# Modeling Recurrent Events in Panel Data Using Mixed Poisson Models

V. Savani and A. Zhigljavsky *

*Abstract*— **This paper reviews the applicability of the mixed Poisson process as a model for recurrent events in panel data. Methods for testing goodness of fit of the mixed Poisson process are compared and the model is applied to consumer purchases of a product.**

***Keywords: mixed Poisson models; panel data analysis; recurrent events; renewal processes; market research***

## 1 Introduction

This paper considers mixed Poisson processes for the analysis of panel data. We define panel data as a collection of sampled individuals from a population observed over a period of time called the analysis period. During the analysis period, each unit has events occurring at random (exponentially distributed) times. We aim to statistically model the occurrence of events for the population as a whole using mixed Poisson processes.

Section 2 reviews the theory of mixed Poisson models. Section 3 considers methods for assessing goodness of fit of the mixed Poisson model.

## 2 Background

This section presents the fundamentals of mixed Poisson process theory (see [4] for a detailed description of mixed Poisson processes).

### 2.1 Mixed Poisson distributions (MPDs)

A random variable $X$ has a mixed Poisson distribution if it has probability mass function (p.m.f.)

$$p_x = \mathbb{P}(X = x) = \int_{0-}^{\infty} \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!} dF(\lambda), \quad x = 0, 1, 2 \ldots . \tag{1}$$

where $F(\lambda)$ is a cumulative distribution function (c.d.f.) of a random variable which takes values in the interval $(0, \infty)$. The distribution with c.d.f. $F(\lambda)$ is often termed the 'structure' distribution. The structure distribution often depends on parameters $\boldsymbol{\theta}$ (unknown in practice) so that $F_\Lambda(\lambda) = F_\Lambda(\lambda; \boldsymbol{\theta})$.

A common structure distribution is the gamma distribution with density function

$$f(\lambda) = \frac{1}{a^k \Gamma(k)} \lambda^{k-1} \mathrm{e}^{-\lambda/a}, \quad a > 0,\, k > 0, \quad \lambda > 0\,;$$

the resulting MPD is the negative binomial distribution (NBD) with probabilities

$$p_x = \frac{\Gamma(k+x)}{x!\Gamma(k)} \left(\frac{1}{1+a}\right)^k \left(\frac{a}{1+a}\right)^x, \qquad \begin{matrix} x = 0, 1, 2, \ldots \\ k > 0,\ a > 0. \end{matrix} \tag{2}$$

In the case of panel data analysis, equation (1) has the following natural interpretation: if the number of events for each individual in a fixed time period follows the Poisson distribution and the mean of this Poisson distribution has c.d.f. $F(\lambda)$ then the number of events for a random individual has the mixed Poisson distribution (note that this holds for any fixed time interval).

### 2.2 Mixed Poisson processes (MPPs)

Let $\boldsymbol{X} = (X(t_1), X(t_2), \ldots, X(t_n))$ be a random vector with $0 = t_0 < t_1 < \ldots < t_n$ representing an increasing sequence of time points, let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ be a vector of non-negative integers with $0 = x_0 \leqslant x_1 \leqslant \ldots \leqslant x_n$ and let $\lambda > 0$ be the intensity of a process, then given the multivariate Poisson distribution

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} | \Lambda = \lambda) = \prod_{i=0}^{n-1} \frac{[\lambda(t_{i+1} - t_i)]^{x_{i+1} - x_i}}{(x_{i+1} - x_i)!} e^{(-\lambda(t_{i+1} - t_i))}, \tag{3}$$

the mixed Poisson process is consequently defined as a process $\{X(t) : t \in \{t_1, t_2, \ldots, t_n\}\}$ whose finite-dimensional distributions are

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) = \int_{0-}^{\infty} \mathbb{P}(\boldsymbol{X} = \boldsymbol{x} | \Lambda = \lambda)\, dF_\Lambda(\lambda). \tag{4}$$

Here $F_\Lambda(\lambda)$ is the c.d.f. of a random variable $\Lambda$ with support $(0, \infty)$.

Note that the mixed Poisson process conditioned upon $\Lambda = \lambda$, so that the value of $\lambda$ is fixed, is simply a pure Poisson process with stationary and independent increments whose finite dimensional distributions are given by equation (3).

*School of Mathematics, Cardiff University, Senghennydd Road, Cardiff, UK, CF24 4AG. E-mail: SavaniV@Cardiff.ac.uk; ZhigljavskyAA@Cardiff.ac.uk. We would like to thank Phil Parker of ACNielsen / BASES for providing household transaction data.

### 2.3 Measures of recurrence

Measures of recurrence are functionals of the MPD that summarise the repeat behavior of the occurrence of an event. Statistical estimators of these measures are commonly used for the analysis of data in fields such as consumer research (see e.g. [2]), insurance (see e.g. [4]) and health care (see e.g. [1]). The measures can be used in any field where the theory of MPPs is applicable.

In consumer research, where an event represents the purchase of a product, the measures of recurrence are more commonly known as repeat buying measures. The repeat buying measures are functionals of the one-dimensional mixed Poisson distribution and represent measures of recurrence within a fixed time period. In consumer research, more accurate forecasts of sales can be achieved by considering aggregate sales broken down into the repeat buying measures (see e.g. [3]).

Assume that the analysis period $(0, t]$ is fixed, the measures defined below are applicable to any time interval of length $t$. Let $X(t)$ be a random variable from the one-dimensional MPD with probabilities $p_x(t)$ ($x \in \{0, 1, 2, \ldots\}$) and let $\mu(t) = \mathbb{E}X(t)$ denote the mean of the MPD. We now define various characteristics of recurrence obtained from synonymous measures used in consumer research.

*Penetration.* The penetration is the probability that at least one event occurs for a random individual.

$$b(t) = 1 - p_0(t) = 1 - \int_{0-}^{\infty} e^{-\lambda t} dF(\lambda), \qquad 0 \leqslant b(t) \leqslant 1.$$

The penetration is a non-linear non-decreasing function of $t$ as $t$ increases. Note that $b(0) = 0$ so that no events may occur at time intervals of length $t = 0$. Additionally $b(\infty) = 1$; thus, given an infinite amount of time, a random individual will have at least one event with probability one.

*Occurrence frequency.* The occurrence frequency is the mean number of events for a random individual who has non-zero number of occurrences.

$$w(t) = \mathbb{E}(X(t)|X(t) \geqslant 1) = \mu(t)/b(t), \qquad w(t) \geqslant 1.$$

*Measured repeat.* The $r$-th ($r = 1, 2, \ldots$) measured repeat is the probability that a random individual is likely to have at least one more event given that the individual has already had $r$ events. The $r$-th measured repeat is

$$\beta_r(t) = \frac{1 - \sum_{j=0}^{r} \mathbb{P}(X(t) = j)}{1 - \sum_{j=0}^{r-1} \mathbb{P}(X(t) = j)}, \qquad 0 \leqslant \beta_r(t) \leqslant 1.$$

*Repeats per repeater.* The $r$-th ($r = 1, 2, \ldots$) repeats per repeater is the mean number of events for a random individual who has at least $r + 1$ occurrences. The mean number of events is usually shifted by the value $r$ so that the minimum possible value is always one. The $r$-th repeats per repeater is

$$\omega_r(t) = \frac{\mu(t) - \sum_{j=0}^{r} j\, \mathbb{P}(X(t) = j)}{1 - \sum_{j=0}^{r} \mathbb{P}(X(t) = j)} - r, \qquad \omega_r(t) \geqslant 1.$$

### 2.4 Panel data

We define panel data as a collection of sampled individuals from a population observed over a period of time called the analysis period. During the analysis period, each unit has events occurring at random times.

When applying the mixed Poisson process to data we use the following model. We assume that each individual $i$ has events that occur according to a pure Poisson process with random intensity $\lambda_i$ that is fixed over time; thus, for a fixed individual, inter-event times are independent and identically exponentially distributed with mean $1/\lambda_i$. The $\lambda_i$ are random variables from a structure distribution with c.d.f. $F(\lambda)$.

Note that an alternative interpretation may be applied (see e.g. [4]) whereby each individual starts with a common fixed $\lambda$. For a fixed individual, the $\lambda$ changes over time according to the number of events observed for that individual. In this paper, we only consider the first interpretation mentioned in the previous paragraph.

## 3 Analysing goodness of fit

The suitability of the mixed Poisson process as a model for panel data is considered by verifying adequacy of i) the mixed Poisson distribution and ii) the mixed Poisson process. In assessing adequacy, we compare model based estimators for the measures of recurrence to the empirical estimators.

The measures of recurrence provide a standard set of measures upon which to compare goodness of fit of different models. When applying the mixed Poisson model to data, estimates for the measures of recurrence are obtained by estimating the vector of parameters $\boldsymbol{\theta}$ (e.g. by using maximum likelihood) and computing the measures as described in Section 2.3. Let $n_j$, ($j = 0, 1, 2, \ldots$) denote the number of individuals with $j$ occurrences during the analysis period and let $N$ be the total number of individuals in the panel; empirical estimators for the measures of recurrence are computed by replacing the probability $\mathbb{P}(X(t) = j)$ with its sample equivalent $n_j/N$.

As a practical application, we apply the mixed Poisson process with a gamma structure distribution to household panel data where individuals are households and events are the time points at which a product is purchased. We are grateful to Phil Parker of ACNielsen BASES for providing the data.

### 3.1 Adequacy of the mixed Poisson distribution

The simplest method of assessing adequacy of the mixed Poisson process is to verify the adequacy of the one dimensional mixed Poisson distribution by comparing observed and expected frequencies of households with a given number of purchase occasions observed during a fixed time interval.

Figure 1 shows observed and expected frequencies when fitting the NBD to panel data using a time interval of length 13 and 52 weeks. It is clear, by observation, that the NBD provides a good fit for the data. The graphs in Figure 1, however, only consider data during the time interval from the start of the analysis to the first 13 and 52 weeks respectively. The mixed Poisson process has the property that the number of events in a fixed time period follows the mixed Poisson distribution irrespective of the start point and the chosen length of the analysis period.

To overcome this problem, we consider plotting ratios of model based estimates for the measures of recurrence to corresponding empirical estimators where the estimation is made using data from multiple time intervals of varying lengths. Figure 2 shows ratios of model based measures of recurrence to empirical measures of recurrence plotted against different lengths of analysis period. For a fixed length of analysis period, each point represents a ratio computed when fitting the model to data in sequential non-overlapping intervals of equal length. Additionally shown are lines for the mean of the estimates together with corresponding 95% confidence intervals. The fact that the ratios are very close to 1 indicate that the NBD provides a good fit for the data over different time intervals.

### 3.2 Adequacy of the mixed Poisson process

When assessing adequacy of the mixed Poisson process, analysing goodness of fit of the mixed Poisson model by considering ratios of model based estimators to empirical estimators does not take into consideration the fact that parameters must remain stationary over time and hence parameter estimates must not differ significantly in different time periods.

Figure 3 shows time series plots of parameter estimates of $b$ and $w$ computed in consecutive non-overlapping intervals of length 13 weeks. The plots, in addition, show dotted lines representing the overall mean of the estimators and a confidence interval for the mean. The confidence intervals for the mean was computed on the assumption of asymptotic normality of the estimators (for more details see [5]). Note that, for the NBD, the joint parameters $b$ and $w$ are unique to the joint parameters $a$ and $k$ and thus it is sufficient to consider stationarity of $(b, w)$ when investigating stationarity of $(a, k)$.

The figures indicate that, although the parameter estimates for $b$ and $w$ are fairly steady, there are two periods for which the estimators significantly differ from one another. Note also that, even though the majority of the estimators do not differ significantly, high estimators of $b$ tend to have high estimators of $w$ (i.e. the two estimators seem to be correlated and indeed are correlated).

The methods considered so far only assess adequacy of the one-dimensional marginal distribution of the mixed Poisson process over a fixed analysis period. These methods do not consider the growth of the parameters as the length of the analysis period varies. Figure 4 shows box plots of empirical estimators of penetration and purchase frequency for varying lengths of time window. (For each length of time period, multiple estimators are obtained by taking consecutive non-overlapping analysis periods.) In addition, a solid line is plotted indicating the expected growth of the measure of recurrence based on parameter estimates of the mixed Poisson process from a single period of length 52 weeks. It is clear that the observed growth and the model based growth of the measures closely match.

The final method of assessing adequacy of the mixed Poisson process uses the results of [5] in which we derive the multivariate asymptotic distributions of statistics and estimators computed over different time intervals using samples observed from mixed Poisson processes. The method of assessing goodness of fit compares the asymptotic distribution of statistics and estimators (under the assumption that the data are generated from a mixed Poisson process) to the observed distribution of statistics and estimators computed in two different time intervals.

Figure 5 shows, in separate plots, estimators of penetration and purchase frequency computed in two consecutive non-overlapping intervals. The points represent estimators obtained from randomly selected sub-groups of the population. In addition, a 95% confidence ellipse is shown which is constructed under the assumption that the data follows a mixed Poisson process and that the estimators are asymptotically normal. The plots again indicate a good fit for the mixed Poisson model.

Assessing adequacy of the mixed Poisson processes by analysing the bivariate distributions of the estimators has the benefits of: testing goodness of fit of the mixed Poisson distribution (since the bivariate asymptotic distributions of the estimators are based on the assumption of the underlying mixed Poisson distribution); testing stationarity of parameter estimates over different time intervals; testing the Poisson process assumption for individuals; detecting changes in intensity for individuals over time periods; and finally assessing adequacy of the mixed Poisson model to varying length analysis periods (achieved by comparing the asymptotic distributions of statistics in two periods of different lengths).

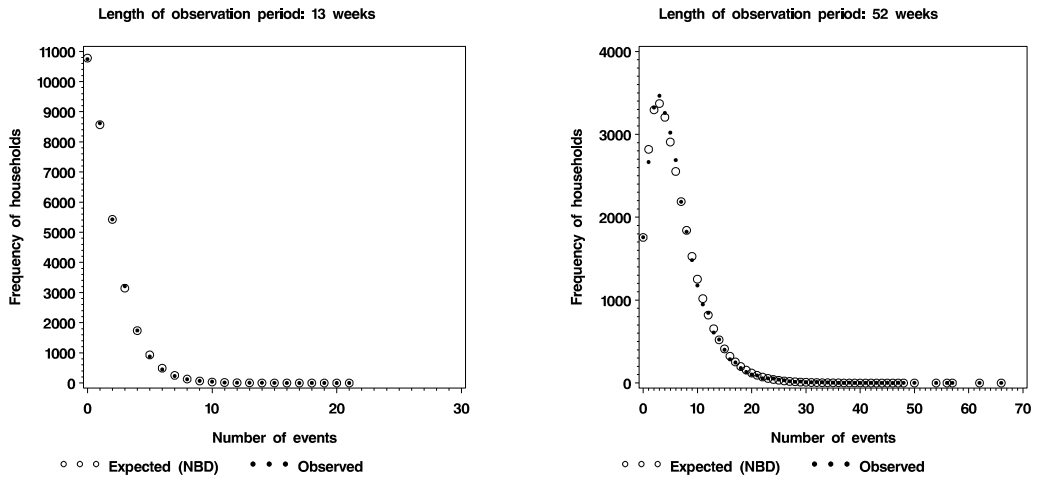


Figure 1: Assessing adequacy of the MPD (single time interval)

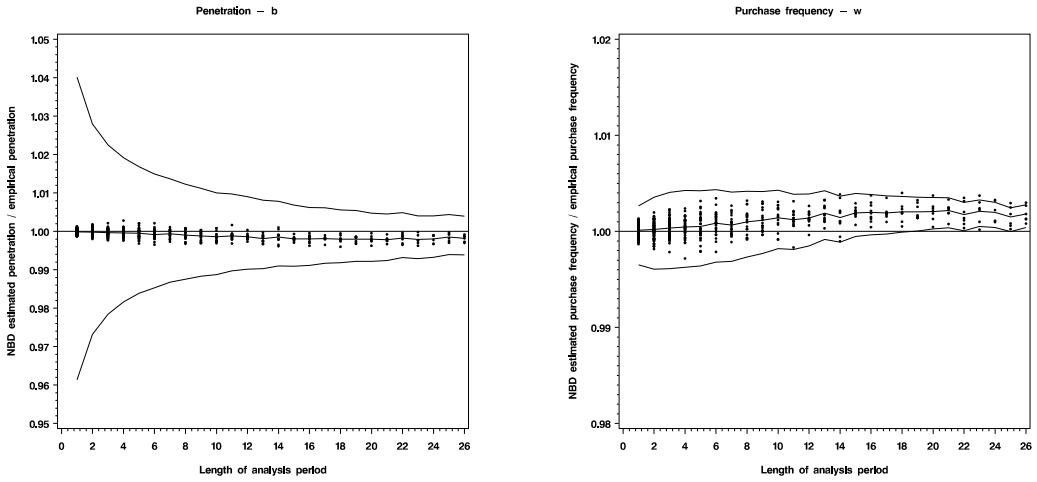


Figure 2: Assessing adequacy of the MPD (multiple time interval)

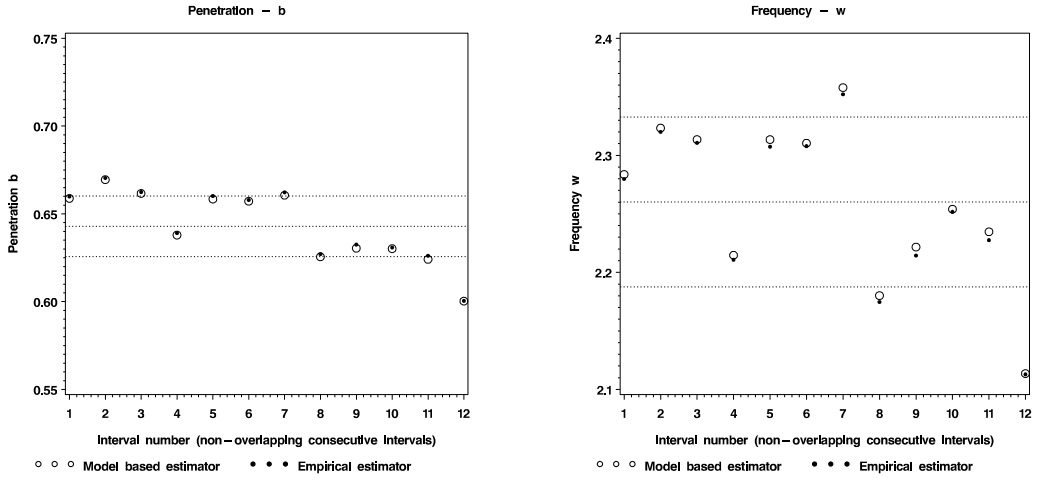


Figure 3: Assessing stationarity of parameters over time

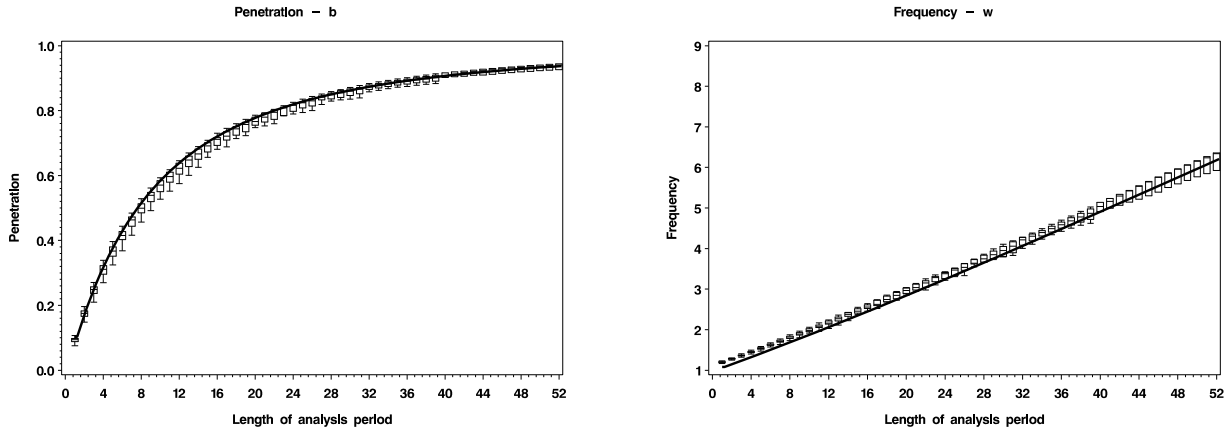


Figure 4: Assessing adequacy of the MPP using growth curves

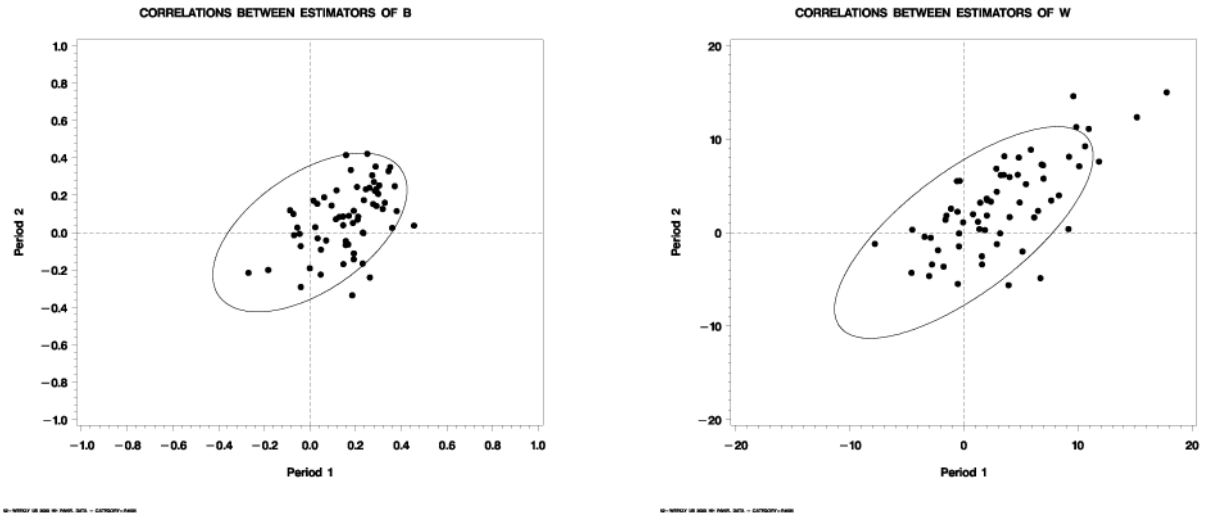


Figure 5: Assessing adequacy of the MPP using confidence ellipses

## Conclusion

The mixed Poisson process has been shown to be an adequate model for the modeling of recurrent events in market research. Numerous methods of assessing adequacy of the mixed Poisson process have been considered. These range from the simplest methods (assessing adequacy of the one-dimensional mixed Poisson distribution) to more complicated methods which take into consideration comparison of the dynamical behavior the data to the model.

Assessing adequacy of the mixed Poisson processes using methods that take the dynamical behavior of the data into account help examine possible causes of deviation from the model (see e.g. [2]). As a result, it is possible to develop more accurate processes for the modeling of data. For example, it is straightforward to de-seasonalize mixed Poisson processes to homogenous Poisson processes when there are trends in the mean. In the case of panel flow, where individuals have recurrent events that occur according to a mixed Poisson process for a random period of time, it is possible to extend the standard mixed Poisson model (see e.g. [6]) to accommodate for panel flow.

## References

[1] R. J. Cook and W. Wei. Conditional analysis of mixed poisson processes with baseline counts: implications for trial design and analysis. *Biostatistics*, 4:479–494, 2003.

[2] A.S.C. Ehrenberg. *Repeat-Buying: Facts, Theory and Applications*. London: Charles Griffin & Company Ltd.; New York: Oxford University Press., 1988.

[3] Peter S. Fader and Bruce G. S. Hardie. The value of simple models in new product forecasting and customer-base analysis. *Appl. Stoch. Models Bus. Ind.*, 21(4-5):461–473, 2005. ISSN 1524-1904.

[4] J. Grandell. *Mixed Poisson processes*, volume 77 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1997. ISBN 0-412-78700-8.

[5] V. Savani and A. Zhigljavsky. Asymptotic distributions of statistics and parameter estimates for mixed Poisson processes. *J. Statist. Plann. Inference*, 137 (12):3990–4002, 2007. ISSN 0378-3758.

[6] A. A. Zhigljavsky and V. Savani. Mixed poisson processes with panel flow. *Present volume*, 2008.