
MOX-Report No. 36/2016

# Marked Point Process models for the admissions of heart failured patients

Mancini, L.; Paganoni, A.M.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

# Marked Point Process models for the admissions of heart failured patients

Luca Mancini[♯] and Anna Maria Paganoni[♯]

October 9, 2016

[♯] MOX– Modellistica e Calcolo Scientifico
Dipartimento di Matematica
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy

anna.paganoni@polimi.it, luca.mancini@mail.polimi.it

**Keywords**: Marked Point Process; Conditional intensity function; Hawkes process; Temporary ground process; Inference; Simulation.

**Abstract**

The aim of this paper is to model the stochastic process of hospitalizations with Marked Point Processes. We examine the longitudinal dataset including the admissions of heart failured patients to Lombardia hospitals on a follow-up period of six years since January 1st, 2006. We analyse four separate groups of patients, which we call HF groups, according to their diagnoses-codes contained in the SDO (dimission hospital discharge form) of their first hospitalizations.
The statistical model links the temporal trend of hospitalization (the ground process) with the length of stay (the mark) at each event. Instead of framing our application in the more theoretical context of the counting measures and processes, we make use of the conditional intensity function, a parametric approach which leads us to deal with Hawkes processes.
Hypotheses are made on the mark concerning its distribution as well as its independence or dependence with the ground process. Independence is better to model and give us significant results while dependence is harder to be dealt with due to computational and modeling issues.
Finally, we provide a general framework for modeling longitudinal data with a MPP as of methods for statistical inference and suggest a specific model for our topic, validating it through a goodness of fit technique.

## 1 Introduction

Nowadays, Marked Point Processes (MPPs) are becoming increasingly relevant not only from a theoretical point of view but also in real applications. We

may find examples of these stochastic processes in finance, queueing theory and telecommunication network and, especially, in seismology to model earthquakes, taking into account their temporal trends together with their magnitudes.
However, there are not significant applications in a biomedical context so far in spite of lots of longitudinal studies concerning clinical researches, therapeutic evaluations and epidemiologies. Here, we aim at modeling a longitudinal dataset involving the hospitalization process of heart failured patients with MPPs. Since it is a quite new approach to this topic, we give details as of some particular parametric models and inferential procedures.
First of all, we should recall that by point process we mean a model of points randomly distributed in some space and indistinguishable for their locations. Points represent times of events or, better, times elapsed since a starting point and will be referred to as a collection of random variables $T_i$, the timepoints at which the i-th recording of an event takes place. Of course every point or statistical unit not only cointains information on times but also secondary features which constitute the so called **marks** of the points and are indeed random variables called $\boldsymbol{Y_i}$.
For instance, when an earthquake occurs, we can collect the time of occurrence $T_i$ but also information $\boldsymbol{Y_i}$ about its magnitude or spatial location. Also, when a patient is admitted to a hospital, we know the starting date of the hospitalization and the related length of stay.
There are two ways of characterizing a marked point process (see Daley and Vere-Jones, 2008). It can be studied in the context of counting processes and measures or through the conditional intensity function $\lambda(t, \mathbf{y}|\mathcal{H}_t)$ which represents the infinitesimal expected rate of events at time $t$ with marks $\mathbf{y}$, given all the observations up to $t$ and is made up of two parts (Harte, 2010):

$$\lambda(t, y|\mathcal{H}_t) = \lambda_g(t|\mathcal{H}_t) f(\mathbf{y}|t, \mathcal{H}_t), \tag{1}$$

where $\mathcal{H}_t$ is the filtration of the process, $\lambda_g(t|\mathcal{H}_t)$ is the intensity of the ground process (i.e. of the times $\{T_i\}$) and $f(\mathbf{y}|t, \mathcal{H}_t)$ stands for the multivariate distribution of the marks $\{\boldsymbol{Y_i}\}$, which generally depend on time.
The most difficult issue is the modeling of the ground process intensity function; if we are able to assign a particular expression for it, we may then focus on specific parametric models known as the Hawkes processes (see Daley at al., 2008). However, it may be difficult to model the mark distribution too, especially due to its relation with time. Then, some assumptions on the mark structure are usually made, leading to unpredictability and independence.
A mark is unpredictable if it does not depend on the past and can be regarded as conditionally i.i.d given the past of the process while the independence hypothesis is stronger and means that the $\{\boldsymbol{Y_i}\}$ are independent of everything else except maybe $\{T_i\}$.
The main advantages of framing a marked point process under this parametric approach concern the statistical inference as well as the simplicity in suggesting some algorithms for parameters' estimation and methods for goodness of fit and

simulation.

The paper is organized as follows. In section 2, we introduce and analyse the dataset. In section 3, we introduce a parametric model for dealing with marked point processes and suggest some inferential procedures for our topic. In section 4, we contextualize the model, assigning particular expressions to the right hand terms in (1). Then, we discuss the results, presenting a simultation method for our hospitalization stochastic process.
All the statistical models and tools have been implemented by using `R` software (see R Core Team, 2014). Precisely, the `R` package we used to model MPPs indexed by time is named `PtProcess` (see Harte, 2010) which provides a structure and environment so as to define and analyse our own MPP models. We therefore implement some specific `R`-functions for the fit of Hawkes processes of different kinds, which could be definitely included in the existing `R`-package.

## 2 Data description

Data comes from a long pre-processing of *Regione Lombardia* database of hospital discharge forms, collecting events of hospitalization from January 1st, 2006 to December 31st, 2012 for a follow-up period of six years.
The dataset consists of a list of events of admissions, containing both demographical and administrative information of a patient at that time. It is also possibile to follow the patients' hospitalizations in an individual way thanks to their encrypted ID. Here, we decide to focus on patients older than 18 which have less than six hospitalizations, whence analysing the 95.10% of all available events (see Ieva et al, 2014). Then, 51,186 patients are considered and their related 83,138 events of hospitalizations are analysed.
We mainly aim at modeling the hospitalization process of heart failured patients, linking its temporal trend with the length of stay through a marked point process. When dealing with these stochastic processes, it is quite common to fix an initial time of observation, i.e. in earthquakes' context where one of the main goals is to continuously monitor their temporal trend and relation with the magnitude for safety and prediction purposes. Thus, we basically focus on the following two variables:

- **Time**: time elapsed since January 1st, 2006.
- **Length Of Stay (LOS)**: difference in days between the date of an admission and date of the relative discharge.

Furthermore, Mazzali et al.(2015) showed that heart failure should not be treated and diagnosed in the same way, leading to a sharper distinction of patients in four subgroups, which will be called HF (Heart Failure) groups, due to the classification of patient's disease. Actually, in order to cluster heart failured people, Mazzali et al. (2015) rely on the type and number of patient's diagnosis coded

with ICD-9-CM (International Classification of Diseases, 9th revision, Clinical Modification) and on two slightly different criteria: AHRQ and HCC (see AHRQ, 2015 and Pope et al. 2004).
Then, the hospitalization process should be studied in a more specific way according to the given HF groups: the most meaningful one is the first which includes patients suffering from evident heart failure condition. In Table 1, we give an overview of some useful summary statistics for every given group.

| **HF Groups** | No. events | No. patients | % Men | % Women | LOS mean and sd [days] |
|---|---|---|---|---|---|
| G1 | 57,622 | 34,866 | 52.97 | 47.03 | 13.86 ($\pm$14.99) |
| G2 | 12,750 | 7,617 | 35.89 | 64.11 | 12.65 ($\pm$16.96) |
| G3 | 12,387 | 8,487 | 53.09 | 46.91 | 16.02 ($\pm$16.82) |
| G4 | 379 | 216 | 50.92 | 49.08 | 14.56 ($\pm$13.86) |

Table 1: Summary statistics for HF groups

Since the length of stay will be one of the mainstays of the hospitalization process, we note that the groups have qualitatively the same shape of distribution (Figure 1) with a mode ranging from three days to one week; then we may suppose the same statistical distribution and validate this hypothesis later, with a more accurate inferential procedure. Also, variability of LOS distribution is affected by outliers in every group, most of which stand for patients suffering from severe diseases or spending long time intensive care.

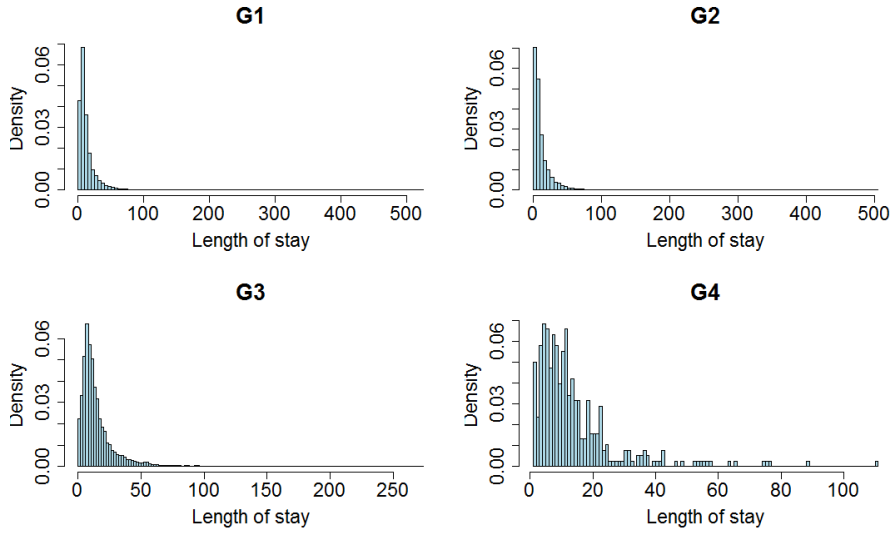


Figure 1: Histograms of length of stay for each HF group.

Finally, a thorny issue is surely the testing of sex influence over the length of stay. First of all, we should point out that the proportion of men and women

is unbalanced and different in percentage in every group (Table 1); it is then natural to check if we should refine our analyses by actually regarding sex of patients as a discriminating factor in every cluster in an ANOVA context. We test for equality between the distribution of length of stay of males and females in every HF group under the null hypothesis that they could not be distinguished. Through a permutational one-way anova (see Pesarin and Salmaso, 2010) and a Kruskall-Wallis test, we find that we do not need to make any distinction based on sex within the first, third and fourth group. As of the second group, since the fitted marked point process is quite similar for men and women, we will not take into account this sex distinction.

# 3 The model

We provide a parametric approach to model the stochastic process of hospitalizations. The conditional intensity function (1) is well defined when assigning specific expressions to $\lambda_g(t|\mathcal{H}_t)$ and $f(\mathbf{y}|t, \mathcal{H}_t)$. The ground intensity function models the temporary trend underneath the marked point process, here governed by Time covariate while the mark distribution describes the length of stay only.
Assuming independence of the mark distribution given the ground process, we can deal with the ground process firstly and with the mark distribution then in a separate way.
In our topic, the ground intensity function is a stochastic process itself and is even regarded as a Hawkes process, having the following functional form:

$$\lambda_g(t|\mathcal{H}_t) = \mu(t) + \eta \sum_{t_i<t} \nu_\theta(t - t_i). \tag{2}$$

Given the assumed left-continuous filtration $\mathcal{H}_t$, the ground intensity function is the sum of a deterministic base intensity $\mu(t)$ called *immigration intensity* which represents the background rate of the process and of a 'self-exciting' term $\eta \sum_{t_i<t} \nu_\theta(t-t_i)$, the so called *memory kernel* that is the convolution of the path of the process with an *interaction kernel* $\nu_\theta$ and gives rise to event clustering through an endogenous feedback (past events contribute to the rate of future events).
In particular, $\nu_\theta : \mathbb{R}^+ \to \mathbb{R}^+$ is called *offspring density*, being taken as a probability density function with a positive support absolutely continuous with respect to the Lebesgue measure, and expresses the positive influence of past events on the current value of the positive intensity process. On the other hand, $\eta$, named *branching ratio*, is a non-negative costant determining the strength of self-excitation and making $\nu_\theta$ a probability density function (see Hardiman et al., 204 and Wheatley at al., 2014).
The branching ratio plays a crucial role in the dynamics of the model. It stands for the fraction of endogenously generated events among the whole population

and it must be less than one for the process to be stationary, due to the autoregressive nature of this last one.
The Hawkes process is then a powerful framework for simulating and modeling the occurrence or arrivals of events which cluster in time, i.e first hospitalizations and consecutive ones.
It also heuristically represents the expected number of events per unit of time where each occurrence increases the probability of other events in the near future or increases the rate of new occurrences momentarily.
While in real applications it is usual to regard the background rate $\mu(t)$ as a constant, there are several choices for the memory kernel.
The choice of the most appropriate memory kernel for the dataset is one of the greatest and appealing issues. In Hawkes process literature, an exponential kernel is usually recommended due to its simple expression and 'numerical' advantages even if it may be not really efficient. Then, we suggest other ways to model the memory kernel (all listed in Table 2), whose goodness depends on the real topic we deal with.

| | Memory Kernel | Branching Ratio $\eta$ |
|---|---|---|
| Exponential | $\alpha e^{-\beta t}$ | $\frac{\alpha}{\beta}$ |
| Gamma | $\frac{\alpha c^{\beta}}{\Gamma(\beta)} e^{-ct} t^{\beta-1}$ | $\alpha$ |
| Weibull | $\alpha \left(\frac{\beta}{\gamma}\right) \left(\frac{t}{\gamma}\right)^{\beta-1} e^{-\left(\frac{t}{\gamma}\right)^{\beta}}$ | $\alpha$ |
| Hyperbolic | $\frac{\alpha}{(t+\beta)^p}$ | $\begin{cases} \frac{\alpha \beta^{1-p}}{p-1} & \text{if } p > 1 \\ \infty & \text{if } p \leq 1 \end{cases}$ |

Table 2: Common analytic expressions for the Hawkes process kernel.

However, we firstly have to estimate the model's parameters by maximizing the loglikelihood of the marked point process (see Daley and Vere-Jones, 2008)

$$\log L = \sum_{i:T_1 \leq t_i \leq T_2} \log \lambda_g(t|\mathcal{H}_t) - \int_{T_1}^{T_2} \lambda_g(t|\mathcal{H}_t) dt + \sum_{i:T_1 \leq t_i \leq T_2} \log f(y_i|\mathcal{H}_t), \quad (3)$$

where $\{(t_1, y_n), \ldots, (t_n, y_n)\}$ is a set of marked point patterns on an observation interval $[T_1, T_2] \times \mathbb{Y}$ with $\mathbb{Y}$ the mark space.
The most difficult term to maximize involves the ground process. Plenty of problems may arise: some optimization routines are very sensitive to poor initial starting values of the parameters while different parameters may take only specific range (Peng, 2003). Then, we use the estimated parameters through the `optim` function with an optimization procedure based here on Nelder-Mead method, which is more robust to poor starting values, as starting values for `nlm` function that is conversely more sensitive to poor initial values but guarantees a faster convergence (Harte, 2010).

We underline that Nelder-Mead algorithm (see Lagarias et al., 1998 for details) turns out to be more efficient than a quasi-Newton method in our application, producing reasonable results in a relatively short time.
After getting the estimates, we test the absolute goodness of fit of the model. Here, we rely on some qualitative methods for the ground process, all being based on the Random Rescaling theorem (Daley and Vere-Jones, 2008) and on the residual process which is a new point process defined as

$$\tau_i = \int_0^{t_i} \hat{\lambda}_g(t|\mathcal{H}_t), \tag{4}$$

where $\hat{\lambda_g}(t|\mathcal{H}_t)$ the fitted ground intensity function.
If the fitted ground intensity function is the true ground intensity function, $\tau_i$, also called transformed times, will form a homogeneous Poisson process of rate on some interval $[0, T]$.
Then, if we plot the event number $i$ versus the transformed time $\tau_i$ in a quarter, we would like to expect the points $(i, \tau_i)$ to follow the diagonal without relevant departures. However, as the dataset's size increases, deviations from the diagonal get no longer sharp; thus, as Page (1954) suggested, we should replace $\tau_i$ with $\tau_i - i$ so as to have a cumsum plot, which is nothing but a zoom of the residual process near the diagonal.
As far as the mark distribution is concerned, it is easier to get its parameters estimates under our initial independence assumption; actually, if the two terms of (1) share no parameters, maximization of (3) can be done separately and we can assess the goodness of fit of the mark distribution through a cumsum plot, in a similar way we do for the ground process.
Finally, when dealing with heart failured patients, it may be useful to monitor the admissions' trends and predict future ones in order to improve the efficiency of clinical facilities and collective welfare. For instance, it may be of a great interest to find the empirical probability distribution of the time to the next event with a defined length of stay. It is indeed a simulation and predictive issue. When a conditional intensity function is specified, it is quite affordable to do simulation; we take Ogata's modified thinning algorithm (Daley and Vere-Jones, 2008) as a starting point and extend it to our application thanks to Harte (2010).

# 4 Results

As we have already underlined, the hardest issue when dealing with Hawkes processes is modeling the ground intensity function (2).
Before giving details, it is useful to remind one of its possible interpretations: the total number of events occurring in the unit of time is given by the background rate $\mu$ and the number of secondary events, that is the number of events triggered by previous events. Each event has a positive probability of generating

an offspring sequence, whose number of events is connected to the time distance between triggering and triggered ones.
In our topic, the first admissions are the triggering events while the consecutive ones are the triggered. Since events of hospitalizations seem to hyperbolically

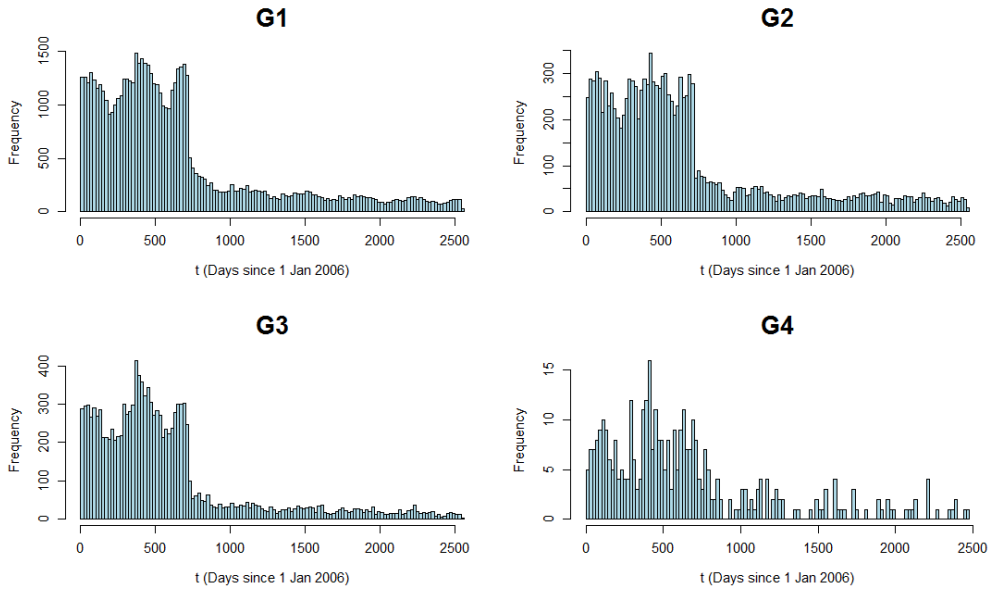


Figure 2: Empirical distributions of Time covariate.

decay in time (Figure 2), we adopt a Hawkes model with hyperbolic interaction kernel in each HF group, expecting long-memory features and long-range interactions to be comparatively more important than for exponential kernels with the same branching ratio. Thus, the ground intensity function we use is

$$\lambda_g(t|\mathcal{H}_t) = \mu + A \sum_{i:t_i<t} \left( \frac{c}{c + (t - t_i)} \right)^p \tag{5}$$

where the parameters $(\mu, A, c, p)$ must be all positive and are estimated through `optim` and `nlm` R functions (note that (5) is a re-parametrization of hyperbolic kernel listed in Table 2). This expression is very similar to Omori's law kernel (Ogata, 1988), except for a function in the sum taking into account their magnitudes.
As far as the mark is concerned, we should note that, primarly, any discrete or continuous covariate may be taken into account; of course, not all make sense since they explore several aspects which look somehow marginal to the hospitalization process. What it may be of a great interest so as to inspect a possible influence over the above temporary process is, as we have already noticed, the length of stay.

Then, we inspect its empirical distribution in every given group (Figure 1) and model the marks $Y_i$ through a Gamma($a$,$s$) distribution with parametrization is

$$f(y|a,s) = \frac{1}{s^a \Gamma(a)} y^{a-1} e^{-\frac{y}{s}} \mathbb{I}_{(0,+\infty)}(y) \tag{6}$$

where $a$ stands for the shape and $s$ for the scale.
Since $\lambda_g(t|\mathcal{H}_t)$ and $f(y|a,s)$ have not any parameter in common, the maximization of $\lambda(t,y|\mathcal{H}_t)$ is easier; so we firstly present the results about the ground process and then we discuss the ones concerning the mark distribution, in a coherent way with Harte's analyses (2010).

### 4.1 The ground process

At the beginning of this section, we have supposed a hyperbolic memory kernel is more appropriate for modeling hospitalizations in time by just inspecting their empirical distributions. In order to validate this hypothesis, we display the plots of the fitted ground intensity function $\lambda_g(t|\mathcal{H}_t)$ which represent the expected number of event per unit of time and make us suppose they underestimate the underneath temporary process (Figure 3).

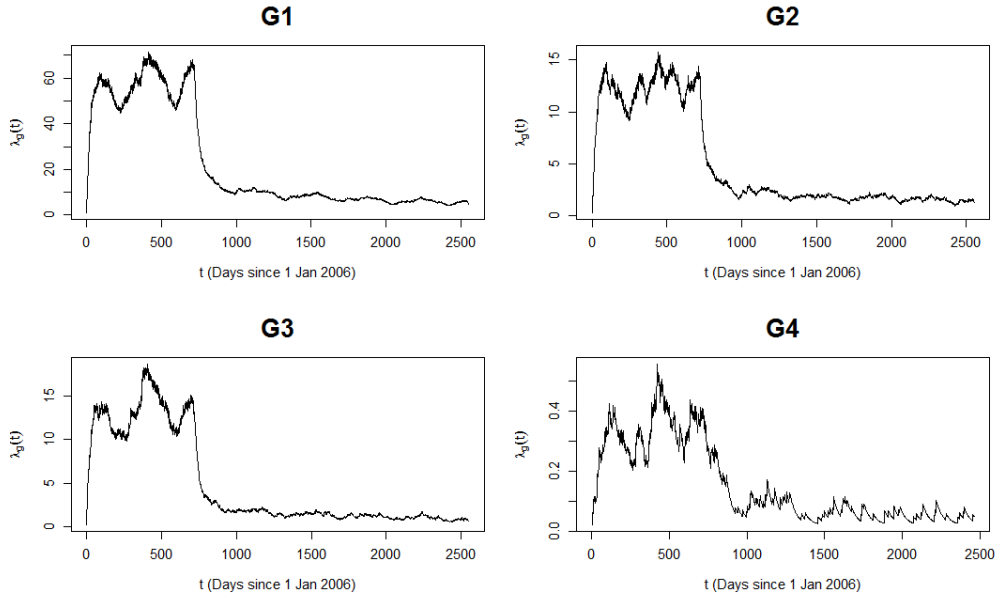


Figure 3: Fitted ground intensity function plots

Moreover, the ground intensity plots highlight some curious features. While we find that the days with maximum number of events are concentrated in the first month of 2007 for every group, as we could expect from medical literature, we also note that each plot points out two different trends in term of the number

of hospitalizations.
Precisely, the first three groups show a 'down step' at December 24th and 25th, 2007, while for the fourth group there is not a sharp distinction concerning the same topic, maybe due to the small number of hospitalizations (and patients). The estimated parameters, which are listed in Table 3, allows us to draw some considerations.

| **Parameters** | G1 | G2 | G3 | G4 |
|---|---|---|---|---|
| $\mu$ | $8.11 \times 10^{-1}$ | $2.34 \times 10^{-1}$ | $1.69 \times 10^{-1}$ | $2.01 \times 10^{-2}$ |
| $A$ | $4.31 \times 10^{-2}$ | $4.07 \times 10^{-2}$ | $4.52 \times 10^{-2}$ | $2.23 \times 10^{-2}$ |
| $c$ | $9.20 \times 10^{11}$ | $1.76 \times 10^{6}$ | $1.92 \times 10^{6}$ | $2.02 \times 10^{6}$ |
| $p$ | $4.11 \times 10^{10}$ | $7.50 \times 10^{4}$ | $8.96 \times 10^{4}$ | $5.14 \times 10^{4}$ |
| $\eta$ | 0.965 | 0.955 | 0.967 | 0.875 |
| $\log L$ | $1.48 \times 10^{5}$ | $1.31 \times 10^{4}$ | $1.41 \times 10^{4}$ | $-9.52 \times 10^{2}$ |

Table 3: Parameter estimates of (5)

Firstly, we note that the parameter $\mu$ determines the intensity of exogenous events, roughly speaking, how many events occur per unit of time and does not affect the stability in the event rate of the process which is entirely governed by the branching ratio. Furthermore, $(\mu, A, c, p)$ determine the clustering of the process and the intra-event dynamics; they substantially give information about the stationarity of the process as well as the proportion of events that are generated inside the model to all events.
The branching ratios of each group are very high, meaning that their dynamics are almost entirely driven by endogenous events and only a small percentage by exogenous ones. At the same time, we may observe that there is some clustering in the ground intensity plot as displayed by the occurrences of spikes in the plots. We can conjecture some main point patterns (primary events and secondary ones) by inspecting the plot of the stochastic process and support these empirical considerations through an inferential procedure. As we have previously underlined, two temporal point patterns are evident standing for first and consecutive hospitalizations; the same two can be indeed found by relying on a cluster analysis based on CLARA (CLustering LARge Applications) and on the *Silhouette Coefficient*, a quality index which allows us to select an optimal number of clusters and whose values are displayed in Figure 4 in an increasing number of clusters (see Strufy et al., 1997).

Finally, while presenting the four common analytical expressions for the ground process (Table 2), we affirm that its choice depends on the specific data to be modeled and a measure of absolute goodness of fit is needed. Here, the discriminating factors which lead us to the most suitable model are the analysis of the residual process and the cumsum plot. We display these kind of plots only

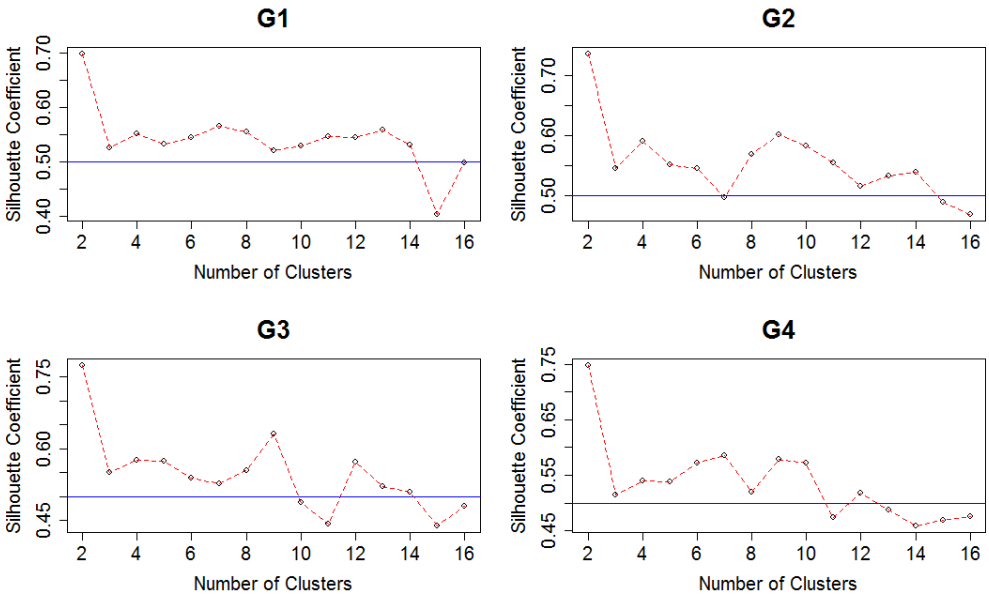


Figure 4: Silhouette Coefficients values in correspondance with an increasing number of clusters. The blue line stands for a value equal to 0.50.

for the first and second group (Figure 4), obtaining similar results for the other groups.
As we can see from these plots, the larger a group, the more the residual process gets close to the diagonal line and deviation from the straight line is negligible. Also, the cumsum plots show that the fitted ground processes underestimate the underneath temporal processes, as we have already expected in the beginning.

## 4.2 The mark distribution

We recall that we assume $\{Y_i\}$ as mutually independent random variables given the ground process. This hypothesis leads to an independently marked point process and make the computations easier. The parameters of the mark distribution can be estimated separately and set as default fixed values within the intensity function expression. This partially justify our previous computational procedure and why we have presented the results on the ground process firstly (see Harte, 2010).
The parameters of a Gamma distribution cannot be found by analitically maximizing its loglikelihood since they do not have a closed form. Simple numerical algorithms are suggested such as the fast conditional likelihood already implemented in `rGammaGamma` R package (Triche, 2013), which is the one we used in our analyses.
However, we check the adequacy of this assumed distribution by plotting the

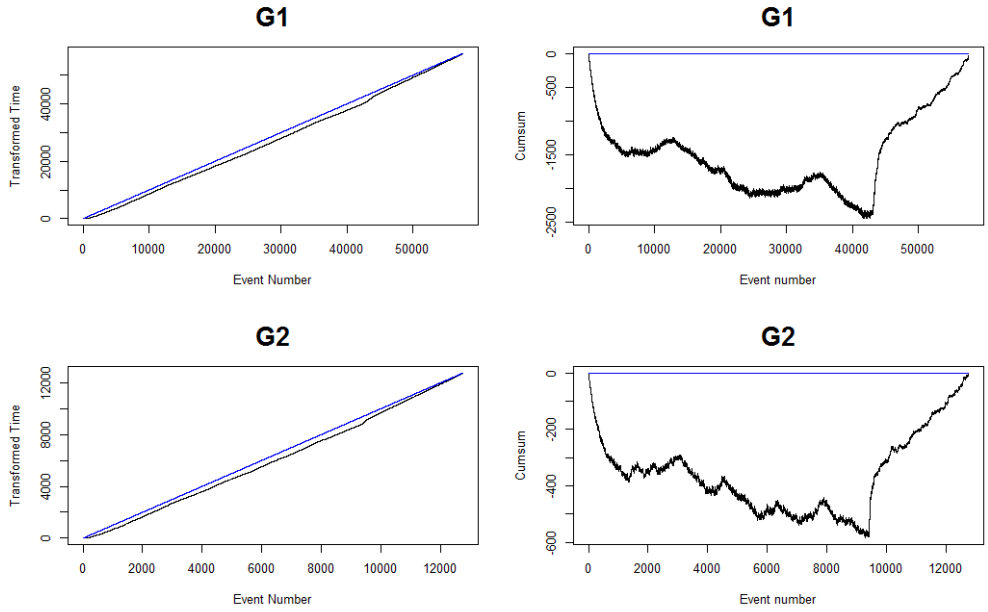


Figure 5: Residual process times on the left and cumsum of residual process times on the right for G1 and G2. The diagonal for the residual process and the x-axis for the cumsum are added in blue colour.

| **Parameters** | G1 | G2 | G3 | G4 |
|---|---|---|---|---|
| $a$ | 1.554 | 1.181 | 1.554 | 1.665 |
| $s$ | 8.920 | 10.706 | 10.308 | 8.752 |
| $\log L$ | $-2.061 \times 10^5$ | $-4.499 \times 10^4$ | $-4.610 \times 10^4$ | $-1.369 \times 10^3$ |

Table 4: MLE parameters for the mark distribution (6).

cumsum of the length of stay over time in the same way we did for the ground process (Figure 6).

Finally, simulation is a useful tool for evaluating some features of our model, being also strictly related to predictive purposes when no explicit numerical algorithms are available (see Daley and Vere-Jones, 2008).
Once the expression of the conditional intensity function is known, simulation of a marked point process is straightforward. We then focus on simulating the time to the next event of a hospitalization with a defined length of stay, determining its empirical distribution and checking some quantitative features through location parameters.
We may suppose that the follow-up period is concluded and a new patient belong-

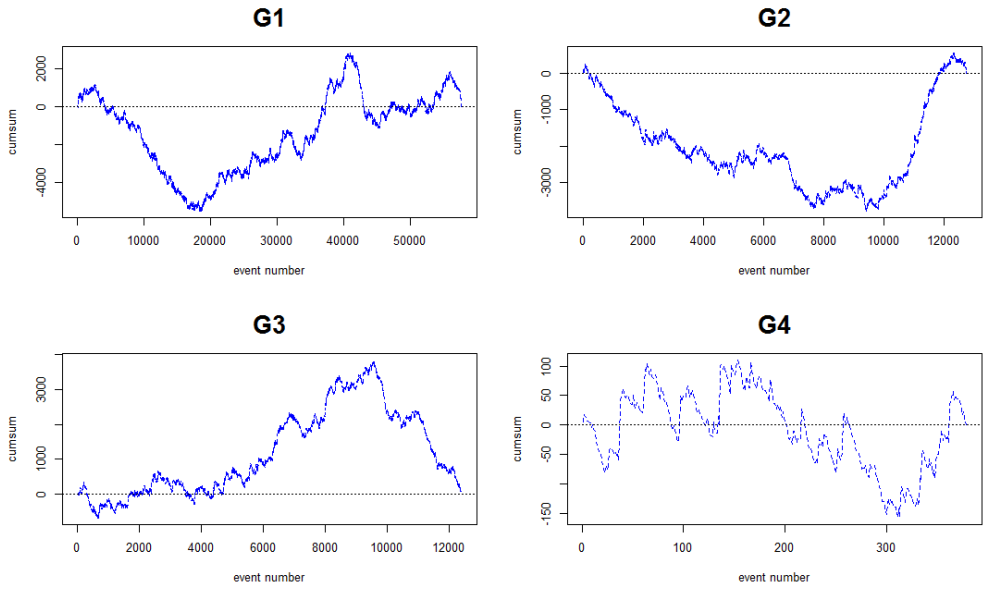


Figure 6: Cumsum of LOS in each HF group.

ing to one of the four given group has to be monitored: what is the probability that he/she will be admitted to hospital for a specific number of days?
This question perfectly translates the need to simulate (and, eventually, predict) times of hospitalization events together with their length of stay under the context of marked point processes. Here, we build a simulation method starting from Harte's algorithm (see Harte, 2010), supposing a Gamma distribution for the length of stay, independent from the ground process.
For a matter of example, we set the above 'defined length of stay' as the 0.90 and 0.95-quantiles of the empirical distribution of LOS, regarding them as extreme values rarely got in each HF group; of course, any reasonable value can be assigned, being fixed according to a particular phenomenon one is interested in.
We start simulating events from the day after the last recorded event in each group and record the time to the first event with length of stay greater than $q_\alpha$. We display the histograms in Figures 7 and 8, noting that these empirical probability distributions show a hyperbolic trend and cover a period (in days) which increases from the first to the fourth group.

| Empirical quantiles $q_\alpha$ | G1 | G2 | G3 | G4 |
|---|---|---|---|---|
| $q_{0.90}$ | 29 | 28 | 33 | 29 |
| $q_{0.95}$ | 40 | 39 | 46 | 41 |

Table 5: Empirical quantiles of LOS distribution for each group (in days).

So far we have dealt with models where the marks are independent of the history of the process. However, it may seem an optimistic and somehow restrictive hypothesis in general. Even if it does not make much sense to suppose length of stay is related with the temporary stochastic hospitalization process, we try to conjecture a particular dependent model with the same probability distribution for the mark as in the independent case. Precisely, we model the mark as

$$f(y|t, \mathcal{H}_t) = Gamma(\alpha, s) \tag{7}$$

where $\alpha$ is the shape, set equal to $1 + ag(t|\mathcal{H}_t)$ (with $g(t|\mathcal{H}_t)=\lambda_g(t|\mathcal{H}_t)^{1/k}$) and $s$ stands for the scale. The 'optimal' $k$ that gives coherent results with the independent case and assures a smaller AIC turns out to be equal to 8.
The plots of the fitted ground and residual process are quite similar to the independent ones while cumsum plots of the mark confirm an independent model is preferable for almost all the groups.

# 5 Conclusions and future developments

In this paper, we framed the admissions of heart failured patients in the context of Marked Point Processes. Patients are divided into four separate groups according to their diagnoses-codes contained in the SDO of their first hospitalizations and the same statistical model was adopted in each group, leading to different parameter estimates. The underneath temporal ground process was hard to model even if making use of a parametric approach such as the Hawkes process while the mark distribution was easy to be dealt with due to its independence hypothesis with the ground process.
We gave details about modeling a longitudinal dataset, chose a particular model and validated it through a specific technique. Besides, we provided a general framework for simulating an independently marked point processes.
All these results are very important and useful for our real application; actually, the fact of monitoring the admissions' trend could allow hospitals to preview the needs of future hospital admissions so as to improve the efficiency of clinical facilities and collective welfare.

Now a greater dataset is available containing more information and events; actually, not only hospitalizations are recorded but also when drugs are prescribed after discharge and when outpatient medical examinations take place. Then, we may extend our analyses and introduce a new mark structure; we may assume a joint probability distribution for it: a Gamma distribution for the length of stay and a discrete one accounting for the number of pharmacological and medical examination's events.
Furthermore, the assumption of independent marking seems strong and should be inspected at the beginning of the analyses; hence, statistical tests for assessing independence based, for example, on likelihood ratio statistics or subsampling

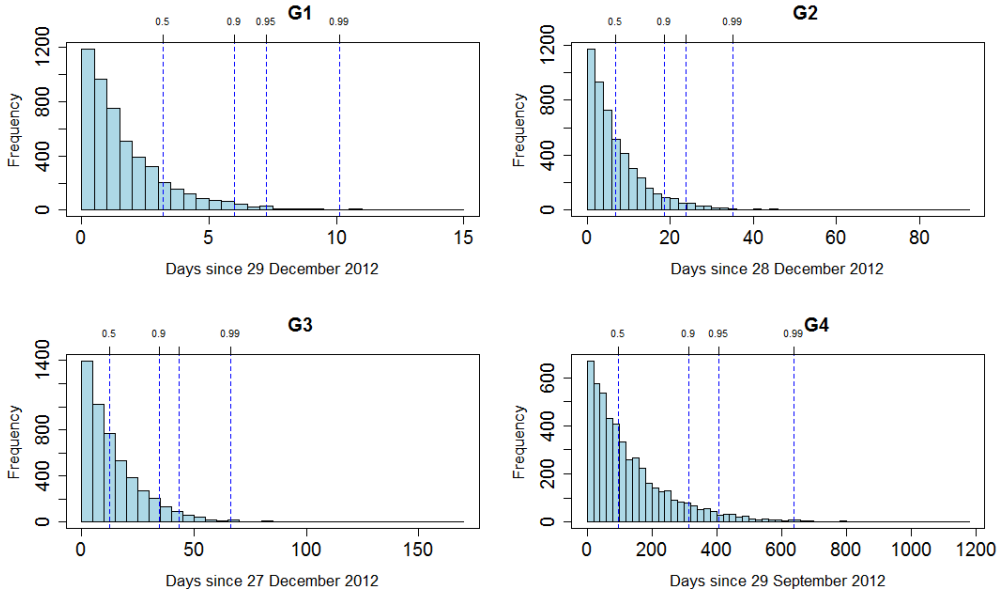


Figure 7: Histograms of the times to the first event with LOS greater than **0.90**-quantile (in days) for each HF group. The blue dash lines stand for the 0.5, 0.9, 0.95 and 0.99 quantiles.

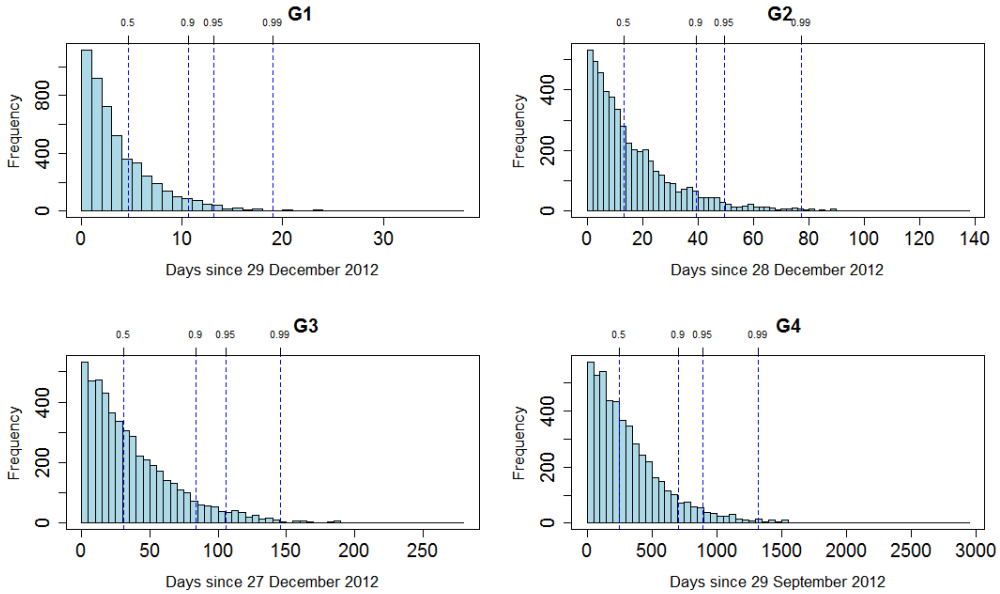


Figure 8: Histograms of the times to the first event with LOS greater than **0.95**-quantile (in days) for each HF group. The blue dash lines stand for the 0.5, 0.9, 0.95 and 0.99 quantiles.

approaches are needed.
Finally, we may make inference on longitudinal data modeled by a marked point processes with non-parametrical and compare the results.

## Acknowledgments

In this work, data are collected from the major project called 'Utilisation of Regional Health Service databases for evaluating epidemiology, short and medium term outcome, and process indexes in patients hospitalized for heart failure' funded by the Italian Ministry of Health and Regione Lombardia - Healthcare division. The authors wish to thank Regione Lombardia - Healthcare division for having funded and supporting the project.

## References


AHRQ: *Agency for Healthcare Research and Quality*, 2015, URL http://www.ahrq.gov/professionals/prevention-chronic-care/decision/mcc.

Crowley, S.: *Point Process Models for Multivariate High-Frequency Irregularly Spaced Data*, 2013.

Daley, D.J., Vere-Jones D.:*An Introduction to the Theory of Point Processes*, Springer, 2008.

Hardiman, S., Bouchaud, J.P:*Branching ratio approximation for the self-exciting Hawkes process*, 2014.

Harris, T.:*The theory of the branching processes*, Rand Corporation, 1964.

Harte, D.:***PtProcess**: An R Package for Modelling Marked Point Processes Indexed by Time*, Journal of Statistical Software, **35**(8), 1-32. URL http://www.jstatsoft.org/v35/i08/, 2010.

Ieva, F., Paganoni, A.M., Pietrabissa, T.: *Dynamic clustering of hazard functions: an application to disease progression in chronic heart failure*, Health Care Management of Science. doi:10.1007/s10729-016-9357-3, 2016.

Lagarias, J., Reeds, J., Wright, M-H., Wright, P-E.: *Convergence properties of the Nelder-Mead simplex method in low dimensions*, SIAM Journal of Optimization,**9**, 112-147, 1998.

Mazzali, C., Maistrello, M., Ieva, F., Barbieri, P.: *Methodological issues in the use of administrative databases to study heart failure.*, Advances in Complex Data Modeling and Computational Methods in Statistics (eds: A.M. Paganoni, P. Secchi), Springer, 2015.

Page, E.S.: *Continuous Inspection Schemes*, Biometrica, **41**(1-2), 100-115, doi:10.1093/biomet/41.1-2.100, 1954.

Peng, R.D.: *Multi-dimensional Point Process Models in R*, Journal of Statistical Software,**8**(16),1-27, 2003.

Pesarin, F., Salmaso, L.: *Permutation Tests for Complex Data: Theory, Application and Software*, Wiley, 2010.

Pope, G.C., Kautter, J., Ellis, R.P., Ash, A.S., Ayanian, J.Z., Lezzoni, L.I.: *Risk adjustment of Medicare capitation payments using the CMS-HCC model*, 2004; 25(4):11941.

R Core Team: ***R**: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna, Austria, 2014, URL http://www.R-project.org/.

Struyf, A., Hubert, M., Rousseeuw, P.J.: *Clustering in an Object-Oriented Environment*, doi:10.18637/jss.v001.i04, 1997.

Triche, T., Jr.: *rGammaGamma: Gamma convolutions for methylation array background correction*, R package version 1.0.12, 2013, URL http://CRAN.R-project.org/package=rGammaGamma.

Wheatley, S., Filimonov, V., Sornette, D. : *Estimation of the Hawkes Process With Renewal Immigration Using the EM Algorithm* , Swiss Finance Institute, Research Paper Series, 14-53, 2014.

# MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

**35/2016** Zonca, S.; Formaggia, L.; Vergara, C.
*An unfitted formulation for the interaction of an incompressible fluid with a thick structure via an XFEM/DG approach*

**33/2016** Antonietti, P. F.; Ferroni, A.; Mazzieri, I.; Quarteroni, A.
*hp-version discontinuous Galerkin approximations of the elastodynamics equation*

**34/2016** Menafoglio, A.; Secchi, P.
*Statistical analysis of complex and spatially dependent data: a review of Object Oriented Spatial Statistics*

**32/2016** Tarabelloni, N.; Schenone, E.; Collin, A.; Ieva, F.; Paganoni, A.M.; Gerbeau, J.-F.
*Statistical Assessment and Calibration of Numerical ECG Models*

**30/2016** Abramowicz, K.; Häger, C.; Pini, A.; Schelin, L.; Sjöstedt de Luna, S.; Vantini, S.
*Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament*

**31/2016** Antonietti, P.F.; Merlet, B.; Morgan, P.; Verani, M.
*Convergence to equilibrium for a second-order time semi-discretization of the Cahn-Hilliard equation*

**28/2016** Antonietti, P.F.; Dal Santo, N.; Mazzieri, I.; Quarteroni, A.
*A high-order discontinuous Galerkin approximation to ordinary differential equations with applications to elastodynamics*

**29/2016** Miglio, E.; Parolini, N.; Penati, M.; Porcù, R.
*GPU parallelization of brownout simulations with a non-interacting particles dynamic model*

**27/2016** Repossi, E.; Rosso, R.; Verani, M.
*A phase-field model for liquid-gas mixtures: mathematical modelling and Discontinuous Galerkin discretization*

**26/2016** Brunetto, D.; Calderoni, F.; Piccardi, C.
*Communities in criminal networks: A case study*