Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

FUNCTIONAL DATA ANALYSIS FOR NON HOMOGENEOUS POISSON PROCESSES

Martín Gastón
Dept. Statistics and O. R.
Campus Arrosadia
Public University of Navarre
Pamplona, 31006, SPAIN

Teresa León
Dept. Statistics and O. R.
Campus Burjassot
University of Valencia
Burjassot, 46100, SPAIN

Fermín Mallor
Dept. Statistics and O. R.
Campus Arrosadia
Public University of Navarre
Pamplona, 31006, SPAIN

ABSTRACT

In this paper we intend to illustrate how Functional Data Analysis (FDA) can be very useful for simulation input modelling. In particular, we are interested in the estimation of the cumulative mean function of a *non-homogeneous* Poisson Process (NHPP). Both parametric and nonparametric methods have been developed to estimate it from observed independent streams of arrival times. As far as we know, these data have not been analyzed as functional data. The basic idea underlying of FDA is treating a functional observation as a single datum rather than as a large set of data on its own. A considerable effort is being made in order to adapt some standard statistical methods for functional data, for instance Principal Components Analysis, ANOVA, classification techniques, bootstrap confidence bands, or outlier detection. We have studied a set of real data making use of these techniques and obtaining very good results.

1 INTRODUCTION

Time-varying arrival processes are an important ingredient of many complex systems in which discrete event simulation is the appropriate analysis tool. We can mention data arrival to telecommunication systems, failure times of repairable systems, arrival of pieces to be processed in an industrial plant, patients arriving to a health care centre, etc. These arrival processes are usually seen as discrete event processes that can be described by using appropriate stochastic point processes. A reasonable choice is the Poisson Point Process, due to the following characterization:
A stochastic event arrival process $\{N(t), t \geq 0\}$ is a *Poisson Process* if:

1. Events arrive one at a time.
2. The number of arrivals in the time interval $(t, t+s]$, $N(t+s) - N(t)$, is independent of the number and times of arrivals taking place from 0 until time t. That is, it is independent of the variable set $\{N(u), 0 \leq u \leq t\}$.
3. The distribution of $N(t+s) - N(t)$ is independent of t for all $t, s \geq 0$.

Let us consider the case of patient arrivals to a health care centre (HCC). Then, properties 1 and 2 can be interpreted as follows. Patients arrive at the HCC on an individual basis, knowing nothing about the patients that have arrived before them (or whatever they know has no influence in their decision about when to go to the health care center) and without anyone coordinating the arrivals of patients according to a pre-established plan. Condition 3 sets the *homogeneity* of the process through time. This condition is more difficult to assume in this kind of arrival process, because arrivals usually peak several times throughout the day. If this third condition is removed from the definition, we get a *non-homogeneous* Poisson Process (NHPP).

Let us denote $\Lambda(t)$ the expected number of arrivals until time t, that is, $\Lambda(t) = \mathrm{E}[N(t)]$ $(t \geq 0)$. When $\Lambda(t)$ can be derived, its derivative is called the arrival ratio function $\lambda(t) = \Lambda'(t)$, which can be interpreted as the instantaneous expected number of arrivals per unit time at time t. In a NHPP this instantaneous expected mean varies through time. Both $\lambda(t)$ and $\Lambda(t)$, completely determine the Poisson Process.

Several methods have been proposed to estimate NHPP, of parametric type in Lee et al. (1991), Kuhl et al. (1997) and Kuhl and Wilson (2000) and of nonparametric type in Leemis (1991, 2004) and Kuhl and Wilson (2001).

In Alexopoulos el al. (2008) it is shown that, in a health care context, the homogeneous Poisson process (HPP) is rarely an appropriate model for random arrivals. Nevertheless, the review made by the authors found that investigators explicitly assumed that the arrival times follow an HPP. Arrival data in our illustrative example (Figure 1), coming from a primary health center in Colombia, clearly show a non-stationary behavior. They were successfully modeled by using a *non-homogeneous* Poisson Process (NHPP) in Azcarate et al. (2008) whose cumula-

tive mean function ($\Lambda(t)$) was estimated smoothing the piecewise linear estimator proposed by Leemis (1991).

Our belief is that high-fidelity probabilistic input models are necessary to perform accurate simulation experiments. In this paper we provide a novel alternative way for analyzing input arrival data, based on Functional Data Analysis, which shows a better performance than the traditional approach for our illustrative example. In particular we consider the problem of testing the homogeneity of the observed arrival processes and, if it is the case, the question of identifying the groups in which the arrival processes can be classified. Each one of these groups is characterized by a different arrival ratio function.

A first general objective of this paper is to introduce FDA as an alternative way of analysing arrival process. We have to mention that Bouzas et al. (2006) use FDA to model the mean of a doubly stochastic Poisson process but the scope of the problems addressed in this paper is broader than theirs. A second objective is to formulate and implement a methodology for the analysis of arrival processes having possibly different patterns, maybe because there exists a seasonal behaviour or because arrival processes come from different populations. The third objective is to show the performance of this methodology by applying it to a set of real data describing the patient arrivals to a HCC.

The remainder of this article proceeds as follows: section 2 describes the methodology for analyzing homogeneity in observations from arrival processes when they exhibit different profiles and for identifying the groups in which the arrival processes can be classified. Section 3 illustrates the application of this methodology to a set of real data. The main findings of this research and other capabilities of the FDA in the input modeling context are summarized in section 4.

2 METHODOLOGY

2.1 Functional Data Analysis and Functional ANOVA

The term Functional Data Analysis refers to the set of statistical methods for analyzing continuous data as curves or images for instance. The starting point for the development of FDA was the functional extension of the Principal Component Analysis (PCA) to stochastic processes (Deville 1973). Nowadays FDA is quite popular mainly since the publications of the monographies by Ramsay and Silverman (1997, 2002) and Ferraty and Vieu (2006) and not only new theoretical developments are introduced but also the applications in disciplines as medicine, econometrics or biostatistics are increasing.

In practice functional data are usually observed and recorded discretely, i.e. a functional datum for replication i arrives as a set of discrete measured values, $y_{i1}, y_{i2}, \ldots .y_{in}$, at times $t_{i1}, t_{i2}, \ldots .t_{in}$, however we can transform these observations into a function $x_i(t)$ by interpolating or smoothing. One of the most familiar smoothing procedures involves representing functions as linear combinations of known basis functions. The polynomial, the spline, the wavelet and the Fourier bases are the more widely used in practice. If the discrete observed values can be assumed errorless, an appropriate process to convert them to a function is interpolation. In this functional context, each observation of an arrival process during a time interval $(0,T)$ provides a functional datum which is built from the pairs of recorded data (t_i, n_i) where t_i denotes the n_i^{th} arrival time.

The classical summary statistics have a counterpart in the context of FDA. For instance the mean function is defined as the average of the functions point wise across replications. A considerable effort is being made in order to adapt some standard statistical methods for functional data.

In this paper we make use of a one-way ANOVA procedure due to Cuevas et al. (2004) to test for differences among two or more independent groups of functional observations. This test can be seen as an asymptotic version of the well-known ANOVA F-test. Actually it is based on the numerator of the classical F statistic. To be more precise the test statistic is defined as: $\sum_{i<j} m_i \left\| X_{i.} - X_{j.} \right\|^2$ where $\overline{X}_{i.}$ stands for the mean function of the i-th group, m_i denotes its size and $\| . \|$ the usual L^2 norm for functions: $\|x\| = \left(\int x(t)^2 \, dt \right)^{1/2}$. The authors propose a numerical Monte Carlo procedure to handle in practice the asymptotic distribution of the test statistic.

2.2 Functional PCA and Cluster Analysis

We have also performed a functional PCA of our data. It is well known that, in the multivariate context, PCA finds the directions in the observation space along which the data have the highest variability. For each principal component, the analysis provides a loading or weight vector which gives the direction of variability corresponding to that component. The functional PCA associated to a functional data set X is defined to have the same optimal properties as in the multivariate case. In the functional case each principal component is specified by a principal component weight function $\xi(t)$ defined over the data range. The score z_i of the individual $x_i(t)$ in the sample is given by $z_i = \int \xi(t) x_i(t) dt$. The principal components provide a decomposition of the curves in terms of their variability components i.e. the basis whose elements are the principal components $\{\xi_1(t), \xi_2(t), \ldots, \}$ allows for a very useful representation of the curves. In particular when the variability explained by the first k components is close enough to

100%, the *k*-tuple with the scores of these principal components constitute an accurate description of the curve. Using these scores we can obtain a classification of the (cumulative) mean-value functions by applying some multivariate clustering technique. For example, we have implemented the cluster procedure called partitioning around medoids, PAM for short (Kaufman 1990).
This method is suitable for clustering problems in which one is interested in the characterization of the clusters by means of typical objects, representing the various structural features of objects under investigation. The algorithm PAM first computes *k* representative objects, called *medoids*. A medoid can be defined as that object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal. After finding the set of medoids, each object of the data set is assigned to the nearest medoid. The dissimilarity matrix can be obtained using the Euclid distance but many different options can be considered.

The average silhouette width allows us to select the optimal number of clusters. Let us recall that the silhouette width for the i-th observation, *s(i)*, is defined as :

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where *a(i)* is the average dissimilarity between the observation *i* and all other points of the cluster to which *i* belongs, meanwhile *b(i)* is the mean dissimilarity between *i* and its neighboring cluster i.e. the nearest one to which *i* does not belong. Clearly, observations with a large *s(i)* are very well clustered, *s(i)* around 0 means that the observation is between two clusters, and observations with a negative *s(i)* are probably placed in the wrong cluster. The average silhouette width is obtained as the average of the silhouette widths of all the observations. This quantity ranges in the interval (-1, 1) and it has been used both to evaluate the quality of a classification and to estimate the "correct" number of clusters: the partition with the maximum average silhouette width is taken as the optimal partition.

3 AN ILLUSTRATIVE EXAMPLE

In this section we apply the described methodology to the data of patient arrivals to a primary health center of the Hospital San Juan de Dios in the city of Pamplona (Colombia). Patients arrive at the center without a scheduled appointment, from Monday to Friday. From 5 a.m. onwards they start to arrive at the invoicing office, where two people attend patients from 7 a.m. to 9 a.m. Patients are not allowed to arrange appointments by telephone.

To check for time-dependence in the arrival pattern, daily patient arrival time data for a period of nine months were analyzed.

Although no monthly pattern emerged, time-dependence in day of the week and hour of the day were detected. A non-*homogeneous* Poisson Process was considered for each day of the week.

3.1 Construction of the Functional Data

Our original data are the observed arrival times of patients to the primary health centre during 150 days. Clearly, the total number of patients depends on the day as Figure 1 shows. Let us briefly describe how we transform these daily data into functions of type $n = f(t)$, where t is the time and n is the total number of patients arrived until that time. As the process is the same for all of them, we are not using sub indices to distinguish between the different days.

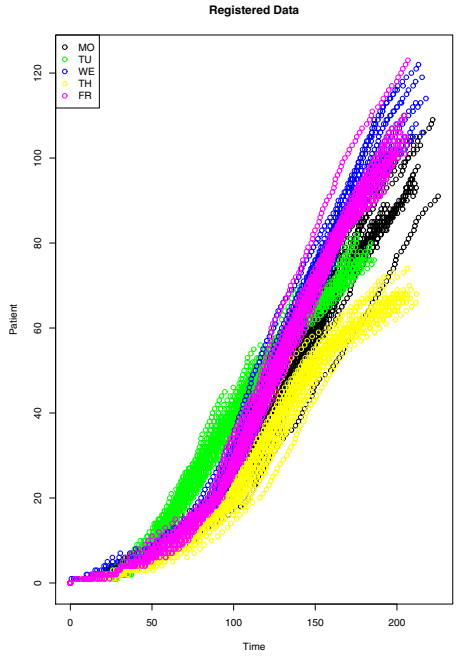

Figure 1: Original observations.

Firstly we have performed a linear interpolation where the nodes are the points (t_i, n_i) obtaining a non decreasing piecewise linear function $f(t)$, such that $f(0) = 0$ and $f(t_i) = n_i$. Interpolation provides a means of estimating the function at intermediate points, then $f(t)$ can be evaluated at a grid of equally spaced (2 seconds) argument values.

Finally, in order to achieve that all the data are defined on the same time interval, the functions are extended from the last registered time until time

$$T = \max\{t_i / (t_i, n_i) \text{ is an arrival datum}\}$$

by setting them as constantly equal to the total number of arrivals in that day. See Figure 2.
This data preprocessing is essential in the FDA approach. From now on all the appropriate techniques can be used.

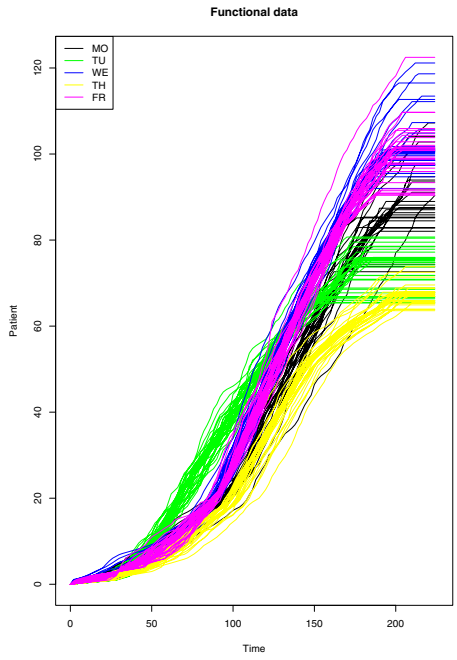

Figure 2: Functional data.

3.2 ANOVA Test and Clustering Input Observations.

First of all we have used the ANOVA test for functional data to check whether or not the observed differences between the average curves of the different days of the week are statistically significant. The results indicated a strong evidence (p-value < 0.005) in favor of the hypothesis that the curves are actually different. Table 1 shows the L^2 distance matrix between the average curves. We can see that the curves corresponding to Wednesdays and Fridays are the closest.

Table 1: L^2 distances between the average curves.

Dist.	Mon.	Tues.	Wed.	Thurs.
Mon.	0.00			
Tues.	10741.32	0.00		
Wed.	10325.09	23036.51	0.00	
Thurs.	11849.07	15873.74	42961.96	0.00
Fri.	7848.40	20112.35	176.77	37742.43

Then we have performed a functional PCA obtaining that the first two principal components explain more than 97% of the variability. Thus, the representation of the curves (daily observed NHPP) in the first factorial plane through their scores reproduces with high fidelity their relationship of proximity.

Next we have used the PAM method to classify our data which consist of bivariate observations, because we identify each functional datum with their expansion in terms of the first two functional principal components.

In our case the average silhouette width is maximum for K=4 groups and it is equal to 0.63, indicating a clear cluster structure. Figure 3 shows the silhouette plot for K=4.

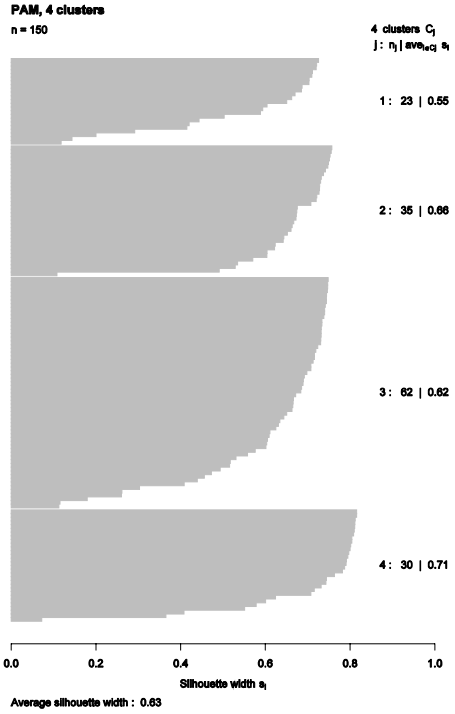

Figure 3: Average silhouette plot.

The cluster analysis suggests that the curves corresponding to Wednesdays and Fridays belong to the same group while the curves corresponding to remaining days of the week form their own cluster. The only exception is that a few of the curves from Mondays are clustered in other days' groups (see Table 2 and Figure 4).

Table 2: Composition of the clusters.

	MON.	TUES.	WEDN.	THUR.	FRI.
CL1	23	0	0	0	0
CL2	1	34	0	0	0
CL3	1	0	31	0	30
CL4	2	0	0	28	0

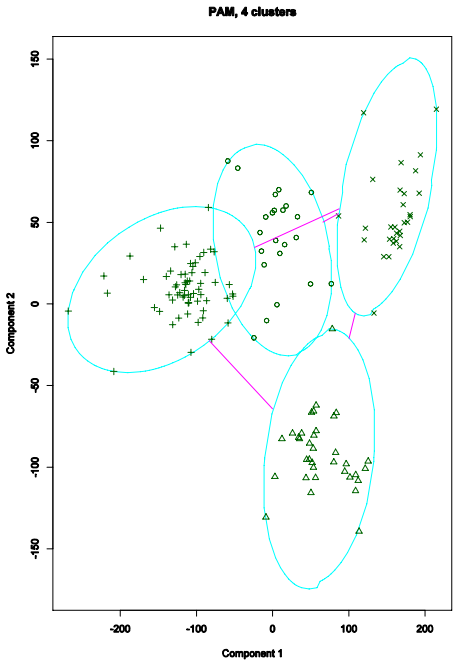

Figure 4: Four clusters: Mondays, Tuesdays, Thursdays and {Wednesdays+Fridays}.

In order to confirm this result we have used the ANOVA procedure to test for differences among the curves of Wednesdays and Fridays. We have obtained a p-value > 0.1, then the observation is consistent with the null hypothesis of equal means. The (cumulative) mean-value functions of Wednesdays and Fridays will not be distinguished when we estimate the average (cumulative) mean-value function.

The power of this approach is that it takes into account all the shape of the curve. For example, in Azcárate et al. (2008) the classification in homogeneous groups was done by only considering the total number of arrivals and then applying a classical ANOVA. The Scheffe's test was employed to identify significantly homogeneous groups among populations. In this case different conclusions were reached (see Table 3). For a significance level of α=0.05, three different patterns are obtained for the arrival process, one for Tuesdays and Thursdays, another for Wednesdays and Fridays and the third one for Mondays.

Table 3: Sheffe´s test for homogeneity applied to the total arrivals per day.

Day	Sample size	subgroups for α = .05		
		1	2	3
Thurs.	28	69,64		
Tues.	34	73,91		
Mon.	27		89,00	
Fri.	30			100,77
Wedn	31			103,19
Sig.		,177	1,000	,723

All the procedures described in this paper have been implemented making use of the capabilities of R (R Development Core Team , 2007). We have used the library "fda" (Ramsay, Wickham and Graves 2007) and we have performed an implementation of the ANOVA test due to Cuevas et al. (2004).

4 FINAL REMARKS AND FURTHER RESEARCH

In this paper we have provided an alternative way to analyze observations from NHPP based on the use of Functional Data Analysis. Although we have centered our developments in testing the homogeneity and classification of the observed processes, FDA provides a useful framework for studying another problems related with NHPPs. For example, the estimation of the cumulative mean function can be done from this perspective. As in univariate point estimation, together with the functional sample mean, other location estimators can be considered in order to get some idea about the "central value" of the population from which a sample of curves has been drawn, for instance a "functional median" or an α-trimmed mean.

Figures 5 and 6 show the functional means and medians, respectively, for each week day. They are quite similar, and we can think of the last ones as an alternative to simulate the arrivals.

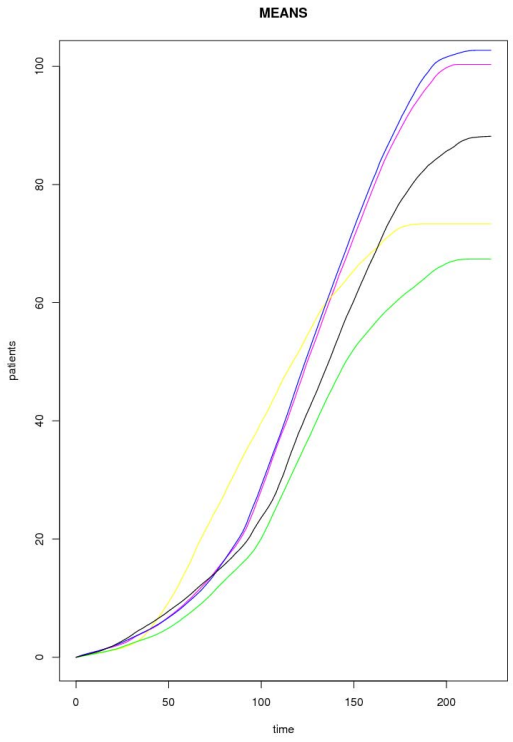

Figure 5: Functional means for the 5 groups of data.

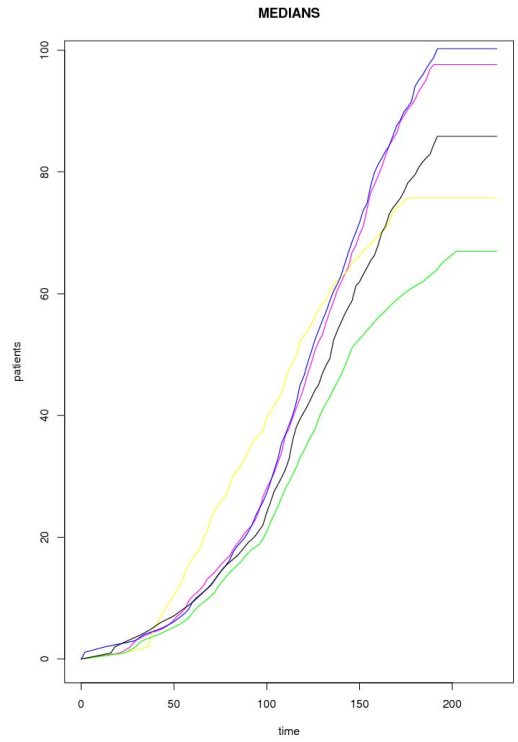

Figure 6: Functional medians for the 5 groups of data.

Even more, in many cases obtaining confidence bands is more interesting. A confidence-interval estimator for the cumulative intensity function is developed in Leemis (2004). From the functional point of view, Cuevas et al. (2006) use the bootstrap methodology for functional data and present a study of the performance of the bootstrap confidence bands (obtained with different resampling methods) of several functional estimators.

In Febrero et al. (2007) a functional outlier detection procedure is proposed**.** The authors consider that a curve is an outlier if it has been generated by a stochastic process with a different distribution than the rest of curves, which are assumed to be identically distributed. The statistic of the procedure measures the deviation of each functional observation to the mean also taking into account the standard deviation. They have implemented a bootstrap procedure for the determination of the "threshold" value for the statistic. The computer codes, written in the R language (R Development Core Team, 2007), are available from the authors and we have used them to analyze our data set.

We have looked for outliers among our data set and we have found out that one of the Mondays' observations can be considered as an outlier, as Figure 7 shows. In Figure 4 this outlier corresponds with the triangle symbol that is located at the top of the lower cluster.

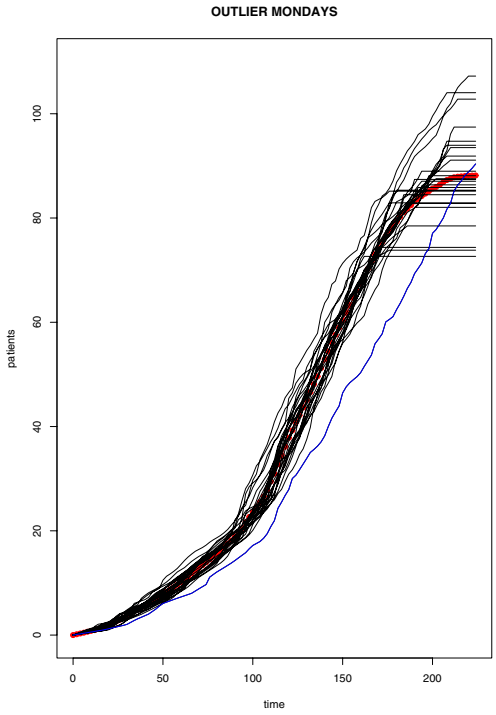

Figure 7: The mean of the Mondays' arrivals is displayed in red and the outlier curve in blue color.

Currently we are working on the use of this functional data approach to curve simulation and particularly in the simulation of NHPPs.

ACKNOWLEDGMENTS

The authors are indebted to the Spanish Ministry of Education and Science for financing this research with grants TIN2006-10134 and DEP2006-56076-C06-06/ACTI.

REFERENCES

Alexopoulos, C., Goldsman, D., Fontanesi, J., Kopald, D. and Wilson, J.R. 2008. Modeling patient arrivals in community clinics. *Omega*, 36: 33-43.

Azcarate, C., Mallor, F. and Gáfaro, A. 2008. Multiobjective Optimization in Health Care Management. A metaheuristic and simulation approach. *Algorithmic Operational Research*. (In press)

Bouzas, P R., Valderrama, M.J., Aguilera, A.M. and Ruiz-Fuentes, N. 2006. Modelling the mean of a doubly stochastic Poisson process by functional data analysis. *Computational Statistics & Data Analysis* 50: 2655-2667.

Febrero M., Galeano, P. and González-Manteiga, W. 2007. A functional analysis of NOx levels: location and scale

estimation and outlier detection. *Computational Statistics*, 22: 411-427.

Cuevas, A., Febrero, M. and Fraiman, R. 2004. An ANOVA test for functional data. *Computational Statistics & Data Analysis* 47: 111-122.

Cuevas, A., Febrero, M. and Fraiman, R. 2006. On the use of the bootstrap for estimating functions with functional data. *Computational Statistics & Data Analysis* 51: 1063-1074

Deville, J.C. 1973. Estimation of the eigenvalues and of the eigenvectors of a covariance operator. Technical Report. Ann L'INSEE,

Ferraty, F. and Vieu, P. 2006. *Nonparametric Functional Data Analysis. Theory and Practice*. Springer.

Kaufman, L. and Rousseeuw, P.J. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.

Kuhl, M.E. and Wilson J.R. 2000. Least squares estimation of nonhomogeneous Poisson processes. *J. Statist. Comput. Simul.* 67: 75-108.

Kuhl, M.E. and Wilson J.R. 2001.Modeling and simulating Poisson processes having trends or nontrigonometric cyclic effects. *European Journal of Operational Research.* 133: 566-582.

Kuhl, M.E., Wilson J.R. and Johnson, M.A. 1997. Estimating and simulating Poisson processes having trends or multiple periodicities. *IIE Transactions*, 29: 201-211.

Lee, S., Wilson, J.R. and Crowford, M.M. 1991. Modeling and simulation of a nonhomogeneous Poisson Process having cyclic behavior. *Communications in Statistics-Simulation and Computation*, 20: 777-809.

Leemis, L.M. 1991. Nonparametric estimation of the cumulative intensity function for a non homogeneous Poisson process. *Management Science* 37: 866-900.

Leemis, L.M. 2004. Nonparametric estimation and variate generation for a nongomogeneous Poisson process from event count data. *IIE Transactions* 36: 1155-1160.

R Development Core Team. 2007. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ramsay, J.O., Wickham, H. and Graves S. 2007. fda: Functional Data Analysis. R package version 1.2.3. http://www.functionaldata.org

Ramsay, J.O. and Silverman, B.W. , 1997. *Functional Data Analysis*. Springer.

Ramsay, J.O. and Silverman, B.W. 2002. *Applied Functional Data Analysis. Methods and Case Studies*. Springer.

AUTHOR BIOGRAPHIES

MARTÍN GASTÓN is a PhD student at the Department of Statistics and Operations Research in the Public University of Navarre (Spain). He also works as researcher at the Renewable Energy National Center (Spain). His present research interests are mainly concerned on Functional Data Analysis, Simulation, Mathematical Morphology and Statistical Learning Theory. His e-mail address is <martin.gaston@unavarra.es>.

TERESA LEÓN is an associate professor at the Department of Statistics and Operations Research in the University of Valencia (Spain). Her present research interests are mainly concerned on Functional Data Analysis, Simulation, Mathematical Morphology, Image Retrieval and Fuzzy Set Theory and applications. Her e-mail address is <teresa.leon@uv.es>.

FERMÍN MALLOR is an associate professor at the Department of Statistics and Operations Research in the Public University of Navarre (Spain). His topics of research cover different areas of probability (Renewal Theory and its application to Reliability Systems) and Operations Research (Simulation, Scheduling and Optimization with Simulation). He has successfully applied his knowledge on these areas in the analysis of complex real problems arisen in several industrial companies and institutions. His e-mail address is <mallor@unavarra.es>.