NAM

A Simulation Study into the Detectability Threshold for Seasonal Variations in Earthquake Occurrence Rates within the Groningen Field

Machine Learning

IBM and Shell Research

T. Park and K. Nevenzeel

Date February 2019

Editors Jan van Elk & Dirk Doornhof

General Introduction

The seismological model (Version 5) currently used in the assessment of hazard and risk for the induced seismicity in Groningen, provides a probabilistic prediction of the seismicity dependent on the local reservoir pressure depletion associated with the gas volume produced. The seismicity is in this model not dependent on the gas production rate. The gas volume extracted determines reservoir pressure depletion, which governs the expected number and magnitude of induced earthquakes. Within the model, the expected number of events depends on the pressure depletion, but not the rate of that depletion. Theoretically, there are processes which potentially could cause the expected event number, for a given incremental volume of gas production to depend on the rate of that gas production. These could be associated with the geomechanical behaviour of faults (e.g. rate and state frictional fault behaviour) or compaction (e.g. a-seismic stress relaxation at production time scales).

However, studies carried out as part of the research program of NAM have not been able to identify whether these processes play a significant role or been able to quantify the impact of gas production rate on seismicity. In an environment of decreasing and more stable gas production rates, ignoring potential production rate dependency of the seismicity will be conservative and lead to a potential over-estimation of hazard and risk.

Given the current state of knowledge, NAM is not in a position to increase the sensitivity of the seismological model to production rate changes as this was so far found to degrade the performance of the model and accepts that as a result the assessment of hazard and risk might be conservative. The current model yields a sensitivity to seasonal depletion rate changes that is thought to be close to the upper bound of sensitivities consistent with the observed catalogue. On the other hand, based on the research to date, seasonal seismicity variations within the catalogue are lower than the detection threshold.

In the operation of the field, NAM will make every effort to reduce fluctuations in gas production. The Minister of Economic Affairs has, on the advice of the regulator SodM, imposed limits to the production fluctuations. NAM will report on any excursions from these set limits.

Over the past years, NAM has carried out several studies into the dependency of the induced seismicity in Groningen on the gas production rate from the field. This included studies into reservoir behaviour (Ref. 1), modelling of the various mechanisms that could induce production rate dependency (Ref. 4 and 5) and analysis of field data using machine-learning and statistical techniques (Ref. 2, 3, 4, 6, 7 and 8). The last study (Ref. 8) investigated, using Machine-learning techniques, whether the earthquake catalogue for the Groningen field shows seasonality resulting from the seasonally changing gas production rate. As an extension to this seasonality study the current report investigates the detectability threshold for seasonal variations in earthquake occurrence rates within the Groningen field.

References

1. Geurtsen, L., P. Valvatne and A. Mar-Or, Optimisation of the Production Distribution over the Groningen field to reduce Seismicity, NAM, December 2017
2. Bierman S.M., R. Paleja, and M. Jones, Statistical methodology for investigating seasonal variation in rates of earthquake occurrence in the Groningen field, January 2016
3. Bierman S.M., Seasonal variation in rates of earthquake occurrences in the Groningen field, August 2017
4. Bourne, Stephen and Steve Oates, The influence of stress rates on induced seismicity rates within the Groningen field, Shell Research, August 2018.
5. DeDontney, Nora, and Suvrat Lele, Impact of Production Fluctuations on Groningen Seismicity – Part 1, Geomechanical Modelling using Rate of State friction, ExxonMobil Upstream Research Company, 2018.
6. Burch D. and B. Symington, Impact of Production Fluctuations on Groningen Seismicity – Part 2, Data Analytics, ExxonMobil Upstream Research Company, 2018.
7. J. Limbeck, F. Lanz, E. Barbaro, C. Harris, K. Bisdom, T. Park, W. Oosterbosch, H. Jamali-Rad and K. Nevenzeel, Evaluation of a Machine Learning methodology to forecast induced seismicity event rates within the Groningen Field.
8. Seasonality analysis for induced seismicity event rate time series within the Groningen Field Machine Learning, IBM and Shell Research, Park T., H. Jamali-Rad, W. Oosterbosch, J. Limbeck, F. Lanz, C. Harris, E. Barbaro, K. Bisdom & K. Nevenzeel, October 2018.

NAM

Title	**A Simulation Study into the Detectability Threshold for Seasonal Variations in Earthquake Occurrence Rates within the Groningen Field** **Machine Learning**		**Date**	February 2019
			Initiator	NAM
Autor(s)	**T. Park and K. Nevenzeel**	**Editors**	Jan van Elk and Dirk Doornhof	
Organisation	IBM and Shell Research	**Organisation**	NAM	
Place in the Study and Data Acquisition Plan	Study Theme: Impact Production Fluctuations Comment: The seismological model (Version 5) currently used in the assessment of hazard and risk for the induced seismicity in Groningen, provides a probabilistic prediction of the seismicity dependent on the local reservoir pressure depletion associated with the gas volume produced. The seismicity is in this model not dependent on the gas production rate. The gas volume extracted determines reservoir pressure depletion, which governs the expected number and magnitude of induced earthquakes. Within the model, the expected number of events depends on the pressure depletion, but not the rate of that depletion. Theoretically, there are processes which potentially could cause the expected event number, for a given incremental volume of gas production to depend on the rate of that gas production. These could be associated with the geomechanical behaviour of faults (e.g. rate and state frictional fault behaviour) or compaction (e.g. a-seismic stress relaxation at production time scales). However, studies carried out as part of the research program of NAM have not been able to identify whether these processes play a significant role or been able to quantify the impact of gas production rate on seismicity. In an environment of decreasing and more stable gas production rates, ignoring potential production rate dependency of the seismicity will be conservative and lead to a potential over-estimation of hazard and risk. Given the current state of knowledge, NAM is not in a position to increase the sensitivity of the seismological model to production rate changes as this was so far found to degrade the performance of the model and accepts that as a result the assessment of hazard and risk might be conservative. The current model yields a sensitivity to seasonal depletion rate changes that is thought to be close to the upper bound of sensitivities consistent with the observed catalogue. On the other hand, based on the research to date, seasonal seismicity variations within the catalogue are lower than the detection threshold. In the operation of the field, NAM will make every effort to reduce fluctuations in gas production. The Minister of Economic Affairs has, on the advice of the regulator SodM,			

	imposed limits to the production fluctuations. NAM will report on any excursions from these set limits. Over the past years, NAM has carried out several studies into the dependency of the induced seismicity in Groningen on the gas production rate from the field. This included studies into reservoir behaviour (Ref. 1), modelling of the various mechanisms that could induce production rate dependency (Ref. 4 and 5) and analysis of field data using machine-learning and statistical techniques (Ref. 2, 3, 4, 6, 7 and 8). The last study (Ref. 8) investigated, using Machine-learning techniques, whether the earthquake catalogue for the Groningen field shows seasonality resulting from the seasonally changing gas production rate. As an extension to this seasonality study the current report investigates the detectability threshold for seasonal variations in earthquake occurrence rates within the Groningen field.
Directliy linked research	(1) Gas Production (2) Machine Learning (3) Reservoir Modelling (4) Geomechanical Modelling (5) Seismological Model
Used data	KNMI Earthquake catalogue Groningen gas production data
Associated organisation	NAM
Assurance	

A Simulation Study into the Detectability Threshold for Seasonal Variations in Earthquake Occurrence Rates within the Groningen Field

by
T. Park (GSNL-PTX/D/S)
K. Nevenzeel (IBM Services)

Executive Summary

Business purpose:

Decades of gas production caused induced seismicity in the Groningen gas field, located in the Northern part of the Netherlands. Any increased understanding of the physical mechanisms governing induced seismicity within the Groningen field will create opportunities to improve the reliability of the Probabilistic Seismic Hazard and Risk Analysis (PSHRA) for the exposed population. Potential seasonality of seismicity might be a useful diagnostic to screen candidate physical mechanisms (Bourne & Oates, 2018). This study is a direct extension of (Park, et al., 2018), who concluded that for magnitude ranges above the concordance magnitude of completeness, little to no evidence for seasonality could be found. (Park, et al., 2018) could not say whether the absence of evidence for seasonality is due to:

a. there is no seasonal pattern in the event rates;
b. there is a pattern, but its amplitude is too small for our tests to detect.

This study further investigates the absence of evidence by determining the amplitude of seasonality required for detection.

Approach:

As the true form of the seismicity event rate function – and any seasonal variations therein – is unknown, seismicity event rate model functions are simulated using a factorial approach based on factors like the mean event rate, seasonal shape and dispersion level. We also crucially vary the amplitude of any seasonal effect. Here we follow (Ader & Avouac, 2013) and define the amplitude as the size of the seasonal variation in earthquake rates as compared to the mean rate. Analysis of the agreement between the simulation models and the observed seismicity allows us to rule out certain factor combinations. Subsequently, five hypothesis tests based on different assumptions and potentially sensitive to different simulation factors are applied to the simulated data: spectral analysis using the Schuster Test or the Discrete Fourier Transform, model fitting using Generalized Additive Models and comparing monthly seismicity either parametrically or nonparametrically. Testing at an increasing level of seasonality amplitude allows us to determine at which threshold amplitude seasonality becomes statistically significantly detectable.

Main Findings:

- The false positive rates of the hypothesis tests employed varies quite a lot. The Discrete Fourier Transform and the Parametric Hypothesis tests have a false positive rate close to the expected level of 5%. The Generalized Additive Models have an elevated false positive rate, in some cases exceeding 15%, positive results of this test should be treated with some caution and subject to additional scrutiny. The nonparametric hypothesis test was in general conservative with a rate closer to 4%. The Schuster test was found to have the expected 5% false positives for the case of independent events but was found to have an elevated false positive rate for counts simulated from an over dispersed distribution. Positive results from the Schuster test should also be scrutinised for aftershocks.
- In terms of detecting seasonal variations, (reassuringly) the magnitude of the seasonal effect is the most important factor. The form of the seasonal function also plays a large role, especially for the Schuster Test and the Discrete Fourier Transform. The same is true for the mean earthquake occurrence rate, which is particularly important for the parametric and nonparametric hypothesis tests.

- Let the minimum seasonality detection threshold be the minimum amplitude of any seasonality variation in the earthquake occurrence rate where a given test has a false negative rate, tested at the 95% confidence level, of less than 20%. For the range of the factors considered, the minimum seasonality detectability threshold is in the range 17.5%-27.5% for a minimum magnitude of 1.2 and in the range of 22.5%-55% for a minimum magnitude of 1.5.

Table of Contents

Table of Figures

Table of Tables

1. Introduction: Overview, Earlier Work & Study Goals

In our previous report, (Park, et al., 2018), we aimed to detect seasonal variations in the occurrence rate of earthquakes in the Groningen field. This is relevant as gas production rates are seasonal, with higher rates in the winter months as compared to the summer, and consequently so might be the occurrence rate. The exact size and shape of the seasonal effect will depend on the physical properties of the reservoir. Studying the seasonal patterns can therefore lead to a greater appreciation of physical earthquake generating mechanism (Bourne & Oates, 2018). For certain potential physical mechanisms, it is possible to reduce the expected number of earthquakes for a given volume of production by reducing the fluctuations in production rate. As part of the Study and Data Acquisition Plan (NAM, 2016) in the context of the Measure and Control Protocol (NAM, 2017), this study aims to answer some key questions following on from (Park, et al., 2018).

1.1. Open questions from Previous Work

Our previous report, (Park, et al., 2018), concluded that for magnitude ranges above the generally agreed magnitude of completeness (the concordance magnitude of completeness), little to no evidence for seasonality in seismicity event rates could be found. Given this result there are two possible explanations,

1) There is no seasonal pattern in the event rates.
2) There is a pattern, but its size is too small for our tests to detect.

The natural next question is to consider is:

- how small does the amplitude of any seasonal effect need to be for our tests to fail to detect it?

Another issue arising from the previous study is around the relative merits of the five tests we applied. Each one has different theoretical properties and expected strengths and weaknesses however it is not immediately clear which, if any, are appropriate for our use case. Each test may also have a different definition of the null and alternative hypotheses. We address this by looking at the following questions:

- What are the false negative rates of each test?
- What are the false positive rates of each test?

In both cases the questions need to be considered for a range of different assumptions about the earthquake generating process. While there is no data set which truly matches the assumptions of a test, the degree to which the assumptions are violated and the sensitivity of the test to those assumptions can dictate if the test results are valid. Different tests may also test different null hypotheses which will again affect their suitability for our application. These points then lead us to another question:

- Which of the tests is most suitable for use on the Groningen catalogue?

1.2. This Study: Finding a Detectability Threshold for Seasonal Variations

In order to answer the above questions, we perform a simulation study. In this study we simulate earthquake count data from a range of plausible simulation scenarios. In this way we are able to control the properties of the simulated data including the amplitude and form of seasonal patterns. Given that we know the truth we also know if the test has returned a positive result when seasonality is absent (type I error or false positive) or if the test has returned a negative result when seasonality is present (type II error or false negative). If we repeatedly simulate from the same model, we can estimate the type I and type II error rates for each test given the simulation model. We can then vary the simulation model to explore how the detectability varies under different model forms and for different amplitudes and functional forms of seasonality.

The tests we will be using are detailed in Section 2 and include the four used in (Park, et al., 2018). In addition to these we include the Schuster test which is based on the work of (Schuster, 1897). This is the original, single hypothesis, test which was later adapted to the multiple hypothesis test known as the Schuster Spectrum test (Ader & Avouac, 2013). The reason for including this test is that it is widely used for the detection of seasonal patterns in earthquake occurrence rates and has previously been applied to the Groningen catalogue (Bierman, 2017) and (Bourne & Oates, 2018). As the true earthquake generating process is unknown, a set of simulation models are used. This set of models was chosen to test the different assumptions made in the tests whilst also remaining consistent with the observed data. The parameters defining the simulation models are described in Section 3 and the consistency between the resulting seismicity event rate functional forms and observations is discussed in Section 4. Results in terms of type I, type II and the detectability thresholds are provided in Section 5 and conclusions and discussions follow in Section 6.

This study was executed in R version 3.5.1 (R Core Team, 2017).

2. Statistical Tests for Seasonal Patterns

As stated in previous section this study makes use of a range of different test for seasonality in order to evaluate their relative merits. The test we have chosen to use are as follows:

- The Schuster Test (SCH);
- Spectral analysis using Discrete Fourier Transform (DFT);
- Seasonal model fitting using Generalized Additive Models (GAM);
- Group Comparison Parametric Hypothesis Testing (HYP);
- Group Comparison Nonparametric Hypothesis Testing (NPHYP).

In our view, these five methods target three different angles to tackle the same problem. More specifically, SCH and DFT decompose the signal in a spectrum with different frequencies looked at individually, GAM explores how well our data fits a model with a periodic component of interest, and HYP and NPHYP aim at discovering the differences among data groups (in our case grouped by month) to detect seasonal behaviour. The output of each method is a p-value, this is the confidence with which we can reject the null hypothesis H_0. In all cases we judge that the test result is significant at the 95% confidence level if the p-value is below a threshold of 0.05. The exact definition of the null and alternative hypothesis varies for each test but all are chosen to be sensitive to yearly seasonal patterns in the earthquake event rate.

2.1. Detrending

For most of the test methods we will use there is an assumption that the rate of earthquake occurrences is constant between years. The Groningen catalogue does not appear to meet this assumption as the rate of event occurrences appears to be increasing over time. To correct for this we can apply a detrending method to the count data. The detrending method we use is the same as (Park, et al., 2018). This method involves estimating the trend in the earthquake count time series using a simple moving average. The window size for the moving average is 24 months. We chose this size as it is a multiple of 12 months and so will avoid adding or removing any yearly seasonal pattern. To avoid inducing any phase shift in the detrended series we apply the average symmetrically about the current month such that the window in fact covers 25 months with the 1st and 25th months having half weights.

This detrending method cannot be applied to the catalogue directly but is instead applied to a time series of counts within time bins, after detrending these counts are no longer guaranteed to be positive or integers. This is not an issue for most of the methods as they are already applied to binned data and do not assume positive integers. The exception is the Schuster test which makes use of the times of individual events and cannot be easily applied to non-integer or negative count data, we therefore do not apply the detrending step when applying this test. We also do not apply the detrending when using the GAM test, this is because this test does not make the assumption of constant event rates between years. A summary of the input data for each test is given in Table 1, note that while the Schuster test does not require binning we are applying it to daily counts, or equivalently event times rounded to the nearest day. This is for ease of simulation and should have negligible influence on yearly seasonal signals.

Table 1: Summary of input data and pre-processing for each test.

Test	Detrending Applied	Input Data
Schuster Test	No	Daily Counts
Discrete Fourier Transform	Yes	Monthly Counts
GAM	No	Monthly Counts
Parametric Hypothesis Testing	Yes	Monthly Counts
Nonparametric Hypothesis Testing	Yes	Monthly Counts

2.2. Schuster Test

The Schuster test is a hypothesis based on the work of (Schuster, 1897). This test was later adapted to a multiple hypothesis test by (Ader & Avouac, 2013). The Schuster test is based on a 2D random walk process which consists of jumps of unit length. The number of jumps is the total number of recorded earthquakes and the direction of each jump is dictated by the occurrence time of the earthquakes. For an earthquake occurring at time t_n the direction is given by,

$$\theta_n = 2\pi \frac{(t_n \bmod L)}{L}$$

Where L is the period of seasonality we are testing for, in our case $L = 1$ year, and $t_n \bmod L$ is the remainder found when dividing t_n by L. The final distance from the origin of the random walk is then given as,

$$D_L = \sqrt{\left[\sum_{n=1}^{N} cos(\theta_n)\right]^2 + \left[\sum_{n=1}^{N} sin(\theta_n)\right]^2}.$$

The is value of D_L can then be used as our test statistic for the following null and alternative hypotheses:

- H_0: The random walk is a uniform random walk, i.e. $(t_n \bmod L)/L$ follows a uniform distribution.
- H_1: There is a seasonal pattern with period L.

The p-value for this test is calculated as,

$$p = e^{-D_L^2/N}.$$

Where N is the total number of earthquakes in the catalogue. This test makes the following assumptions,

- Events occur independently of each other, e.g. no aftershock process.
- The distribution of event times may vary within a year but is constant between years, e.g. no trend.

Note that the Schuster test can equally be applied to binned count data, in this case all events in a bin are assumed to have the same event time.

2.3. Discrete Fourier Transform

The Discrete Fourier transform (DFT) is the discrete version of the general Fourier transform. It converts a sequence of equally-spaced samples (in our context from the time domain given the time-series we are dealing with) into an equivalent representation in the Fourier domain as follows:

$$\text{DFT:} \quad X_k = \sum_{t=0}^{T-1} x_t \,.\, e^{-i2\pi kt/T}$$

$$\text{Inverse DFT:} \quad x_t = \frac{1}{T} \sum_{k=0}^{T-1} X_k \,.\, e^{i2\pi kt/T}$$

where n and k are integer indices of the data x_n and its discrete Fourier domain representation X_k. Using this representation, X_k will be complex valued. Therefore, it is more usual to analyse the Fourier periodogram, $S_k = |X_k|^2$. Figure 1 helps to illustrate how Fourier Transforms work. Far left we see a simple sine function, where middle left a noise term is added – the combined pattern is shown middle right. Far right the Fourier Transform is shown, where each peak corresponds to a repeated pattern: the more left the longer the period, the higher the peak the stronger the repeated pattern. Fourier transform is a well-known concept in signal processing with a plethora of techniques related to its resolution and accuracy; the interested reader is referred to (Oppenheim, et al., 1983) and (Kay, 1993) for more details.

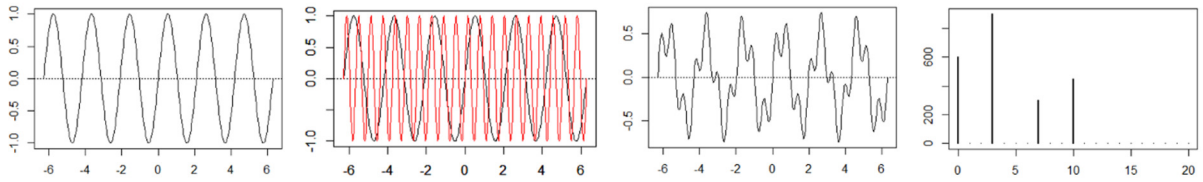

Figure 1: Visual illustration of DFT (right figure) of the time series in the middle-right figure. The time series is a superposition of a higher frequency signal (red, middle-left figure) and a lower frequency signal (left figure).

When applying this transform to the Groningen catalogue we must convert the continuous event times into a discrete time series of counts within equally spaced bins. In principle the bins can be any size, smaller bins give higher temporal resolution but at the risk of having a large number of zero counts. We choose to use monthly bins as this should give sufficient resolution to observe yearly periodicity whilst minimising zero counts. Our sampling time is therefore $T_s = 1$ month and thus we can calculate the Fourier periodogram for frequencies up to 0.5 month^{-1} or 6 year^{-1}, this is known as the Nyquist frequency for this sampling rate. A time series which shows a peak in the periodogram at a frequency $f = 1$ year^{-1} corresponds to behaviour with a period L $= 12$ months, i.e. yearly seasonal behaviour. The null and alternative hypotheses for this test are as follows,

- H_0: The expected value of the periodogram component, S_k, corresponding to a frequency of 1 year^{-1} is equal to the expected value across all frequencies.
- H_1: The expected value of S_k is larger than the expected value of the periodogram across all frequencies indicating the presence of a sinusoidal seasonal pattern with frequency 1 year^{-1}.

Under the null hypothesis it can be shown that,

$$Pr(S_k > s) = e^{-s_k/c}.$$

Here $c = T^{-1} \sum_{k=0}^{T-1} S_k$, is the mean value of the periodogram across all frequencies. This relationship can therefore be used to calculate a p-value indicating the probability of observing a given value of S_k or higher given H_0. This test has the following assumptions:

- The distribution of event counts varies within a year but is constant between years.
- Within a year the pattern of the mean event rate has a Sinusoidal component.
- Events occur independently of each other, e.g. no aftershock process.

2.3.1. Comparison of DFT and Schuster Tests

While the Schuster and DFT tests are generally presented in a different way, one makes use of event times and the other binned counts, we can show that under certain circumstances the tests are very similar. Let us assume an idealised case where earthquakes only occur at discrete time points which are equally spaced with separation T_s. Without loss of generality we can pick our time units such that $T_s = 1$. Following Section 2.2 we can say that,

$$\theta_n = 2\pi \frac{(t_n \bmod L)}{L} \text{ for } t_n \in \{0, 1, 2, \ldots, T-1\},$$

$$D_L^2 = \left[\sum_{n=1}^{N} cos(\theta_n)\right]^2 + \left[\sum_{n=1}^{N} sin(\theta_n)\right]^2,$$

$$= \left[\sum_{n=1}^{N} cos\left\{2\pi \frac{(t_n \bmod L)}{L}\right\}\right]^2 + \left[\sum_{n=1}^{N} sin\left\{2\pi \frac{(t_n \bmod L)}{L}\right\}\right]^2.$$

Since $\cos(\theta + 2\pi) = \cos(\theta)$ and $\sin(\theta + 2\pi) = \sin(\theta)$ we can say that,

$$\cos\left\{2\pi \frac{(t_n \bmod L)}{L}\right\} = \cos\left\{2\pi \frac{t_n}{L}\right\},$$

$$\sin\left\{2\pi \frac{(t_n \bmod L)}{L}\right\} = \sin\left\{2\pi \frac{t_n}{L}\right\}.$$

We now introduce $x_t = \sum_n \delta_{tt_n}$ where δ_{ab} is the Kronecker delta function. Equivalently x_t is the count of the number of earthquakes occurring at time t. Putting this together we can say that,

$$D_L^2 = \left[\sum_{t=0}^{T-1} x_t \, cos\left\{2\pi \frac{t}{L}\right\}\right]^2 + \left[\sum_{t=0}^{T-1} x_t \, sin\left\{2\pi \frac{t}{L}\right\}\right]^2$$

$$= \left|\sum_{t=0}^{T-1} x_t e^{-i2\pi \frac{t}{L}}\right|^2.$$

The period L is related to the index of the Fourier spectrum, k, as $L = T/k$. Therefore we can say that,

$$D_L^2 = \left|\sum_{t=0}^{T-1} x_t e^{-i2\pi \frac{kt}{T}}\right|^2 = S_k.$$

The differences between the DFT and Schuster Tests, as we have applied them, can be summarised as,

- The p-value calculation makes use of either the mean spectrum, c, for the DFT test or the number of earthquakes, N, for the Schuster test.
- The DFT involves grouping the times of earthquakes into discrete bins while the Schuster test makes use of the continuous event times.
- The binning of events allows us to detrend the count series before applying the DFT test which is not done for the Schuster test.

2.4. Model fitting using Generalized Additive Models (GAMs)

An alternative way to test for the presence of seasonality is by model comparison. This is the process of comparing the relative fit of different models, in our case those which contain seasonal terms and those which do not. The models we compare are part of a simple class of Generalised

Additive Models (GAM). A GAM is a flexible class of parametric models, more information can be found in for example (Hastie & Tibshirani, 1990). The exact form of model we choose aims to model the log of the earthquake occurrence rate in the ith month of the jth year of the time series, $\lambda_{i,j}$, as,

$$\log(\lambda_{i,j}) = \alpha_j + s(i).$$

Here α_j is a term which depends on the year, j, only. The second term, $s(i)$, is a smooth spline function which depends only on the month of the year, i, only. The addition of the yearly term is in place of the detrending which is not needed for the GAM test. We chose this model form as it clearly separates the within year seasonality from any trends or longer-term variation. It also makes few assumptions about the form of any seasonal behaviour except that there should be some smoothness between adjacent months within a year. This model form is also used in (Bierman, et al., 2016). Again following (Bierman, et al., 2016) we use the quasi-Poisson likelihood. This is similar to the usual Poisson likelihood but allows for an inflation of the variance caused by possible non-independence of earthquake events, see (Verhoef & Boveng, 2007). An alternative distribution commonly used for this is the negative binomial. In using the quasi-Poisson likelihood we aim to reduce the sensitivity of our test to the assumption of independent earthquake events.

We estimate the values of $\lambda_{i,j}$ time series of monthly counts and the gam function in the R package mgcv, details of the fitting algorithms can be found in (Wood, 2004) and (Wood, 2011). The hypothesis test we are applying has the following null and alternative hypotheses,

- H_0: $s(i) = 0$ for $i = 1, 2, \ldots, 12$.
- H_1: $\exists\, i = 1, 2, \ldots, 12 : s(i) \neq 0$.

Or in other words, we are testing if the function $s(i)$ is nonzero for at least one month. In order to test for the significance of the seasonal effect we examine the significance of the spline term. The p-value for this test is taken directly from the R implementation and is calculated during the model fitting process.

2.5. Group Comparison Parametric Hypothesis Testing

Group Comparison Hypothesis testing can be used to decide whether several groups are statistically significantly different. Here, we test whether the earthquake count in any one month is significantly different to at least one other month. If this is the case, the null hypothesis of no seasonality, H_0, can be rejected. This is done using a multiple hypothesis post-hoc test (multiple comparison test). Post-hoc tests gain insight via pairwise comparison between all the possible combinations of different groups. For example, if we have M different groups of data, multiple comparison post-hoc tests conduct $M(M-1)/2$ different pairwise tests. In our case we group together earthquake counts which correspond to the same month in different years. Therefore, each group will have as many data points as there are years in the data set. If we do not have a whole number of years in our data set then some months will have more data points than others, this is taken into account by the test we are applying. In order to use a post-hoc test we must first use a portmanteau test to test if there is any pair of groups which are different, the post-hoc test is then used to identify which pair are significantly different.

An important point to highlight is that the more groups we compare, the higher the chance of erroneous inference to occur, as each individual post-hoc comparison has a Type I error rate, where a true null hypothesis is rejected, equal to the significance level. Several statistical methods try to prevent this from happening by requiring a stricter significance threshold for individual comparisons to compensate for the number of inferences being made. This is called p-adjustment for post-hoc tests and well-known examples of that are Bonferroni and Holm tests, see for example (Miller, 1981).

We have chosen to apply the test to monthly earthquake counts, $M = 12$, but could have grouped the events in other ways such as weekly, quarterly or any arbitrary time bin. The choice of time bin is a trade-off, large bins give greater statistical power as M is reduced. If the bins are too large we reduce the temporal resolution which may, depending on the phase and functional form of any seasonality, reduce the separation between bins. We could also use larger bins and a range of starting points. We chose not to do this as it would increase the number of individual tests done and necessitate additional p-value adjustment.

For the parametric test we first apply an ANOVA test to test if there are any differences between at least one pair of months, if this gives a positive result we move on to the post-hoc test. The parametric post-hoc test we use is Tukey's honest significant difference test, this is very similar to a two-sample unpaired t-test with a build in p-adjustment. The p-values are inflated to keep the Type I error rate equal to the significance level. As with a standard t-test this test assumes that the data follow a normal distribution and are independent. For this test we consider the mean of the detrended event count in each month, denoted as C_i for month i. The null and alternative hypotheses for this test are as follows:

- H_0: $C_i = C_j$ for $i, j = 1, 2, \ldots, 12, i \neq j$.
- H_1: $\exists\ i, j = 1, 2, \ldots, 12;\ \ i \neq j: C_i \neq C_j$.

Or alternatively we are testing if the mean of the detrended event rate count is the same for every month or if there is at least one pair of months with different means. We applied this test using the standard R implementations of the ANOVA test with Tukey's HSD.

2.6. Group Comparison Nonparametric Hypothesis Testing

The final test we apply is the nonparametric version of the parametric test described in Section 2.5. We again do this using monthly earthquake counts. The portmanteau tests is the Kruskal-Wallis hypothesis test and the post-hoc test is the Dunn's test with Bonferroni p-value adjustment. This test does not require the data to be normally distributed. The Kruskal-Wallis H-test (sometimes also called the "one-way ANOVA on ranks") is a rank-based nonparametric test that can be used to determine if there are statistically significant differences in the ranking between multiple groups. If X_i is a single sample of the detrended counts from the ith month then the null and alternative hypotheses for this test are as follows:

- H_0: $\Pr(X_i > X_j) = 0.5$ for $i, j = 1, 2, \ldots, 12; i \neq j$.
- H_1: $\exists i, j = 1, 2, \ldots, 12;\ \mathrm{i} \neq \mathrm{j}: \Pr(X_i > X_j) \neq 0.5$.

We applied this test using the standard R implementations of the Kruskal-Wallis and Dunn's test.

3. Earthquake Simulation Model

In order to evaluate our hypothesis testing methods we must be able to simulate earthquake catalogues from different models. This allows us to test the ability of each method to detect different levels of seasonality under a range of different assumptions. In all cases we will simulate count data within a bin. For a given model we define the mean count in time as, $\mu(t)$. This mean function is used as the mean parameter for the distribution we simulate from, either Poisson or Negative Binomial. We define this mean function in terms of the different simulation factors we wish to include and we define it on the time range 1995-01-01 to 2016-12-01 inclusive. For the DFT, GAM, parametric and nonparametric hypothesis tests we simulate earthquake counts in discrete monthly bins. The Schuster test is different from the others as it does assume discrete time steps but rather uses continuous event times. For this test we therefore simulate counts per day by first scaling the mean function by a factor of 12/365. While this still does not give a continuous scale, we believe that daily bins are sufficiently small compare the expected yearly seasonal period. The validity of this simulation model in terms of its match to observed data will be explored in Section 4. The general form of the simulation model is,

$$\mu(t) = \mu_0[(1-\beta) + \beta x(t)][1 + \alpha s(t+\varphi)].$$

In the following sections we will define the different parts of this model.

3.1. Overall Mean

The first factor we can vary is the overall mean number of events per month in the catalogue, μ_0. In its simplest form we can define the mean function as,

$$\mu(t) = \mu_0.$$

3.2. Trend

We can see from the observed earthquake counts that there may be a trend in the event count over time. We can include this into our simulation model by introducing both a trend function, $x(t)$, and a trend multiplier $\beta \in [0,1]$. We can use these to modify the mean function,

$$\mu(t) = \mu_0[(1-\beta) + \beta x(t)].$$

We place the following restrictions on the trend function,

- $x(t) \geq 0 \text{for } t = 1, 2, \ldots, T,$
- $\sum_{t=1}^{T} x(t) = 1.$

With these restrictions it is clear that defining the trend in this way ensures that $T^{-1}\sum_{t=1}^{T}\mu(t) = \mu_0$ and $\mu(t) > 0 \forall t \in \{1, T\}$. The trend functions we consider are shown in Figure 2.

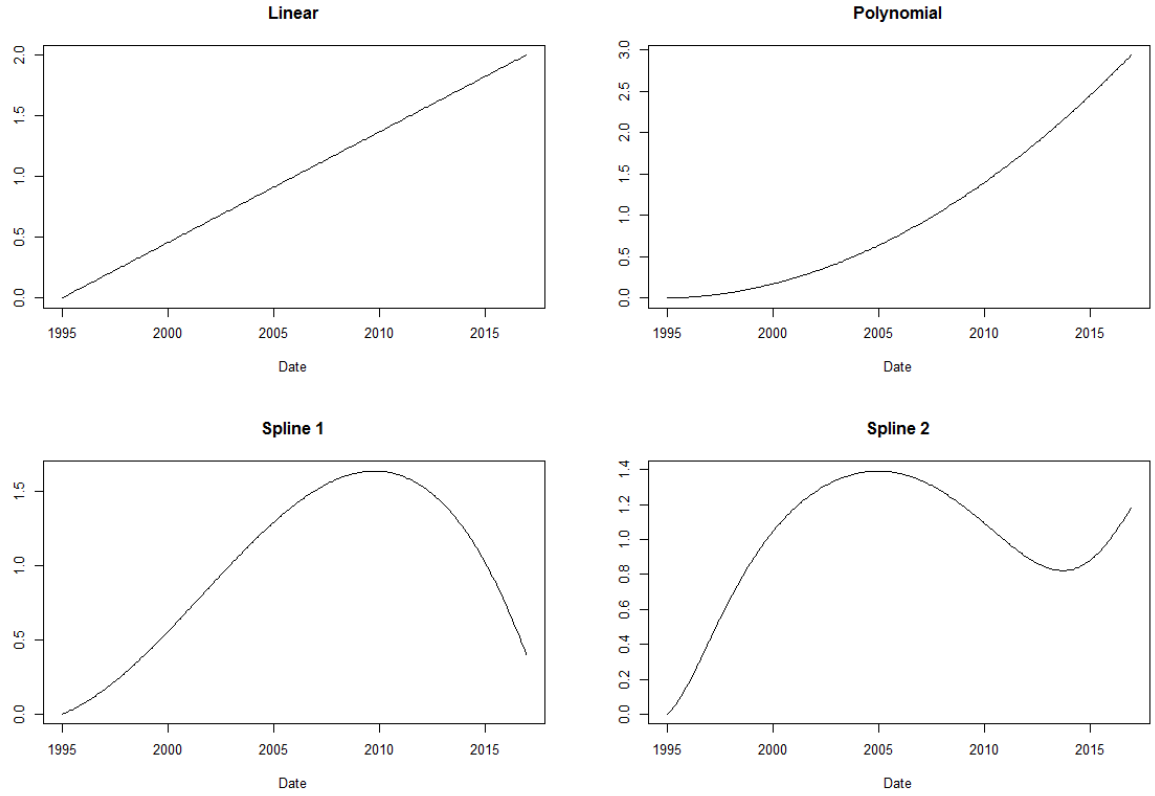

Figure 2: Trend functions considered in the simulations.

3.3. Seasonality

In order to determine the threshold value for detecting seasonality we must be able to vary the amplitude of seasonal variation within the simulated counts. In a similar way to the trend we introduce the seasonal function $s(t)$, and the seasonality amplitude, $\alpha \in [0,1]$. We then introduce seasonal variation in the following way,

$$\mu(t) = \mu_0[(1-\beta) + \beta x(t)][1 + \alpha s(t+\varphi)].$$

We place the following restrictions on the seasonal function,

- $\min_{t\in\{1,T\}} s(t) = -1$,
- $\sum_{t=1}^{12} s(t) = 0$,
- $s(t) = s(t+12)\ and\ \exists t: s(t) \neq s(t+\delta)\ \forall\ \delta < 12.$

This again ensures that $T^{-1}\sum_{t=1}^{T}\mu(t) = \mu_0$ if $\beta = 0$. For the cases where $\beta \neq 0$ the rate will deviate slightly due to the fact that the seasonal amplitude will increase following the trend function. The minimum value of $s(t)$ ensures that $\mu(t) > 0$ for $t = 1, 2, \ldots, T$. The final bullet point above ensures that the function has a period of 12 months and is not seasonal for periods smaller than this. This form of seasonality gives a clear separation between the amplitude and the trend. We expect both of these factors to influence the outcome of the tests and by separating them we are able to explore their individual influence on the test outcomes.

There are an infinite number of choices of $s(t)$ which satisfy the above constraints. We have chosen to consider six different functions, these are shown in Figure 3. In all cases the function is characterised by one large peak and one large trough. We can then further classify the functions in terms of the separation between peak and trough and also the absolute ratio between the sizes of the peak and trough. The ratio between the peak and trough sizes has a physical basis as there may

be an asymmetry between how the reservoir responds to an increase in production rate as opposed to a decrease. The physical explanation for the change in separation between peak and through is the case where any lag between a change in production rate and a change in earthquake rate is again dependent on if that change is an increase or a decrease. Our baseline case is a Sinusoidal pattern where the peak and trough are separated by 6 months and their ratio is 1. The other 5 cases are constructed so that the peak and trough separation is on of {2 months, 6 months} and the absolute ratio between peak and trough values is one of {0.5, 1, 2}. We can also vary the phase of the seasonal function by varying φ. Most of our tests do make any assumptions about the phase of any seasonal pattern. This should in theory make a negligible difference to the detectability however due to monthly binning an offset of a non-integer number of months will affect the size of the largest average rate in any month. The possible exception is the GAM test which does allow for discontinuities in the event rate between years and so may be influenced by the phase, this should be kept in mind when using this test.

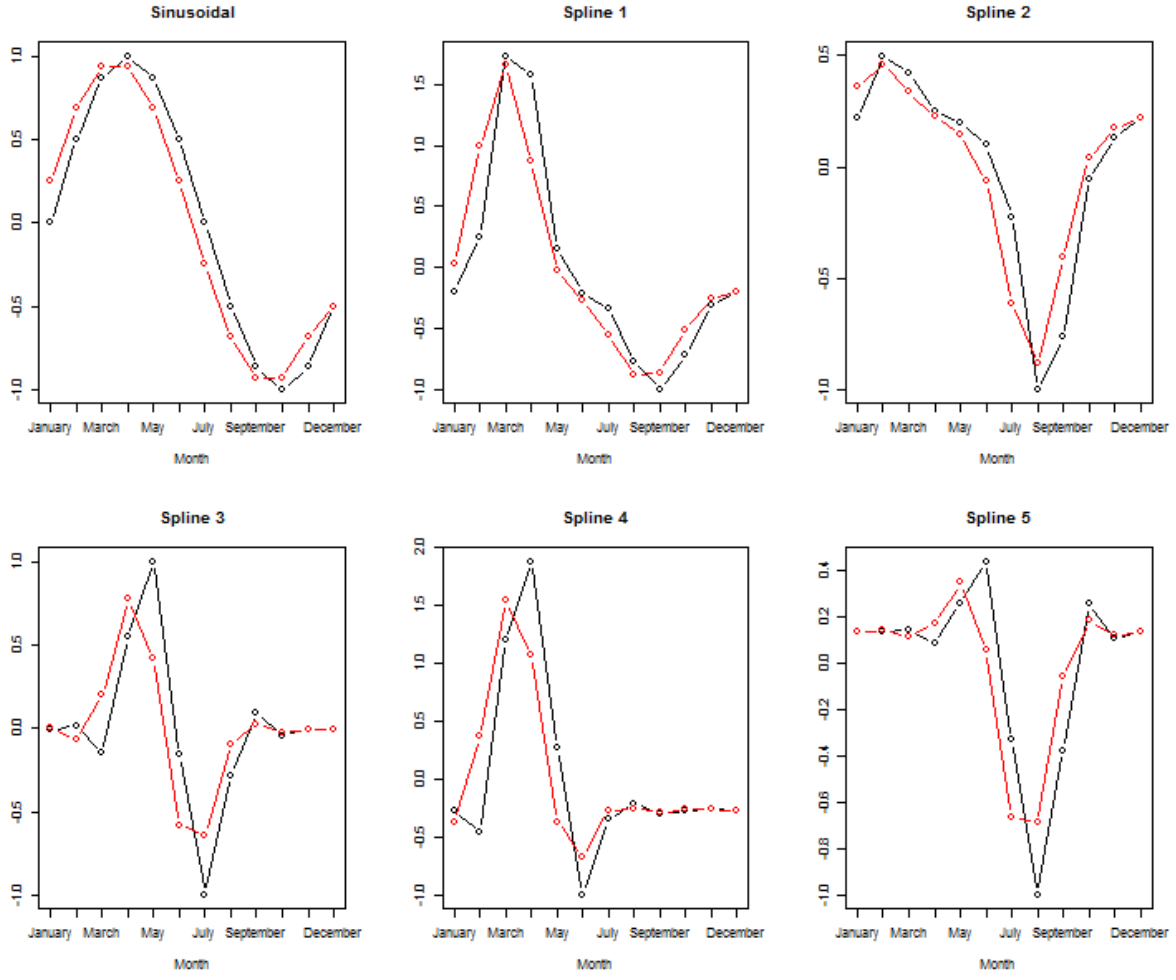

Figure 3: Seasonal Functions, those on the top row have a separation between peak and trough of 6 months while those on the bottom row have a separation of 2 months. Moving from left to right the functions have an absolute peak value to trough value of 1, 2, and 0.5. In each case the black lines show the original function and the red lines show the functions offset by φ = 0.5 Months.

3.4. Dispersion

If we assume that earthquakes occur independently of each other then we would expect the count of earthquakes within a given period to follow a Poisson distribution. However, it is likely that due

to aftershock processes real earthquake data will not follow this distribution. We represent this by introducing a dispersion parameter, d. This parameter controls the ratio between the variance and mean of the distribution of counts. A higher dispersion makes it more likely to have a large cluster of events in the same month, such an event can also be seen as an outlier. We note that this is not a physical aftershock process as it does not introduce dependency between events which fall into different bins. In Section 4 we will explore the fit of this distribution to the data. For a dispersion of 1 we simulate counts from a Poisson distribution. For dispersions greater than 1 we use a negative binomial distribution. The Probability Mass Function, PMF, from which we simulate the earthquake count for month t is then given by,

$$P(X = x) = \frac{\mu(t)^x e^{-\mu(t)}}{x!} \text{ for } d = 1,$$

$$P(X = x) = \binom{x + r - 1}{x} (1 - p)^r p^x \text{ for } d \neq 1.$$

Here d is the dispersion parameter. For the Negative Binomial case the parameters can be expressed in terms of d and $\mu(t)$ as,

$$r = \frac{\mu(t)^2}{(d - 1)},$$

$$p = \frac{\mu(t)(d - 1)}{\mu(t)^2 + \mu(t)(d - 1)}.$$

The PMFs of the distributions we use are shown in Figure 4. Defining the parameters in this way allows for us to vary d whilst keeping $\mu(t)$ fixed.

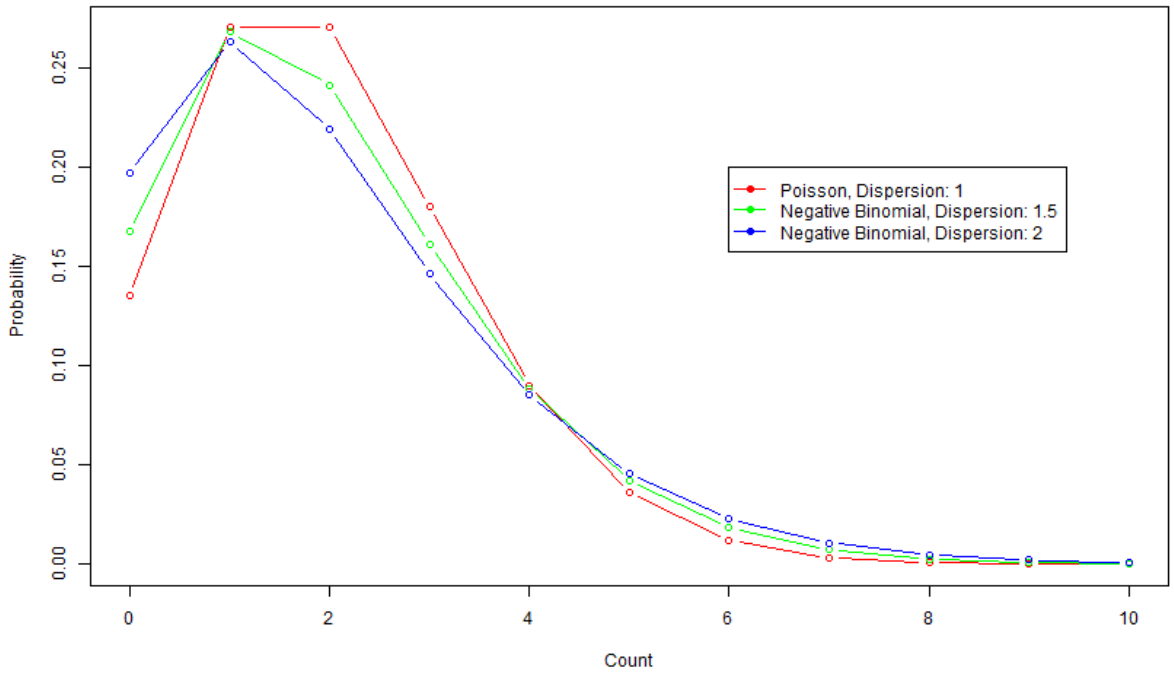

Figure 4: PMFs of Poisson and Negative Binomial distributions, all have a mean count of 2 and dispersions ranging from 1 to 3.

4. Consistency Between Simulated and Observed Earthquakes

In the previous section we detail the range of different factors which go into building the range of different simulation models. It is then natural to ask how closely each of these models matches to the observed data. Clearly results from models which closely match the observations are of greater relevance than those which do not. To analyse this, first earthquake observations are shortly described in Section 4.1. Subsequently, the consistency between observations and simulations is discussed in Section 4.2.

4.1. Observed Earthquakes

4.1.1. Earthquake measurements

The KNMI (the Royal Netherlands Meteorological Institute) has seismicity monitoring stations throughout the Netherlands and specifically in Groningen[1]. The network is described in more detail in e.g. (Dost, et al., 2012) and (Dost & Haak, 2002). Measurements from this network are automatically processed by KNMI and earthquakes detected are formally published in a catalogue[2], which we use as source for seismic events. The induced seismicity catalogue has a straightforward structure as shown in Table 2: The data is provided in tabular form with each row representing an event, with event date and time, location, latitude, longitude, depth, magnitude and evaluation mode. Most of these fields are self-explanatory, possibly except for the location field[3] but that field is not used in our analysis.

Table 2 KNMI induced earthquake catalogue data structure

Date	Time	Location	Lat	Lon	Depth	Mag	Eval mode
1986-dec-26	07h47m51s	Assen	52.992	6.548	1	2.8	Manual
1987-dec-14	20h49m48s	Hooghalen	52.928	6.552	1.5	2.5	Manual
…	…	…	…	…	…	…	…

4.1.2. Uncertainties

The number of sensors in the seismic sensor network, their locations and the data processing procedures used influence detection sensitivities and location uncertainties. As the network has been extended over time, detection sensitivity and location uncertainties vary over time. Table 3 provides an overview of sensitivities as reported by the KNMI, see e.g. (Dost, et al., 2012), (Kraaijpoel, et al., 2015), (Dost, et al., 2017), (Spetzler & Dost, 2017) and the overview of stations referred to above. In general, the horizontal location uncertainty is around 1 km and the vertical uncertainty is between 1-2 km. Given the large vertical uncertainty, vertical locations are pre-set to 3 km for nearly all events.

[1] For an overview of these stations, see https://www.knmi.nl/nederland-nu/seismologie/stations.

[2] Catalogue available at https://www.knmi.nl/kennis-en-datacentrum/dataset/aardbevingscatalogus.

[3] Up to November 30, 2016 the location field described the city or village centre nearest to the event, whilst as of December 1, 2016 the municipality border within which the event took place is registered.

Table 3 KNMI Seismic Sensor Network developments over time

Time	Detection	Localisation	Comments
Since 1995	≥ 1.5	≥ 2.3-1.5	Network installed (8 borehole stations in Northern Netherlands)
±2010	Processing software upgrade, real-time continuous data transmission		
2009-2010	≥ 1.0	≥ 1.5	6 additional borehole stations in Northern Netherlands
2015-2017		≥ ~0.5	Major extension: 64 additional borehole stations in Northern Netherlands

4.1.3. Choice of min. magnitude M_{min}, temporal interval and temporal aggregation period T_{agg}

The magnitude of completeness M_c of a sensor network is usually defined as the lowest value of the moment magnitude of an event for it to be detected with 100% reliability. Event counts with a moment magnitude below M_c are incomplete. With the detection sensitivity increasing over time, an increase in the detection of earthquakes is a combination between a change in seismicity and a change in detection sensitivity. As this effect is strongest for low magnitude seismicity a minimum magnitude cut-off M_{min} is chosen, only earthquakes with a magnitude equal to or higher than M_{min} are taken into account. A sensible choice for M_{min} is the magnitude of completeness M_c – this choice would ensure that all signal picked up comes from seismicity instead of sensor network sensitivity changes. Given the improvements in the sensor network over time, the choice of M_c and the start of the temporal interval T_{start} are coupled: a later T_{start} might allow for a lower M_c and vice versa. The choice for both parameters is, of course, driven by the desire to use as much of the data as possible, while avoiding the introduction of bias.

Following the extensive analysis of M_{min} in section 3.2 of (Limbeck, et al., 2018) we proceed with the following choices for M_{min}, T_{start} and T_{agg}:

- Following both KNMI reported M_c values and the PSHRA default (Bourne & Oates, 2017) $M_{min} = 1.5$ with $T_{start} = 1995$ are used with a $T_{agg} = 1$ months.
- In line with the discussion in (Limbeck, et al., 2018) we find $M_{min} = 1.2$ to be worth considering as an alternative to $M_{min} = 1.5$, whilst acknowledging the possibility that M_c could exceed this choice of M_{min}. This would mean, in turn, that the detectability function would differ between training and forecasting period. Again we take $T_{start} = 1995$ and $T_{agg} = 1$ months.

4.2. Agreement between Observations and Simulations

In order to decide on the range of values to use for the factors in our simulation study we fit the simulation model to the observed monthly earthquake counts. We can then use the uncertainty ranges for each of the parameters as the basis for our simulations. The exact model we are fitting follows that described in Section 3 and takes the form,

$$\mu(t|\mu_0, \beta, x(t)) = \mu_0[(1-\beta) + \beta x(t)].$$

Here we have excluded any seasonal variation, this is because we are already interested in simulating for the full range of all of the factors associated with seasonality. The purpose of this section is to set suitable ranges for the other parameters. In addition to the parameters in the above formula we also fit the dispersion parameter, d. This parameter is present in the likelihood which is either a Poisson if $d = 1$ or a Negative Binomial if $d \geq 1$, this follows on from Section 3.4.

4.2.1. Bayesian Model Fitting

We fit this model using a Bayesian framework and uninformative priors. The priors chosen for each parameter are as follows,

- Mean Rate, μ_0: Gamma(shape = 1, rate = 0.001).
- Trend Multiplier, β: Uniform(0, 1).
- Trend Type, $x(t)$: Uniform Probability across all four functions shown in Figure 2.
- Dispersion, d: Gamma(shape = 1, rate = 0.001).

The posterior distributions for each of the parameters is estimated using MCMC. At each iteration of the MCMC scheme we propose new value for each of the parameters as follows,

- Mean Rate: $\mu_0' = \mu_0 + \delta$, $\delta \sim \mathrm{N}(0, 0.01)$.
- Trend Multiplier: $\beta' \sim U(0,1)$.
- Trend Type: $x(t)'$ is sampled with equal probability from the function in Figure 3.
- Dispersion: $d' = d + \delta$, $\delta \sim \mathrm{N}(0, 0.01)$.

The MCMC ran for 11,000 iterations, the first 1000 were treated as burn-in and discarded leaving 10,000 posterior samples.

4.2.2. Posterior Distributions

The results of this are shown in

Figure 5 for M≥1.2 and

Figure 6 for M≥1.5. Looking at these plots we see that for all factors there is some information in the data to constrain our choices. For the mean rate the 95% quantiles for the posteriors were [1.59, 1.98] for M≥1.2 and [0.86, 1.12] for M≥1.5, unsurprisingly these are different and fairly well constrained. The dispersion parameter is also informed by the data, the 95% quantiles for the posteriors were [1.17, 1.61] for M≥1.2 and [1.13, 1.63] for M≥1.5. This indicates that the dispersion parameters are not significantly different for the two choices of minimum magnitude. The 95% quantiles of the trend multiplier were [0.55, 0.93] for M≥1.2 and [0.5, 0.92] for M≥1.5. Again this is a similar range for both choices of minimum magnitude. Finally, we see that for the trend function the linear and polynomial functions are the only two which have a significant posterior probability. Based on these results we can make a choice of what factors to use in our simulations. These are shown in Table 4. In all cases we have extended the range to include the 95% quantiles of the posteriors. For the trend multipler the value 0 is included as well even though this is well outside of the range of posterior values. This is because this is the case of no trend which should fit the assumptions of all of the tests and so is a useful case to consider.

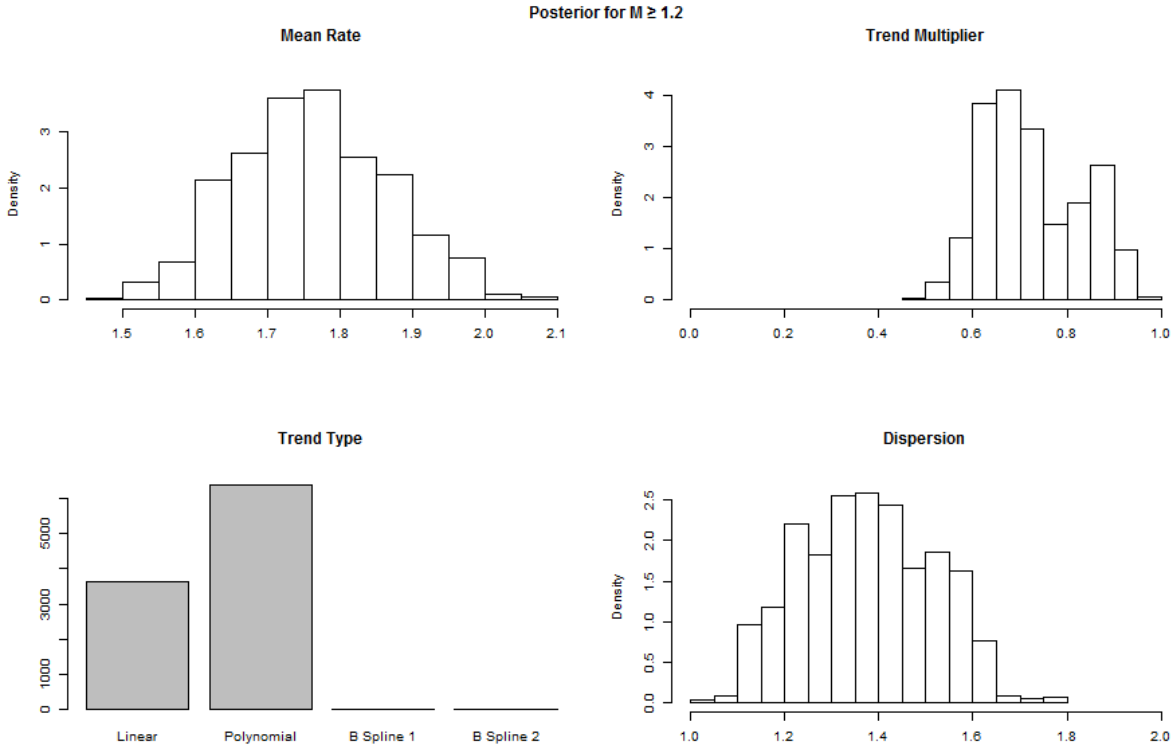

Figure 5: Posterior Distributions from fitting to counts of earthquakes with M≥1.2.

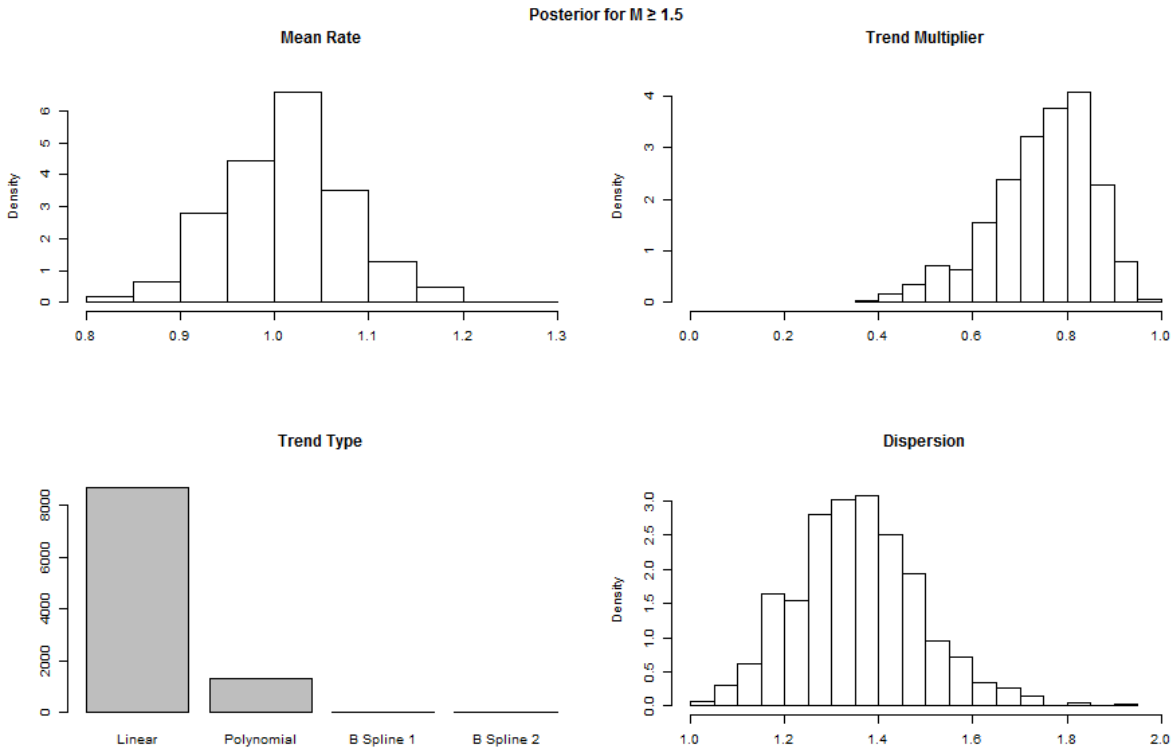

Figure 6: Posterior Distributions from fitting to counts of earthquakes with M≥1.5.

Table 4: Factors used in Simulation Model.

Factor Name	Symbol	Factor Levels Used in Simulation
Mean Rate	μ_0	{0.5, 1, 1.5, 2, 2.5}
Trend Function	$x(t)$	{Linear, Polynomial}
Trend Multiple	β	{0, 0.5, 1}
Dispersion	d	{1, 1.5, 2}
Seasonal Function	$s(t)$	All six from Figure 3.
Seasonal Amplitude	α	{0, 0.025, 0.05, …, 0.975, 1}
Seasonal Offset	φ	{0, 0.5} Months

4.2.3. Inter Event Times

So far in this section we have only considered the cases of simulating monthly counts. When applying the Schuster test we will be simulating daily counts using the same factors detailed in Table 4. It is therefore important that the catalogue is well represented by these daily counts simulations as well as the monthly counts. We do this by comparing the distribution of inter-event times of the simulated data with those from the observed catalogue. The inter-event times are important as they are linked to both the rate of earthquake occurrences and also any aftershock process. As stated in Section 3.4 we use the dispersion parameter to represent an aftershock

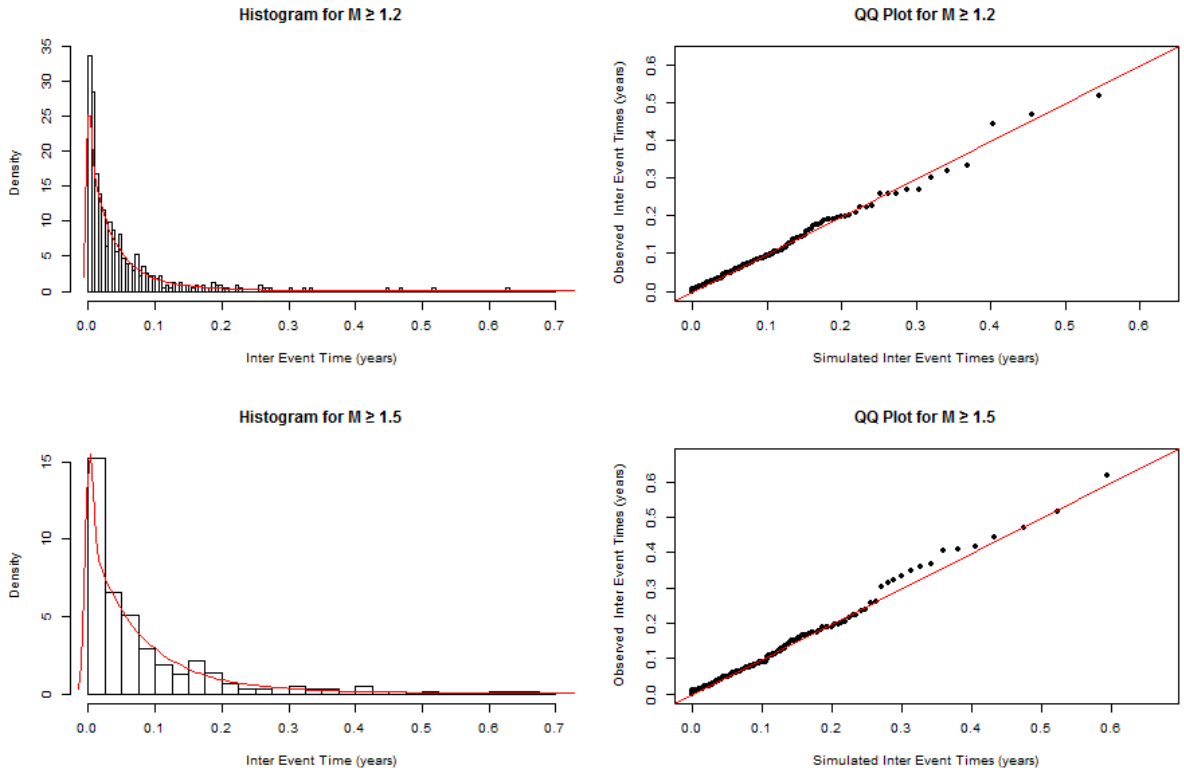

Figure 7: Inter Event times comparisons. The plots on the left show a histogram of the observed inter event times (black boxes) and a density of the simulated inter event times (red lines). The plots on the right show qq-plots comparing the quantiles of the observed and simulated times. The red line shows the expected relationship of Y=X.

process. This is not a physical aftershock process as it does not allow for events from different bins (either months or days) to be correlated. However, comparing the inter-event times from this model with those in the observed catalogue should reveal any significant differences. The results of this comparison are shown in Figure 7. Here the observed inter event times are taken directly from the catalogue without binning. The simulated inter event times are found by simulating daily earthquake counts from the model,

$$\mu(t|\mu_0, \beta, x(t)) = \mu_0[(1-\beta) + \beta x(t)].$$

The parameters used are the mean posterior parameters found in Section 4.2.2. We see that for both choices of minimum magnitude there is a good agreement between the simulated inter event times and the observations. This is an indication that there is a good match between our simulated counts and the observed catalogue in terms of the mean earthquake rate and correlations between events.

5. Simulation Results

In this section we present the results of the simulation study into the detectability of seasonal variations in earthquake occurrences. The factors used for the simulation are those outlined in Table 4. We have a total of 5×2×3×3×6×41×2=44,280 different factor combinations. For each combination of factors, we simulate 1000 time series of event counts, both daily and monthly. Each of the five tests is then applied to each of the 1000 count series, in each case we calculate the proportion of tests which give a positive result. This proportion, for each test and factor combination, is then used in the following sections.

5.1. False Positive Rates

In this section we examine the false positive rate of the different methods. Under ideal conditions we would expect each method to have a false positive rate of 5% since we are evaluating the tests at the 95% significance level. In practice since the data will violate some of the test assumptions there will be some deviation from this value. In our simulation study we can examine the false positive rate by looking at the 1080 factor combination where the amplitude of the seasonal effect, α, is 0. For these cases we would expect all tests to return a positive result in 5% of simulations. The results of this are shown in Figure 8. Looking at this plot we see that there are some deviations from the expected level. In particular the GAM test has a higher rate and, in some cases, the false positive rate is greater than 15%. This indicates that for this test the p-value by itself is not sufficient to judge the presence of seasonality. To apply this test in practice we should also look at other ways to assess the fit of the GAM to the data. Going forward we will omit results from the GAM test as we believe that simply taking the p-value will not give reliable results.

The other test which has a higher than expected false positive rate is the Schuster test. This test is highly influenced by the dispersion parameter, d. For the Poisson case, where the dispersion parameter is equal to 1, the false positive rate is close to the expected level of 5% however for higher values of the dispersion the false positive rate increases dramatically. This issue is already identified by (Ader & Avouac, 2013) who explore simulations from aftershock processes. They suggest that aftershock processes can be distinguished from seasonal processes using the Schuster Spectrum test. Aftershock processes will produce significant results across a range of frequencies whereas seasonality will only produce significant results for a narrow range of frequencies. This does not apply to our single frequency Schuster test and so we cannot distinguish aftershock processes from seasonality without further exploration. The Fourier and parametric hypothesis tests give a similar level of performance with a median rate of just above the expected 5% level and all cases below 10%. The non-parametric hypothesis test has, in general, a lower than expected false positive rate around 4%. This is an indication that this test is perhaps overly conservative.

5.1.1. Disparity Between the Schuster and DFT False Positive Rates

It is interesting to note the difference between the Schuster and DFT tests. The DFT is not sensitive to the dispersion parameter whereas the Schuster test is highly sensitive. The reason for this can be traced back to the test statistics used for each test. In the Schuster test our test statistic is equal to $e^{-D^2/N}$ where N is the total number of earthquakes. For the DFT test our test statistic is $e^{-s/c}$ where c is the mean value of the Fourier periodogram. The Schuster test statistic is therefore linked to the expected number of earthquakes within the time span whereas the DFT test statistic is linked to the variance of the monthly earthquake count series. In this way the DFT test implicitly accounts for the increased variance of the count series whereas the Schuster test cannot make this correction as it is based on the event times rather than counts.

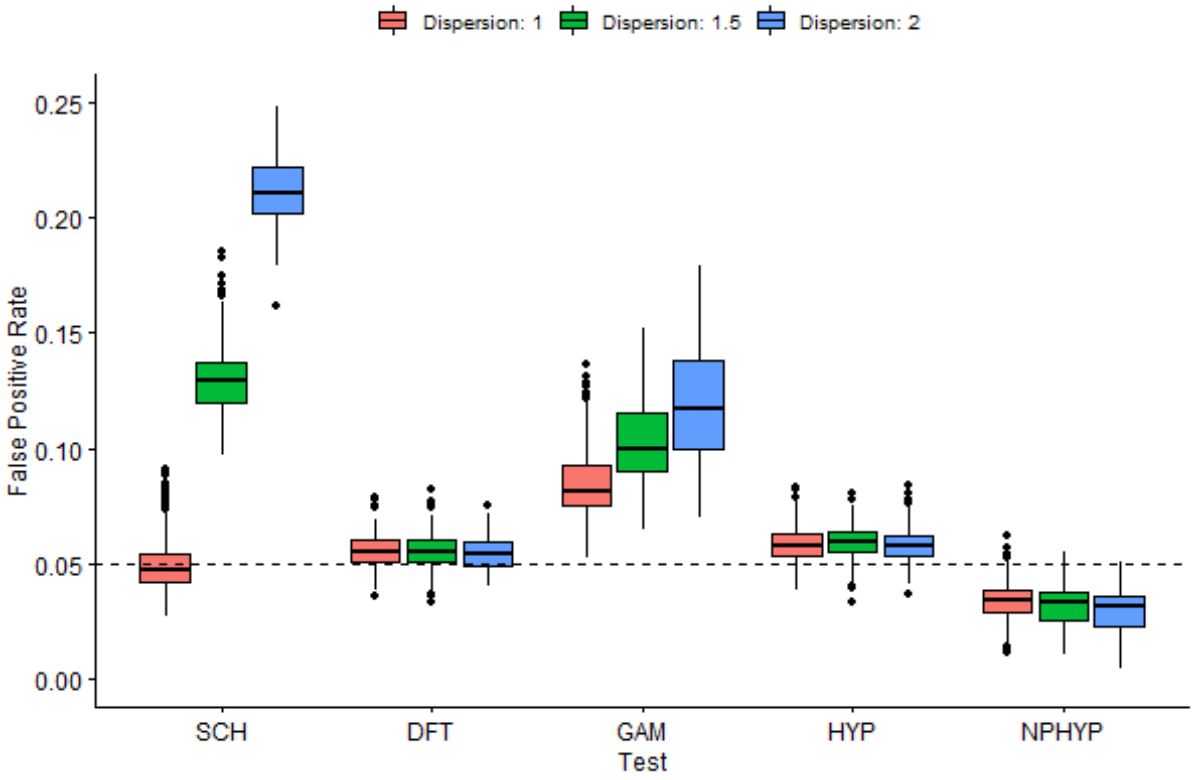

Figure 8: Box plot of false positive rates for each combination of features, broken down by test method. The horizontal dashed line shows the expected 5% level.

5.2. False Negative Rates

In this section we present the results of the 43,200 factor combinations which contain seasonality with an amplitude ranging from $\alpha = 0.025$ to $\alpha = 1$. For each set of factors, we calculate the false negative rate, this is the proportion of simulations where the test incorrectly rejects the null hypothesis. Given the number of different factors we first perform an exploratory analysis to determine which of the factors is the most influential. We do this by performing an ANOVA analysis on the false negative value of each test for the different simulation factor values. This analysis quantifies how much of the variance can be explained by each factor. This is done by fitting a simple linear model for each of the factors separately and looking at the difference between the original sum of squared difference from the mean and that of the residuals of each model. Factors which are highly influential will have a high value to indicate that they explain a large proportion of the variation in false negative rates. The results of this analysis should not be over interpreted as there are several strong assumptions being made such as a linear functional form of the relationships and a lack of interaction terms. It is however useful as an exploratory tool to identify the main influential factors. The results are shown in Table 5 alongside a 95% confidence interval. The confidence intervals have been obtained by bootstrap resampling the simulation results with replacement. The ANOVA results are then recalculated for each resample.

Table 5: Influence on detectability of each factor.

Factor	DFT Variance Explained (%)	HYP Variance Explained (%)	NPHYP Variance Explained (%)	SCH Variance Explained (%)
Seasonality Amplitude	58.5 (57.6, 59.3)	78.5 (77.7, 79.2)	77.0 (76.2, 77.7)	68.3 (67.5, 69.2)
Seasonality Type	32.9 (32.0, 33.7)	7.51 (7.01, 8.02)	5.89 (5.45, 6.34)	23.2 (22.4, 24.0)
Mean Rate	6.96 (329, 382)	11.1 (10.5, 11.7)	14.7 (14.1, 15.4)	7.95 (7.40, 8.54)
Dispersion	1.57 (1.33 1.83)	2.24 (1.96, 2.54)	1.64 (1.41, 1.89)	0.411 (0.295, 0.554)
Seasonality Offset	0.006 (0.000, 0.031)	0.610 (0.471, 0.771)	0.667 (0.523, 0.832)	0.001 (0.000, 0.018)
Trend Multiplier	0.101 (0.048, 0.177)	0.041 (0.012, 0.092)	0.080 (0.037, 0.146)	0.057 (0.021, 0.120)
Trend Type	0.009 (0.000, 0.037)	0.005 (0.000, 0.028)	0.028 (0.006, 0.068)	0.007 (0.000, 0.033)

Looking at the results we see that for all tests the largest influence on the detectability is the amplitude of the seasonal signal. This is as expected but it is encouraging to see that this is by far the largest influence on the likelihood of a positive test result. For both the parametric and non-parametric hypothesis tests we see that the next largest influence comes from the mean rate. As discussed previously we do expect this to have an influence as larger mean count rates will have a distribution which is closer to a Normal distribution and it is less likely to have zero values. It is interesting to note that the effect is similar for both the parametric and non-parametric tests as the non-parametric test does not make the assumption that the data is Normally distributed. The Fourier and Schuster tests shows a different behaviour, these tests are influenced to a lesser, although still significant, extent by the mean rate. The other factor which plays a role in all tests is the dispersion factor. We also see that the dispersion has the smallest influence on the Schuster test. This is perhaps surprising as the dispersion was a major factor in the false positive rate for this test however this points to an interaction term where the dispersion is only influential for certain value of the Seasonality Amplitude.

The Seasonality Offset has an influence in the Parametric and Non-Parametric Hypothesis tests however it has no significant influence on the DFT and Schuster tests. This is not surprising as these two tests rely on a large difference between the highest and lowest monthly counts whereas the DFT and Schuster test look for seasonality through all events. Finally, we see that the trend plays a very small role in the detectability, likely this is in part due to the use of a detrending step although the Schuster test does not include detrending and is still not influenced by the trend.

Figure 9 shows a plot with the median false negative rates for each of the different tests broken down by the three most influential factors. In this plot we can see the relative influence of each factor as described in

Table 5. Here we clearly see that false negative rate decreases with increasing seasonal amplitude and also with increasing mean rate. We can also see from this plot that the Schuster and DFT tests struggle to detect seasonality of the forms of Spline 3 and Spline 5. Figure 9 also shows that for lower values of the seasonal amplitude the Schuster test has a generally lower false negative rate than the DFT test. This result can be explained by the dispersion parameter causing a general increase in the number of positive tests as demonstrated in Section 5.1.

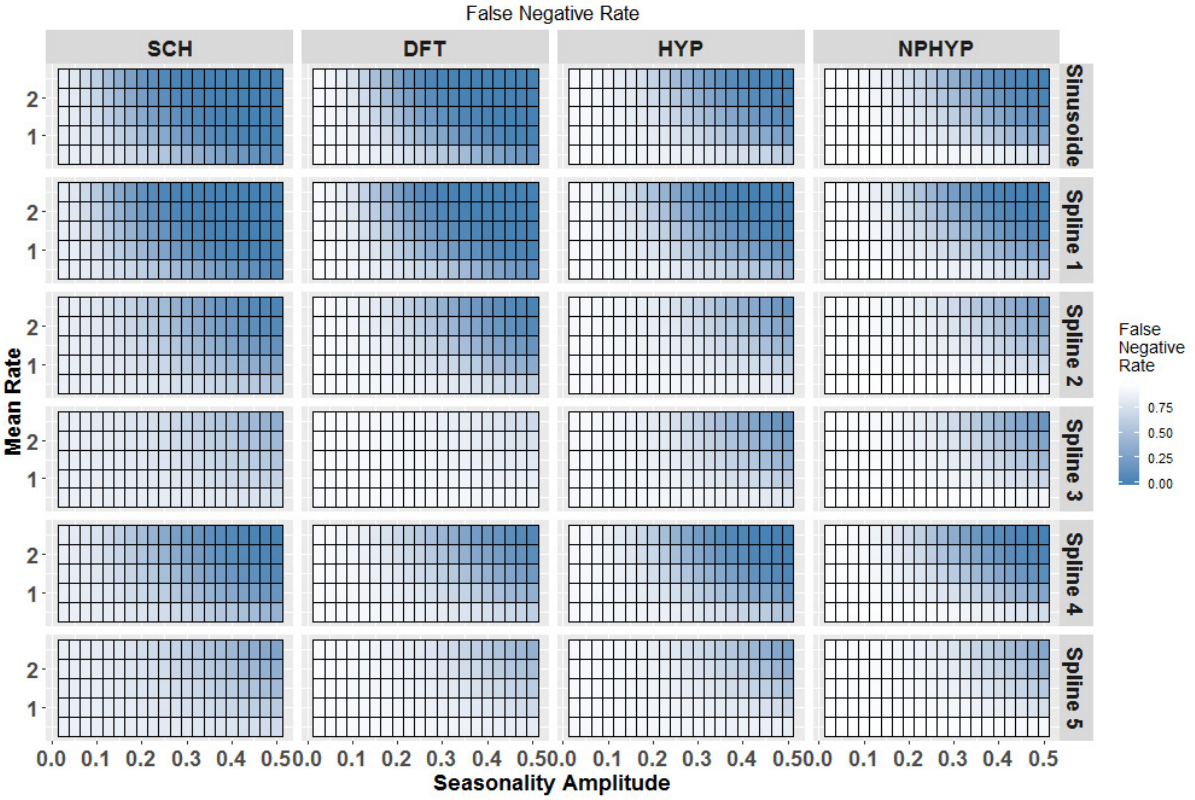

Figure 9: False negative rates for the main influential factors. The colour shows the false negative rates for the simulation runs when testing at the 95% confidence level. The main columns show the different tests, the main rows show the different seasonal functions. The sub columns show the amplitude of the seasonal effect and the sub rows show the mean rate.

5.3. Detectability Threshold

In this section we present the results of estimating the detectability threshold for seasonality. We define this threshold as the minimum value of α needed to achieve a false negative value of less than 20% when applying the tests at the 95% confidence level. In other words, we are estimating the seasonal amplitude where there is an 80% probability that a given test will return a positive result at the 95% confidence level. The value of 20% is chosen as at this level there is a reasonable chance that any true seasonality will be detected. Given that in our previous work we were not able to detect seasonality we can conclude that if seasonality is present then it is likely to be smaller than this threshold. In (Ader & Avouac, 2013) the authors derive a formula for the threshold which is,

$$\alpha_{95} = \frac{2}{\sqrt{N}}\sqrt{2 + ln\left(\frac{T}{L}\right)}.$$

Where T is the time span, L is the period of interest and N is the total count of events. We note that the authors derive this threshold by considering the case that the expected value of their p-

value is equal to 0.05 as opposed to considering false negatives directly. Their approach is analogous to looking at a false negative rate of 50% meaning there is still a 50% probability that seasonality of this level will not give a positive test result. We consider this level to be overly optimistic. Another difference is that this formula is derived for the Schuster Spectrum test which is a multiple hypothesis test, we are only interested in yearly seasonality and so are using single hypothesis tests. It is therefore expected that our results will differ from this threshold. For the Groningen catalogue the Ader and Avouac thresholds are $\alpha_{95} = 27.8\%$ for M≥1.5 and $\alpha_{95} = 20.9\%$ for M≥1.2. We present our results separately for the two minimum magnitudes. For each case we use the results of Section 4 to focus on the simulation runs where the factors are reasonable given the observations.

5.3.1. Threshold for M≥1.2

For this minimum magnitude we restrict ourselves to only runs where the mean rate, μ_0, is in the interval [1.5, 2.5] based on the results in Section 4. For comparison the observed rate above this magnitude is 1.8 events per month. Figure 10 shows a boxplot of the threshold results broken down by test method and by type of seasonality. We can already see from this plot that again the type of seasonality plays a large role here. Spline 1, which has a larger peak and 6-month spacing has a consistently lower threshold. The Sinusoidal shape is also easier to detect for the DFT and Schuster tests. Spline 4, which has a higher peak and 2-month separation is also relatively easy for the two hypothesis tests to detect. Figure 11, Figure 12, Figure 13 and Figure 14 show the results for each test separately. These are broken down by seasonality type and mean rate. Here we see that the mean rate also has a large impact on the threshold for detecting seasonality. If we focus on the easier to detect seasonal functions which are sinusoidal and Spline 1, we see that the Schuster and DFT tests both have a detectability threshold of between 17.5% and 27.5%. The parametric and nonparametric hypothesis tests have threshold in the range of 22.5% and 55% for these functions. If we look across all types of seasonality the upper limits of these ranges increase substantially, up to 100% for the DFT test.

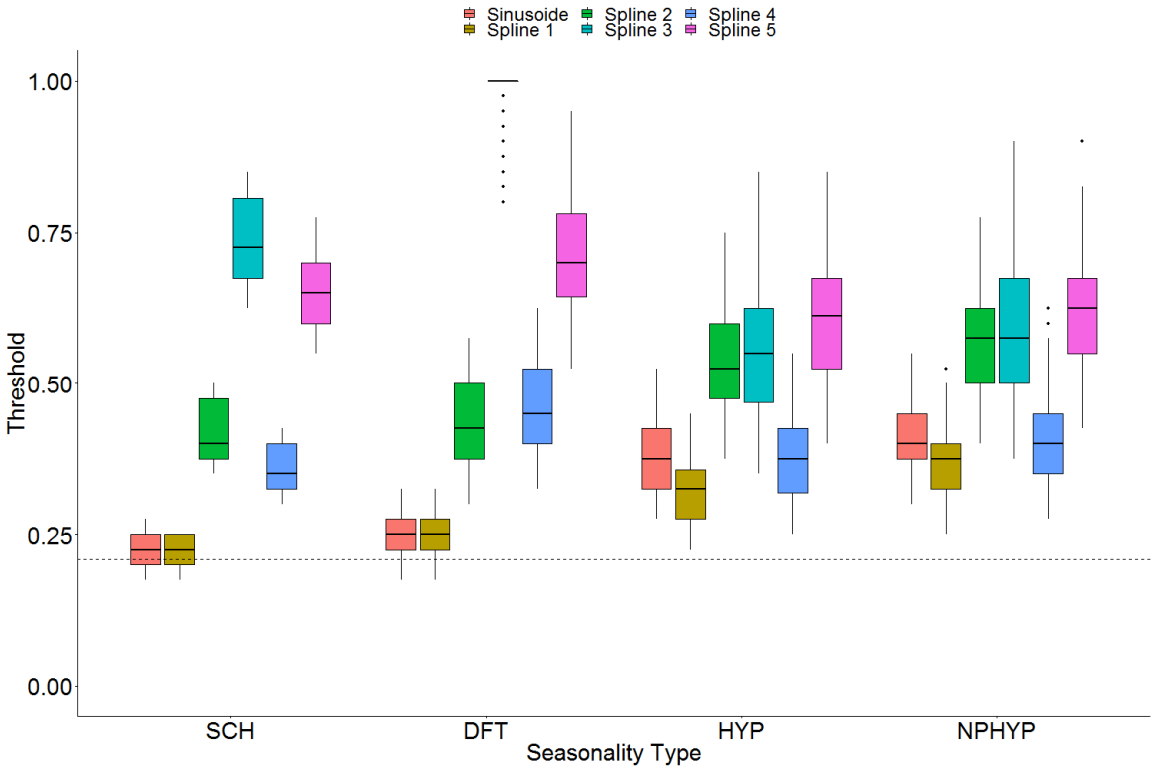

Figure 10: Detectability threshold for M≥1.2 broken down by test method and form of seasonality. The horizontal line shows the Ader and Avouac level of 20.9%.

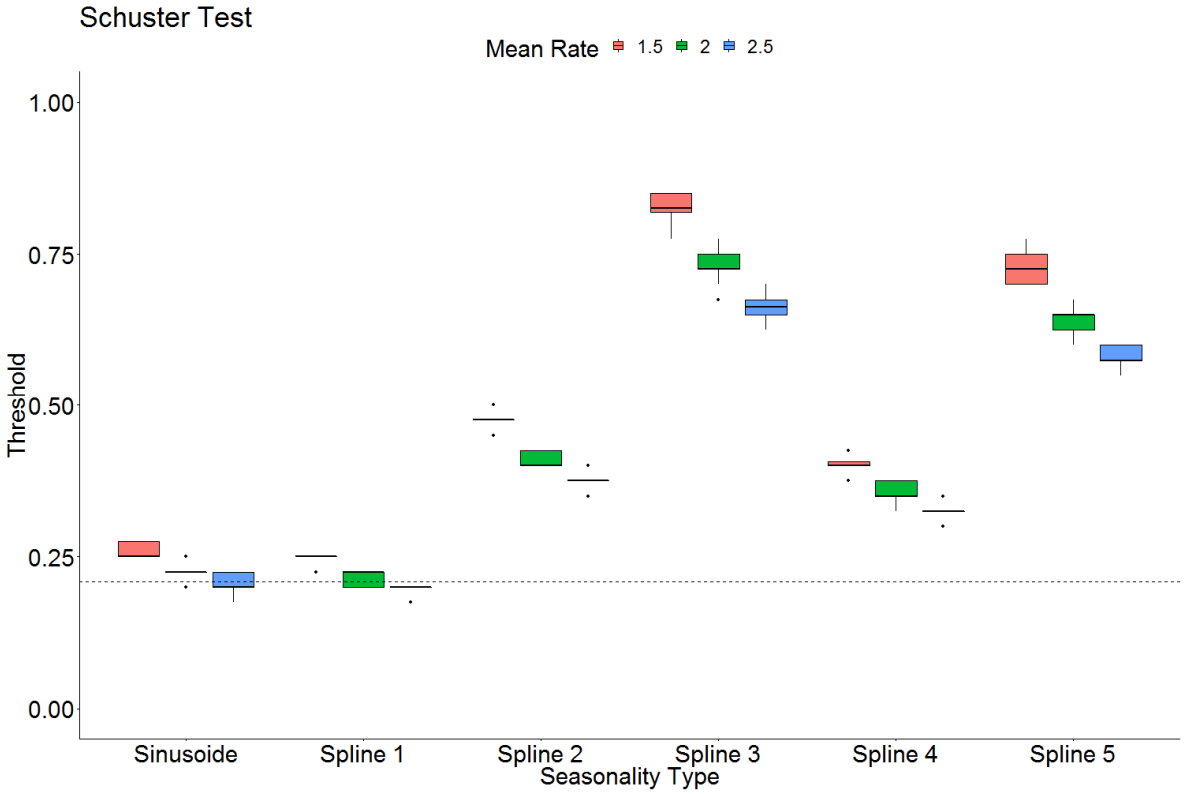

Figure 11: Detectability of the Schuster Test broken down by both seasonality type and mean rate for M≥1.2. The horizontal line shows the Ader and Avouac level of 20.9%.

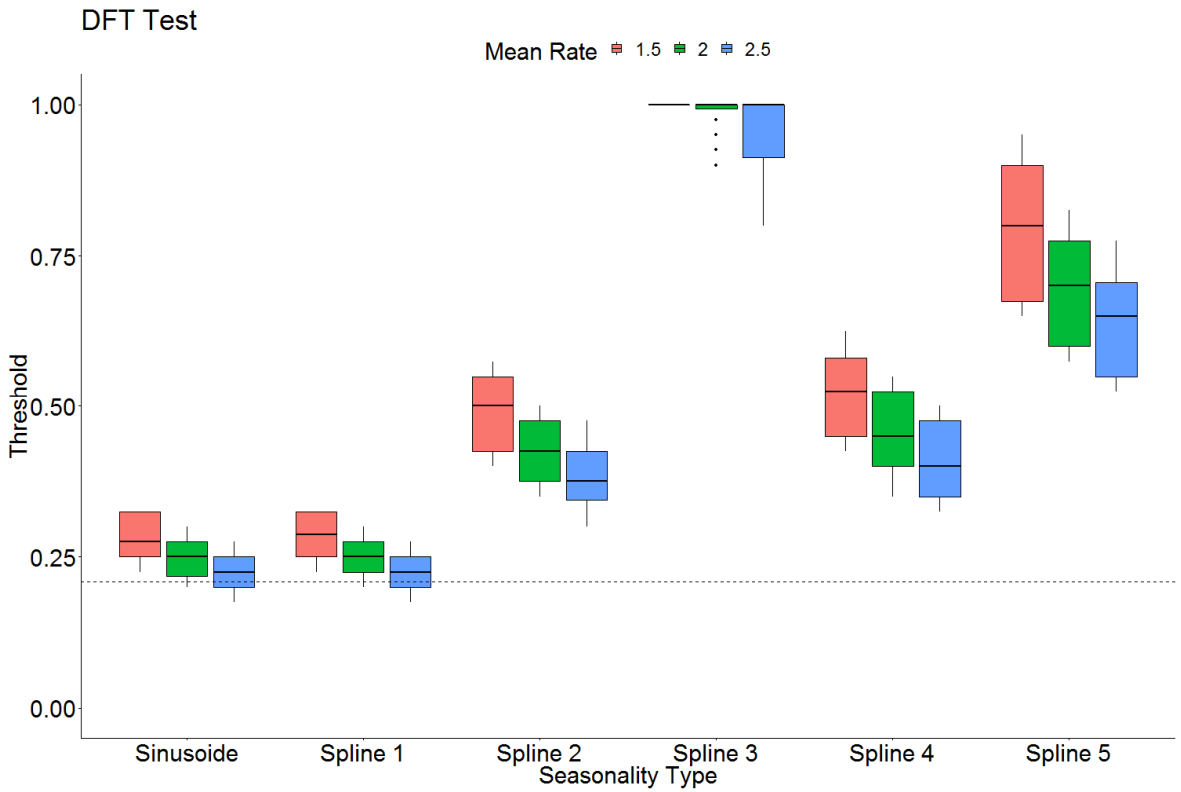

Figure 12: Detectability of the DFT Test broken down by both seasonality type and mean rate for M≥1.2. The horizontal line shows the Ader and Avouac level of 20.9%.

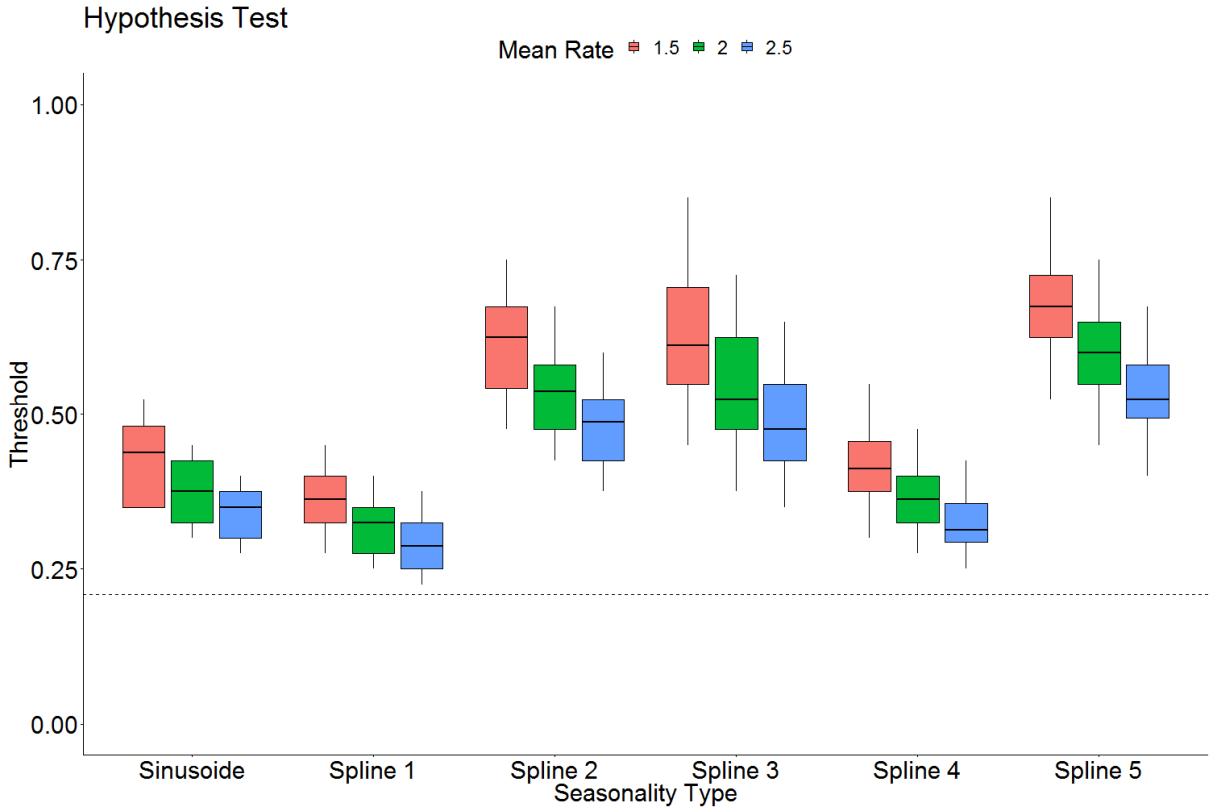

Figure 13: Detectability of the Parametric Hypothesis Test broken down by both seasonality type and mean rate for M≥1.2. The horizontal line shows the Ader and Avouac level of 20.9%.

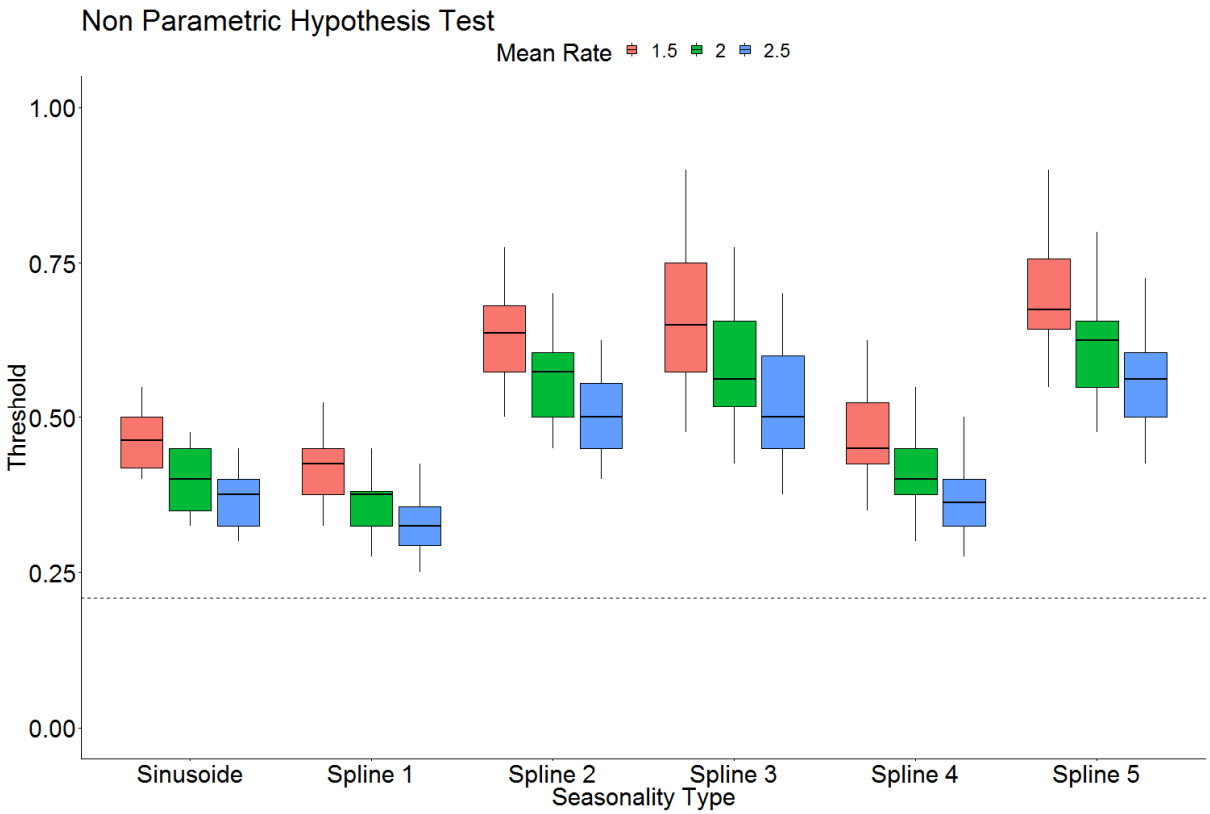

Figure 14: Detectability of the Non Parametric Hypothesis Test broken down by both seasonality type and mean rate for M≥1.2. The horizontal line shows the Ader and Avouac level of 20.9%.

5.3.2. Threshold for M≥1.5

For this minimum magnitude we restrict ourselves to only runs where the mean rate, μ_0, is in the interval [0.5, 1.5], for comparison the observed rate above this magnitude is 1.0 events per month. Figure 15 shows the threshold results broken down by test method and by type of seasonality. From this plot we can already see that the general picture is similar to that for M≥1.2. The main difference is that there is generally more variability in the threshold for a given seasonal function. Figure 16, Figure 17, Figure 18 and

Figure 19 show the results for each test broken down by seasonal function and mean rate. From these plots we see that for the lower rates the detectability threshold is generally higher, this in part explains the additional variability. For this minimum magnitude the Schuster and DFT tests applied to the Sinusoidal and Spline 1 functions, again have the lowest thresholds which range from 22.5% to 55%.

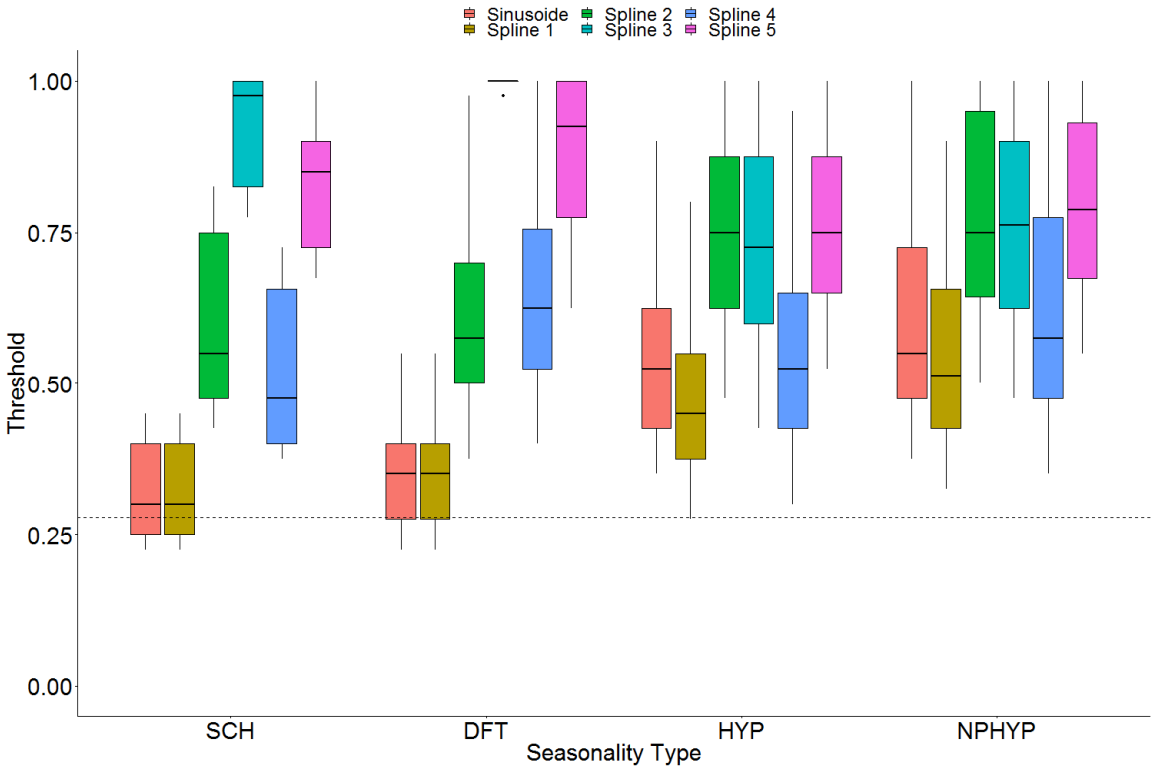

Figure 15: Detectability threshold for M≥1.5 broken down by test method and form of seasonality. The horizontal line shows the Ader and Avouac level of 27.8%.

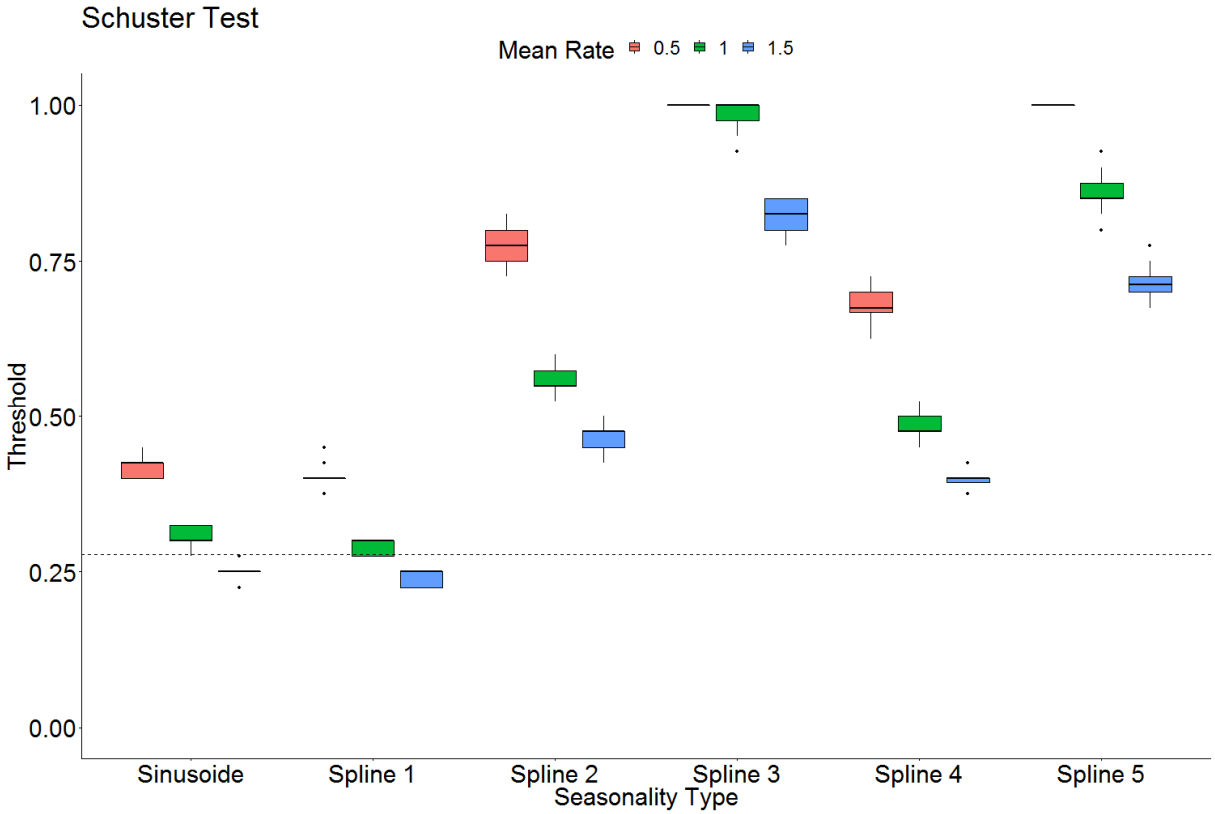

Figure 16: Detectability of the Schuster Test broken down by both seasonality type and mean rate for M≥1.5. The horizontal line shows the Ader and Avouac level of 27.8%.

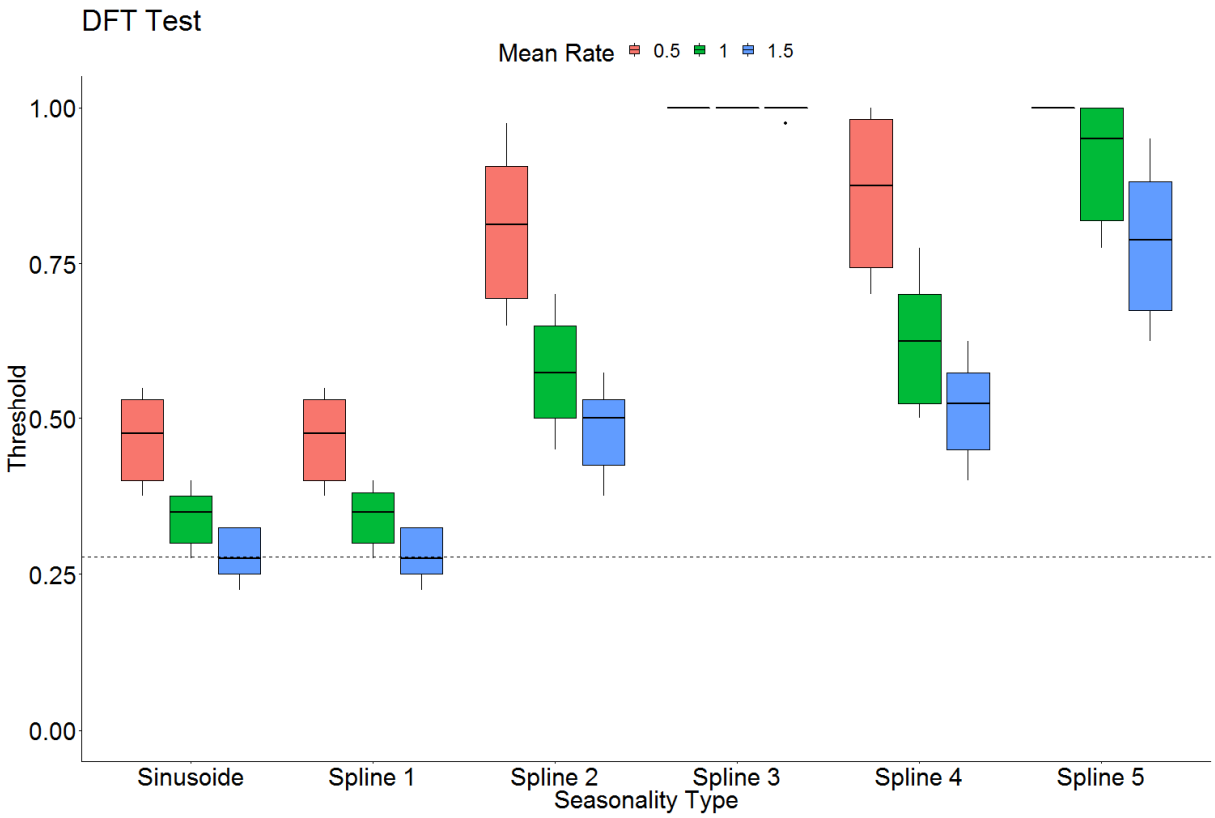

Figure 17: Detectability of the DFT Test broken down by both seasonality type and mean rate for M≥1.5. The horizontal line shows the Ader and Avouac level of 27.8%.

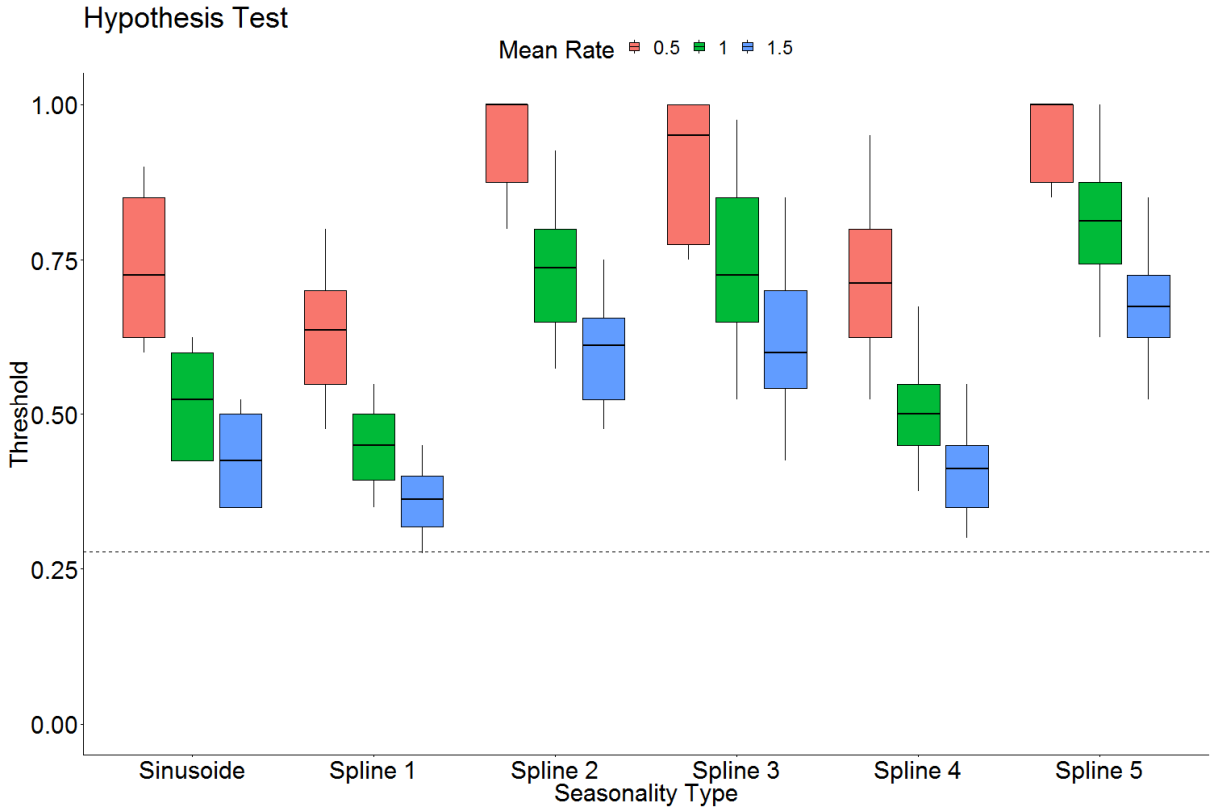

Figure 18: Detectability of the Parametric Hypothesis Test broken down by both seasonality type and mean rate for M≥1.5. The horizontal line shows the Ader and Avouac level of 27.8%.

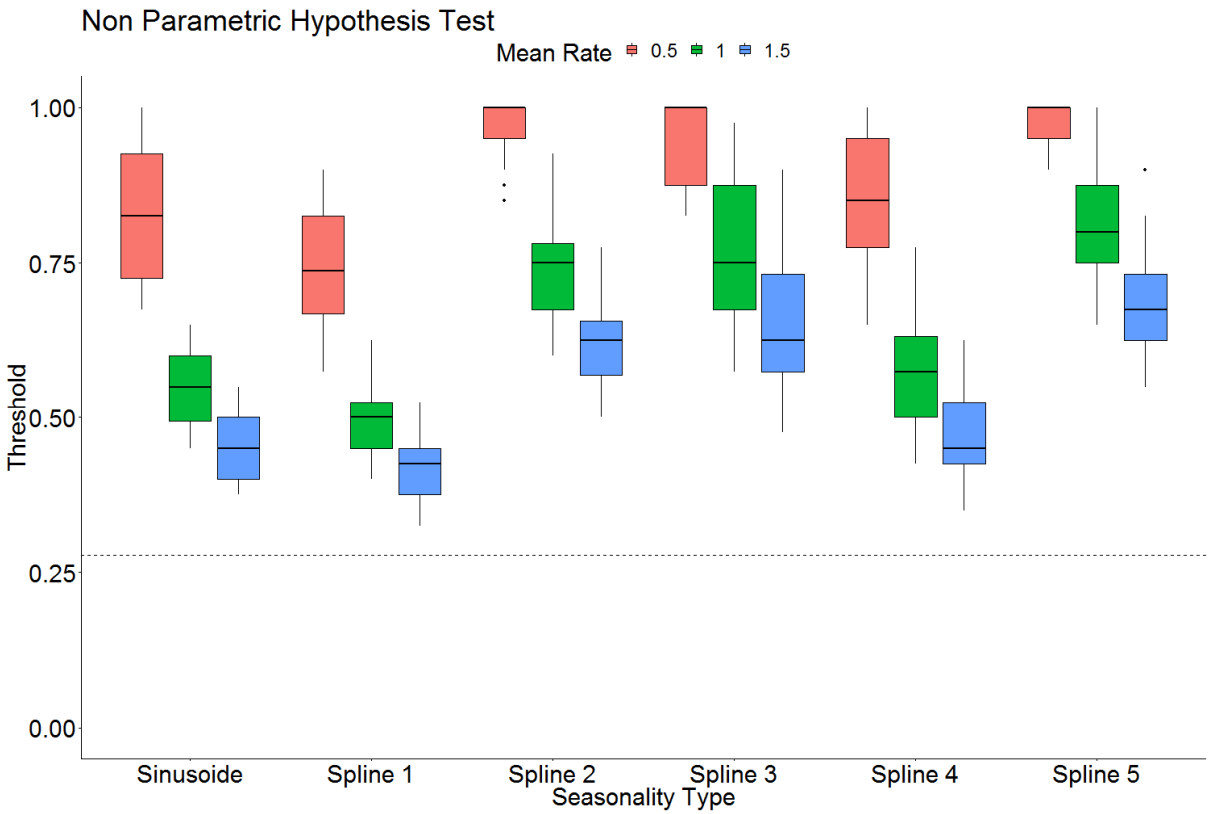

Figure 19: Detectability of the Non Parametric Hypothesis Test broken down by both seasonality type and mean rate for M≥1.5. The horizontal line shows the Ader and Avouac level of 27.8%.

5.3.3. DFT Bin Size

In this report we have applied the DFT test using monthly bins. As detailed in Section 2.3 the DFT must be applied to binned data however there is no restriction on the size of bins. Section 2.3.1 showed the equivalence of the DFT and Schuster tests under certain assumptions, one of these is that the bin size is equal to the minimum recordable time between earthquakes. This raises the possibility of removing the binning assumption from the DFT test by setting the bin size to be very small and so a close approximation of continuous time. We explore this by looking at our default of monthly bins as compared to daily bins. Since this will result in a large number of bins with zero count we do not apply the detrending as this will result in nonzero values for every day. Figure 20 and Figure 21 show the detectability thresholds as calculated for the DFT when either daily or monthly bins are used. These plots are analogous to Figure 10 and Figure 15. Looking at these plots we see that there is very little difference between the two methods for all seasonal functions where the peak and trough are evenly spaced. For the other seasonal functions there is some difference, in particular for Spline 4 which is the most difficult to predict. This shows that changing the bin size from monthly to daily has little effect, in terms of the threshold, for the most physically plausible forms of seasonality we have considered.

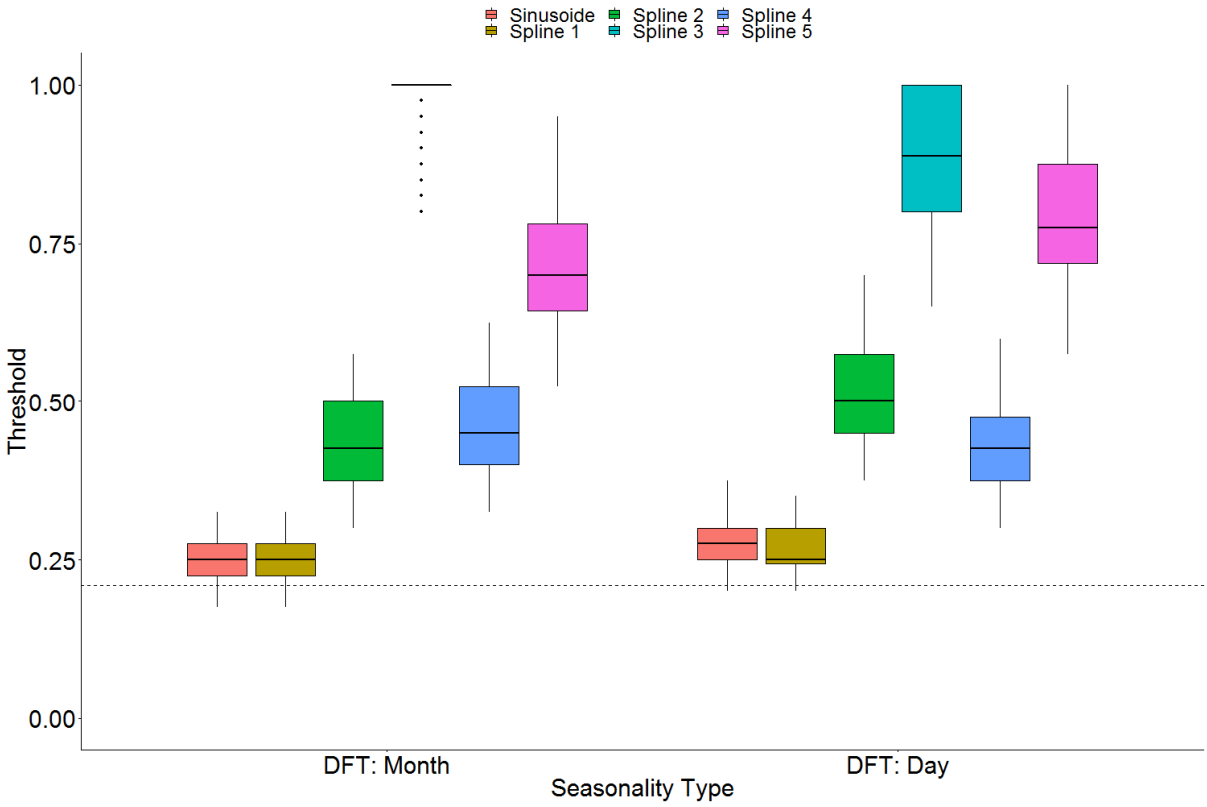

Figure 20: Detectability threshold for M≥1.2 comparing the DFT test used with Monthly and Daily bins.

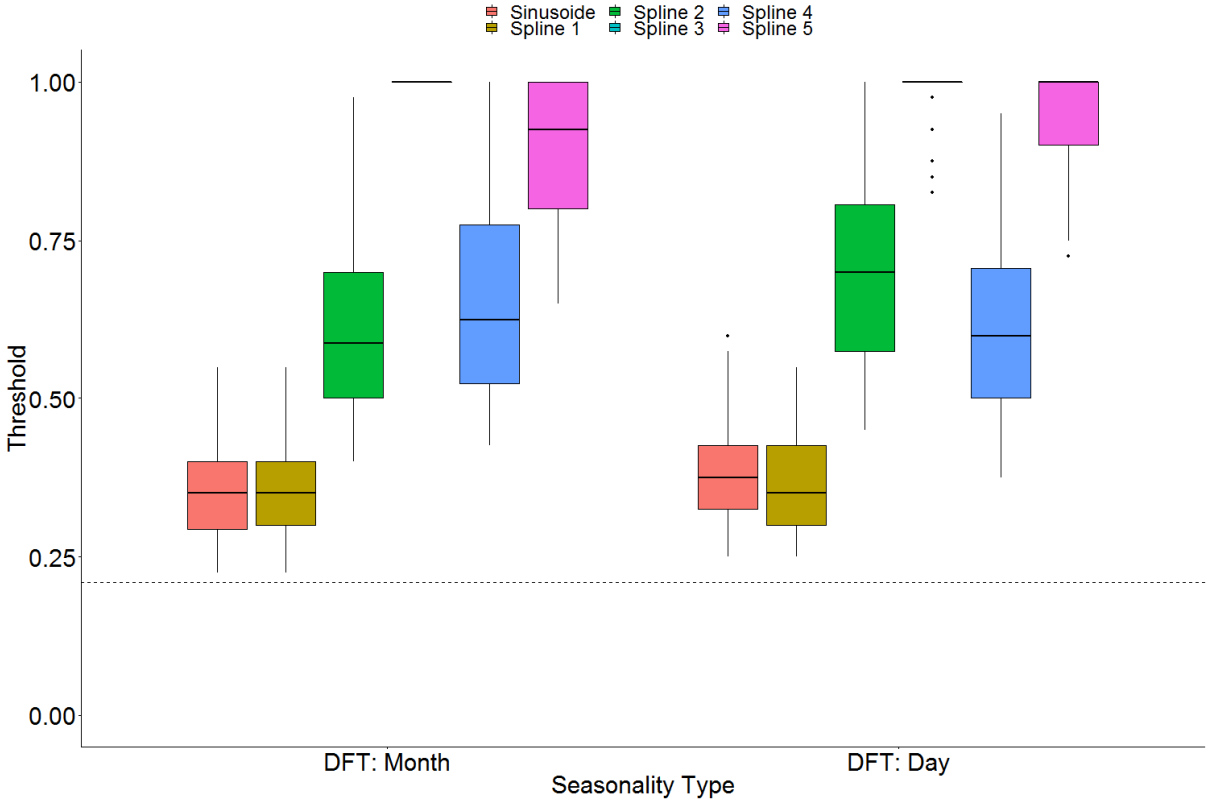

Figure 21: Detectability threshold for M≥1.2 comparing the DFT test used with Monthly and Daily bins.

6. Conclusions & Discussion

In conclusion this report presents the results of a simulation study into the detectability threshold of yearly seasonal variation in earthquake occurrence rates for a range of different testing methods, shown in Table 1, and under a range of different simulation factors, shown in Table 4 . Our main findings are as follows.

6.1. Testing Methods

In addition to describing a range of different test methodologies in Section 2.3.1 we explore the relationship between the Schuster and DFT tests and show that, under certain assumptions, the test statistics D_L and S_k are identical.

6.2. False Positives

We found that the DFT and Parametric Hypothesis tests have a false positive rate which shows good agreement with the expected rate of 5%. The nonparametric hypothesis test was slightly conservative with a false positive rate around 4%. The GAM test shows an elevated false positive rate and so any positive results from this test should be treated with some caution and looked at in the wider context of the fitted model form by looking at additional model comparison diagnostics. The Schuster test shows an elevated false positive rate when the dispersion parameter, *d*, is greater than 1. We therefore recommend steps are taken to ensure that any positive result is due to seasonality and not aftershocks. These steps may include looking at the full Schuster Spectrum or a separate test for aftershocks. The DFT did not suffer from this issue as it takes into account the variance of the earthquake counts when calculating the p-value which makes use of the mean spectrum, c, rather than the number of events, N.

6.3. False Negatives

We found that the probability of detecting seasonal variation in earthquake rates is highly influenced by the magnitude of the seasonal effect. This is perhaps obvious but is still reassuring. The form of the seasonal function played a large role in the detectability. In particular this was an important factor for the DFT and Schuster test which both make assumptions about the form of the seasonality. The parametric and nonparametric hypothesis tests were also influenced by this factor, but this is likely to be linked to differences in the maximum separation in rates between any two months. Another important factor was the rate of earthquake occurrence, this is to be expected as fewer data points generally makes any inference more difficult. The dispersion parameter in the distribution has the next largest effect. This parameter is linked to both the assumption of independence of events and the presence of outliers in the monthly counts. We also found that the amplitude and shape of any trend or an offset in the phase of seasonality did not play a significant role in the detectability.

6.4. Detectability Threshold

Using the detectability results we were able to calculate the minimum detectability threshold for each of the choices of factors for the simulations. We defined this as the minimum amplitude of seasonality where a given test had a false negative rate, when testing at the 95% significance level, of less than 20% of the simulations. We did this separately for choices of minimum magnitude 1.2 and 1.5. For magnitude 1.2 we found that the threshold appears to be in the range 17.5% and 27.5% and for magnitude 1.5 and above we found a threshold in the range of 22.5%-55%. In both cases this is valid under the assumption that the seasonality follows the form of a sinusoid or Spline 1 which has a peak which is larger than the trough. None of the test methods explored are able to test this assumption and so we rely on physical arguments to constrain the expected form of

seasonality. The most difficult to detect seasonal forms tested were those with a short separation between the peak and trough, Spline 3, Spline 4 and Spline 5. There is also less physical justification for these as they would rely on the lag between a change in production rate and a change in earthquake rate being highly dependent on the direction of the change. There is more physical justification for a seasonal form similar to Spline 1 or Spline 2 where the peak and trough are asymmetric. This is particularly relevant for Spline 2 which is generally more difficult to detect and so has a higher detectability threshold.

6.5. General Discussion

Overall in this report we have shown how it is important to consider all assumptions on a test when applying it to real data. This is not always simple as some assumptions, such as in this case trend, can be easily corrected for or simply have little impact when violated. Other assumptions may be very influential only for some tests, such as the independence of events. This highlights the importance of considering the results of the test beyond simply the p-value. This may include additional diagnostic plots or tests, such as testing for aftershocks when using the Schuster test. Related to this we show that the Schuster and DFT tests generally perform the best on the most physically plausible seasonal shapes considered. Each however has drawbacks, the Schuster test relies on a lack of aftershocks and the DFT relies on a good choice of bin size. This therefore opens the possibility of adapting one of these tests to address these weaknesses. The parametric and nonparametric hypothesis tests were more robust to the choice of seasonal function and performed best on the less physically plausible functions.

Finally, we note that the validity of the results of this simulation study are heavily reliant on the choice of factors we have simulated from. While we have motivated these choices based on the assumptions of the test, the observed earthquake catalogue and known physical processes, we acknowledge that this is a limitation of this type of study.

Acknowledgements

The authors are indebted to Jan van Elk (NAM) and Taco den Bezemer (NAM) for their strong and continued support for this study – without their enthusiasm and trust this study would not be here. We are indebted the Review Committee for insightful discussions throughout this study, much of which has been incorporated in this study. In alphabetic order:

- Stijn Bierman (Shell P&T);
- Stephen Bourne (Shell P&T);
- Franz Király (University College London).

The general approach of this study was defined in a Framing Session involving those mentioned above as well as (in alphabetic order) Kevin Bisdom (Shell P&T), Matthew J. Jones (Shell P&T), Jan Limbeck (Shell P&T) and David Randell (Shell P&T) – many thank for the discussions and creativity on which this study was built.

We would like to thank Paul Gelderblom (Shell P&T), Maggie Wenham (Shell P&T), Harry van der Burg (Shell SITI), Mando Rotman (IBM Services), Jonito Douwes Dekker (IBM Services) and Phaedra Kortekaas (IBM Services) for their organizational support in realizing this study.

Bibliography

Ader, T. J. & Avouac, J.-P., 2013. Detecting periodicities and declustering in earthquake catalogs using the Schuster spectrum, application to Himalayan seismicity. *Earth and Planetary Science Letters.*

Bierman, S., 2017. *Seasonal variation in rates of earthquake occurrences in the Groningen field,* s.l.: Shell Global Solutions International.

Bierman, S., Paleja, R. & Jones, M., 2016. *Measuring seasonal variation in rates of earthquake occurrence in the Groningen field - Improved methodology following independent external review,* s.l.: Shell Global Solutions International.

Bourne, S. J. & Oates, S. J., 2017. Extreme Threshold Failures Within a Heterogeneous Elastic Thin Sheet and the Spatial-Temporal Development of Induced Seismicity Within the Groningen Gas Field. *Journal of Geophysical Research: Solid Earth,* Volume 122, pp. 10299-10320.

Bourne, S. J. & Oates, S. J., 2018. *Tentative: the influence of stress rates on induced seismicity rates within the Groningen gas field,* s.l.: NAM Technical Report (draft, under development).

Dost, B., Goutbeek, F., Van Eck, T. & Kraaijpoel, D., 2012. *Monitoring induced seismicity in the North of the Netherlands: status report 2010,* De Bilt: KNMI.

Dost, B. & Haak, H., 2002. *A comprehensive description of the KNMI seismological instrumentation,* De Bilt: KNMI.

Dost, B., Ruigrok, E. & Spetzler, J., 2017. Development of seismicity and probabilistic hazard assessment for the Groningen gas field. *Netherlands Journal of Geosciences,* pp. s235-s245.

Hastie, T. & Tibshirani, R., 1990. *Generalized Additive Models.* s.l.:Chapman and Hall.

Kay, S. M., 1993. *Fundamentals of statistical signal processing, volume I: estimation theory (v.1).* Englewood Cliffs: PTR Prentice-Hall.

Kraaijpoel, D., Caccavale, M., Van Eck, T. & Dost, B., 2015. *PSHA for seismicity induced by gas extraction in the Groningen Field.*
s.l.:http://www.seismo.ethz.ch/en/static/schatzalp/2015/Kraaijpoel.pdf.

Limbeck, J. et al., 2018. *Evaluation of a Machine Learning methodology to forecast induced seismicity event rates within the Groningen Field,* The Netherlands: Nederlandse Aardolie Maatschappij.

Miller, R. G., 1981. *Simultaneous Statistical Interference.* 2nd ed. s.l.:Springer Verlag.

NAM, 2016. *Study and Data Acquisition Plan Induced Seismicity in Groningen, Update Post-Winningsplan 2016,* s.l.: NAM.

NAM, 2017. *Groningen Measurement and Control Protocol (Translated into English from the Original Dutch document),* s.l.: s.n.

Oppenheim, A., Willsky, A. & Nawab, H., 1983. *Signals and Systems.* 2nd ed. s.l.:Pearson Education Limited.

Park, T. et al., 2018. Seasonality analysis for induced seismicity event rate time series within the Groningen Field. *NAM Report.*

R Core Team, 2017. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Schuster, A., 1897. On Lunar and solar periodicities of earthquakes. *Proc. R. Soc. Lond.,* Volume 61, pp. 455-465.

Spetzler, J. & Dost, B., 2017. Hypocentre estimation of induced earthquakes in Groningen. *Geophysical Journal International,* pp. 453-465.

Verhoef, J. M. & Boveng, P. L., 2007. Quasi-poisson versus negative binomial regression: how should. *Ecology*.

Wood, S. N., 2004. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association ,* pp. 99:673-686.

Wood, S. N., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B),* pp. 73(1):3-36.

Bibliographic information

Title	A Simulation Study into the Detectability Threshold for Seasonal Variations in Earthquake Occurrence Rates within the Groningen Field
Author(s)	T. Park (GSNL-PTX/D/S) K. Nevenzeel (IBM Services)
Keywords	Seismicity, earthquakes, event rates, seasonality, statistics, detection thresholds, simulation study, NAM, Groningen
Date of Issue	March 2019
Period of Work	July 2018 to February 2019
US Export Control	Non US - Non Controlled
Issuing Company	Nederlandse Aardolie Maatschappij B.V. Upstream International Schepersmaat 2 9405 TA Assen The Netherlands