A PICTURE OF FOREIGNERS INTEGRATION IN ITALY: METHODOLOGY AND EXPERIENCES

MARTA BLANGIARDO[1] AND GIANLUCA BAIO[2]

[1]Department of Epidemiology, Public Health and Primary Care
Imperial College London
St. Mary's Campus, Norfolk Place London W2 1PG, UK
m.blangiardo@imperial.ac.uk

[2]Department of Statistical Science
University College London
Gower Street, London WC1E 6BT, UK
gianluca@stats.ucl.ac.uk

1. Introduction

Social integration is arguably a very subtle concept, difficult even to define uniquely, let alone to measure. However, aiming at promoting political strategies for a thorough involvement of immigrants into the native social life, we need to take advantage of capable indicators to describe their overall condition. The importance of an adequate information system is largely acknowledged: recently, much attention has been paid to the measurements of foreigners' integration in Europe.

In Italy, the law 40/1998 identifies *social integration* with "a process of non discrimination and of differences inclusion, contamination and experimentation of new relations and behaviours" (Zincone 2000). The major attempt to define a set of suitable indicators for foreigners' integration in the Italian literature is that of Golini et al. (2004). However, this study mainly provides a theoretical framework based on the use of aggregated data from official sources, due to the inadequateness of the information system, and advocates ad hoc surveys. On the other hand, in the Italian literature there are no significant works using individual data, which are time and money consuming.

In the need of carrying out sampling surveys, the first relevant contribution is that of Fondazione ISMU (Milan), which have promoted and realised a survey involving about 30 000 interviews, with the aim of evaluating the characteristics and the needs of the foreigners' population in Italy, with particular reference to the Southern Regions. The availability of detailed individual data calls for more advanced statistical analysis, which can model properly all the sources of variability and provide a more precise account of the problem at hand. The model that we propose in this study is based on Bayesian latent class analysis technique.

The idea is to use some relevant individual characteristics to gain information about the non measurable level of "social integration". The underlying assumption that we carry on throughout the paper, is that integration is determined by the process of possessing, or aiming at possessing, some individual features that are perceived as prerequisites for gaining stability in the native society. Some examples are represented by working or juridical status. By no means could this definition be considered as definitive, nor does it apply to all contexts. However, we reckon that it is pragmatically acceptable and therefore we use it as a basis for our analysis.

According to the Bayesian approach, each quantity subject to uncertainty (either *experimental* - due to a sampling procedure, or *structural* - due to our ignorance

or even incapability of ever observing its actual value) is associated with a suitable probability distribution. This uncertainty is modified by the observation of the available evidence. Bearing these assumptions in mind, the paper is structured as follows: in section 2 we present the data and the survey carried out by Fondazione ISMU; in section 3 we introduce the methodology adopted; section 4 presents the results and section 5 the discussion.

2. Data

A survey has been conducted by Fondazione ISMU in 2005, covering the national territory, with the aim of collecting data on different characteristics of the immigrants. The whole foreigners' population present in Italy and coming from the Developing Countries (DCs) has been targeted as the reference population, without any restriction on residence or juridical status.

The survey has interested a total number of 30 thousand units: 22 thousand were divided within the 30 Provinces forming the six Southern Regions (Campania, Apulia, Basilicata, Calabria, Sicily, Sardinia), while the remaining 8 thousand were distributed in 10 Northern Provinces (Turin, Milan, Bergamo, Brescia, Mantua, Verona, Vicenza, Bologna, Florence, Rome). Although according to the inter-Province variability of foreigners' distribution, the sampling scheme ensures a minimum and maximum presence for each territorial unit – from 400 for each of the two provinces of Oristano (Sardinia) and Enna (Sicily), to 1 600 of Naples. For each Province, the total number of cases has been divided within an appropriate number of Municipalities (first order units).

The second order units have been selected from the over-18 year old population from DCs. The sampling scheme follows the methodology of centre-sampling (Blangiardo 2004) and the term "centre" refers to a partial list or a place where units congregate. A collection of centres is then identified in order to ensure an adequate coverage of the population under the assumption that every unit belongs to, or regularly visits, at least one centre.

3. Methodology

3.1. **Latent Class Analysis.** Latent Class Analysis (LCA) methods were first introduced by Lazarsfeld (1950), as a tool for clustering. Later on, Goodman (1974) developed an algorithm for obtaining maximum likelihood estimators of the parameters and Haberman (1979) showed the connections with log-linear models. Finally, the 1990s witnessed many works on Bayesian classification and mixture models (Pearl 1988, Neapolitan 1990, Whittaker 1990), concepts that are closely related to LCA and that helped broaden its scope.

The basic idea underlying LCA is that some of the parameters of the model we choose to describe the phenomenon under study may differ across unobserved subgroups of individuals. The unobserved (latent) groups may be associated with a categorical variable. In other words, suppose that we observe J "manifest" variables $\mathbf{Y} = (Y_1, \ldots, Y_J)$, and that we can assume the existence of an unobservable variable X, say taking over C possible values (classes), accounting for the heterogeneity within the subjects.

The number of possible states C of the latent variable X is obviously associated with the layers of heterogeneity that we presume to exist in the study population. In this work we focus on situations where the experimenter can define this number from their prior knowledge of the problem (for the case where also the number of layer is estimated by the data, the classical methodology is that discussed in Richardson and Green 1997).

Simple Probability Calculus allows us to express the probability distribution of the manifest factors in terms of the latent variable as:

$$p(\mathbf{Y} = \mathbf{y}) = \sum_{x=1}^{C} p(X = x)p(\mathbf{Y} = \mathbf{y} \mid X = x). \tag{1}$$

A useful simplifying assumption is that of *local independence*, i.e. that the J manifest variables are mutually independent within each latent class. In other words, we assume that the correlation that might be observed among the manifest factors is entirely accounted for by their common dependence on the latent variable. This condition can be formalized as:

$$p(\mathbf{Y} = \mathbf{y} \mid X = x) = \prod_{j=1}^{J} p(Y_j = y \mid X = x), \tag{2}$$

where y is a generic value for the $j - th$ factor.

Combining (1) and (2), we then obtain that the joint probability distribution of the manifest factors is:

$$\begin{aligned} p(\mathbf{Y} = \mathbf{y}) &= \sum_{x=1}^{C} p(X = x)p(\mathbf{Y} = \mathbf{y} \mid X = x) \\ &= \sum_{x=1}^{C} p(X = x) \prod_{j=1}^{J} p(Y_j = y \mid X = x), \end{aligned} \tag{3}$$

or, equivalently, that the joint probability distribution of the manifest factors *and* the latent variable is:

$$p(\mathbf{Y} = \mathbf{y}, X = x) = p(X = x) \prod_{j=1}^{J} p(Y_j = y \mid X = x) \tag{4}$$

- recall from Probability Calculus that $p(B) = \sum_A p(A, B)$, and therefore (3) can be obtained from (4) marginalizing out X. The condition in (2) can be released, obtaining more complicated models (see Zhang 2004), but here we consider only the simpler case of conditional independence.

A convenient alternative way to represent the model of (4) is in terms of a corresponding directed acyclic graph (DAG, directed for the link between each pair of nodes, acyclic for the impossibility of turning on the same node after leaving it following the direction of the arrows, Gilks et al. 1996). In a DAG, the circles denote random quantities, while the arrows between the nodes represent a stochastic dependence. Figure 1 shows the graphical representation of the model. As one can appreciate, the assumption of local independence is encoded in the absence of direct links connecting the manifest factors $Y_1, \ldots, Y_J$. This model is often referred to as "naïve Bayesian classifier" (Friedman et al. 1997).

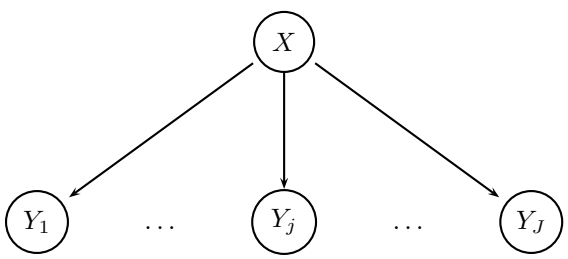


FIGURE 1. Latent Class Analysis in terms of a DAG

The objective of the LCA model is to observe a sample of data for $\mathbf{Y}$, in order to estimate its probability distribution, as factorised in (3), i.e. to infer the two "ingredients":

- $p(X = x)$, which is the *prevalence* of each class (subgroup) in the general population;
- $p(Y_j = y \mid X = x)$, which is the conditional response probability for the $j - th$ factor.

Let us assume that the manifest factors are all discrete; for example the variable Y_j can take on the values $[y_{j1}, \ldots, y_{jl}, \ldots, y_{js_j}]$, where s_j is the number of its possible states. If we consider a total of J manifest factors, each of which may take on s_j states ($j = 1, \ldots, J$), the possible observable profiles, i.e. combinations of all the values of these variables, are:

$$\mathcal{P} = \begin{bmatrix} \mathcal{P}_1 \\ \vdots \\ \mathcal{P}_k \\ \vdots \\ \mathcal{P}_K \end{bmatrix} = \begin{bmatrix} Y_1 = y_{11} & \ldots & Y_j = y_{j1} & \ldots & Y_J = y_{J1} \\ \vdots & & \vdots & & \vdots \\ Y_1 = y_{1h} & \ldots & Y_j = y_{jl} & \ldots & Y_J = y_{Jm} \\ \vdots & & \vdots & & \vdots \\ Y_1 = y_{1s_1} & \ldots & Y_j = y_{js_j} & \ldots & Y_J = y_{Js_J} \end{bmatrix}, \tag{5}$$

where $K = \prod_{j=1}^{J} s_j$ is the number of possible configurations. What we actually measure (the evidence), is the number of times, i.e. the frequency, that each profile is observed in our dataset:

$$\mathcal{E} = \begin{bmatrix} \#\mathcal{P}_1 = n_1 \\ \vdots \\ \#\mathcal{P}_k = n_k \\ \vdots \\ \#\mathcal{P}_K = n_K \end{bmatrix},$$

where the symbol # means "number of occurrences of" and $N = \sum_{k=1}^{K} n_k$ is the total number of units surveyed.

3.2. **Bayesian modelling.** The main feature of the Bayesian philosophy is that probability represents numerically the degree of belief (uncertainty) in the occurrence of an event, rather than a relative frequency derived from a hypothetical large number of experiments repeated under the same conditions. Individuals can express their own evaluation, often referred to as *subjective probability*. According to the evidence that becomes sequentially available, the individuals tend to update their belief. Several studies in psychology and cognitive science suggest that rational human reasoning is in fact based on these principles (Dayan et al. 2000, Gold and Shadlen 2001).

Under these premises, probability models are defined not only for observable variables, as also happens in the frequentist approach, but for the parameters too. Formally, within the Bayesian framework, the experimenter needs to assign a *prior probability distribution* for the parameter of interest, which encodes their uncertainty over its unknown value. Once the experiment is performed and the evidence $\mathcal{E}$ is observed and summarised by means of the *likelihood function*, the experimenter can update their prior opinion into the *posterior distribution*, by means of Bayes Theorem. In our case the objective is to compute the posterior distributions for X and each $(Y_j \mid X)$, which we indicate respectively by $p(X \mid \mathcal{E})$ and $p(Y_j \mid X, \mathcal{E})$.

The model is specified as follows. The observed factors in $\mathbf{Y}$ and the unobservable latent variable X are jointly modelled as $(\mathbf{Y}, X) \sim \text{Multinomial}(\boldsymbol{\theta}, \mathbf{n})$, where:

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_{11} & \dots & \theta_{1x} & \dots & \theta_{1C} \\ \vdots & & \vdots & & \vdots \\ \theta_{k1} & \dots & \theta_{kx} & \dots & \theta_{kC} \\ \vdots & & \vdots & & \vdots \\ \theta_{K1} & \dots & \theta_{Kx} & \dots & \theta_{KC} \end{bmatrix} \quad \text{and} \quad \mathbf{n} = \begin{bmatrix} n_1 \\ \vdots \\ n_k \\ \vdots \\ n_K \end{bmatrix}$$

According to (4), the elements of $\boldsymbol{\theta}$ can be written as:

$$\theta_{k\,x} = p(X = x)\left[p(Y_1 = y_{1h} \mid X = x) \times \ldots \times p(Y_J = y_{Jm} \mid X = x)\right],$$

which represent the probability that the $k-th$ configuration in $\mathcal{P}$ belongs to the latent class x, for $x = 1, \ldots, C$. The elements of $\mathbf{n}$ are the observed number of occurrences of each profile in the dataset.

This parametric assumption for the likelihood function seems natural, as we are modelling a vector of counts. Moreover, in line with the Bayesian paradigm, along with the Multinomial likelihood, we also need to specify a prior distribution for the parameter $\boldsymbol{\theta}$, or, equivalently, for its two components $p(X = x)$ and $p(Y_j = y \mid X = x)$. In this case, a reasonable choice that guarantees a great flexibility in the data fitting, is to assign X and each $(Y_j \mid X)$ a dispersed Dirichlet prior probability distribution.

This model, so called Multinomial-Dirichlet, has now become quite standard in the statistical literature (Bernardo and Smith 1999, Congdon 2001) and has several applications in many research fields from marketing to genetics. Using this formulation, it is possible to determine the posterior distributions $p(X \mid \mathcal{E})$ and $p(Y_j \mid X, \mathcal{E})$, for example using an iterative algorithm, such as the following.

- (i) Simulate a value for each θ_{kx} using the prior distributions on X and on each $(Y_j \mid X)$. Assuming a disperse prior essentially means that each class of the latent variable and each state of the manifest factors are equally likely *a priori*, i.e. $p(X = x) = \frac{1}{C}$, for $x = 1 \ldots, C$ and $p(Y_j = y \mid X = x) = \frac{1}{s_j}$, for $j = 1, \ldots, J$;
- (ii) Simulate a value for $(\mathbf{Y}, X)$ from their joint Multinomial distribution, using the value of the parameter $\boldsymbol{\theta}$ obtained in (i), and the observed frequency of each profile, n_k. This step amounts to randomly allocate the number of occurrences of a profile $\mathcal{P}_k$ into the C classes of X, according to the value of $\boldsymbol{\theta}$;
- (iii) Simulate a value from the posterior distributions of X and each $(Y_j \mid X)$, updating the priors of (i) by means of the simulated observation of $(\mathbf{Y}, X)$ obtained in (ii). Since this latter is a function of the observed frequencies n_k, *a posteriori* the states of the manifest factors are no longer equally likely, as some profiles will generally be more frequent than others;
- (iv) Simulate a value for each θ_{kx} using the posterior distributions for X and each $(Y_j \mid X)$ obtained in (iii). This amounts to update the probability of the $k-th$ profile being in class x due to the random variability imposed by the simulated observation of $(\mathbf{Y}, X)$ and by the observation of the frequencies;
- (v) Repeat (ii) to (iv) until convergence is reached (suitable procedures can be used to check convergence – see for instance Gelman and Rubin 1996); use the simulations to produce the estimations required.

In effect, this algorithm finds an optimal way to allocate the observed frequencies in $\mathbf{n}$ into a C dimensional table; the allocation is produced by a combination of a

random process, i.e. the model assumed under (4) and the evidence provided by the observation of the data and their intrinsic variability.

3.3. **Probability propagation.** Beyond the estimation of the relevant parameters, Bayesian LCA offers other interesting features. In fact, once the distributions for $(X \mid \mathcal{E})$ and each $(Y_j \mid X, \mathcal{E})$ have been derived from the observed data, it is also possible to proceed with the estimation of the value of the latent variable X for a specific observed profile $\mathcal{P}_k = [Y_1 = y_{1h}, \ldots, Y_j = y_{jl}, \ldots, Y_J = y_{Jm}]$. This goal can be reached directly by means of the application of Bayes Theorem, which obviously is specific to the Bayesian approach.

First, we instantiate (set) the observed factors, according to a certain realised configuration. Then, we propagate the evidence, in order to update the probability distribution of the unobserved latent variable, applying the theorem:

$$p(X = x \mid \mathcal{P}_k, \mathcal{E}) \quad = \quad \frac{p(\mathcal{P}_k \mid X = x, \mathcal{E}) p(X = x, \mathcal{E})}{\sum_x p(\mathcal{P}_k \mid X = x, \mathcal{E}) p(X = x, \mathcal{E})}. \tag{6}$$

All the quantities in the right hand side of (6) are directly available from the Bayesian model. Again, thanks to the local independence assumption, the conditional response probability can be calculated as the product of the single conditional distributions of the J variables involved:

$$p(\mathcal{P}_k \mid X = x, \mathcal{E}) = p(Y_1 = y_{1h} \mid X = x, \mathcal{E}) \times \ldots \times p(Y_J = y_{Jm} \mid X = x, \mathcal{E}).$$

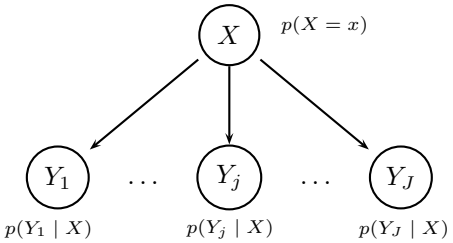
A summary of the modelling is presented in Figure 2 using the graphical representation of DAGs.

3.4. **Post stratification.** This latter feature of the Bayesian method is quite interesting in terms of allowing post stratification and analysis. Let us suppose that the original dataset is $\mathcal{D} = \{\mathbf{Y} \cup \mathbf{Z}\}$, i.e. it consists of the J manifest factors used for the Bayesian LCA: $\mathbf{Y}$, and of some other Q observable attributes: $\mathbf{Z} = (Z_1, \ldots, Z_Q)$ not used in the model. After the analysis is carried out, it is possible to *augment* $\mathcal{D}$ with the probability that an individual is associated with each of the latent classes, as estimated from (6). An example is shown in Table 1. We use the notation $[u]$ (read "of the unit u") to map the observed profile of a generic unit u to one and only one of the possible profiles in $\mathcal{P}$; $\mathcal{P}_{[u]} = [Y_1(u), \ldots, Y_j(u), \ldots, Y_J(u)]$ represents the observation of the manifest variables on the $u-th$ unit $(u = 1, \ldots, N)$ of the sample in terms of the K profiles listed in (5) - therefore $[u]$ takes on the values $1, \ldots, K$.

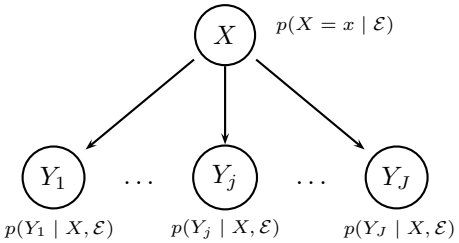
Id	Y_1	...	Y_J	Z_1	...	Z_Q	$p(X = 1 \mid \mathcal{P}_{[u]}, \mathcal{E})$	...	$p(X = C \mid \mathcal{P}_{[u]}, \mathcal{E})$
1	y_{12}	...	y_{J2}	z_{11}	...	z_{Q1}	0.9715	...	0.0285
...	...	...	...	...	...	...	...	...	...
u	y_{11}	...	y_{J1}	z_{11}	...	z_{Q3}	0.9321	...	0.0004
...	...	...	...	...	...	...	...	...	...
3218	N/A	...	y_{J1}	z_{13}	...	z_{Q2}	0.9950	...	0.0050
...	...	...	...	...	...	...	...	...	...
N	y_{11}	...	y_{J3}	z_{11}	...	z_{Q3}	0.0005	...	0.0810

TABLE 1. Data augmentation. The posterior probability that a given individual is associated with each latent stratum is included in the data set, along with the original variables $\mathcal{D}$

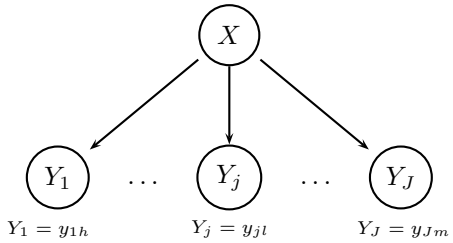
Notice that the Bayesian approach is able to associate an integration level also with those cases that present some missing values in $\mathbf{Y}$. For instance, consider in Table 1 the row indexed by the identifier 3128; the value of the factor Y_1 is not known to the experimenter, as the individual did not respond to the relevant



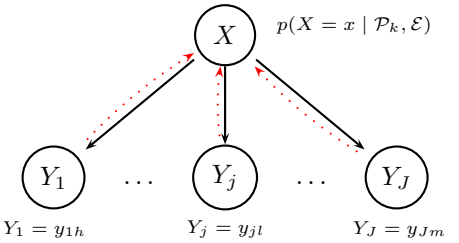
(1) Assume local independence and the number of layers of heterogeneity, and set the prior distributions



(2) Update uncertainty by means of the observation of $\mathcal{E}$ and obtain posterior distributions



(3) Set the manifest variables to the values of the observed profile $\mathcal{P}_k$; uncertainty remains only on the variable X



(4) Evidence propagation and probability inversion by means of Bayes Theorem: information travels "against the arrows"

FIGURE 2. Graphical summary of the model: the evidence is entered in the manifest variables $Y_1, \ldots, Y_J$ and by means of Bayes Theorem it is propagated against the arrows, in order to update the distribution of the unobservable variable X. Steps (1) and (2) represent the estimation process, whereas steps (3) and (4) describe the analysis

question. However, the complete conditional probability distribution of Y_1 (the variable affected by the missing data) is available from the LCA model and therefore Bayes theorem can be applied nevertheless. Consequently, we can still estimate $p(X = x \mid \mathcal{P}_{[3128]}, \mathcal{E})$, averaging over the probability distribution of the missing variable. In this case, we could calculate:

$$\begin{aligned} p(\mathcal{P}_{[3218]} \mid X = x, \mathcal{E}) &= p(Y_1 = N/A \mid X = x, \mathcal{E}) \times \ldots \times p(Y_J = y_{J1} \mid X = x, \mathcal{E}) \\ &= \mathrm{E}\left[p(Y_1 \mid X = x, \mathcal{E})\right] \times \ldots \times p(Y_J = y_{J1} \mid X = x, \mathcal{E}) \\ &= \frac{1}{s_1} \sum_{h=1}^{s_1} p(Y_1 = y_{1h} \mid X = x, \mathcal{E}) \times \ldots \times p(Y_J = y_{J1} \mid X = x, \mathcal{E}) \end{aligned}$$

for $x = 1, \ldots, C$ and use this estimation to compute (6).

Notice that this estimation is possible only when the hypothesis of *Missing At Random* (MAR, Little and Rubin 1987) is sustainable and the posterior mean (based on all the similar observed cases) can be used as a valid substitute for the missing value. As compared to a completely observed unit, in such a situation the estimate of (6) will be subject to higher variability, due to lesser available information. Of course, in case the MAR hypothesis does not hold, it is impossible to predict the missing data without the addition of (often non testable) further assumptions.

4. Bayesian model for immigrants social integration: some results

We applied the methodology described above to the data of the Fondazione ISMU survey of section 2, in order to estimate an index of social integration within the Italian society for a generic immigrant showing a profile $\mathcal{P}_k$. We assume here that the target population is characterised by two unobservable layers of heterogeneity, i.e. each individual is either "integrated" or not within the native population. However, this attribute can never be directly measured and therefore we estimate it on the basis of a set of observable factors that describe some relevant aspects of social life.

Consequently, following the notation introduced earlier we considered $C = 2$ classes for the latent variable (i.e. X = integrated, or X = not integrated) and $J = 4$ manifest variables that we used as proxies of relevant aspects associated with the integration level: these are *Gender*, *Juridical Status*, *Working Status* and *Housing and Family Arrangement*. Table 2 shows the possible states of each factor. Since these are $s_1 = 2, s_2 = 7, s_3 = 4$ and $s_4 = 15$, the total number of profiles is $K = 840$.

Variables	Modalities
Y_1 = Gender	$y_{1\,1}$ = Male
	$y_{1\,2}$ = Female
Y_2 = Juridical Status	$y_{2\,1}$ = Italian Citizenship
	$y_{2\,2}$ = Permanent Permit of Stay
	$y_{2\,3}$ = Temporary Permit of Stay/Registered in the Birth Record
	$y_{2\,4}$ = Temporary Permit of Stay/Not Registered in the Birth Record
	$y_{2\,5}$ = Currently renewing the Permit of Stay
	$y_{2\,6}$ = Expired Permit of Stay
	$y_{2\,7}$ = Never Had a Permit of Stay
Y_3 = Working Status	$y_{3\,1}$ = Stable Worker
	$y_{3\,2}$ = Unstable Worker
	$y_{3\,3}$ = Inactive (Student or Housewife)
	$y_{3\,4}$ = Unemployed
Y_4 = Housing and Family Arrangement	$y_{4\,1}$ = Single Property Owner
	$y_{4\,2}$ = Owns a Property with Family of Origin
	$y_{4\,3}$ = Owns a Property with Partner and Children (if any)
	$y_{4\,4}$ = Owns a Property with Children only
	$y_{4\,5}$ = Owns a Property with Friends
	$y_{4\,6}$ = Single Tenant
	$y_{4\,7}$ = Renting a Property with Family of Origin
	$y_{4\,8}$ = Renting a Property with Partner and Children (if any)
	$y_{4\,9}$ = Renting a Property with Children only
	$y_{4\,10}$ = Sharing a Property with Friends
	$y_{4\,11}$ = Unstable Housing Arrangement (on their own)
	$y_{4\,12}$ = Unstable Housing Arrangement with Family of Origin
	$y_{4\,13}$ = Unstable Housing Arrangement with Partner and Children (if any)
	$y_{4\,14}$ = Unstable Housing Arrangement with Children only
	$y_{4\,15}$ = Unstable Housing Arrangement with Friend

Table 2. Possible states of the 4 observed factors

By means of (6), we calculated the index of social integration in the Italian society as:

$$i(u) = p(X = \text{integrated} \mid \mathcal{P}_{[u]}, \mathcal{E}),$$

for the $u - th$ immigrant in the dataset. Some comments are due at this point.

- Strictly speaking, the model is only able to analyse the alleged heterogeneity structure. Each individual is associated with a probability that describes how likely it is that they belong to each latent stratum. In fact, the states of the variable X merely represent the layers within the population and it is only the experimenter that can label each stratum, by interpreting the results *a posteriori*, for instance as "integrated" or "not integrated".
- Although the index i is defined for each unit u, it can be associated with other individuals (even not yet observed) sharing the same features in terms of the manifest factors used to define the probability of integration. Using this property, other individuals that have not been surveyed can be analysed.
- Being in fact a probability, i is naturally normalised, as it is defined in the interval $[0, 1]$. It is therefore easy to interpret: the closer to 1, the higher the level of estimated social integration.
- Besides the individual index, it is also possible to calculate the *average index of integration* in the overall population:

$$I(\mathcal{D}) = \mathrm{E}_{\{\mathcal{D}\}}\left[i(u)\right]$$

  (we introduce the notation I to highlight the fact that this is an average of the individual values $i(u)$, calculated over a suitable set - the entire sample $\mathcal{D}$, in this case).
- More interestingly, it is possible to compute any kind of *conditional average index*, i.e. among units sharing a particular set of characteristics, possibly in terms of (some of) the variables in $\mathbf{Z}$. For instance:

$$I(\mathbf{z}^*) = \mathrm{E}_{\{\mathbf{Z}(u)=\mathbf{z}^*\}}\left[i(u)\right]$$

  can be calculated simply taking into account the values of $i(u)$ for the units possessing the required values $\mathbf{z}^*$ of the variables in $\mathbf{Z}$. This can be particularly helpful in the analysis.

Back to Fondazione ISMU data, the most integrated type of immigrants are women who possess Italian citizenship or a permanent permit of stay, are inactive and live with their partner (and children, if any) in an owned house. On the one hand it might seem odd that the people associated with the highest level of social integration have inactive working status. However, it is often the case that inactive women come to Italy to rejoin their family of origin, or their partner. For this reason, upon arrival they typically find an already stable environment. Moreover, their permit of stay is usually granted for rejoining purpose, in which case they are not allowed to work. This possible explanation is supported by the finding that the most integrated profiles within males include the same features, but with a stable job as for the working status.

Post stratification also allows characterising the index of integration within specific groups of immigrants. In particular, we focused the attention on the length of stay in Italy and on the geographical distribution, in terms of which Italian regions are more likely to be associated with higher levels of integration. The average index of social integration by year of arrival in Italy is presented in Figure 3. Clearly, the longer the presence in Italy, the higher the value of the index of integration; this finding seems to make sense, as some time is needed to settle down in a new environment and to gain stability. In particular, Southern Regions consistently show a lower value of the index I, with the largest difference registered among the most recently arrived individuals, i.e. in the period 2004-2005 (cfr. Figure 3).

The index of integration has been calculated also by area of origin for immigrants (in this case we consider 5 macro-areas: East Europe, Asia, North Africa, Central-South Africa and Latin America). As is depicted in Figure 4, immigrants coming from Asia are mainly at the top of the rank. At the bottom there are individuals from Central-South Africa. However, if we focus on the latest arrivals in Italy, people from East Europe, North Africa and Latin America have a lower index of integration.

Figure 5 presents the distribution of the index of integration for specific groups of immigrants with respect to education, religion and civil status. The level of education shows women always in advantage. Among women, different levels of education are not associated with substantial different levels of the index of integration, whereas this appears to happen among men. The estimated index of social integration is 0.16 for men with no tile, while it increases up to 0.67 for men with a degree.

The distribution of religion shows a peak of the index of integration for the Buddhists, regardless on the sexes; Hindustani women are associated with the highest integration level, while Hindustani men have the lower value of the index. Muslim women have high levels of integration, whereas Muslim men and Catholics in general are ranked in the middle.

Finally, the civil status shows how being married is highly associated with a high index of integration for both the sexes. On the other hand, women not married or divorced have higher integration than men in the same status.

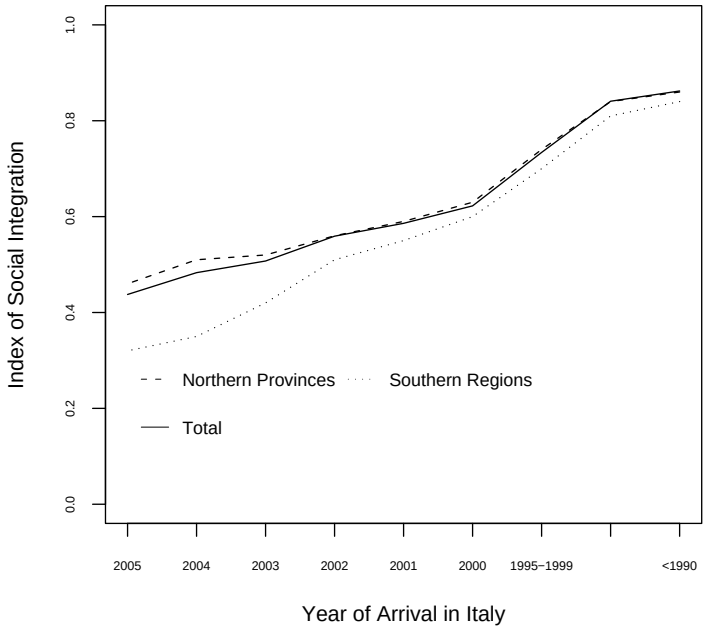


FIGURE 3. Average index of social integration by time of arrival and Italian area where the immigrants live

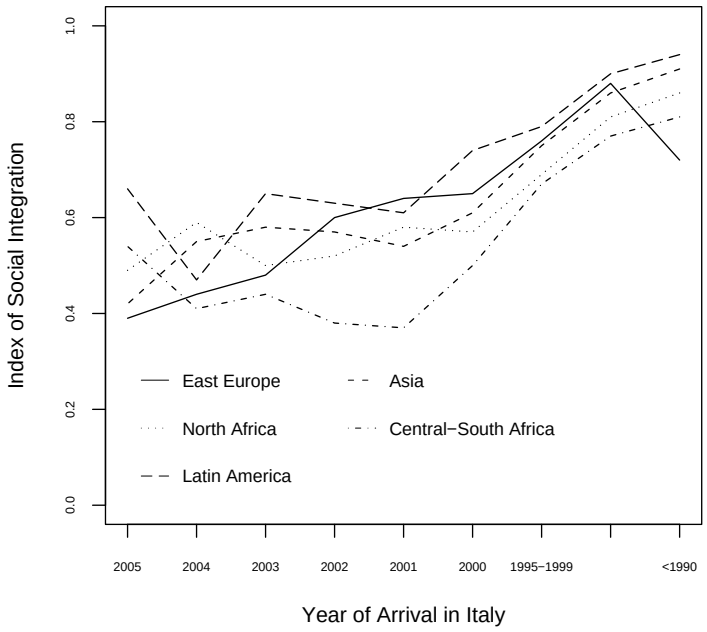


FIGURE 4. Average index of social integration by time of arrival in Italy and area of origin of immigrants

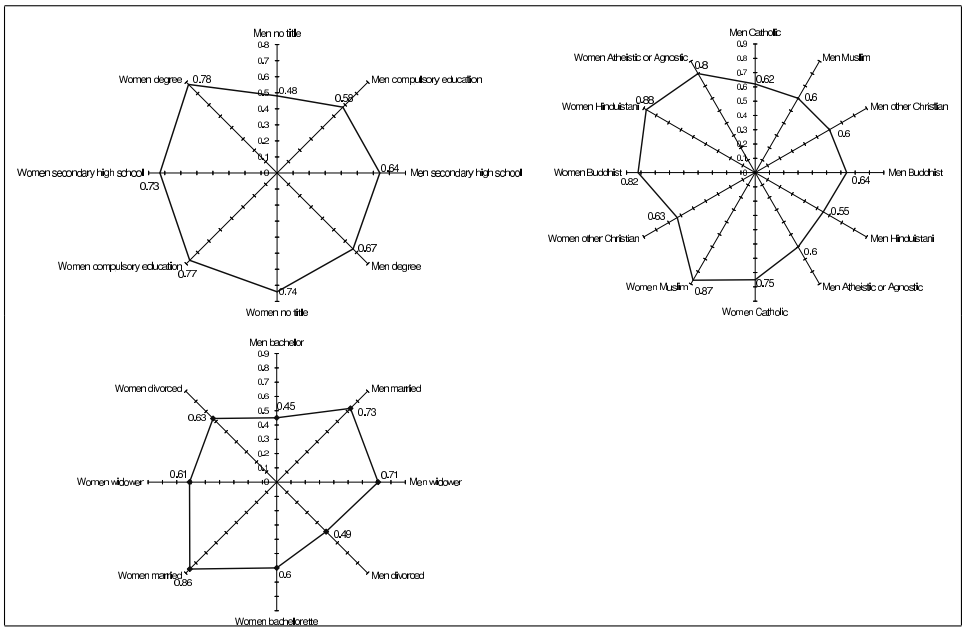


FIGURE 5. Radar plots depicting the average index of social integration by level of education, religion and juridical status of immigrants

5. Discussion

In this paper we used the Bayesian approach to LCA to estimate a social integration index for immigrants in Italy. The main advantage of this methodology is that, under a set of assumptions (stated above), it is possible to infer about quantities that we will never be able to observe. This characteristic makes the use of standard statistical techniques (such as multivariate regression analysis) non applicable in this case. We are fully committed to the Bayesian approach, as we reckon that both theoretically and pragmatically it provides a more natural framework where all the available information can be properly taken into account. Moreover, the findings of the current research can be used as prior knowledge to inform similar future studies.

A limiting prerequisite for the implementation of a latent class analysis is the availability of a large dataset. This is due to the fact that the number of possible profiles might be large and increasing with the complexity of the design (i.e. the possible configurations of the manifest variables). In this case, we had available a very large database from an observational study, which makes our results quite robust. Moreover, the data were sufficiently representative of the whole Italian territory. However, in case the researcher could not access such a comprehensive dataset, it would be necessary to choose either a smaller number of variables or a fewer possible states for each of them.

The future extensions of this work include relaxing the hypothesis of conditional independence among the manifest factors. More complex models that include correlations among the observed variables and that allow an empirical estimation of the number of latent strata in the reference population would be highly valuable. Nevertheless, this issue can be quite controversial, because the latent sub-groups that the researcher presumes in the target population has also to be easily interpretable to be meaningful.

References

Bernardo, J. and Smith, A. (1999), *Bayesian Theory*, John Wiley and Sons, New York, NY.

Blangiardo, G. (2004), Campionamento per centri nelle indagini sulla presenza straniera in Lombardia: una nota metodologica, *in* M. Pelagatti, ed., 'Studi in ricordo di Marco Martini', Giuffrè, Milan, Italy, Milan, Italy.

Congdon, P. (2001), *Bayesian Statistical Modelling*, John Wiley and Sons, Chichester, UK.

Dayan, P., Kakade, S. and Montague, P. (2000), 'Learning and selective attention', *Nature Neuroscience Supplement* **3**, 1218–1223.

Friedman, N., Geiger, D. and Goldszmidt, M. (1997), 'Bayesian Network Classifiers', *Machine Learning* **29**, 131–163.

Gelman, A. and Rubin, D. (1996), 'Markov chain Monte Carlo methods in Biostatistics', *Statistical Methods in Medical Research* **5**, 339–355.

Gilks, W., Richardson, S. and Spiegelhalter, D. (1996), *Markov Chain Monte Carlo in Practice*, Chapman Hall, London, UK.

Gold, J. and Shadlen, M. (2001), 'Neural computations that underlie decisions about sensory stimuli', *Trends in Cognitive Science* **5**, 10–16.

Golini, A., Strozza, S., Basili, M., Ribella, N. and Reginato, M. (2004), L'immigrazione straniera: indicatori e misure di integrazione, Technical report, FIERI - International and European Forum of Research on Immigration and Department of Demographic Sciences, University of Rome "La Sapienza", Italy.

Goodman, L. (1974), 'Exploratory latent structure analysis using both identifiable and unidentifiable models', *Biometrika* **61**, 215–31.

Haberman, S. (1979), *Analysis of Qualitative Data*, Academic Press, New York, NY.

Lazarsfeld, P. (1950), The logical and mathematical foundation of latent structure analysis, *in* S. Stouffer, ed., 'Measurement and Prediction', Princeton University Press, Princeton, NJ.

Little, R. and Rubin, D. (1987), *Statistical analysis with missing data*, Wiley, New York, NY.

Neapolitan, R. (1990), *Probabilistic Reasoning in Expert Systems*, John Wiley and Sons, New York, NY.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kauffman, San Mateo, CA.

Richardson, S. and Green, P. (1997), 'On Bayesian analysis of mixtures with unknown number of components (with discussion)', *Journal of the Royal Statistical Society,* B **59**, 731–792.

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, John Wiley and Sons, New York, NY.

Zhang, N. (2004), 'Hierarchical Latent Class Models for Cluster Analysis', *Journal of Machine Learning Research* **5**, 697–723.

Zincone, G. (2000), *Primo Rapporto sull'integrazione degli immigrati in Italia*, Il Mulino, Bologna, Italy.