predictive approach.

3. THE SYSTEM MODEL

Audience profiling predictions is made possible by the availability of big data storage of viewing actions captured from set-top boxes or other connected consumer devices, used by subscribers to view content broadcast by a service operator platform. The ingested viewing actions are processed to identify the channel, the content, the time of viewership, and other parameters of interest. Broadcast event viewing actions are analysed to identify viewing activities that have been watched by the subscriber for a minimum amount of time; this is designed to capture only full content engagement activities and filter out channel surfing activity that will negatively bias the dataset.

Actually, the service operator has a list of the primary account holders of a household that have subscribed but typically no information on the other residents which consume the service. To ensure the direct association between viewing activity and household composition of the primary account holder to the viewers in the household, we will consider households with only one resident and referred

in the following as *1-membership household*. The data used to validate the model will include only 1-membership household and will be used to generate a complete household viewing model as a mixture of distributions of 1-membership household profiles. In Figure 1, the distribution of residents per households captured for three months of audience measures by a French IPTV operator is presented. The households with only one member represent 47% and justifies in this case the assumption to use 1-membership household mixture composition to model multi user households. While the data used to construct the model set only includes single users and the classification performance is validated on single users, multi users households will be included in the model combining single users and their corresponding viewing impression counters.

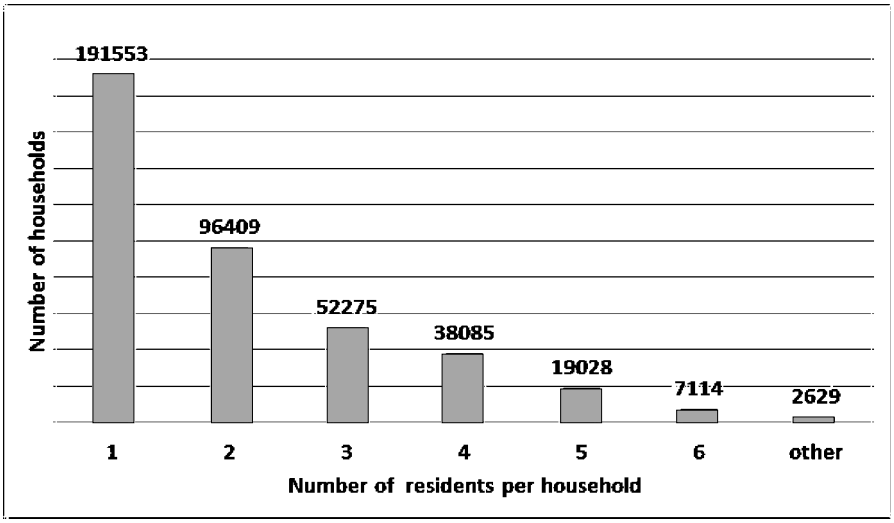

Figure1: Panelhouseholdresidentsdistribution(sourceFrenchIPTVprovider2013)

Within the 1-membership household model, the service operator provides a list of 1-membership household accounts together with the declared demographic attributes for age and gender indexed by the household resident unique identifier (HHID). The corresponding 1-membership household viewing activities are captured and grouped by subscriber account and aggregated into demographic guidebooks. A minimum time period of three months of captured activity is required, although more calendar coverage yields usually to improved prediction accuracy. Seasonal models can be added to encapsulate viewing activities occurred correlating seasonal calendar months. The demographic guidebook provides viewing

event counts of demographic attributes (e.g. Gender, Age-Band) for individuals who simultaneously choose the corresponding viewing activity attributes (Content, Genre, Channel, Time of Day, Day of Week) for a certain viewing event. The viewing event counts are derived from measurements of k nown demographic attributes viewership to viewing activity attributes across a representative sample of the subscribers base. This particular guidebook generation technique depends on the availability of ground-truth data for a sample of 1-membership households across the subscriber base. The ground-truth data can contain the gender and age band of the individual resident of the 1 membership household and other demographic attributes like education, profession, income, etc.

4. THE DATASETS USED

A number of different datasets are acquired through various forms of collaborations and processed to provide the audience profiling service. The Audience Profiling Service principal datasets are listed in Table 1 with a short description. However, the most important are the *Household Panel Dataset* and the *Audience Panel Dataset*.

Table 1: Datasets used by the Audience profiling service

Database	Description
Viewer Action	Usage reports viewer actions
VOD Catalogue	Schedule and content metadata for VOD events
Linear TV Catalogue	Schedule and content metadata for live viewing events
Household Panel	User account attributes including subscription identifier, household composition and demographic segments.
Audience Panel	Ground truth data for a sample of the service provider audience including viewing impression counters

The *Household Panel Dataset* (Table 2) is used to build the demographic guidebooks to provide for each household unique identifier (HHID), the household composition and the household demographic segment. The HHID is anonymized and not directly associated with the actual householder subscription identifier to enforce privacy.

Table 2: The Household Panel Dataset

HHID	HOUSEHOLD COMPOSITION	DEMOGRAPHIC SEGMENT
Household unique identifier	Total number members, socio/economic segment, number of males, number of females, presence of teenagers, presence of kids, number of TV sets, smartphnes, tablets, laptops, etc.	Number of females/males grouped per age bands

The *Audience Panel Dataset* (Table 3) is used to create the ground truth datasets to validate the audience prediction model and build the Viewing Event Attributes for descriptive profiling. The demographic guidebooks provide the measured viewing events for the viewing events attributes for a particular HHID in the form of an indicator matrix.

Table 3: Viewing Events Attributes

VIEWING ATTRIBUTES (categorical variables)	**MODALITIES**	**DESCRIPTION**
CONTENT	List provided by the service operator	Indicator matrix of a specific content title for viewing preferences of specific live viewing content titles
GENRE	List provided by the service operator	Indicator matrix of a specific content genre for viewing preferences such as sport, drama, movie. music, etc
CHANNEL	List provided by the service operator	Indicator matrix of a specific channel for input into viewing preferences of selected channels
DAY	Mon, Tue, Wed, Thu, Fri, Sat, Sun	Indicator matrix for watching content on a day of the week for input into detecting preferred week of the day
TIME	H00, H01,. . . , . . . , H22, H23	Indicator matrix for watching content for hourly preconfigured time segments for input into detecting preferred time of day

4.1. THE BURST SUB-TABLE

During a typical surveying campaign, subscribers are requested to answer to questions about their viewing experience and provide the corresponding demographic attributes. Each of the m answers is collected into an indicator matrix of individuals-by-categories indexed by HHID, partitioned into n columns, one for each of the

modalities in which the viewing categories for the viewing guidebooks (Content, Genre, Channel, Day, Time) have been instantiated, together with other p columns one for each of the modalities corresponding to the demographic attributes (household composition and demographic attributes). The overall table can be represented by a complete matrix $\mathbf{Z}_v$ obtained stacking by row the indicator matrices of the viewing attributes, and the indicator matrix $\mathbf{Z}_d$ of the demographic attributes, as indicated in Table 4. The Burt table can be readily obtained as follows:

$$[\mathbf{Z_v} \quad \mathbf{Z_d}]^T [\mathbf{Z_v} \quad \mathbf{Z_d}] = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix} \tag{1}$$

However, we are interested just into a sub-table of the complete Burt table, derived as follows:

$$\mathbf{B}_{21} = \mathbf{Z_d}^T \mathbf{Z_v} \tag{2}$$

Table 4: The audience indicator matrix

	$\mathbf{Z_v}$ (m × n) Viewing Attributes					**Zd** (m × p) Demographic Attributes	
	CONTENT	**GENRE**	**CHANNEL**	**DAY**	**TIME**	**GENDER**	**AGE**
m	C_{max}	G_{max}	Ch_{max}	7	24	2	3
	-	-	-	Mon, Tue, . . . , Sun	H00, . . . , H23	M,F	A25, A50, A5
	n = C_{max} + G_{max} + Ch_{max} + 7 + 24					p=5 for Gender and Age	
	C_{max} : total number of Content titles					p=6 for compound Gender-Age	
	G_{max} : total number of Genres						
	Ch_{max} : total number of channels broadcast by the operator						

The Burt sub-table $\mathbf{B}_{21}$ has some attractive properties that models quite well how advertisers or service providers use audience profiling services. Actually, advertisers or service providers are not interested in any specific user, but rather in getting a global view of how many people actually use the service. Here we are facing two contradictory requirements: from one side, if users are not tracked they cannot enjoy a personalised viewing experience; and from the other side, if users accept to expose private information to the service provider, they must be assured that nobody will exploit their private information, service preferences or personal habits. One way to enforce individuals privacy protection is to collect all user reports viewing activity and de-identify by aggregation. We assume also that the profiles in a same group of individuals are similar and that the aggregation is a way to increase association and not to remove valuable information according

to the *Principle of Distributional Equivalence* (PDE)[2]. Eventually, all these requirements are nicely satisfied by the Burt sub-table $\mathbf{B}_{21}$ that enforces privacy by aggregation with respect to pre-assigned demographic groups $\mathbf{Z_d}$ and provides the best discrimination among the viewing attributes $\mathbf{Z_v}$.

For the demographic attributes we may use two categorical variables Gender=(M,F) and Age=(A25, A50, A5P)[3] with $p = 5$ levels as defined in Table 4 that naturally comes from the survey reports, or we may use the single compound categorical variable *Gender-Age*=(F25, F50, F5P, M25, M50, M5P)[4] with $p = 6$ levels. In this paper, we will always use the compound categorical variable *Gender-Age* which is the natural choice when the main objective is to investigate the variations among the viewing audience based on both gender and age more than the specific variations on gender or age (Redfern, 2012)

5. THE METHODOLOGY

The methodology used for audience profiling and demographic predictions can be described in four steps:

1. Simultaneous Analysis of Contingency Tables (SA) is applied to the aggregated contingency tables (viewing guidebooks) of viewing event counts
2. Multinomial logistic regression is used to the SA factor space variables computed by the partial analysis on each guidebook to estimate the households compositions in the limit of 1-membership household
3. Multimember household demographic composition is estimated by a generative model as a mixture distribution of 1-membership household profiles
4. Finally, the demographic class percentage for each demographic group is calculated as a weighted sum of the actual class percentage of the single users and the actual class percentages of the multi user groups.

---

[2] PDE is unique to Correspondence Analysis and states that if two row profiles are identical then the corresponding two rows can be replaced by their sum without affecting the geometry of the column profile (Greenacre and Lewi, 2009).

[3] A25: individual aged between 16 and 25 years; A50: individual aged between 26 and 50 years; A5P: individual aged above 50 years.

[4] F25: female aged between 16 and 25 years; F50: female aged between 26 and 50 years; F5P: female aged above 50 years; M25: male aged between 16 and 25 years; M50: male aged betwen 26 and 50 years; M5P: male aged above 50 years.

The use of a mixture of 1-membership profiles to estimate multi member composition is very effective but suffers of two main design limitations that will become our working assumptions:

1. A resident of the multi member household may have multiple distinct viewing activity patterns that can be associated with multiple clusters in the learnt model and consequently to multiple virtual 1-membership household contribution. We will assume that 1-membership profiles have no joint relationships and that a multi member profile is the aggregation of elementary 1-membership profiles (*profiles independence assumption*).
2. Some multi member profiles not used in the 1-membership household training phase will not be identified. In particular, this applies to residents not present in a pure 1-membership household such as minors aged less than 15 years living with their parents that are present in the ground truth dataset. We will limit the analysis to multi member households without minors (*in-sample profiles assumption*) .

5.1. STEP ONE: SIMULTANEOUS ANALYSIS

Simultaneous Analysis (SA) (Zárraga and Goitisolo, 2006) is a factorial method developed for the joint treatment of a set of several data tables without modifying the internal structure of each table, especially frequency tables whose row margins are different to each other. It is typically employed for different attributes representing the same group of individuals but unlikely to MFACT the row margins may be different. This is an attractive property of the method because audience profiling reports may be collected at different time points, as it happens when data refers to audience surveys collected at different epochs. Moreover, new tables can be added and used to perform joint analysis of variables even with different measurement scales (Zarraga and Goitisolo, 2009). In the following, we summarize how SA on multiple contingency tables is carried out, addressing the interested reader to the specialized references for the implementation details (Zarraga et al., 2013).

Let $\mathbf{E} = \{1, \ldots, e, \ldots, E\}$ be the set of $E = card(\mathbf{E})$ contingency tables representing the viewing guidebooks (Content, Genre, Channel, Day, Time) described in Section 4.1. Each of them classifies the aggregated answers of $a_{..e}$ Gender-Age groups with respect to two categorical variables: the demographic variable for rows and the viewing guidebooks variable for columns. All the tables have the row variable with demographic attributes $\mathbf{C_d} = \{F25, F50, F5P, M25, M50, M5P\}$

in common, while the column variable with viewing attributes $\mathbf{C_v} = \{1,\ldots,J_e\}$ depends to the viewing guidebook e observed. After concatenating all the guidebooks, a joint set of columns $\mathbf{J} = \{1,\ldots,j,\ldots,J\}$ represented by the matrix $\mathbf{B}_{21}$ of size $6 \times n$ (2) is obtained, where the element a_{cje} corresponds to the total number of individuals in the demographic group $c \in \mathbf{C}_d$ who choose the viewing attribute $j \in \mathbf{C_v}$, for the viewing guidebook $e \in \mathbf{E}$. Upon these definitions, SA proceeds as follows:

- *Separate analysis*: A classical CA on each of the E tables is performed to check for the existence of structures common to the different viewing categories. CA on table e is done computing the SVD of the matrix of residuals $\mathbf{S}^e$ of size $p \times J_e$ (see also Section A.1) with $J_e = (C_{max}, G_{max}, Ch_{max}, 7, 24)$, and actually $\mathbf{S_b} = \left[\mathbf{S}^1 \ldots \mathbf{S}^e \ldots \mathbf{S}^E\right]$, Section A.1. We compute the first squared singular value λ_1^e, the row margins $\mathbf{r}^e$, and the column margins $\mathbf{c}^e$ for each table $e \in \mathbf{E}$, which will be used in the next steps:

- *Joint analysis*: it is performed comparing the points corresponding to the same row in the different tables by computing the weighted PCA of the matrix $\mathbf{S}$ obtained concatenating the tables $\sqrt{\alpha_e}\mathbf{S}^e$ by rows:

$$\mathbf{S} = \left[\sqrt{\alpha_1}\mathbf{S}^1 \ldots \sqrt{\alpha_e}\mathbf{S}^e \ldots \sqrt{\alpha_E}\mathbf{S}^E\right] = \mathbf{U}\mathbf{D}_k\mathbf{V}^T \quad \forall e \in \mathbf{E} \tag{3}$$

  where, $\mathbf{D}_k$ is the diagonal matrix of singular values, $k = p-1$ is the rank of matrix $\mathbf{S}$, the weight α_e is included to balance the influence of each table. Different values of the weighting α_e are possible and are discussed in Zarraga et al. (2013). The most frequently used, and adopted in this work, is $\alpha_e = 1/\lambda_1^e$, where λ_1^e is the first eigenvalue (square of first singular value) of table $\mathbf{S}^e$.

- *Partial row analysis*: The projection along the s^{th} principal axis of all partial rows for table $\mathbf{S}^e$, denoted by $\mathbf{F}_s^e$, is computed as follows:

$$\mathbf{F}_s^e = (\mathbf{D}_r^e)^{-\frac{1}{2}} \left[\mathbf{0} \ldots \frac{\mathbf{S}^e}{\sqrt{\lambda_1^e}} \ldots \mathbf{0}\right] \mathbf{v}_s \tag{4}$$

  where, $\mathbf{D}_r^e = diag(\mathbf{r}^e)$ is the diagonal matrix of size $p \times p$ of row masses for table e, and $\mathbf{v}_s$ is the s^{th} eigenvector (column) of matrix $\mathbf{V} = \left[\mathbf{v}_1 \ldots \mathbf{v}_s \ldots \mathbf{v}_k\right]$. Actually, the column vector $\mathbf{v}_s$ can be written $\mathbf{v}_s = \left[\mathbf{v}_s^1 \ldots \mathbf{v}_s^e \ldots \mathbf{v}_s^E\right]^T$ to show the contribution of the various tables.

- *Global Analysis*: the projection $\mathbf{F}_s$ for the aggregated demographic profiles in the barycentric space is the weighted average of the projections of the partial rows $\mathbf{F}_s^e$. It takes into account the balanced contribution of each guidebook and is computed as follows:

$$\mathbf{F}_s = (\mathbf{D}_w)^{-1}\mathbf{S}\mathbf{v}_s = (\mathbf{D}_w)^{-1}\sum_{e\in\mathbf{E}}(\mathbf{D}_r^e)^{\frac{1}{2}}\mathbf{F}_s^e \tag{5}$$

where, $(\mathbf{D}_w)^{-1} = diag\left(1/\sum_{e\in\mathbf{E}}\sqrt{\mathbf{r}^e}\right)$

When the partial row margin is equal in all tables, as in our case, (5) simplifies to the average of the projections of the partial projections $\mathbf{F}_s = \frac{1}{|E|}\sum_{e\in\mathbf{E}}\mathbf{F}_s^e$

- *Inertia*: It is interesting to evaluate the projected inertia I_s^e of columns of table e on the s^{th} axis using the principal coordinates $\mathbf{G}_s^e$ of columns:

$$\mathbf{G}_s^e = \left(\mathbf{D}_c^e\right)^{-\frac{1}{2}}\mathbf{v}_s^e\sqrt{\lambda_s} \tag{6a}$$

$$I_s^e = \mathrm{Tr}\left(\left(\mathbf{G}_s^e\right)^T\mathbf{D}_c^e\mathbf{G}_s^e\right) \leq \lambda_s \tag{6b}$$

where, $\mathbf{v}_s^e$ is the right eigenvector for table e on the s^{th} axis, $\sqrt{\lambda_s} = (\mathbf{D}_k)_{ss}$, and $\mathbf{D}_c^e = diag(\mathbf{c}^e)$. The influence of tables is well balanced as $I_s^e/\lambda_s \leq 1$, $\forall e \in \mathbf{E}$ due to the weighting $\alpha_e = 1/\lambda_1^e$

5.2. STEP TWO: SINGLE VIEWER PREDICTION

We follow the proposal described in (Saporta and Niang, 2006) to perform logistic regression (LR) on the factors obtained by MCA that we will adapt to the partial analysis obtained by SA. LR can easily incorporate categorical predictors as covariates and, according to Saporta and Niang (2006), performs better than linear discriminant analysis when the conditional distributions are not normal or have different covariances. The proposed method involves training the model to estimate the coefficients of the LR model for gender and age, followed by a n-fold cross validation and predictions on out-of-sample feature vectors. The combination of demographic information and viewing event counts is used as input to the LR, collected by service providers as usage reports, and transformed into predictive features. Training is performed using the principal coordinates of matrix $\mathbf{F}^e$ of size $p \times k$, stacked by row for each of the demographic classes $c \in \mathbf{C_d}$ to gener

ate the covariates. Prediction involves using out-of-sample subscriber profile and it is done as follows:

- the impression counters for each of the corresponding guidebooks (Content, Genre, Channel, Day, Time) form the partial profiles (feature vector);
- The partial profiles are projected as supplementary variables for table e on the principal axis calculated as principal coordinates $\mathbf{F}^e = \left[\mathbf{F}^e_1 \dots \mathbf{F}^e_s \dots \mathbf{F}^e_p\right]^T$;
- The projections $\mathbf{F}^e$ are stacked by row (i.e. by demographic class), where each of the p rows is used to compose the feature vector of predictors for class c.

5.3. STEP THREE: MULTI VIEWER PREDICTION

Multi-viewer prediction will be derived adapting the method described in De Bock and Van den Poel (2010) that analyses the relationship between online preferences, browsing behaviour and demographic characteristics of website audience, including age and gender. In our case, the browsing behaviour and the click streams collected are replaced by household viewing events counts for linear TV programs, and website audience is replaced by IPTV network audience. We assume that multi-member row profiles $\mathbf{XM}$ of size $(1 \times n)$ for all guidebooks viewing attributes (see Table 4) are generated from a mixture of 1-membership barycentric row profiles $(\mathbf{P_b})_c$ of size $(1 \times n)$ for each demographic household class $c \in \mathbf{C_d}$, as follows :

$$\mathbf{XM} = \frac{1}{f_s} \sum_{c \in \mathbf{C_d}} h_c \frac{(\mathbf{P_b})_c}{(r_b)_c} = \frac{1}{f_s} \mathbf{h} \, (\mathbf{D_b})_r^{-1} \mathbf{P_b} \tag{7}$$

where, $\mathbf{C_d} = \{F25, F50, F5P, M25, M50, M5P\}$, $(\mathbf{P_b})_c$ is the barycentric row vector profile for class c of the barycentric matrix $\mathbf{P_b} \in \mathbb{R}^{p \times n}_{\geq 0}$ for p=6 defined in Section A.1, the vector $\mathbf{h} = \{h_c\}$ of size $1 \times p$ is the household members composition, and $f_s = \sum_{c \in \mathbf{C_d}} h_c$ is the family size that may be obtained by a direct question included in the survey.

We have an overdetermined problem for which we seek to minimize the Euclidean distance among the multiuser profile projections and the single user factor scores in the barycentric space. To situate the multi-user row profile in the barycentric space we will use the transition equation (Greenacre and Blasius, 2006) and project the row profiles as supplementary rows using (12c) applied to both members of (7). This optimization problem can be formulated as follows:

$$\begin{cases} \underset{\mathbf{h}}{\text{argmin}} & \frac{1}{2}\left\| \frac{\mathbf{h}}{f_s}\mathbf{F_b} - F_b(\mathbf{XM}) \right\|_F^2 \\ \text{subject to} & \sum_{c \in C_d} h_c = f_s \\ & \mathbf{h} \geq \mathbf{0}. \end{cases} \tag{8}$$

Eq.(8) completely describes the mixture problem in the unknowns $\mathbf{h}$, and the barycentric factor score matrix $\mathbf{F_b}$, defined in Eq.(10b), as a non-negative least square problem (NNLS) (Chen and Plemmons, 2009) that can be solved efficiently by quadratic programming optimization methods (Branch and Grace, 2002). Actually, in the implementation, the factor $\mathbf{F_b}$ was replaced by the global projections $\mathbf{F}_s \; \forall s \in \{1 \ldots k\}$ for the aggregated demographic profiles in the barycentric space, i.e. $\mathbf{F_b} = (\mathbf{D}_w)^{-1}\mathbf{SV}$, see Section 5.1 and Equation (5).

5.4. STEP FOUR: AUDIENCE PREDICTION

As already stated in Section 4.1, we are not interested to know the exact composition of households, but instead we want to estimate the IPTV network audience demographic class percentages P_p. We consider five multi-member household groups with 2 to 6 members (Figure 1) and for each group we compute the predicted P_p demographic class membership percentage as a weighted sum of the predicted class percentage of the single users P_S and the predicted class percentage of the multi user P_M groups for each class $c \in \mathbf{C_d}$, as described in De Bock and Van den Poel (2010):

$$P_p(c) = \beta\, P_S(c) + (1-\beta) \sum_{f_s=2}^{6} m_g\, P_M(c) \quad \forall c \in \mathbf{C_d} \tag{9}$$

where:

- P_S : the percentage of single users for demographic class c
- P_M : the average percentage of multi-user groups for demographic class c
- m_g : family distribution size for multi-member households in the panel
- β : the percentage of 1-membership households in the panel.

The class percentage P_M of the multiuser group with f_s members is the average of predicted class membership probabilities for each group, i.e. $P_M(c) = \mathbb{E}(h_c/f_s)$. The weights m_g, are used to combine multi-user class percentages of the different multi-user groups and signal the relative importance of each multi-

user group in the final audience profile and approximated by the distribution in family size among respondents which identify themselves as members of multi-user groups (De Bock and Van den Poel, 2010).

6. ILLUSTRATIVE EXAMPLE

The proof-of-concept was requested by a French IPTV operator that provided the survey reports for subscribers usage collected for one month of VOD and for three months of linear TV programs. The survey collected 400K observations and we kept the 200K observations concerning 1-membership householders for the five guidebooks reporting viewing impression counters. Households with less than 50 answers were filtered out. The viewing categorical variables with the corresponding modalities (levels) are indicated in the following table:

CONTENT	GENRE	CHANNEL	DAY	TIME
C_{max}=5979	G_{max}=105	Ch_{max} = 349	7	24

The aggregated contingency table has 6 rows and 6464 columns for each of the 6 explicative variables: 5 viewing activity attributes (Content, Genre, Channel, Day of the week, Time of the day) $\times$ 6464 modalities and one demographic compound attribute (Gender-Age) $\times$ 6 modalities (F25, F50, F5P, M25, M50, M5P). This contingency table was analysed with SA giving partial row $\mathbf{F}^e$ solution with 5 factors for each table or 25 factors after stacking. The global factors can be estimated with the general formula or by averaging the corresponding partial factors for the five tables. It is easy to verify that this is true only when the table margins are the same, as in our case.

Factor	λ	τ (%)	Cumulated (%)
F1	4.48	46.4	46.4
F2	1.78	18.5	64.9
F3	1.75	18.2	83.1
F4	0.97	10.0	93.1
F5	0.66	6.9	100.0

In the 5-dimensional factor space, we have 6 centroids (one for each of the 6 modalities expressed by the demographic variable Gender-Age). The contribution

of each of the Guidebook is represented by a satellite of points around the centroid as part of the separate and joint stages described in Section 5.1. The biplot for Factor 1 vs. Factor 2 (see Figure 2 - left) gives the principal two directions and explains already 65% of the overall dataset variance.

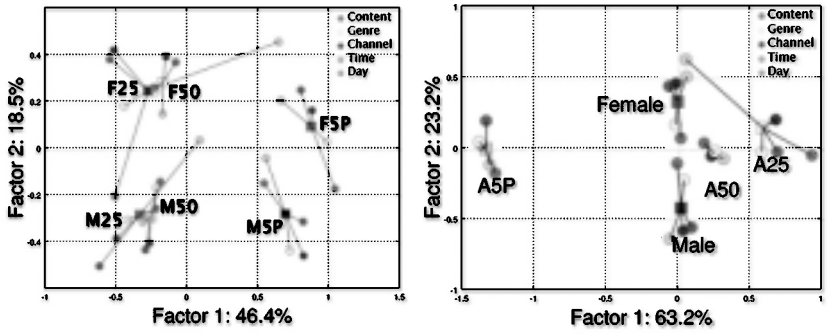

Figure 2: Biplotsgender-age(left)andgender/age(right)

Gender separation is quite evident as well as the association between individuals above 50 years (F5P and M5P) and Female or Male below 50 years and below 25 years respectively (F25,M25). However, viewers above 50 years make different choices for content selection. Females below 25 years (F25) have the greatest variance for day of the week and time. For comparison, on Figure 2-right we present the analysis using two demographic variables Gender (M,F) and Age (A25,A50,A5P). As expected the inertia explained by the two main factors is greater than the inertia explained using the combined variable Gender-Age: it represents the 86% of the total variance. The gender group is contrasted and shows the Female group and the Male group on opposite side on the vertical axis that is representative of Gender for the 23% of the total inertia. The group A25 has the highest variance among the viewing attributes as opposed to the group A50 that has the lowest variance. Overall, the horizontal a xis is representative of the Age group and explains the 63% of the total inertia. The 25 factor scores computed with the partial analysis have been used as a predictor feature vector for the multinomial logistic regression stacking the contribution of the five guidebooks. We have used WEKA® (Hall et al., 2009) for running the logistic regression using a ridge parameter of 10^{-8} and 20-fold cross-validation test mode. We classified correctly 2517 (79%) out of 3186 total instances (1-membership householders). The best partial score was obtained for the class of female below 25 years (F25) with an F-score of 81%. The overall classification results are found in the following table:

Class	Precision	Recall	F-Score	ROC Area
F25	0.85	0.78	0.81	0.99
F50	0.76	0.82	0.79	0.92
F5P	0.83	0.76	0.79	0.96
M25	0.83	0.76	0.79	0.99
M50	0.79	0.78	0.79	0.93
M5P	0.85	0.74	0.79	0.97

Multimember audience segmentation was tested using a ground truth dataset that complied with the in-sample profiles assumption described in Section 5, i.e profiles without minors (F15 and M15). The percentage of 1-membership household for the panel composition shown in Figure 1 was estimated to β=0.47, the distribution in family size among multi-member household is $m_2 = 0.45, m_3 = 0.25, m_4 = 0.18, m_5 = 0.09, m_6 = 0.03$. The model accuracy for audience predictions was on average 70% for the six demographic classes realizations.

7. CONCLUSIONS AND FUTURE WORK

The results obtained have shown the importance of Simultaneous Analysis (SA) as an intermediate step to estimate coefficients used in classification for concatenated contingency tables. The use of SA factors as predictors for classification not only improves the performance of the method, but permits its application to categorical predictors as already shown for MCA (Saporta and Niang, 2006). The primary use of the methodology described in this paper is to deliver audience demographic predictions to support service providers audience measurements or personalized advertisement. The method can be used in the estimation of the demographic composition of program audiences and in the classification of a generic household with unknown member composition as a mixture of 1-membership households under the assumptions of *profiles independence* and *in-sample profiles* (Section 5) that we plan to remove in a future work. Moreover, the methodology can be appli ed to other classification problems that can take advantage of Correspondence Analysis and of the Principle of Distributional Equivalence, like systems faults isolation and detection (Detroja et al., 2011).

A. OBSERVATION AND BARYCENTRIC SPACES

Given the matrix of observations $\mathbf{P_o}$ of size $m \times n$ and the associated matrix of standardized residuals $\mathbf{S_o}$ of size $m \times n$, the singular value decomposition (SVD) of a matrix $\mathbf{S_o}$ is $\mathbf{U_o}\,\mathbf{D_o}\,{\mathbf{V_o}}^T$, where the orthogonal matrix $\mathbf{U_o} \in \mathbb{R}^{m\times m}$, the diagonal matrix of positive singular values $\mathbf{D_o} = diag(\sigma_1, \sigma_2, \ldots, \sigma_{min(m,n)}) \in \mathbb{R}^{m\times n}$, with $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_{min(m,n)-1} \geq 0$, and the orthogonal matrix $\mathbf{V_o} \in \mathbb{R}^{n\times n}$. The rank of $\mathbf{S_o}$ is the number of non zero singular values, i.e. $k_o = min(m,n) - 1$ (Greenacre and Blasius, 2006) .

A.1. BARYCENTRIC ANALYSIS

We want to see how the coordinates of rows and columns profiles in the principal axis change when the rows (profiles) are aggregated by the indicator matrix ${\mathbf{Z_d}}^T$ as described on Section 4.1. We rewrite Eq.(2) as $\mathbf{P_b} = {\mathbf{Z_d}}^T\,\mathbf{P_o}$, by defining $\mathbf{Z_v} \triangleq \mathbf{P_o}$ and $\mathbf{B}_{21} \triangleq \mathbf{P_b}$, where $\mathbf{P_b} \in \mathbb{R}^{p\times n}$ is the correspondence matrix of barycentre. It is clear that the column masses do not change, i.e. $\mathbf{c_b} = \mathbf{c_o} = \mathbf{c}$, the diagonal matrices of column masses is $(\mathbf{D_b})_c = (\mathbf{D_o})_c = \mathbf{D}_c$, the matrix $\mathbf{S_b}$ of barycentric standardized residuals has rank $k_b = min(p,n) - 1$, which can be written as $\mathbf{S_b} = \mathbf{B}\,\mathbf{S_o}$ with $\mathbf{B} \triangleq (\mathbf{D_b})_r^{-1/2}\,{\mathbf{Z_d}}^T\,(\mathbf{D_o})_r^{1/2}$ of size $k_b \times m$, where $(\mathbf{D_o})_r$ and $(\mathbf{D_b})_r$ are the diagonal matrices of row masses. From the SVD decomposition for the matrices $\mathbf{S_o}$ and $\mathbf{S_b}$ we have, $\mathbf{U_b}\,\mathbf{D_b}\,{\mathbf{V_b}}^T = \mathbf{B}\,\mathbf{U_o}\,\mathbf{D_o}\,{\mathbf{V_o}}^T$. The matrix $\mathbf{Q} \triangleq \mathbf{B}\,\mathbf{U_o}$ of size $k_b \times m$ is rectangular and satisfies $\mathbf{Q}\,\mathbf{Q}^T = \mathbf{I}_m$ ($\mathbf{I}_m$ is the identity matrix of order m) hence the rows of $\mathbf{Q}$ are orthonormal and $n \geq m$, but $\mathbf{Q}^T\,\mathbf{Q} \neq \mathbf{I}_m$ and the right side is not a regular SVD decomposition. To have $\mathbf{Q} = \mathbf{U_b}$ in the image subspace, the ranks will be equal ($k_o = k_b$). This condition is only possible if the rows profiles that have been aggregated are equal or that the rows are linearly dependent. In this case, $\mathbf{V_o} = \mathbf{V_b}$, i.e. the geometry of column profiles is not affected.

A.2. ROW FACTOR SCORES

We can write the principal coordinates $\mathbf{F_o}$ of rows in the observation space and $\mathbf{F_b}$ in the barycentric space, as follows (Greenacre and Blasius, 2006):

$$\mathbf{F_o} = (\mathbf{D_o})_r^{-\frac{1}{2}}\,\mathbf{S_o}\,\mathbf{V_o} = (\mathbf{D_o})_r^{-1}\,\mathbf{P_o}\,\mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}_o \qquad \in \mathbb{R}^{m\times k_o} \tag{10a}$$

$$\mathbf{F_b} = (\mathbf{D_b})_r^{-\frac{1}{2}}\,\mathbf{S_b}\,\mathbf{V_b} = (\mathbf{D_b})_r^{-1}\,\mathbf{P_b}\,\mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}_b \qquad \in \mathbb{R}^{m\times k_b}. \tag{10b}$$

The row factor scores for the observation space is defined on a domain corresponding to the first k_o singular values (image subspace), the same for the row factor score in the barycentric space defined on the first k_b singular values. The equations (10a) and (10b) provide also the transition equations (barycentric relationship) (Greenacre and Blasius, 2006) that can be used to situate supplementary points in the observation and in the barycentric map. However, if we want to situate the row profiles defined in the observation space into the barycentric space, we need the factor score $\mathbf{F_{ob}}$ to project from the observation space into the barycentric space, as follows:

$$\mathbf{F_{ob}} = (\mathbf{D_o})_r^{-1}\,\mathbf{P_o}\,\mathbf{D}_c^{-1/2}\mathbf{V_b} \quad \in \mathbb{R}^{m\times k_b} \tag{11a}$$

$$\mathbf{F_b} = (\mathbf{D_b})_r^{-1}\,\mathbf{Z_d}^T\,\mathbf{D_o}_r\,\mathbf{F_{ob}}. \tag{11b}$$

From (10b) and (11a) it follows (11b), i.e. the factor scores of barycentres are the barycentres of the factor score projections (Williams et al., 2010)

A.3. TRANSITION EQUATION

We can interpret the transition equations (10a), (10b), (11a) given a supplementary row point $\mathbf{X}$, as the transformation $f : \mathbb{R}^n \to \mathbb{R}^k$, that maps the row profile into the row factor score. Actually, the following relations hold in general:

$$F_{\cdot}(\mathbf{X}) = \frac{\mathbf{X}}{\|\mathbf{X}\|_1}\,\mathbf{D}_c^{-1/2}\mathbf{V}_{\cdot} \quad \forall \mathbf{X} \in \mathbb{R}^n_{\geq 0} \tag{12a}$$

$$F_{\cdot}(a\mathbf{X}) = F_{\cdot}(\mathbf{X}) \quad \forall a\ const \geq 0 \tag{12b}$$

$$F_{\cdot}\Big(\sum_j w_j \mathbf{X}_j\Big) = \frac{\sum_j w_j\,\|\mathbf{X}_j\|_1\,F_{\cdot}(\mathbf{X}_j)}{\|\sum_j w_j \mathbf{X}_j\|_1} \quad \forall w_j \geq 0 \tag{12c}$$

where $\|\bullet\|_1$ is the Manhattan norm, $\mathbf{D}_c$ is the diagonal matrix of column masses, $\mathbf{V}$ is the right orthogonal column matrix, and k is the rank of the corresponding matrix of residuals $\mathbf{S}$.

REFERENCES

Ballantine, D., Chow, A., De La Rosa, D., Piché, R. and Xu, L. (2006). Probabilistic estimation and prediction of television viewer demographics. In *Canadian Applied Mathematics Quarterly,* 16 (2): 2008.

Bornman, E. (2009). Measuring media audiences. In *Media Studies, Media Content and Media Audiences,* vol. 3. Cape Town, Juta.

Branch, M.A. and Grace, A. (2002). *Optimization Toolbox User's Guide, Version 2.* 24 Prime Park Way, Natick, MA 01760-1500.

Chen, D. and Plemmons, R.J. (2009). Nonnegativity constraints in numerical analysis. In *The Birth of Numerical Analysis,* 10: 109–140.

De Bock, K. and Van den Poel, D. (2010). Predicting website audience demographics for web advertising targeting using multi-website clickstream data. In *Fundamenta Informaticae,* 98 (1): 49–70.

Detroja, K.P., Gudi, R.D. and Patwardhan, S.C. (2011). Data reduction and fault diagnosis using principle of distributional equivalence. In *Advanced Control of Industrial Processes (ADCONIP), 2011 International Symposium on,* 30–35. IEEE.

Greenacre, M. and Blasius, J. (2006). *Multiple Correspondence Analysis and Related Methods.* Chapman and Hall/CRC, Boca Raton, FL.

Greenacre, M. and Lewi, P. (2009). Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements. In *Journal of Classification,* 26 (1): 29–54.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009). The WEKA data mining software: an update. In *ACM SIGKDD explorations newsletter,* 11 (1): 10–18.

Redfern, N. (2012). Correspondence analysis of genre preferences in UK film audiences. In *Participations,* 9 (2): 45–55.

Saporta, G. and Niang, N. (2006). Correspondence analysis and classification. In M. Greenacre and J. Blasius, eds., *Multiple Correspondence Analysis and Related Methods,* 371–392. Chapman and Hall/CRC, Boca Raton, FL.

Williams, L.J., Abdi, H., French, R. and Orange, J.B. (2010). A tutorial on multiblock discriminant correspondence analysis (MUDICA): A new method for analyzing discourse data from clinical populations. In *Journal of Speech, Language, and Hearing Research,* 53 (5): 1372–1393.

Zarraga, A. and Goitisolo, B. (2006). Simultaneous analysis: A joint study of several contingency tables with different margins. In M. Greenacre and J. Blasius, eds., *Multiple Correspondence Analysis and Related Methods,* 327–350. Chapman and Hall/CRC, Boca Raton, FL.

Zarraga, A. and Goitisolo, B. (2009). Simultaneous analysis and multiple factor analysis for contingency tables: Two methods for the joint study of contingency tables. In *Computational Statistics and Data Analysis,* 53 (8): 3171–3182.

Zarraga, A. and Goitisolo, B. (2013). *SimultAnR: Correspondence and Simultaneous Analysis. R package.* https://cran.r-project.org/web/packages/SimultAnR/SimultAnR.pdf.