Statistics Netherlands

2017 Scientific Paper

Supplementary material to CBS reports on earthquake frequencies

Frank P. Pijpers

Contents

Nederlands Dit supplement is opgesteld om enkele punten en begrippen toe te lichten die in de CBS rapporten betreffende de aardbevingen in Groningen worden gebruikt (Pijpers, 2014, 2015, 2016a,b). In deze rapportage komen drie onderwerpen aan de orde. Het eerste onderwerp betreft de essentiële karakteristieke eigenschappen van Poisson processen. Het tweede onderwerp betreft de noodzaak van het gebruik van bootstrap procedures voor het bepalen van onzekerheidsmarges die gebruikt worden voor het testen van hypotheses. Het derde onderwerp betreft de keuze van tijdsinterval om tellingen van aardbevingen te aggregeren voor het beoordelen van het tijdsverloop van de aardbevingsfrequentie.

English This supplement is intended to provide further explanatory detail on some points and conventions used in the Statistics Netherlands reports on the earthquakes in the Groningen province (Pijpers, 2014, 2015, 2016a,b). It deals with three separate topics. The first topic concerns the defining characteristics of Poisson processes. The second topic concerns the need for bootstrapping in the reported hypothesis testing. The final topic concerns the choice of time interval over which to aggregate earthquake counts in order to asses and follow the earthquake rate behaviour with time.

1 Event statistics

In the CBS reports (Pijpers, 2014, 2015, 2016a,b) there is a very brief discussion regarding the process that describes the earthquake rate in Groningen, as well as the cumulative distribution function of earthquakes as a function of magnitude: the Gutenberg-Richter plot. Reference is made a number of times to Poisson processes.

Different sources in the scientific literature and statistical textbooks refer to Poisson processes in different ways. In some cases a Poisson process is identified as the unique process with a stationary rate λ of events for which the events are independent. In such a process the waiting time between successive events is then exponentially distributed with a probability distribution function (pdf) P:

$$P(t;\lambda) = \lambda e^{-\lambda t} dt \qquad t > 0 \tag{1}$$

However, this is occasionally also referred to as a homogeneous Poisson process, where the distinction is made with heterogeneous Poisson processes for which the rate λ can be time dependent. For the latter the pdf for the waiting time between successive events is:

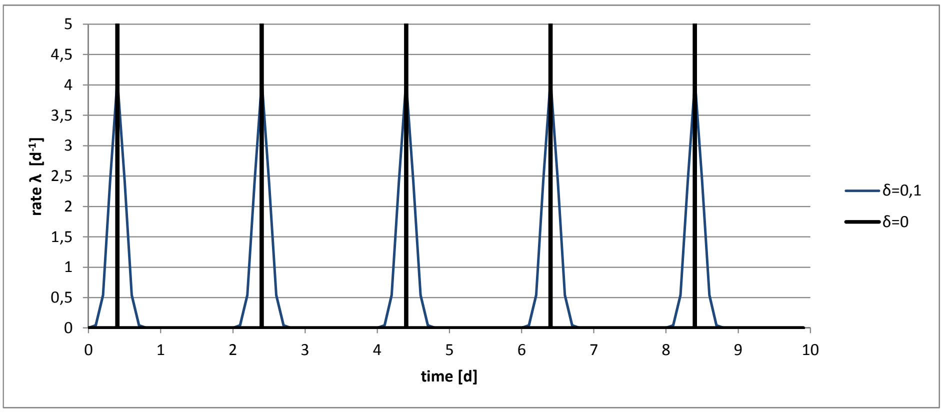
$$P(t;\lambda) = \lambda(t) e^{-\int_0^t \lambda(t')dt'} dt \qquad t > 0 \tag{2}$$

The latter definition has some issues however. For instance one could suppose a rate $\lambda(t)$ of the form:

$$\lambda(t) \equiv \sum \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{(t-n\Delta-t_0)^2}{2\delta^2}} \tag{3}$$

In the limit of $\delta \downarrow 0$ this becomes a comb function (see fig. 1.1) where the rate λ is 0 at all times except at the times $t_0 + n\Delta$ for all integer values of n, for which the rate λ is infinite . While conforming to the prescription for a heterogeneous Poisson process, this process is evidently not stochastic at all, but deterministic. Clearly the lack of stationarity of the process has also affected

Figure 1.1 Graph of the time dependent rate λ for two different choices of the peak width parameter δ. The unit of time is taken as days $[d]$ but could also be any other unit, where evidently the unit for λ would have to be the reciprocal of that same unit



the independence of events: as soon as the time for one event is known, the timing of all successive and previous events is also known.

Even for finite small values of δ the independence of events is questionable since the predictive power of a single event for the timing of all other events is very high, as would also be quantifiable for instance by using the autocorrelation function of the process. While the stationarity of a stochastic process and the statistical independence between events in that process are distinct concepts, there is interdependence. One cannot simply abandon stationarity and take preservation of event independence as self evident.

It is possible to additionally impose a condition on λ such that an autocorrelation function becomes practically indistinguishable from a process in which the events are independent and therefore uncorrelated. The condition is that the expectation value for the interevent time must be much less that the typical time-scale over which λ changes value:

$$\left|\frac{1}{\lambda(t)}\right| \ll \left|\frac{1}{\lambda(t)}\frac{d\lambda}{dt}\right|^{-1} \tag{4}$$

which can be rewritten as:

$$\left|\frac{1}{\lambda(t)^2}\frac{d\lambda}{dt}\right| \ll 1 \tag{5}$$

In this case the evolution of the process is sufficiently slow that over successive finite epochs it can be interpreted as a succession of processes which are indistinguishable from Poisson processes with different (average) rates for the different epochs.

The above example demonstrates that time-dependent events rates are incompatible with (complete) event independence. This means that even if event independence is considered to be the only requirement or essential characteristic of a Poisson process, lack of stationarity of the event rate implies that such a process does not satisfy this requirement. Therefore it is incorrect to state that for a stochastic event process to be designated as a Poisson process, it is sufficient to assume or demonstrate independence between events, and (by implication) that any kind of time dependent event rate would be permissible. For this reason it is preferable to limit the use of the designation ”Poisson process” to a process with independent events **and** a stationary event rate.

Under some conditions a time-dependent event rate can yield a process which has low-order correlation functions that are statistically indistinguishable over finite time-intervals from those of a Poisson process and in that sense such a process can be **approximately** Poissonian. In the CBS reports the wording Poisson process is always reserved to mean a stochastic process that has independent events and also a stationary event rate. This means that any stochastic process for which the event rate is (assumed to be) not stationary, or the events are not fully independent, by this definition is not a Poisson process.

From the analyses of the earthquake rate in Groningen it appears clear that this rate is not constant over time. It is therefore not a Poisson process, given the reasoning above. It is also not clear that the generation rate of earthquakes in Groningen changes sufficiently slowly with time for the process to be even approximately Poissonian. In addition, the possible occurrence of aftershocks, triggered by prior events, implies that the events are not all fully independent. For these reasons, the earthquakes in Groningen do not satisfy the statistics of a Poisson process.

2 bootstrapping

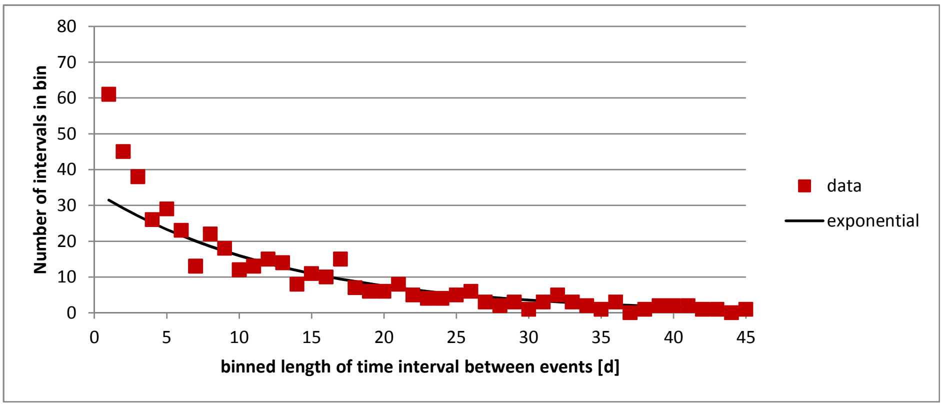
The CBS reports (Pijpers, 2014, 2015, 2016a,b) use a bootstrapping procedure in which the earthquake catalog is used to draw samples in order to calculate likelihoods and confidence limits for earthquake counts. It should be specified that this bootstrapping procedure is done in order not only to enable comparing counts or count rates in different regions and at different epochs, but also to enable comparing distribution functions of earthquake magnitudes. In the tabular material in the CBS reports only count rates are needed. For that purpose alone it would have sufficed to use the exact solution: a multinomial distribution for the counts in each region and epoch. However, if also the distribution with respect to earthquake magnitude is of interest (eg. fig. 2.4 of Pijpers (2016b)), a bootstrapping procedure is more convenient.

By using a bootstrapping procedure it is not necessary to make assumptions regarding the processes responsible for producing the shape of the distribution function (or its cumulative ie. the Gutenberg-Richter plot). This means it is not required to know how to model the geophysical processes responsible for generating earthquakes of various strengths nor what precisely determines the sensitivity limitations of the detection process. The bootstrapping procedure ensures that for the synthetic data the shape of the distribution function for the full real catalog overall is preserved, while at the same time allowing for regional or temporal variations.

It is only in this limited sense that no assumptions are made regarding the earthquake process. An assumption that is implicit in the use of the bootstrapping procedure is the independence between events. The bootstrap samples are drawn independently. If there is an inter-event correlation present in the true events, then the hypothesis testing could be affected if this is not accounted for in the bootstrapping procedure. This would also be the case if a multinomial distribution were to be used to calculate expectation values and confidence limits for counts in the designated regions and epochs, since event independence is also implicit in its application.

It is not uncommon for earthquakes to be clustered in space and time. Physically one would expect that one earthquake can act as a trigger for further earthquakes if the system is in a critical or metastable state. The issue is then to determine to what extent the earthquake catalog as a record of independent events is ’contaminated’ by aftershocks, triggered by previous events. If the timing of earthquakes were to conform to those of a (stationary) Poisson process, one

Figure 2.1 Graph of the interevent timing for all earthquakes in the catalog with magnitudes $M > 1$, that occurred within the zone designated as large in the CBS reports, since 2003



would expect the distribution function of interevent timings for successive events to conform to an exponential distribution. This distribution can be determined from the data: all earthquakes are selected with $M > 1$ in the zone of interest (ie. zone 'large' of the CBS reports) and occurring after 1 Jan. 2003. Using a maximum likelihood procedure, an exponential function is fitted to these timing interval data. At least visually it appears from this that there is an excess of quite short intervals: with less than 4 days between events. Such a result could arise if within the full time range over which data is collected there are intervals where the event rate is higher than the overall average, so that the complete distribution is the summation of several exponential functions with different parameters. An excess can also be produced if there are aftershocks present, ie. an excess of (triggered) events soon after a 'normal' event. If one were to assume that all of the excess events are due to aftershocks, then one obtains a very generous upper limit of 15% for the contamination of the earthquake catalog with aftershocks. There is work by van Thienen-Visser et al. (2015), using the Reasenberg algorithm to identify aftershocks, suggesting that in fact the proportion of earthquakes in the catalog for Groningen is closer to about 3 %.

The issue therefore is whether a maximum of 15% contamination with aftershocks would be enough to affect the conclusions from statistical measures that assume event independence. To assess this in the CBS reports an algorithm is used to identify those shocks which are most likely to in fact be aftershocks. The analyses concerning earthquake rates are then repeated but after removal of these (potential) aftershocks. One such algorithm for the identification of aftershocks is that of Baiesi and Paczuski (2004). In this algorithm the magnitude of earthquakes plays a role in determining the likelihood of aftershocks and hence the identification of a particular shock as an aftershock of a previous earthquake. Also the fractal dimension of the spatial distribution of earthquakes and the timing difference come into this identification process. Adjusting a threshold level in order to identify no more than the upper limit of 15% of earthquakes as aftershocks, results in a cleaned catalog. Evidently there is no absolute certainty that by this process there are no longer any aftershocks in the cleaned catalog. Also it seems likely that if as many as 15% of the shocks are removed, quite a few earthquakes have then been removed from the catalog even though they were not aftershocks. For this reason it is considered to be of importance to report results both with and without applying such filtering for aftershocks. The results of the analyses concerning varying earthquake rates are the same, regardless of whether earthquakes that could potentially be aftershocks are removed or not. Therefore the assumption of complete independence that is assumed in the bootstrap procedure, does not appear to be a major issue when analysing the timing behaviour of earthquakes in Groningen.

3 aggregation time intervals

The Statistics Netherlands reports (Pijpers, 2014, 2015, 2016a,b) also report time series of earthquake counts. In this case the total time span covered by the earthquake catalog is sectioned in successive intervals and the count in each interval is plotted against the central time of that interval. In the choice of the length of each interval, or equivalently the number of such intervals to use, there are two requirements that need to be balanced. The first is the requirement that a sufficiently large number of events be available in each interval in order to reduce the confidence intervals for the underlying rates as much as feasible. Increasing the length of the time interval, ie. reducing the number of such intervals, is evidently helpful for this requirement. The second requirement is that one wishes to have sufficient time resolution to asses variability of the earthquake rate with time. If in particular one surmises that the annual periodicity of the gas production might be reflected in a periodicity of the earthquake rate it is necessary to use a sampling rate for the time series that would allow a detection of annual variation.

In the earlier reports (Pijpers, 2014, 2015), the emphasis is on establishing whether or not there has been a general upward trend in earthquake rate, with a break at, or soon after, the time at which production at a central production cluster was drastically reduced. For that type of analysis, the time resolution is not as crucial as it would be to establish (annual) peridocity. A length of interval of roughly 1.5 yrs, ie. 8 intervals appears sufficient to establish a level, a slope, and a difference before and after January 2014; the time of the gas production adjustment. Using even fewer intervals will not reduce the uncertainty substantially and is not advantageous for establishing the time behaviour, whereas using many more is not essential for establishing a trend or a break and would have costs in terms of uncertainty. In the process of the analysis several variants with different interval lengths were explored without yielding essentially different results from those reported.

A linear time dependence of the earthquake rate over these epochs could have been used as a fitting function but would yield an extrapolation to negative rates in the past, before the beginning of the catalog. While this would be an evident misuse of the fitting function, and not the intention or purpose for which it is constructed, an exponential time dependence which does not suffer from such a shortcoming seems better motivated. In practice when performing weighted least-squares and maximum likelihood fits to the data, the quantitative measures for quality of fit, such as an R^2-measure, support the choice of an exponential time dependence to use in setting up a hypothesis rather than a linear time dependence. In the most recent report (Pijpers, 2016b) an alternative is also explored, which is to use the cumulative gas production as a 'clock' which implies a somewhat different time dependence of the earthquake rate.

If it is necessary to test whether the production adjustment has had any kind of effect on the earthquake rate, the usual way to proceed is to build a prediction based on a null hypothesis that there is no change at all. The aim of the time series fits in the early reports is to establish what trend in the rate can be used with some justification up to the date of the production adjustment. With that in hand, one can then attempt to answer the question whether that trend continues after the production adjustment date. This is distinct from assuming that the true event rate must follow an upward trend with time, be it exponential, linear or otherwise.

In the first few reports it is established that at, or soon after, the time of the production adjustment, the earthquake rate has changed in character. Therefore it is worthwhile to explore whether there is a closer correspondence in time behaviour between the gas production rate and the earthquake rate. One example of that would be to establish whether there is an annual variation in the earthquake rate, just as there is in the gas production rate, possibly with some fixed phase delay between the two. As is also stated in the reports, the presence of such a correlated behaviour does not prove a causal relationship, but if a correlation is absent then some forms of a direct relationship can be discounted immediately. The reason to focus on the annual behaviour is that the full spectrum of time variations present in the gas production data is dominated by the annual periodicity. For the lower amplitude variations at different frequencies/periods any correlation between that and earthquake rate variations will be much harder to measure.

In order to correlate the time series of earthquake rate and gas production rate, with a dominant period of one year, the sampling of the time series needs to have a cadence such that such variations are not 'destroyed' by the smoothing implicit in using time intervals over which to accumulate earthquake counts. The intervals of 1.5 years used in the early reports are too long for this purpose. If one adheres to the requirements of Fourier theory and the Nyquist sampling theorem (cf. Bracewell (1965)) the intervals would need to be shorter than 0.5 years. Unfortunately such short intervals would lead to very few counts in a number of the intervals, with associated large uncertainties for the rates. However, if one uses intervals with a length intermediate between 0.5 years and 1 year, the signature of periodicity is preserved although the amplitude of such a signal is reduced. In terms of Fourier transforms, the frequency corresponding to 0.75 years, which is $1.5\times$ the Nyquist frequency, is 'folded' back into the resolvable range of frequencies between 0 and the Nyquist frequency (Bracewell, 1965) so that it appears with a reduced amplitude at the position of $0.5\times$ the Nyquist frequency.

From this reasoning it appears that the best compromise between sufficiently high event counts and sufficiently high sampling cadence is to aim for an interval length at or near 0.75 years: ≈ 274 days. For the practical reason of using as much of the full extent of the catalog as possible, intervals with lengths of up to 300 days have also been employed. The precise value may have some impact on the signal-to-noise but this is minor compared to the intrinsic noise from the stochasticity of the process generating earthquakes. The conclusions of the reports do not depend on the precise value of this choice.

References

Baiesi, M. and M. Paczuski (2004). Scale-free networks of earthquakes and aftershocks. *Phys. Rev. E 69(6)*, 066106--066114.

Bracewell, R. (1965). *The Fourier transform and its applications*. McGraw-Hill. (paperback edition: 1999).

Pijpers, F. (2014). Phase 0 report 2 : significance of trend changes in tremor rates in Groningen. Technical report, Statistics Netherlands.

Pijpers, F. (2015). Trend changes in tremor rates in Groningen : update november 2015. Technical report, Statistics Netherlands.

Pijpers, F. (2016a). Trend changes in tremor rates in Groningen : update may 2016. Technical report, Statistics Netherlands.

Pijpers, F. (2016b). Trend changes in tremor rates in Groningen : update nov 2016. Technical report, Statistics Netherlands.

van Thienen-Visser, K., D. Sijacic, M. Nepveu, J.-D. van Wees, and J. Hettelaar (2015). Response of induced seismicity to production changes in the Groningen field. Technical report, TNO.

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress: Statistics Netherlands, Grafimedia
Design: Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
verkoop@cbs.nl
Fax +31 45 570 62 68
ISSN 1572-0314