

How to Profile your Customers  
Using Collaborative Database Profiling:  
an Application to Age Estimation

**Arnaud De Bruyn\***  
*ESSEC Business School*

**Nathalie Tramonte\***  
*HEC-ULG Management School*

**Paper submitted to the Marketing Research track  
European Marketing Academy 2008 Conference**

**Abstract** – Firms use external data sources (e.g., Census Bureau data) to infer the most likely age, income, housing situation or marital status of their customers, and better portray their profiles and needs. In this paper, we show that the methodology commonly used in the industry is severely biased towards average national figures. We discuss a maximum likelihood approach to correct for these biases. Taking the example of first name analysis to infer customers' age, we show that the approach we suggest gives much more precise and unbiased estimates, allowing marketers to better profile their customers and tailor their marketing strategy more precisely to their customers' characteristics.

**Keywords** – Collaborative database profiling; Socio-demographic profiling; First name; Age estimation

---

\* Contact author. Assistant Professor of Marketing, ESSEC Business School, Avenue Bernard Hirsch, 95000 Cergy, France. Email: [debruyn@essec.fr](mailto:debruyn@essec.fr). Tel.: 33 (0)1 34 43 32 46. Fax: 33 (0)1 34 43 32 11.

\* Research Assistant, Doctoral Candidate, HEC-ULG Management School, Boulevard du Rectorat 7, Bâtiment B31 Local 3.38-42, 4000 Liège, Belgium. Email: [ntramonte@ulg.ac.be](mailto:ntramonte@ulg.ac.be).

## INTRODUCTION

Over the years, marketing has shifted from a product focus to a customer relationship focus. In this context, customer profiling (that is, the ability for a firm to understand who its customers are) became of paramount importance. Better knowing one's customers in terms of age, income, level of education, geo-demographic environment, cultural or household characteristics, is essential to better target marketing offers, fine tune marketing communications, identify the best positioning strategy, or simply understand consumers' behavior and drivers. Applications in the academic literature are numerous as well<sup>1</sup>.

On the other hand, customers are increasingly aware –and wary– of that constant quest to learn more about who they are. This is particularly tangible on the internet, where the industry has observed a surge of online services and software packages that provide anonymity to their users on the web (Wall Street Journal, 2006). That trend is also confirmed by the rising quantity of phony information provided by customers to company's surveys – especially when the information requested is not clearly linked to potential customers' benefits. For instance, the last time one of the authors' went to purchase a toy in a department store, the vendor asked for a zip code. Obviously, this information was valuable to the department store, but of no added value to us, and it made us uncomfortable. Many customers increasingly share the feeling.

While the need for high-quality information about one's customers is increasing, it is also more difficult and costly to acquire that information directly from customers. It is often perceived as an invasion of one's privacy. Consequently, marketers have developed indirect ways to obtain that information, such as *database profiling*.

The guiding principle of database profiling is to match information from existing customers (the *target population*, about whom little is known) with a larger database (the *reference population*) containing information of interest to the firm, in order to indirectly infer customers' key characteristics. For instance, in the United States, customers' zip codes can be matched with the data from the U.S. Census Bureau (2000) to qualify customers' profiles in terms of most likely age, race, family relationships, household types, housing occupancy, educational attainment, marital status, employment status, income, etc. Databases of that sort exist in most countries (e.g., United Kingdom, Belgium, France, etc.).

This information is valuable to guide firm's marketing strategy in two major areas:

- **Profiling.** To identify key customers' drivers provides valuable insights to fine-tune marketing communications; insights such as how many married people are in the database, or the extent of young people in a specific customer segment.
- **Targeting.** To better understand the key characteristics of a target population can be used to further refine targeting and customers' acquisition strategy – whether to identify cross-selling opportunities with existing customers who share the same profile than a responsive target group, or to identify bought-in lists (or specific individuals in those lists) with corresponding lifestyle characteristics.

---

<sup>1</sup> For instance, Cooil et al. (2007) include age, income and education as moderators of customer loyalty and share of wallet; Kim and Street (2004) use customer characteristics to develop a customer targeting model; Palakurthi and Parks (2000) identify age, income, occupation and gender as the key customers' characteristics that drive housing demands; Lightner (2003) links age, education and income to consumers' preferences for certain characteristics of online shopping experiences.

## CLASSIFICATION OF AGE BY NAME

Although zip codes are largely used to conduct database profiling, other bits of information about customers can be matched with external sources of information to gain insights about who they are. For instance, and of particular illustrative interest for this paper, customer's first name can be used to infer their age.

Age has been reported in the academic literature to be relevant to important concepts such as brand loyalty and repurchase behaviors (Lambert-Pandraud et al., 2005), information processing (Phillips and Sternthal, 1977), or political opinions (Davidson, 2005). In practice, age is also widely used as a predictor, especially in direct marketing. For instance, Crié and Micheaux (2006) report that age is one of the most discriminant variables in targeting and scoring models in the insurance industry.

First names experience up and down fads over the years. In the United Kingdom, Ernest was a very popular first name in the 1930's; Kelly was very popular in the 1980's; and as an extreme example, Adolph as a given name almost disappeared after the Second World War. Appendix 1 reports a few example names with their most likely age patterns in the U.K.

To cite a marketing brochure, "Age is one of the most important factors for understanding more about your customers – and communicating with them more successfully as a result. [First name] classification can help you to identify the age of the people on your database by looking at the likely age profile of their first names. It can also help you to target new data sources which match your distinct customer age profile." (CACI, 2002)

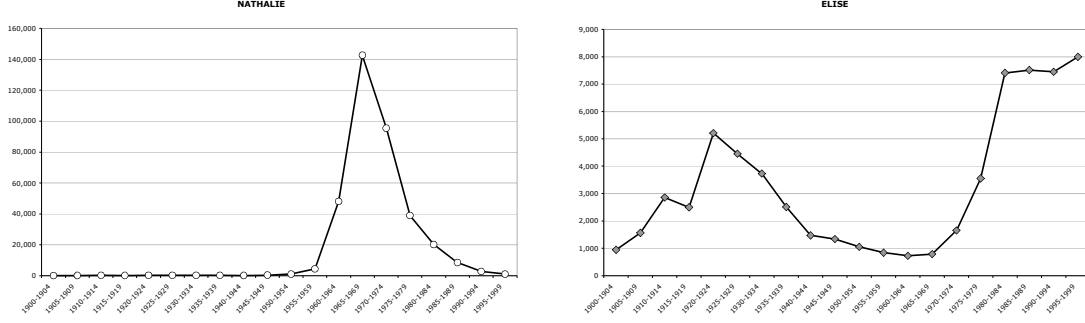
To estimate the most likely age distribution of a customer list (or of specific segments in that list), the typical age patterns of all customers' first names are summed up to obtain a global age distribution of the customers in that list. This is common practice in the industry, and we will show an illustration of that method later (see *Empirical Illustration*). We will refer to this age estimation approach as the *naïve model* throughout. Later, we will show the extent of the biases induced by this approach.

## COLLABORATIVE DATABASE PROFILING

Problems arise when a given name has been popular for many years, and cannot be used to precisely infer someone's age; or when a name experiences ups and downs in popularity. As an illustration, Figure 1 reports age distributions of women with two popular first names in France<sup>2</sup>: Nathalie and Elise. Nathalie has been an extremely popular name in the late sixties, and an astounding 39.1% of all the Nathalie's living in France today were born between 1965 and 1969, allowing very precise inferences. Elise, on the other hand, is a name that has known great variations. Popular for almost 30 years between 1910 and 1939 (with a pike between 1920 and 1924), this name stayed latent for 40 years, and then became in vogue again in the early eighties.

---

<sup>2</sup> We will use France as a test bed for the remaining of this paper. The *National Institute for Statistics and Economic Studies* (INSEE) provides first name tables and age distribution (per 5-year periods) for the *entire* population of France, hence allowing us to not worry about the statistical properties of the sample used to build reference tables.



**Figure 1** – Age distributions (age pyramids) in France, women with the given names *Nathalie* (left) and *Elise* (right), by date of birth (Source: INSEE).

Suppose the *naïve approach* is applied to a list of elderly people. First names would be extracted, compared with their known age pyramids, and summed up to form an average age distribution of the customer list as a whole. It is likely that several Elise’s will be in that list: many of the Elise’s living in France today were born between 1910 and 1939. Taking Elise’s age pyramid, the *naïve model* would largely overestimate the number of customers born after 1975, when the first name Elise became popular again.

We suggest that the customer list’s age pyramid should rather be estimated using a log-likelihood maximization approach. In other words, *what is the age distribution that would most likely generate the first name frequencies found in the customer list?* We call this approach *collaborative profiling*, because this age distribution is fit on the entire list of first names, and could be used later to refine age estimation of each individual in that list (a facet we will not discuss in this paper in the interest of space).

Although space limitation prevents an extensive development of the likelihood function, it can be written as (after some assumptions discussed hereafter):

$$\max \sum_{i=1}^I \ln \left( \sum_{j=1}^J p(\text{name}_i | \text{reference}_j) \cdot p(\text{reference}_j) \cdot p(\text{target}_j | \text{reference}_j) \right) \quad (1)$$

where:

- $i = 1..I$  Refers to the individual customers in the target list.
- $j = 1..J$  Refers to the categories into which the reference population’s age pyramid is categorized (e.g., [1980-1984]).
- $p(\text{name}_i | \text{reference}_j)$  The likelihood that an individual belonging to the *reference population* and being born during the time period  $j$  were given the first name  $i$ . This information is available from the reference tables.
- $p(\text{reference}_j)$  The age pyramid of the reference population, which is known.
- $p(\text{target}_j | \text{reference}_j)$  The relative weight of the target population in the reference population, to be estimated (see below).

Assuming independence,  $p(\text{target}_j) = p(\text{target}_j | \text{reference}_j) \cdot p(\text{reference}_j) \cdot g$  (where  $g$  is a normalizing constant), and one might find more natural to maximize the likelihood function by estimating  $p(\text{target}_j)$  directly (the age distribution of the target population to be recouped). However,  $p(\text{reference}_j)$  naturally captures complex patterns and historical bumps in birth and mortality rates (e.g., due to wars, epidemics, the baby boom), while  $p(\text{target}_j | \text{reference}_j)$  refers to a much smoother weighting function that can be estimated more easily.

Notice that Equation 1 is obtained by making several assumptions. One of them is critical: we assume that  $p(\text{name}_i | \text{reference}_j) = p(\text{name}_i | \text{target}_j) : \forall i, j$ , that is, we assume that at any period of time, the popularity of a name was similar among the reference population and among the target population. This assumption should be questioned. For instance, Lieberman and Bell (1992) report that parental characteristics (noticeably education and race) influence children naming patterns, showing that highly educated parents and less educated parents tend to give different first names to their babies. This approximation is nonetheless necessary, because while  $p(\text{name}_i | \text{reference}_j)$  is known,  $p(\text{name}_i | \text{target}_j)$  is not.

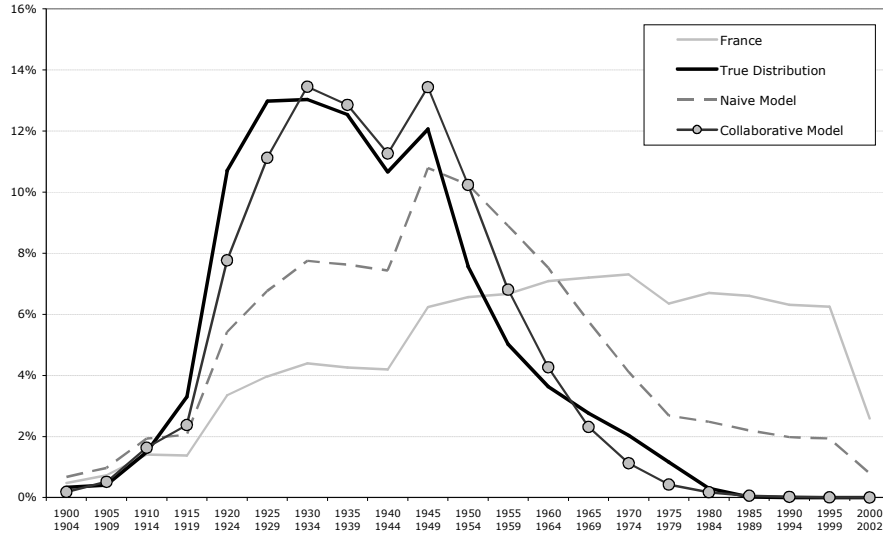
## EMPIRICAL ILLUSTRATION

As an illustration, we will test our model on a customer list of 9,046 individuals, for whom both first names and dates of birth are known. Names will be used to estimate the age distribution of the target list, using different models, while actual dates of birth will only serve as a benchmark to measure models' accuracy. The customer list comes from a database of customers known to come from a highly educated, wealthy, and rather aged population. The identity of the source will be kept anonymous for confidentiality reasons.

The models we will compare are as follow:

- **France** is the age pyramid of the reference population. If the target population were evenly spread across the population, this distribution would provide unbiased estimates.
- **Naïve model** is the age distribution of the customer list estimated directly from their first names only, using the methodology common in the industry.
- **Collaborative model** is the model we suggest in this paper: it performs a log-likelihood estimation procedure to estimate the underlying age pyramid of the target population that is the most likely to generate the above list of first names.

In this example, we assume that  $p(\text{target}_j | \text{reference}_j) \approx N(\mu, \sigma)^\omega$ . Figure 3 provides a graphical representation of the age distributions (reference population, target population, and estimated age distribution using the *naïve* and *collaborative* models). The age distribution estimated by our collaborative model is much closer to the true distribution, and much less biased than the one obtained using the naïve approach common in the industry. For instance, while 59.9% of the customers were born between 1920 and 1944, the *naïve* model estimates this figure at 35.0% (underestimated by 42%). The collaborative model estimates that same figure at 56.4%.



**Figure 2** – Age distribution (age pyramids) for France and for the target population, as well as statistical estimations of the latter using different models. The naïve model (common in the industry) is heavily biased toward the national age pyramid, and does not recoup the true distribution very well.

	Age	R <sup>2</sup>
Customer list (true distribution)	62.9	
France	39.8	.00
Naïve Model	52.4	.63
Collaborative Model	57.9	.95

**Table 1** – The naïve model is biased toward the national age pyramid, and underestimates the average age of the customers’ list by more than 10 years. The error is twice smaller using the collaborative model, and the age pyramid is much closer to the true distribution (as indicated by Pearson’s R<sup>2</sup>).

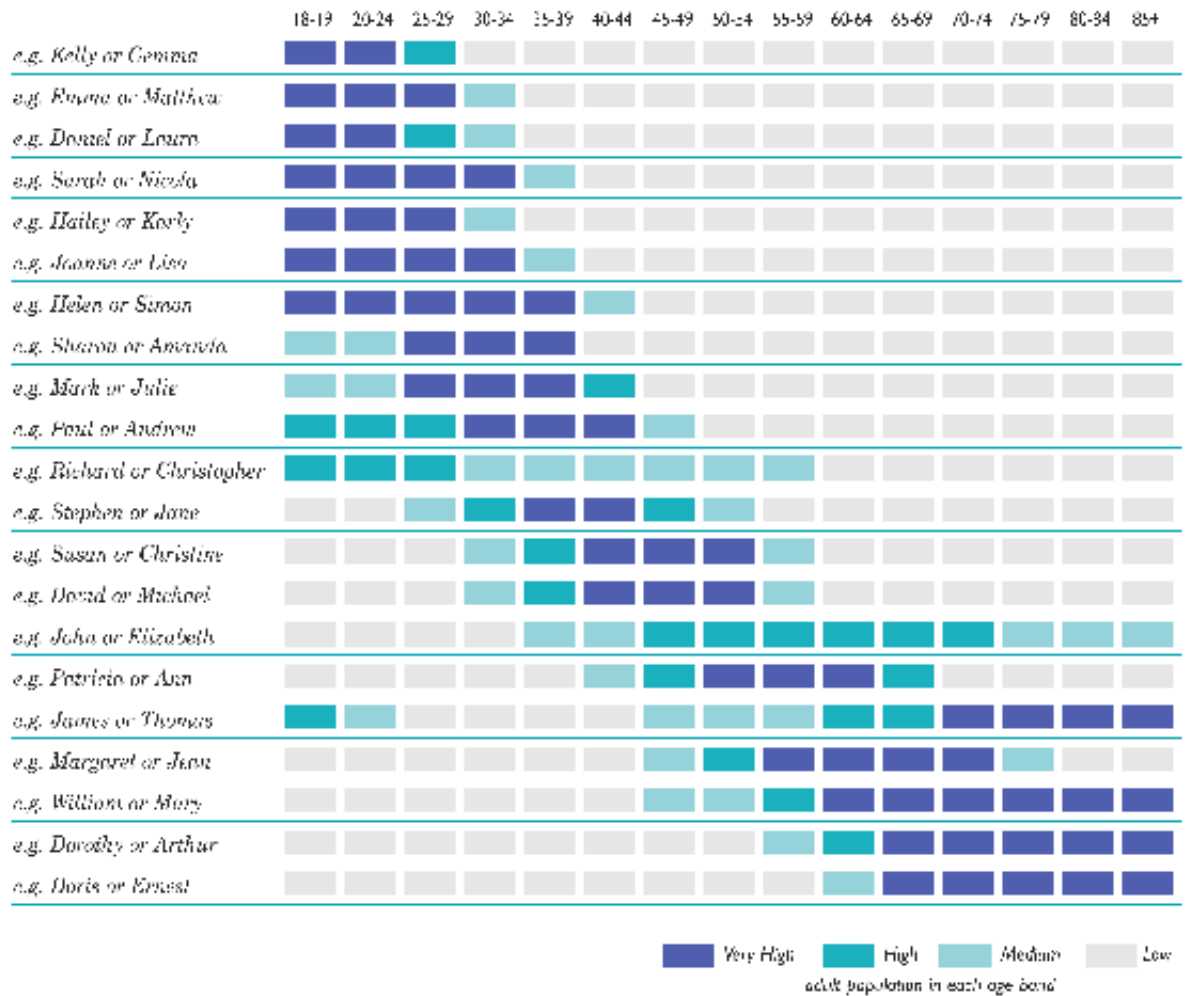
## CONCLUSIONS AND DISCUSSION

In their quest to tailor their marketing strategy to their customers’ specific profiles and needs, many firms use external data sources to infer the most likely demographics characteristics of their customers. In this paper, we showed that the methodology commonly used in the industry is severely biased towards average national figures. We discussed a maximum likelihood approach to correct for these biases, and illustrated this collaborative profiling method to age estimation using customers’ first names. We show that the approach we suggest gives much more precise and unbiased estimates. The same methodology can be used to provide unbiased estimates of customers’ income, housing situation or marital status from Census Bureau data, allowing firms to portray much more precisely their customers.

## REFERENCES

- CACI Information Solutions (2002), "MONICA: Classification of Age by Name for Marketers," marketing brochure. <http://www.caci.co.uk>. London, UK.
- Cooil, Bruce, Timothy L. Keiningham, Lerzan Aksoy, and Michael Hsu (2007), "A Longitudinal Analysis of Customer Satisfaction and Share of Wallet: Investigating the Moderating Effect of Customer Characteristics," *Journal of Marketing*, 71(January), pp. 67-83.
- Crié, Dominique, and Andrea Micheaux (2006), "From Customer Data to Value: What Is Lacking in the Information Chain?," *Database Marketing & Customer Strategy Management*, 13(4), pp.282-299.
- Davidson, Scott (2005), "Grey Power, School Gate Mums and the Youth Vote: Age as a Key Factor in Voter Segmentation and Engagement in the 2005 UK General Election," *Journal of Marketing Management*, 21, pp.1179-1192.
- Kim, YongSeog, and W. Nick Street (2004), "An Intelligent System for Customer Targeting: A Data Mining Approach," *Decision Support Systems*, 37, pp. 215– 228.
- Lambert-Pandraud, Raphaëlle, Gilles Laurent, and Eric Lapersonne (2005), "Repeat Purchasing of New Automobiles by Older Consumers: Empirical Evidence and Interpretations," *Journal of Marketing*, 69(April), pp.97-113.
- Lieberson, Stanley, and Eleanor O. Bell (1992), "Children's First Names: An Empirical Study of Social Taste," *The American Journal of Sociology*, 98(3), pp. 511-554.
- Lightner, Nancy J. (2003), "What Users Want in E-Commerce Design: Effects of Age, Education and Income," *Ergonomics*, 46(1-3), 153-168.
- Palakurthi, Radesh Rao, and Sara J. Parks (2000), "The Effect of Selected Socio-Demographic Factors on Lodging Demand in the USA," *International Journal of Contemporary Hospitality Management*, 12(2), 135-142.
- Phillips, Lynn, and Brian Sternthal (1977), "Age Differences in Information Processing: A Perspective on the Aged Customer," *Journal of Marketing Research*, 14(November), pp.444-57.
- U.S. Census Bureau (2000), "Profile of General Demographic Characteristics: 2000 Census of Population and Housing", U.S. Department of Commerce. Washington, USA.
- Wall Street Journal (2006), "Privacy for People Who Don't Show Their Navels," Jonathan D. Glater. Published: January 25, 2006.

## APPENDIX 1



Most likely age of typical first names in the United Kingdom.  
 Copyright © 2002, CACI Information Solutions. Reproduced with permission.