

Customer base segmentation of a telecommunication company by using Bayesian semi-parametric analysis of traffic data

Bruno Scarpa

Abstract Customer segmentation is one of the most important purposes of the customer base analysis for telecommunication companies. Since companies accumulate very large amount of data on customer behavior, segmentation is typically achieved by profiling and clustering traffic behavior jointly with demographic data and contracts characteristics. Unfortunately most algorithms and models applied for segmentation does not take into account the longitudinal characteristics of data. In particular in telecommunication traffic analysis, it is well known the importance of decreasing pattern of traffic in customer life and it is relevant to aggregate all clients with such a pattern, while other unknown clusters may be of interest of the marketing manager. Our approach for analyzing such a problem is based on specifying the distribution of functions as a mixture of a parametric hierarchical model describing the decreasing pattern segment and a nonparametric contamination that allows unanticipated curve shapes in subjects traffic. The parametric component is chosen based on prior knowledge, while the contamination is characterized as a functional Dirichlet process.

Key words: Statistical methods for customer relationship management, Bayesian nonparametrics, functional data analysis, telephone traffic, customer profiling, functional clustering

1 Introduction

To compete with other providers, a telecommunication company needs to have a very good knowledge of its own customers, in particular by knowing what each of them wants and needs (MacDonald and Dunbar, 1998). With this purpose, com-

Bruno Scarpa
Dipartimento di Scienze Statistiche, Università di Padova, via Cesare Battisti, 241 e-mail: scarpa@stat.unipd.it

panies divide customers in segments and profile them, by analyzing data available from the usage of the services. To segment a customer base, a variety of statistical and data mining techniques have been proposed and used, by considering different customer features such as geographic, cultural and ethnic characteristics, economic conditions, values attitudes and beliefs, lifestyle and traffic behavior (see for example Janusz, 2003).

Many paper and books introduces and describes in detail statistical methods for profiling customer base. The book by Berry and Linoff (2004) is a typical example of introduction on how data mining and statistical methods are used in marketing and CRM (Customer relationship management) problems, while for example Azzaolini and Scarpa (2004) and Rossi et al. (2005) give more technical descriptions of the statistical models used in marketing segmentation. More recently Ngui et al. (2009) present an up to date literature review of data mining techniques used in CRM.

However, in marketing segmentation of customers of a telecommunication company, data are available on the usage of services over time (outgoing and incoming traffic, number of text or multimedia messages in cellular phones are often available in monthly aggregations), while none of the papers we have reviewed, in segmenting customers, takes into account the longitudinal nature of the traffic variables considered. In this paper we propose an approach that considers the longitudinal nature of data, by thinking to them as curves in time. We also observe that strong prior information is available on the shapes of the curves, at least for some segments, with this information being important to include.

Potentially, one could view traffic on the different months in a customer lifetime as having a multivariate normal distribution, with both the mean and covariance estimated empirically (Ramsay and Silverman, 1997). However, this approach does not allow one to separate measurement errors in traffic from variability in the underlying smooth trajectories (Behseta et al., 2005). An alternative approach is to characterize the traffic measurements as error-prone realizations of functions expressed in terms of linear combinations of basis functions. By assuming normally distributed basis coefficients, one can allow variability in the functions. Related approaches have been considered, among others, by Ke and Wang (2001) and James (2002). Bigelow and Dunson (2007) and Thompson and Rosen (2006) developed Bayes approaches using adaptive splines. These approaches have the disadvantage of using a parametric distribution (normal) to characterize variability among the functions. A goal in modeling of traffic curves is to discover unknown behavioral segments. In these cases normality assumptions are not ideal. As a more flexible approach, Bigelow and Dunson (2009) proposed placing a Dirichlet process (DP) prior (Ferguson, 1973, 1974) on the distribution of the basis coefficients, following previous authors who used DP priors for random effects distributions (e.g. Bush and MacEachern, 1996). Rodriguez, Dunson and Gelfand (2008) avoid the basis function representation through the use of dependent DP mixtures. Because of the almost sure discreteness property of the DP, these approaches group subjects into functional clusters.

These approaches are useful for cases in which there is limited prior information about the shapes of the functions under study and one wants a highly flexible

model. However, in many cases, such as in the traffic curve case motivating this paper, there is abundant prior information available. It is not at all straightforward to include such information within the flexible semi-parametric approaches currently available for modeling of distributions of curves. Potentially, one could sacrifice some of this flexibility in favor of a parametric model. However, although a known parametric model may provide a good characterization of many curves, it is often the case that other curves are irregular. In the marketing segmentation application, one may be able to choose a parametric model for decreasing curves for problematic subjects, but it is much more difficult to parametrically characterize all the possible curves that may occur other subjects. To address a similar problem of studying basal body temperature of women during menstrual cycles, Scarpa and Dunson (2009) proposed to model the collection of distributions for different subjects, as a mixture of a parametric component and a nonparametric contamination. In particular, the nonparametric component is characterized as a functional DP. For a well chosen parametric model, curves drawn from the nonparametric component are considered irregular and likely unhealthy. In this paper we use a similar approach to cluster subjects in a telecommunication customer base, by grouping subjects with a decreasing pattern with a parametric cluster, but allowing for new non parametric clusters to include the remaining subjects.

Data available have been extracted from the Customer Base of a mobile telecommunication company. For each Subscriber Identity Module (SIM card) the company records monthly number, duration and value of the phone calls outgoing, incoming, an number of short messages (SMS) and multimedia message (MMS) sent.

For the analysis we use a random sample on 1000 SIM cards form the prepaid subscribers of the company. Traffic is available for 18 months since November 2004 until April 2006, even if in April 2005 there has been some default in the datawarehouse that prevent the recording of all data. Before extracting the random sample some data cleaning operation has been performed by eliminating completely silent SIM's or different behaviors particularly strange.

2 A model for traffic

We consider the simple case where only one variable of traffic is recorded or is of interest. Let $y_i(t)$ be the measured level of traffic at time t of the SIM card i , we follow a general framework for functional data analysis by writing

$$\begin{aligned} y_i(t) &= \eta_i(t) + \varepsilon_i(t), \\ \varepsilon_i(t) &\sim N(0, \sigma^2) \quad \eta_i \sim G, \end{aligned} \tag{1}$$

where η_i is a smooth function getting the traffic behavior and $\varepsilon_j(t)$ is a measurement error. We suppose that G is a prior for the distribution of functions η_i .

A common strategy is to simplify the problem by considering a basis representation:

$$\eta_i(t) = \sum_{h=1}^k \theta_{ih} b_h(t), \quad \forall t \in \mathcal{T}, \quad (2)$$

where $\mathcal{T}$ is the subset of $\mathfrak{R}^+$ indicating time, $\{b_h\}_{h=1}^k$ is a prespecified basis and $\theta_i = (\theta_{i1}, \dots, \theta_{ik})'$ are SIM-specific basis coefficients and we choose a normal model for the basis coefficients $\alpha_i \sim N_k(\alpha, \Omega)$, where α_i provide SIM-specific coefficients, $\Omega = \text{diag}\{\omega_1, \omega_2\}$ measures between SIM variability and α are global mean basis coefficients.

A Bayesian specification is completed with priors,

$$\alpha \sim N_{k+1}(\alpha_0, \Sigma_\alpha) \quad \sigma^{-2} \sim \mathcal{G}(c, d) \quad \omega_h^{-1} \sim \mathcal{G}(a_h, b_h) \quad \text{for } h = 1, 2$$

where $\mathcal{G}(a, b)$ denotes the Gamma distribution with mean a/b and variance a/b^2 . Here $\alpha_0, \Sigma_\alpha, c, d, a_h, b_h$ are prespecified hyperparameters.

Following Scarpa and Dunson (2009), we specify a model for η_i by combining a parametric model with a functional Dirichlet process (FDP) through a mixture specification. Our prior expectation is that there is a group of SIM cards for which traffic trajectories can be well characterized by a simple parametric model, but that there is a number of other SIMs for which this model does not apply. Thus, our model is specified as

$$\begin{aligned} \eta_i &\sim \pi_i \delta_{\mathbf{X}_i} \theta_i + (1 - \pi_i) G, \\ G &= \sum_{h=1}^{\infty} p_h \delta_{\Theta_h}, \quad \Theta_h \sim GP(\mu, \mathcal{C}), \end{aligned} \quad (3)$$

where $\mathbf{X}_i$ and θ_i characterize the parametric model, π_i is the probability that SIM card i falls in the parametric component, and $(1 - \pi_i)G$ is a nonparametric contamination, with G assigned an FDP prior.

The nonparametric component

The FDP prior implies that G is discrete, with probability masses $\{p_h\}_{h=1}^{\infty}$ placed at functions $\{\Theta_h\}_{h=1}^{\infty}$. The probability masses can be expressed in stick-breaking form (Sethuraman, 1994) as $p_h = V_h \prod_{l < h} (1 - V_l)$, with $V_h \sim \mathcal{B}(1, \alpha)$ and $\mathcal{B}(\cdot)$ denoting the beta density. In addition, the Θ_h s are generated independently of the weights from a Gaussian process having mean function μ and covariance function $\mathcal{C}$.

In order to specify G which is a probability measure with support on a functional space, we make a simplification by considering G as characterizing the distribution of curves through a discrete specification that places probability masses on different shapes, leading to functional clustering. In particular, if $\eta_i \sim G$, then the prior probability of allocation of SIM card i to functional cluster h is $\Pr(\eta_i = \Theta_h) = p_h$. Variability among the curves is controlled by the weights $\{p_h\}_{h=1}^{\infty}$ and by the covariance $\mathcal{C}$.

Samples from a GP can take a very wide variety of shapes that have limited sensitivity to the mean function. For this reason, by using a Gaussian process priors for unknown functions we allow an unknown, fixed mean, to avoid sensitivity to the scale of the traffic curves. We choose the exponential covariance function for the covariance function $\mathcal{C}$, controlling the types of shapes observed. The concentration α is estimated by assuming a gamma hyperprior (Escobar and West, 1995).

The parametric model

A cluster of SIM cards that marketing managers already know is characterized by a decreasing behavior. These SIM cards are probably moving toward some competitor and they are going to be churners soon. In order to consider a general shape for the decreasing pattern, we decided to consider a linear decreasing curve. However since the start of the decreasing could differ by SIM, we consider a piecewise linear function with three components where the first and third are horizontal and the second one has negative slope.

Let n be the number of months available for analysis and let $\mathbf{y}_i = [y_{i1}, \dots, y_{in}]^T$ denote the traffic of the i -th SIM card. The simple piecewise linear model used is

$$y_{it} = \begin{cases} \theta_{1i} + \varepsilon_{it} & 1 \leq t \leq k_i \\ \theta_{1i} + \theta_{2i} \left(\frac{t-k_i}{r_i} \right) + \varepsilon_{it} & k_i < t \leq k_i + r_i \\ \theta_{1i} + \theta_{2i} + \varepsilon_{it} & k_i + r_i < t \leq n \end{cases} \quad (4)$$

$$\varepsilon_{ijt} \sim N(0, \sigma^2),$$

where $k_i + 1$ and r_i indicate the month when decrease starts and the number of days in which traffic decreases, respectively; θ_{1i} is the level of traffic at the beginning of observation time, θ_{2i} is the rate of decrease in traffic level, and σ^2 is the measurement error variance. Note that, as explained in Scarpa and Dunson (2009), model 4 may be easily rewritten in a linear form, given r_i and k_i .

Priors for the parameters of such a model are $k_i \sim \mathcal{U}(1, 14)$ and $r_i \sim \mathcal{U}(1, 15)$ where $\mathcal{U}(a, b)$ indicates the Discrete uniform distribution between a and b .

3 Posterior computation

Posterior computation has been performed by following the modification of the Pólya urn Gibbs sampling algorithm (Bush and MacEachern, 1996) described in Scarpa and Dunson (2009).

The algorithm alternates between (1) a data augmentation step for allocating subjects to the parametric or nonparametric component, (2) standard Gibbs sampling steps for updating the parametric component model unknowns from their condi-

tional distributions; and (3) Pólya urn Gibbs sampling steps for updating the non-parametric component model unknowns.

4 Results and Discussion

Our focus was on segmenting the customer base by characterizing interesting features of the traffic curves for SIM cards. We are also interested in identifying and studying heterogeneity in trajectories which do not follow the decreasing pattern.

Weakly informative prior distributions are chosen for each of the parameters. We set $\alpha \sim N([50, 1]^T, \text{diag}(10, 0.1))$ as the hyperprior for the overall mean of the traffic during the first phase and for the increment, with the variance selected to allow a plausible amount of uncertainty. Based on descriptive statistics for data from a different sample, we chose the hyperpriors $\omega_h^{-1} \sim \mathcal{G}(1, 1)$ and $\pi \sim \text{Beta}(1, 5)$. Based on expert elicitation, we chose $\sigma^{-2} \sim \mathcal{G}(0.1, 1000)$, set the GP mean to $\mu(t) = 50$ and chose an exponential correlation function, with $\mathcal{G}(1, 1)$ on the two parameters. A Gamma prior was chosen for the Dirichlet parameter $\alpha \sim \text{Gamma}(0.1, 1)$.

We ran the algorithm for 3,500 iterations after a 500 iteration burn-in period. Convergence was assessed by examining the stationarity of the distribution of the parameters for the parametric component and for the mixture weights π_i , the overall variance σ^2 , and the means of the groups for the nonparametric component. Trace-plots of these parameters provided no evidence against convergence.

The posterior probability of allocation to the parametric component was greater than 90% for about 20% of the SIM cards under study, by identifying the first segment of customer base, the one with SIM cards with likely some problem. There were about 80% of SIM cars that had a posterior probability higher than 90% of being allocated to the nonparametric component. These SIM cards do not fit the decreasing pattern and will be aggregated in different clusters.

In order to identify the segments of customer base in the nonparametric component, we show in the left panel of Figure 1 the posterior distribution of the number of nonparametric clusters.

However, only two clusters for each iteration contain a sufficient number of subjects to be considered as a segment of the customer base. In fact, Figure 2 shows the distribution of SIM cards allocated to the single nonparametric clusters. Many clusters contain a very small number of subjects, while we observe an interesting concentration around 120 and 480.

In order to describe the segments obtained from the nonparametric components we plot in Figure 2 the simulated traffic of two SIM cards randomly chosen among the ones included in the first and second relevant nonparametric clusters.

The description of the trajectories of the parametric cluster is described in Figure 3. This provides histograms showing the variability among SIM cards in different features of the traffic trajectories, including month when decrease starts, the number of months of fall, starting level of traffic, and the final level traffic.

References

1. Azzalini, A., Scarpa, B.: *Analisi dei dati e data mining*. Springer, Milano (2004)
2. Berry, M. J. A., Linoff, G.: *Data mining techniques: for marketing, sales, and customer support*. John Wiley & Sons, New York (1997)
3. Behseta S., Kass R.E., Wallstrom G.L.: Hierarchical Models for Assessing Variability Among Functions. *Biometrika* **92**, 419–434 (2005)
4. Bigelow, J.L., Dunson, D.B.: Bayesian Adaptive Regression Splines for Hierarchical Data. *Biometrics* **63**, 724–732 (2007)
5. Bigelow, J.L., Dunson, D.B.: Bayesian semiparametric joint models for functional predictors. *JASA* **104**, 26–36 (2009)
6. Bush C.A., MacEachern S.N.: A Semiparametric Bayesian Model for Randomised Block Designs. *Biometrika* **83**, 275–285 (1996)
7. Escobar, M. D., West, M.: Bayesian density estimation and inference using mixtures. *JASA* **90**, 577–588 (1995)
8. Ferguson, T.S.: A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics* **1**, 209–230 (1973)
9. Ferguson, T.S.: Prior Distributions on Spaces of Probability Measures. *The Annals of Statistics* **2**, 615–629 (1974)
10. James, G.M.: Generalized linear models with functional predictors. *JRSS-B* **64**, 411–432 (2002)

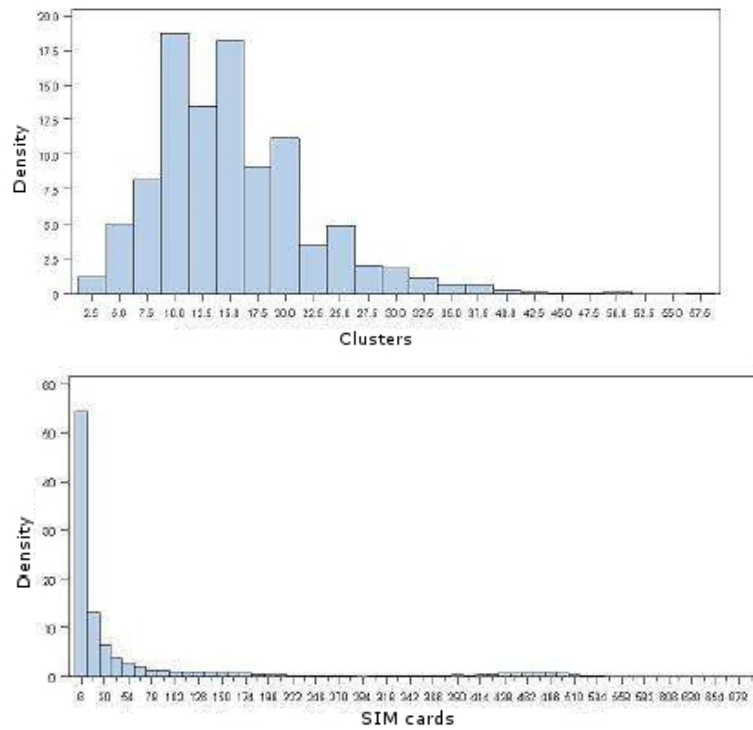


Fig. 1 Posterior distribution of the number of clusters and of the number of SIM cards in each cluster

11. Janusz, G.: Data mining and complex telecommunications problems modeling. *J. Telecomm. Inform. Technol.* **3**, 115–120 (2003)
12. Ke, C., Wang, Y.: Semiparametric Nonlinear Mixed-Effects Models and their Application (with Discussion). *JASA* **96**, 1272–1298 (2001)
13. McDonald, M., Dunbar, I.: Market segmentation. How to do it, how to profit from it. Palgrave Publ., Boston (1992)
14. Ngui, E.W.T., Xiu L., Chau, D.C.K.: Application of data mining techniques in customer relationship management: A literature review and classification. *Expert System with Applications*. **36**, 2592–2602 (2009)
15. Ramsay J.O., Silverman B.W.: *Functional Data Analysis*. Springer, New York (1997)
16. Rodriguez A., Dunson D.B., Gelfand A.E.: Bayesian nonparametric functional analysis through density estimation. **96**, 149–162 (2009)
17. Rossi, P.E., Allenby, G.M., McCulloch, R.: *Bayesian Statistics and Marketing*. John Wiley & Sons (2005)
18. Scarpa B., Dunson D.B.: Bayesian hierarchical functional data analysis via contaminated informative priors. *Biometrics* **65**, 772–780 (2009)
19. Sethuraman J.: A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650 (1994)
20. Thompson, W.K., Rosen, O.: A Bayesian model for sparse functional data. *Biometrics*. **64**, 54–63 (2008)

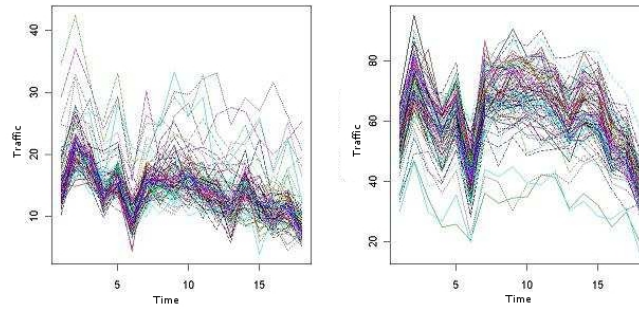


Fig. 2 Simulated behavior for two SIM cards included in the first two nonparametric clusters

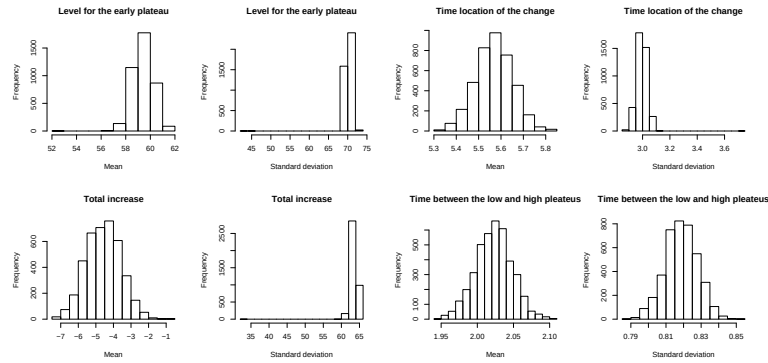


Fig. 3 Distributions among all the SIM cards of the SIM-specific posteriormean for the parametric model coefficients.