NAM

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

Date October 2018

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.

In recent 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 and 6).

This report investigates whether the earthquake catalogue for the Groningen field shows seasonality resulting from the seasonally changing gas production rate using Machine-learning techniques (Ref. 7).

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,

NAM

Title	**Seasonality analysis for induced seismicity event rate time series within the Groningen Field**		**Date**	October 2018
			Initiator	NAM
Autor(s)	**Park T., H. Jamali-Rad, W. Oosterbosch, J. Limbeck, F. Lanz, C. Harris, E. Barbaro, K. Bisdom & 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. In recent 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, modelling of the various mechanisms that could induce production rate dependency and analysis of field data using machine-learning and statistical techniques. This report investigates whether the earthquake catalogue for the Groningen field shows seasonality resulting from the seasonally changing gas production rate using Machine-learning techniques.
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	

Seasonality analysis for induced seismicity event rate time series within the Groningen Field

by

T. Park (GSNL-PTX/D/S)

H. Jamali-Rad (GSNL-PTX/S/IA)

W. Oosterbosch (IBM Services)

J. Limbeck (GSNL-PTX/D/S)

F. Lanz (IBM Services)

C. Harris (SUKEP-UPO/W/T)

E. Barbaro (IBM Services)

K. Bisdom (GSNL-PTX/S/RM)

K. Nevenzeel (IBM Services)

Copyright Nederlandse Aardolie Maatschappij B.V. 2018.

1 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. For example, the seismological model used for PSHRA is a statistical physics model based on elastic mechanisms and Coulomb friction and as such the expected number of seismic events depends on reservoir depletion dP but not on the rate of depletion dP/dt. There are physical mechanisms for aseismic stress relaxation that could, in principle, mean that the expected number of events depends on both dP and dP/dt. Furthermore, explicit incorporation of any seasonality of seismicity might improve seismicity event rate forecast performance. If such performance increases can be established, it might allow statements on a trade-off faced in production planning: volume reduction or reducing the amount of fluctuations. Concretely, the report addresses the following three questions:

1. Do earthquake event rates in the Groningen Field have a measurable seasonal pattern?
2. Can we distinguish forecast performance between models with and without access to seasonal information?
3. Which potential control option, either volume reduction or fluctuation reduction, results in lower seismicity?

Approach:

Potential seasonality of seismicity is investigated with a consensus-based aggregation following a factorial experimental approach, including amongst others earthquake detection magnitude of completeness, epoch, aftershock handling and pressure delay correction to account for propagation of any production changes through the reservoir. Four statistical test methodologies are used: (i) spectral analysis using Discrete Fourier Transform; (ii) seasonal model fitting using Generalized Additive Models; (iii) season-groups Parametric Hypothesis Testing; (iv) season-groups Nonparametric Hypothesis Testing.

To distinguish forecast performance between models with and without a seasonal component as well as to investigate the trade-off between volume reduction or fluctuation reduction machine learning based data driven models are used. This choice is motivated by the desire to minimize the number of physical assumptions made. The data is de-seasonalized by subsampling one data point per year, followed by cubic spline smoothing. Subsequently, performance of models on standard and de-seasonalized data is compared to a baseline and (for selected cases beating the baseline) to each other. This gives a view on the information seasonality can provide to the models. To provide insights in the trade-off between volume reduction or fluctuation reduction, seismology event rate forecasts for production scenarios representing either strategy are compared.

Main findings:

We report the following answers to the questions stated above:

1. The evidence for seasonality in seismic event rate measurements strongly depends on the experimental setup chosen. For magnitude ranges above the concordance magnitude of completeness[1], measured seasonality can likely be ascribed to earthquake occurrence rates. For this range, we find little to no evidence for seasonality. When the magnitude threshold is lowered to include an increasingly large range below the concordance magnitude of completeness, observation bias might play an increasingly large role whilst the statistical power of the tests applied increases. Here, we find increasing evidence for seasonality with decreasing magnitude threshold. Our analysis doesn't allow to say whether this evidence is due to true seasonal variations in the earthquake occurrence rates or observation bias due to including events below the magnitude of completeness. No seasonality was found for the latest epoch considered (2010-2016), which has on average the lowest magnitude of completeness but also has low statistical power due to small sample size. For epochs containing earlier periods, the evidence for seasonality in seismic event rate observations strongly depends on the experimental setup chosen.
2. In general, we cannot conclude that data driven models with access to seasonal information outperform (in terms of forecast errors) models without such access. Looking at the comparisons individually we find some cases where models with access to seasonal information perform better than models which did not, yet in other cases de-seasonalizing the data improves model forecast performance. If however we consider the comparisons collectively and apply multiple testing correction we do not find the results to be significant.
3. Potentially related to the nuanced answer to question 2 above, at a 5% significance level no statistically significant difference between seismicity event rates for the volume reduction or fluctuation reduction production scenario could be found.

Future Work:

Three interesting avenues for future work could be:

1. Assessing the detection threshold of our testing methods for seasonal rate variations would establish an upper bound for the largest possible seasonal rate effect.
2. Another question which we talk about in this report is aftershock detection and removal. There are various methods to identify which events are aftershocks and the results presented in this study may be sensitive to our particular choice. It is therefore useful to conduct a sensitivity analysis of our results under different aftershock identification methods.
3. This study only considered two possible future production scenarios, in principle there are many other choices for a production strategy. We could consider a more extreme range or do a designed experiment to test the strength of different factors.

[1] Various articles written on seasonality in the Groningen field provide different magnitude of completeness estimates but there is consensus that the magnitude of completeness is at most 1.5. Hence all authors agree that choosing $M \geq 1.5$ will avoid observation bias due to incomplete spatial coverage, we therefore call $M_c = 1.5$ the concordance magnitude of completeness M_c^{con}.

Table of Contents

Table of Figures

Table of Tables

2 Introduction

Discovered in 1959 with an initial recoverable reserve estimate of 2900 billion m^3 gas, the Groningen gas field is amongst the largest gas fields in the world (TNO, Geology Service Netherlands, sd). Production commenced by NAM in 1963, by 2015 around 2000 billion m^3 have been produced. The reservoir of the Groningen field is the Upper Rotliegend Group of Early Permian age, consisting of porous sandstone and located at a depth between 2600m and 3200m, with the water zone around 3000m deep. The gas in the reservoir is sealed by a thick impermeable salt and anhydrite layer of the overlying Zechstein Group, as depicted in Figure 1. The Groningen field has several fault systems with around 1500 known faults, whose existence doesn't impact permeability in a significant way.

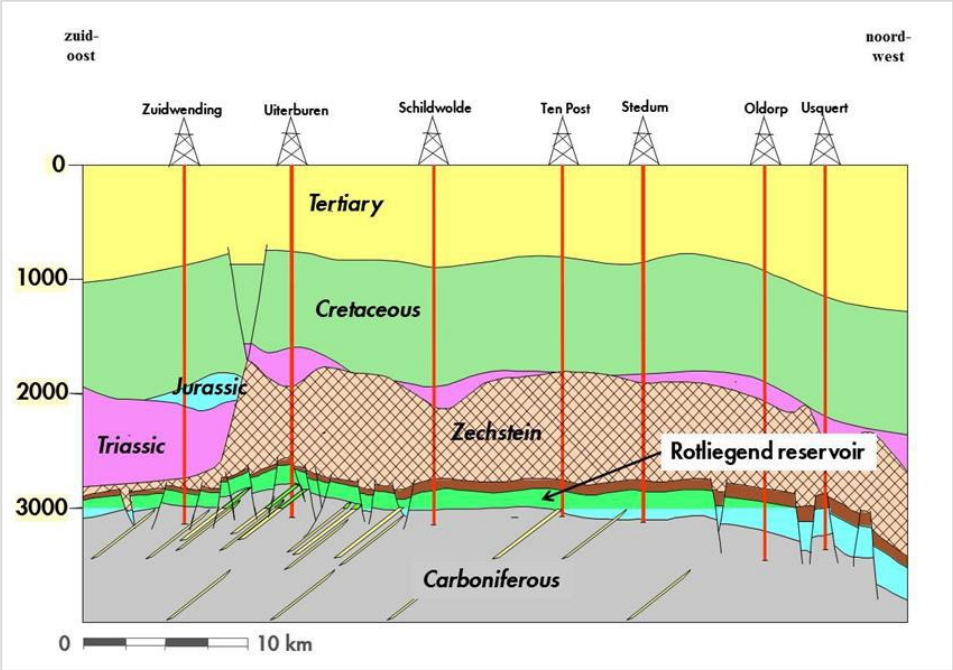

Figure 1: Geological cross-section of the Groningen Field (NAM, 2016)

Following decades of gas production, the historically aseismic region experienced induced earthquakes for the first time in 1991. The frequency and intensity of earthquakes increased steadily to around ten or more earthquakes pear year with a magnitude equal or larger than 1.5 as of 2003, see Figure 6 in Section 3.1. Following an earthquake of magnitude 3.6 on the Richter scale with an epicenter in the village of Huizinge in 2012, a Study and Data Acquisition Plan (NAM, 2016) was put in place to better understand how gas production at reservoir depth affects safety at the surface, and to test the effectiveness of mitigation measures. This led to an integrated Probabilistic Seismic Hazard and Risk Analysis (PSHRA) starting from gas production, sequentially followed by compaction, seismicity, ground motion, exposure, building strengthening and finally risk and safety of inhabitants, see Figure 2.

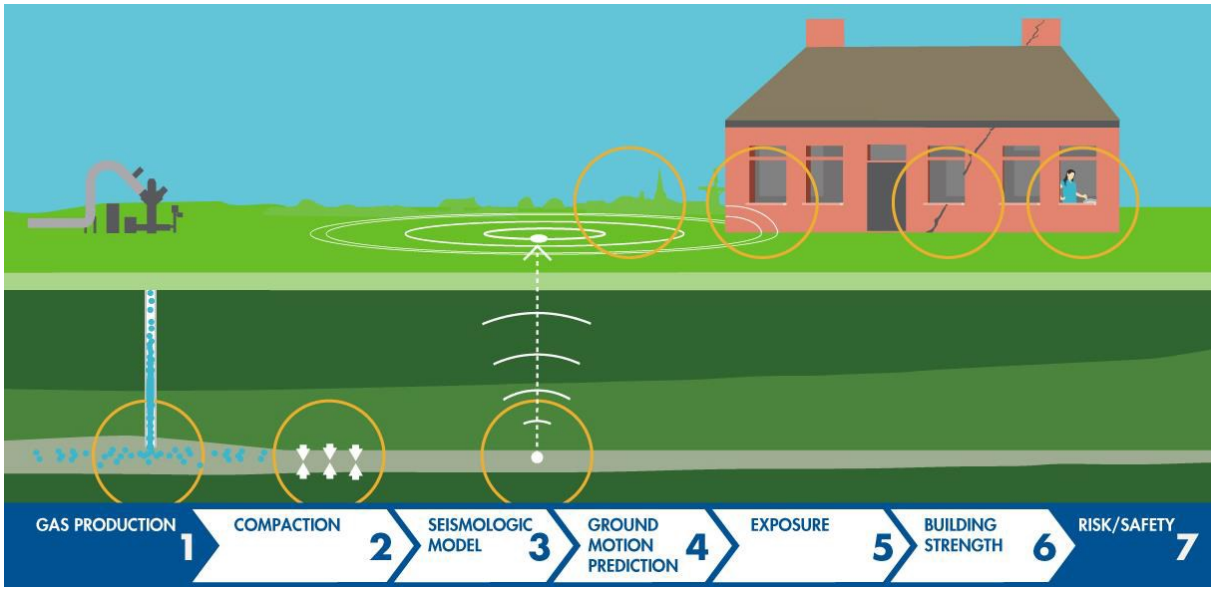

Figure 2: Causal chain from gas production to safety of people in or near a building (NAM, 2016)

PSHRA is realized via an extensive Study and Data Acquisition Plan (NAM, 2016) which encompasses this study in the context of the Measure and Control Protocol (NAM, 2017). This study revolves around the potential propagation of seasonal patterns in gas production (element 1 of PSHRA) to similar patterns in seismicity event rates (element 3 of PSHRA). The Study and Data Acquisition Plan explicitly mentions the assessment of hazard changes due to swing production in chapter 8 (NAM, 2016).

Historically, the amount of gas produced in Groningen showed a strong seasonal pattern as production from the Groningen field is largely demand driven and demand in winter is much larger than demand in summer, see e.g. Figure 10. As production drives seismicity, a natural question is whether a seasonal production pattern results in a seasonal seismicity pattern. Figure 7 shows the seismicity in Groningen aggregated per month, especially for the larger magnitudes a seasonal pattern is not directly evident.

Insights in the seasonality of seismicity matters as it might provide a test for potential aseismic stress relaxation mechanisms present in the field, which in turn might tell whether the frame-rate effect[2] is applicable to Groningen. The NAM seismological model used for the Groningen PSHRA (Bourne & Oates, 2017) assumes an elastic behaviour of the field. SodM [Netherland's State Supervision of the Mines] suggests (Staatstoezicht op de Mijnen, 2016) that aseismic stress relaxation mechanisms potentially play an important role and that consequently, NAM's seismicity estimates are a worst-case scenario. As a final verdict on the importance of aseismic stress relaxation mechanisms of the Groningen field might not be available at this point (Vlek, 2018), a trade-off needs to be made:

[2] Frame-rate seismicity suggests that for a given depletion increment there will be a given number of earthquakes, independently of the production strategy employed. This in analogy to playing a movie: for a given scene a given number of frames will been shown, independently of how fast or slow the movie was played. If Groningen seismicity follows the frame-rate effect, for a given depletion increment different production strategies might change the number of events per unit of time but not the total number of events. Conversely, if the frame-rate effect doesn't hold a production strategy might influence the total number of events for a given depletion increment.

- The total amount of production is minimized to match (strongly varying) demand, which would require (relatively sizable) fluctuations in production. Hence, in case of a non-elastic field this might increase seismicity.
- The amount of fluctuations is minimized to negate the potential effect above. Consequently, production during summer will be higher than demand and the inverse for winter. To ensure sufficient supply during a potentially cold winter, the year-average might need to be higher than would have been the case when the winter turns out to be relatively mild. Hence, a higher production rate is known to lead to higher (short term) seismicity.

This report aims to provide insights in this trade-off by analysing the following questions:

1. Do seismic event rates in the Groningen Field have a measurable seasonal pattern?
2. Can we distinguish forecast performance between models with and without access to seasonal information? [Assuming that under question 1 some degree of seasonal seismic behaviour was identified.]
3. Which of two production scenarios, either leaning towards one of the trade-off choices, results in lower seismicity event rate? [Assuming that under question 2, models with and without seasonal component could be distinguished.]

A sizable body of literature on seismic seasonality in the Groningen field exists, Section 2.1 provides an overview. Building on the insights of earlier work, the factorial consensus based approach to analyse seasonality as taken in this study is detailed in Section 2.2. Subsequently, Section 2.3 briefly outlines the data driven seismicity event rate forecast methodology used to forecast event rates. Section 2.4 outlines how this report will use the forecast methodology to answer the second and third question mentioned above. The temporal and area aggregation used are briefly discussed in Section 2.5. This study was executed in R (R Core Team, 2017) and used the R package MLR (Bischl, et al., 2016) for data driven forecasting.

2.1 Earlier work on induced seismic seasonality in Groningen

Studies searching for relations between seasonal production and seasonal seismicity include (Bierman, Paleja, & Jones, 2015) and (Bierman, Paleja, & Jones, 2016), who consider underlying Poisson and Quasi-Poisson distributions for event counts. Both find strong indications for seasonality for earthquakes with $M \leq 1.0$ with a delay with respect to production of 3-4 months, some indications for seasonality for earthquakes with $1.0 \leq M < 1.5$ (same delay) and no statistically significant evidence for seasonality when $M \geq 1.5$. These conclusions were reached by comparison of two generalized linear models (GLMs) which have as base term an average event rate per month. The first model builds on the base model by adding a smoothed cyclical function of month s, based on penalized regression splines. The second model builds on the base model by adding deviations from the average year effect as a log-linear function of lagged average daily field-wide gas production per month with GLM-coefficient β. To estimate seasonality or a relationship with production four directions were taken: (i) the estimated standard errors of the estimated deviations from the annual average rates (parameters s and β) are used to test whether there is any evidence that they are significantly different from zero; (ii) the relative ability of the models to explain the data is tested using the estimated standard errors of the parameters and the Akaike Information Criterium; (iii) a sampling distribution of the yearly slope β_y average $\bar{\beta}$ is determined; (iv) Pearson's rank correlation coefficient is calculated to determine the association between monthly event counts and monthly lagged gas production.

Further advancements are made in (Bierman S. , 2017) in two ways: (i) in addition to previous methods, Schuster's spectrum method is applied to test for a range of periodicities (such as daily and monthly); (ii) the analysis is split in two epochs: (a) from January 2003 to January 2014 and (b) from January 2014 to January/July 2017 (depending on analysis). Analogously to earlier work, it is

concluded that strong evidence for seasonality exists for earthquakes with $M \leq 1.0$, some for $1.0 \leq M < 1.5$ and none when $M \geq 1.5$. Furthermore, if only data post January 2014 is used no sign of seasonality remains regardless of magnitude.

The correlation between production and seismicity is also investigated by (Nepveu, Van Thienen-Visser, & Sijacic, 2016), who compute the cross-correlation between the change in production and seismicity, both aggregated on a monthly basis, from 2003 to 2012. They find a correlation for all events, for events with $M \geq 1.0$ and for events with $M \geq 1.5$, with the correlation decreasing with increasing minimum magnitude. The statistical significance of the correlations found are not reported. Typical time delays of 5-7 months between production and seismicity are found.

Production leads to changes in reservoir pressure P. Taking into account propagation of pressure changes through the reservoir, (Pijpers, 2016) showed an annual modulation of relative pressure changes $\Delta P/P$ at the locations of seismic events. The effect becomes smaller with increasing minimum magnitude. (Pijpers, 2017) found a correlation between reservoir pressure changes and seismicity event rates with a minimum magnitude of 1, the significance of the correlation is not reported.

2.2 This study: seasonality testing using a consensus-based factorial setup

As is clear from the literature overview above, choices in the experimental setup like epoch, magnitude range, aftershock handling, pressure delay correction and even the test methodology applied might affect (the statistical significance of) the answer. As such, in this study we employ a factorial approach of 320 tests over a wide range of combinations of potential values of the factors mentioned, see Table 1 for an overview.

Factor	Values	Section ref.	# Values
Target	Earthquake count	3.1	1
Minimum magnitude	≥0, ≥1, ≥1.2, ≥1.5	3.1	4
Epoch	[1995, 2003], [2004, 2010], [2011, 2016], [2004, 2016], [1995, 2016]	3.1	5
Aftershock handling	Applied, Not Applied,	4.1	2
Pressure delay correction	Applied, Not Applied	4.1	2
Test methodology	DFT, GLM, PHT, NHT[3]	4.2	4
Nr. of tests			**320**

Table 1: Factorial setup for seasonality testing, resulting in 320 experiments.

[3] These acronyms are defined and explained in Section 4.2.

Each test results in a yes/no answer regarding the statistical significance of seismic seasonality for the specific combination of factor values and a significance level of 5%. The results for different test methodologies are aggregated in consensus-based visualizations to provide an indication of the degree of certainty that seasonal seismic behaviour is present for the chosen combination of factor values. A schematic overview of this approach is shown in Figure 3.

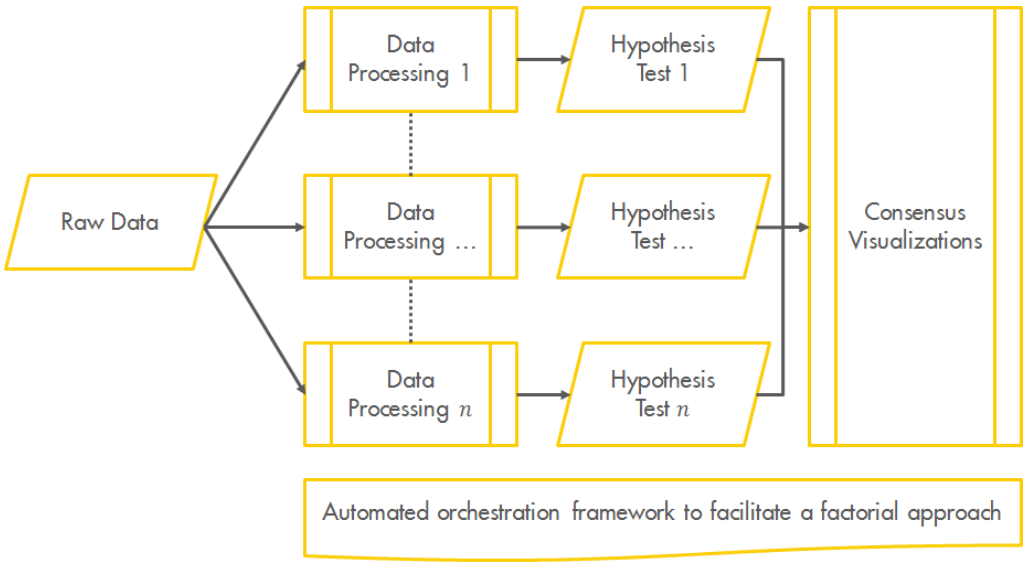

Figure 3: Schematic overview of the consensus-based factorial approach to test for seismic seasonality in the Groningen field.

We note that the 320 test we apply are not independent, this is because there is significant overlap in the data used. For example, the tests applied to the same data are likely to have some degree of correlation. Furthermore, e.g. a specific test applied to the epoch [1995, 2004] will be correlated to the same test applied on the epoch [1995, 2016] as roughly 50% of the data points for the second epoch also appeared in the first epoch. It should therefore be noted that the effective number of independent tests is likely lower than 320.

2.3 Overview of Machine Learning based seismicity event rate framework used

The previous section provided an overview of the consensus based factorial seasonality testing setup used in this study. The outcomes of the setup provide insights into the first question this report aims to answer: does seismicity in the Groningen Field have a seasonal pattern?

To answer the second question (can we distinguish forecast performance between models with and without a seasonal component?) and the third question (which of two production scenarios results in a lower seismicity?), seismicity forecast models are required. Several seismicity event rate forecast models are available, e.g. the default statistical physics based PSHRA model (Bourne & Oates, 2017), an analogous model using numerical 1D fault failure estimations (Dempsey & Suckale, 2017) and a machine learning based data-driven model (Limbeck, et al., 2018). By nature of the approach, the (statistical) physics based models make assumptions on the functional relationship between physical parameters, hence these choices might influence outcomes on the second and third question. Therefore, within this study we use the machine learning based data driven model, as it allows probing of a wide variety of possible linear and non-linear combinations and interaction terms of potential predictor variables (features), without assuming a priori knowledge on the nature of the relationships between the features. Another aspect of the chosen methodology that works to our advantage in the current setup is that its performance on short term timescales (1-3 months) has been benchmarked. We refer to (Breiman, Statistical Modeling: The Two Cultures, 2001) for

an illustrative overview of the foundations behind data driven models and to (Jordan & Mitchell, 2015) for a general overview of the application of such models. We note that the data driven model still partially relies on physics in terms of the input set of potential predictors (features).

The methodology employs a two-step approach: a factorial experimental setup followed by meta-analysis (analysis of the effectiveness of the experimental setup) is used to select robust and relatively well performing models and meta parameters. Important meta parameter choices include a minimum magnitude of 1.2 or 1.5, zero time delay between potentially predicting physical variables and seismicity, as well as the selection of in particular the following machine learning models: Support Vector Machines (SVMs, (Cortes & Vapnik, 1995)), Generalized Linear Model variants (GLMs (Nelder & Wedderburn, 1972), in particular GLM Top and GLM Net, (Friedman, Hastie, & Tibshiranie, 2010)), Random Forests (RFs, (Breiman, Random Forests, 2001)) and K Nearest Neighbours (KNNs, (Hu, Huang, Ke, & Tsai, 2016)).

Using these meta parameters and models for forecasts, (Limbeck, et al., 2018) find that the forecasts of the mentioned machine learning models for the Winningsplan 2016 [Production Plan 2016] (NAM, 2016) are in line with high-level physical expectations. Forecasts for the average production scenario announced by the Ministry of Economic Affairs and Climate in March 2018 (Ministry of Economic Affairs and Climate, 2018) highlight the limitations of validity of the described machine learning based methodology: for future production scenarios which are sufficiently different from past production strategies (e.g. the new average scenario announced by the Minister in March 2018), models without extrapolation capabilities produce forecasts not in line with physics motivated expectations. We note that this limitation is not applicable to the current study, as this study only uses models with extrapolation capabilities when comparing different future production scenarios. Additionally, the methodology and setup described by (Limbeck, et al., 2018) certify forecast performance for short term forecast intervals as required for seasonality testing.

A high-level visualization of the machine learning based framework is shown in Figure 4, step by step:

- **Data sources** are selected and potential predictors (features) are generated from these data sources. The following data sources are incorporated: earthquakes, production, reservoir pressure, faults, compaction and subsidence.
- **Meta-parameters** define the experimental setup within which models are trained and do forecasts. The meta-parameters can be divided in two sets: (i) those related to the prediction target like minimum magnitude; (ii) those describing the experimental setup, like potential time delays.
- **The model evaluation strategy** uses a Wilcoxon signed-rank test based on short term (1-3 months ahead) walk-forward errors in two standard evaluation metrics[4], including the associated standard error estimates. To allow a machine learning model in the final pool of models, additionally a minimum R^2 explanatory power threshold and stability under small changes in meta parameters is imposed.
- **Machine learning model** selection is loosely based on empirical performance studies and includes at least SVMs, GLM variants, RFs and KNNs.
- **Meta-Analysis** is employed on top of factorial runs of experiments to analyse the impact of model and meta-parameter choices on predictive performance. Based on the meta-

[4] The evaluation metrics used are the Mean Absolute Error (MAE), a standard choice in machine learning, and the Root Mean Square Logarithmic Error (RMSLE), particularly useful for count data with a large low-end tail (as is the case here).

analysis robust models with meta-parameter sets are selected for each target. These models are subsequently trained and used for seismicity event rate forecasts.

- **Automated orchestration framework** facilitates the factorial runs of experiments by automating experiment generation, experimental runs and (meta) data collection.

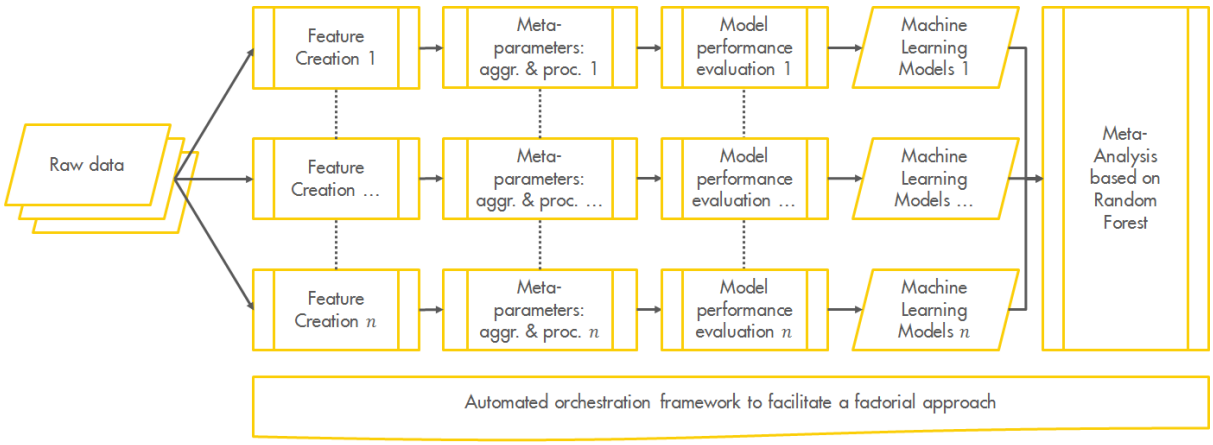

Figure 4: High-level overview of machine learning based forecast methodology used in this study. Reproduced from (Limbeck, et al., 2018).

2.4 This study: seasonality of seismicity event rate forecasts

We use the framework as outlined in the previous section to investigate whether we can distinguish forecast performance between models with and without a seasonal component. Concretely, we do this in three steps:

1. The data is de-seasonalized by subsampling one data point per year, followed by cubic spline smoothing. This to ensure no hidden seasonal signals are picked up by the models.
2. On the seasonal and de-seasonalized data an ensemble of models is trained. If both type of models beat the baseline statistically significantly, both type of models contain useful information and cannot be discounted.
3. We do a direct comparison between seasonal and non-seasonal models and test if seasonal models statistically significantly outperform non-seasonal models. Here, we restrict ourselves to a like-for-like comparison, i.e. we only compare models of the same model type that are built on the same input data, except that one model is trained on seasonal and the other on de-seasonalized data.

These steps give a comprehensive view on the additional information seasonality might provide to the models. Subsequently, the models trained on seasonal data are used to generate forecasts for a reduced volume and a reduced fluctuation production scenario. Comparison of the forecasts might provide insights in which strategy provides the lowest seismicity event rates.

2.5 This study: areal and temporal aggregations used

This study is about gas production induced seismicity on the Groningen field, so the areal aggregation we choose is delineated by the outline of the Groningen Field Outline (GFO), see Figure 5 below for a geographical overview. This region is the same as used in the studies of Shell Statistics and TNO references in section 2.1 – except this study doesn't use a 1000 meter buffer radius around the reservoir as Shell Statistics uses. CBS used multiple smaller circular regions in GFO to analyse the correlation between gas production and seismicity. We could bin GFO in piece-wise constant subregions as well, but this leads to more bins with fewer events per bin, hence decreases statistical power and in all likelihood forecast performance. Given the relatively limited number of events (earthquakes) as will be further outlined in Section 3.1, we restrain ourselves in this study to GFO only. Although this choice follows from practical statistical and machine learning considerations, we note that as the Groningen gas field is considered to be a communicating vessel: differences in dynamic reservoir properties between regions are expected to be limited.

Figure 5: Groningen Field Outline (GFO) geospatial view Google Maps (2018)[5]

Regarding the temporal aggregation: to detect seasonality clearly within-year temporal resolution is required. In case of an ideal sinusoidal seasonal pattern quarterly aggregation might be sufficient, but there is no reason to assume any seasonal pattern is an ideal sinusoid. Therefore, we choose a slightly finer temporal aggregation period of a month.

[5] Map data © GeoBasis-DE/BKG (© 2009) Google. Google Maps image retrieved from: http://maps.googleapis.com/maps/api/staticmap?center=53.5,7&zoom=9&size=640x640&scale=2&maptype=roadmap&language=en-EN&sensor=false

3 Data Sources

To analyse the impact of seasonal production on the potential seasonality of seismicity, two data sources are of obvious importance: earthquake data as described in Section 3.1 and production data as described in Section 3.2. For both data sources we will explore the data, describe the measurements and elaborate on data uncertainties and limitations. This will also allow us to motivate some of the data related choices in factor values of Table 1.

For completeness we note that the machine learning based seismicity event rate model as described in Section 2.3 requires additional data, like dynamic reservoir data, compaction and subsidence. For an overview of that data we refer to (Limbeck, et al., 2018).

3.1 Earthquake Data

3.1.1 Earthquake data exploration

The first earthquake was detected in 1986 nearby the city of Assen. Since then, the number of earthquakes increased both in frequency and intensity, as can be seen in Figure 6. In total, our data set (up to December 2016) contains 1387 earthquakes, of which 973 are within the outline of the Groningen field. Of these, respectively 270, 479 and 634 have a magnitude equal or larger than 1.0, 1.2 and 1.5.

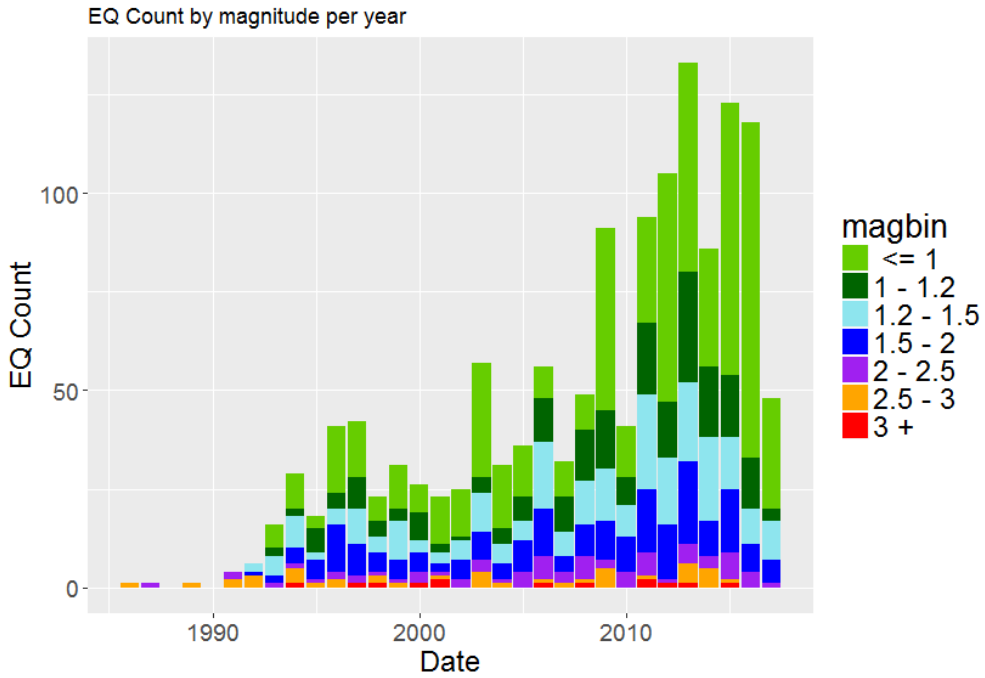

Figure 6: Earthquakes between 1986 and 2016 by magnitude bin

To get an impression of potential seasonality in seismicity, Figure 7 shows the cumulative number of earthquakes per month up to and including the last year which was fully available at the start of this study (2016). On visual inspection, for smaller earthquake magnitudes the months around e.g. February seem to have more earthquakes than the months around September. For larger earthquakes it is more difficult to form a picture.

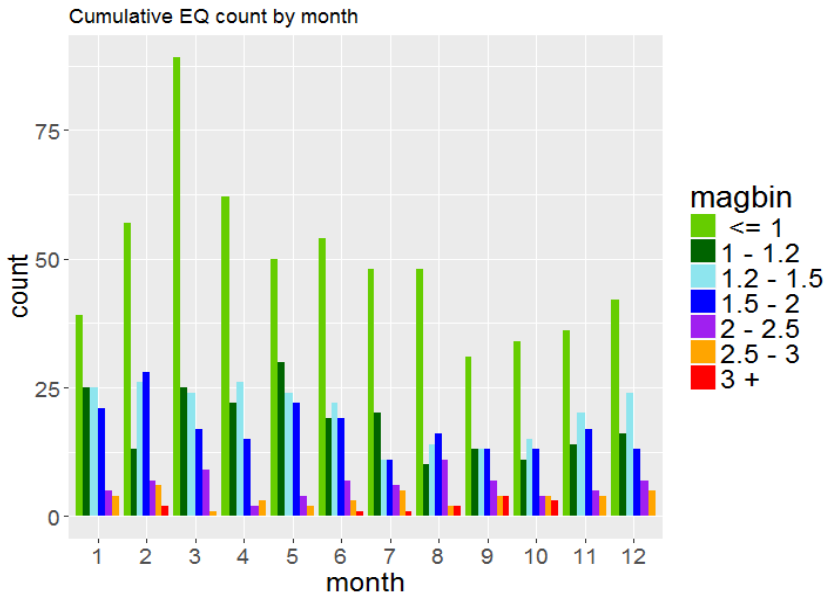

Figure 7: Earthquakes (December 26th, 1986 to December 31st, 2016) aggregated per month.

On a side note we also observe diurnal changes in seismicity: earthquakes seem to occur more at night than during day. (Bierman S. , 2017) observe that diurnal patterns are not immediately apparent for larger magnitudes ($M > 1$) and seem to disappear also for smaller magnitudes post January 2014. They conclude that the most likely explanation of the diurnal fluctuations observed is diurnal fluctuations in the noise floor and therefore detection capabilities of the geophone network. The disappearance of diurnal fluctuations post January 2014 is probably due to improved detection sensitivity of the network.

3.1.2 Earthquake measurements

The KNMI (the Royal Netherlands Meteorological Institute) has seismicity monitoring stations throughout the Netherlands and specifically in Groningen[6]. The network is described in more detail in e.g. (Dost, Goutbeek, Van Eck, & Kraaijpoel, 2012) and (Dost & Haak, A comprehensive description of the KNMI seismological instrumentation, 2002). Measurements from this network are automatically processed by KNMI and earthquakes detected are formally published in a catalogue[7], which we use as source for earthquake detections. The induced earthquake catalogue has a straightforward structure as shown in Table 2. The data is provided in tabular form which each row representing an event. Of each event its date and time, location, latitude, longitude and depth as well as magnitude and evaluation mode are given.

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
…	…	…	…	…	…	…	…

Table 2: KNMI induced earthquake catalogue data structure

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

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

Here most fields are self-explanatory, possibly except for the location field[8] but that field isn't used in our analysis.

3.1.3 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, Goutbeek, Van Eck, & Kraaijpoel, 2012), (Kraaijpoel, Caccavale, Van Eck, & Dost, 2015), (Dost, Ruigrok, & Spetzler, 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.

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

Table 3: KNMI reported Seismic Sensor Network developments over time

3.1.4 Choices of minimum magnitude M_{min} and epoch

When analysing the earthquake data we will not use the entire recorded catalogue but rather only consider those events which are within a certain time epoch and above a certain minimum magnitude, M_{min}. Here we discuss some of the physical considerations behind this choice.

Any measured seasonality in seismic event rates might be attributable to three key causes: (i) seasonality in earthquake occurrence rates, (ii) seasonality in the smallest detectable event magnitude in the presence of seasonal noise variations and (iii) randomness in event occurrence. Our hypothesis testing methods are designed to exclude (iii) as a cause by calculating the probability of observing the measured seasonality under the null hypothesis of no seasonality, this is the p-value. A key factor to distinguish between (i) and (ii) is the magnitude of completeness M_c, usually defined as the lowest value of the moment magnitude of an event for it to be detected with 100% reliability. If seasonality is measured in events above the magnitude of completeness, it can plausibly be ascribed to seasonality in earthquake occurrence rates. If seasonality is measured below the magnitude of completeness it can be caused by either of the two key causes. There are various methods to estimate M_c, we refer to (Mignan & Woessner, 2012) for an overview. Given the improvements in the sensor network over time, the choice of M_c and the start of the temporal epoch T_{start} are coupled: a later T_{start} might allow for a lower M_c and vice versa.

[8] 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.

In literature different authors made various estimates for M_c:

- Following the KNMI reported values (see references in Section 3.1.2), the default PSHRA seismological model (Bourne & Oates, 2017) uses $M_c = 1.5$ from 1995 onwards.
- A probabilistic method based on empirical detection probabilities (Van Thienen-Visser, Sijacic, Van Wees, Kraaijpoel, & Roholl, 2016) leads to the M_c contour plots shown in Figure 8. The plots suggest that for the Groningen field prior to 2010 $M_c = \sim 1.5$, whereas between 2010 and 2014 $M_c = \sim 1.3$.
- Using a Hill-plot (Post, 2017) estimates $M_c = 1.3$ between 1995-2010 and $M_c = 1.1$ between 2010 and 2017.
- Employing both the maximum curvature method (inclined to underestimate M_c) and the b-value stability method (inclined to be conservative) (Limbeck, et al., 2018) estimate $M_c = 1.2$ from 1995 onwards.
- Based on the maximum curvature method (Paleja & Bierman, 2016) estimate $M_c \leq 1.2$ from 2003 onwards and indicate that given the limited number of events prior to 2003, an estimate for M_c is statistically not possible.

We observe that although various authors disagree on the exact value of the magnitude of completeness, there is consensus that the magnitude of completeness is at most 1.5. Hence all authors agree that choosing $M \geq 1.5$ will avoid observation bias due to incomplete spatial coverage, we therefore call $M_c = 1.5$ the *concordance magnitude of completeness* M_c^{con}.

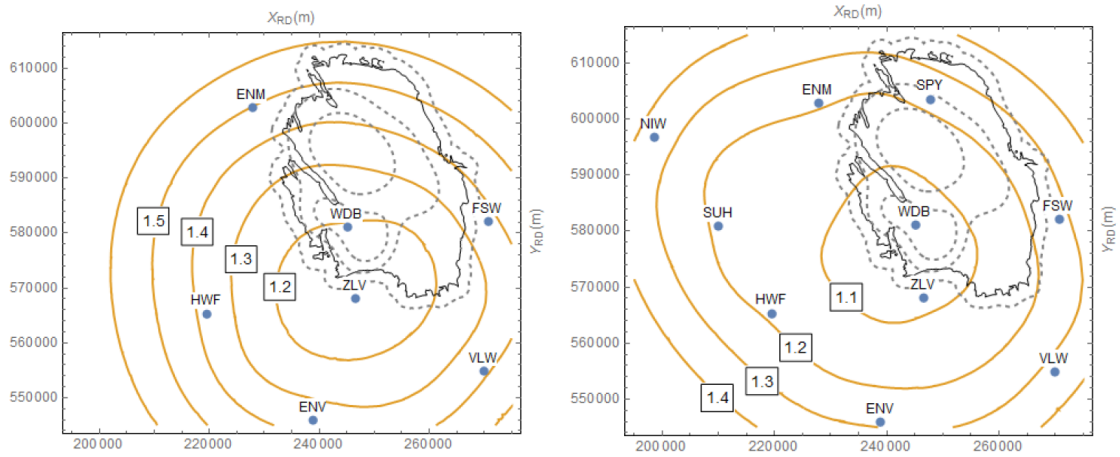

Figure 8: Magnitude of completeness contours for the Groningen borehole network in the period 1996-2010 (left) and 2010-2014 (right) based on a probabilistic model for event detection (Van Thienen-Visser, Sijacic, Van Wees, Kraaijpoel, & Roholl, 2016). For this model the magnitude of completeness is defined as lowest magnitude that has a 95% probability of being detected in 3 or more borehole stations. Figures © TNO.

An important factor in our experimental setup is the choice of minimum magnitude M_{min}, a sensible choice for it is the magnitude of completeness. This choice would ensure that all signal detected comes from seismicity rather than sensor network sensitivity changes. Based on the above, we proceed with the following choices of M_{min}:

- Following the concordance magnitude of completeness, in line with KNMI reported M_c values and following the PSHRA default we choose $M_{min} = 1.5$;
- In line with the M_c estimate of (Limbeck, et al., 2018) and (Paleja & Bierman, 2016) 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};

- For comparative reasons with other reports (e.g. (Pijpers, Interim report: correlations between reservoir pressure and earthquake rate, 2017), (Van Thienen-Visser, et al., 2015) and (Paleja & Bierman, 2016)) we take $M_{min} = 1.0$;
- For completeness, we also analyse nearly all earthquakes, e.g. $M_{min} = 0.0$.

Related to previous, we choose the following epochs:

- The full data set starting from reliable detection in 1995 to the final year available at the start of this work: [1995, 2016];
- The subset starting at the moment M_c can be reliably estimated: [2004, 2016];
- The subset following the sensor network upgrade in 2010: [2011, 2016]
- The complement of previous points splitted at the moment M_c can be reliably determined, giving [1995, 2003] and [2004, 2010].

3.1.5 Choice of target quantity

Three possible target quantities have been generated for this study:

- Earthquake count: the number of earthquakes equal or larger than the minimum magnitude within the temporal and areal intervals;
- Earthquake rate: earthquake count divided by the length of the temporal interval (equivalent with earthquake count for uniform temporal intervals);
- Energy released: let $M_i \geq M_{\min}$ be the magnitude of earthquake *i*, then the energy released is given by: $E_c \sum_i 10^{1.5M_i}$, where the sum is over all earthquakes within a temporal and areal interval and $E_c = 1.259 \times 10^9$ Nm (Sornette & Sornette, 1999).

Although the processing for the methodology developed in this study is largely automated, analysis and results interpretation for a given target quantity remains a time intensive human endeavour so far. As such, for this study we focus on earthquake rate.

3.2 Production Data

3.2.1 Production data exploration

Gas production commenced in 1956 and peaked in the 1970s. Recent years have seen a steep decline in gas production due to production caps. Figure 9 below shows the year to year gas production amounts from 1960 to 2017.

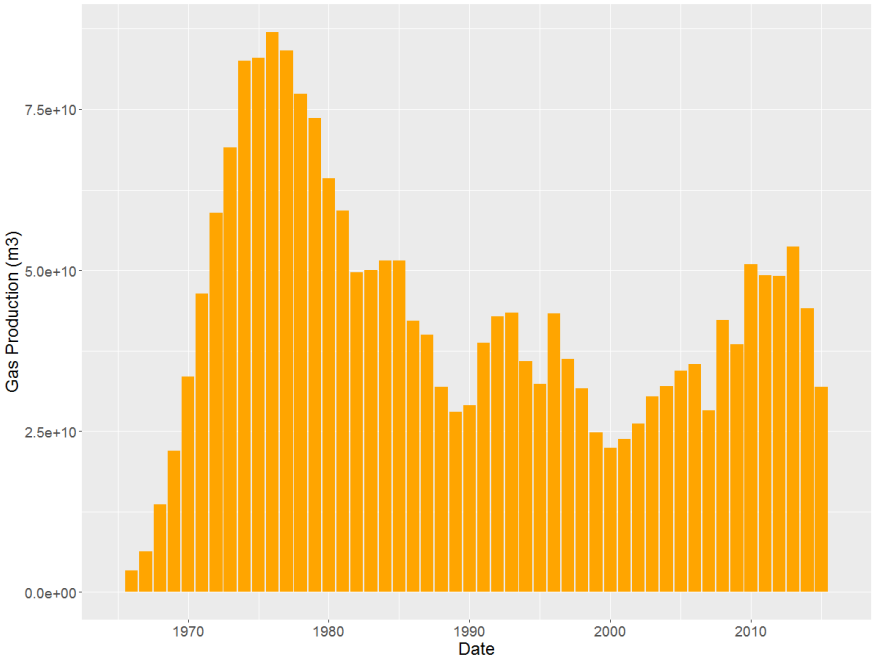

Figure 9: Yearly gas production in the Groningen field, 1960-2017.

A zoom-in on the period under consideration in this study is shown in Figure 10. The month to month variations highlight the historical demand driven seasonal production pattern. In recent years NAM followed alternative production strategies with less seasonal variation. Following the trade-off described in chapter 2, two production scenarios will be compared:

- Reduced volume scenario: a reduction of the production volume w.r.t. the baseline Winningsplan 2016 production scenario to 19.2 bcm, with seasonal swing the same as for the default scenario (± 20%).
- Flat scenario: cancellation of seasonal swing w.r.t. the baseline Winningsplan 2016 production scenario but with the same production volume as for the baseline plan (21.6 bcm).

A zoom-in on the production period used in this study with future production plans as outlined by either scenario is shown in Figure 10.

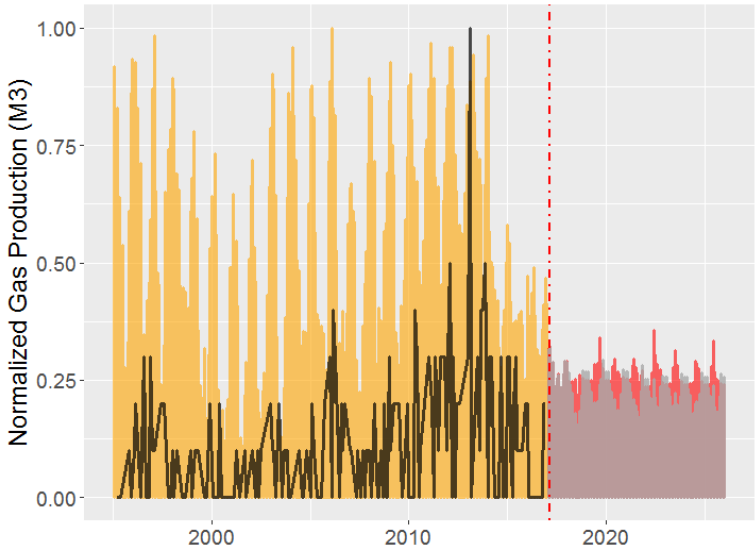

Figure 10: Normalized gas production (yellow) and earthquake rate ($M \geq 1.5$, black) per month in the Groningen field. Left from the red vertical line the historical values (1995-2016), right the production according to the reduced volume production scenario (2017-2025) in grey and the flat production scenario in red.

3.2.2 Production measurements

NAM has 20 production facilities spread over Groningen. Most of these facilities have multiple production wells (around ten to twenty). The total volume produced is measured at well level and aggregated to daily production cluster level. The general structure of the data is shown in Table 4: for each production well (associated to a production cluster) located in a certain region and area for a certain date the amount of gas produced (in m^3) and the amount of water produced (in m^3) are included, just as the BHP and THP in bar.

Well name	Prod. Cluster	Region	Area	Date	Gas (m^3)	Water (m^3)	BHP (bar)	THP (bar)
WAMR1	AMR	East	Central	1956-feb-01	0	0	345.442	0
WAMR1	AMR	East	Central	1956-mar-01	0	0	345.448	0
…	…			…	…	…	…	…
WLRM12	LRM	Loppz	Northwest	2015-nov-1	2.45E6	30.514	87.647	0
…	…			…	…	…	…	…

Table 4: Production data structure

3.2.3 Uncertainties

Of the physical quantities in Table 4, in this study only the gas production is used. There are few uncertainties regarding the historical gas production data since the values for amount of gas extracted are measured at the well level using precise sensors. Future production scenarios depend on policy. While this study was in progress substantial policy changes occurred. By themselves the policy changes do not impact this study, as the key question under consideration with respect to the future production scenarios is the relative impact of volume reduction or fluctuation reduction on seismicity event rates. For the purposes of this study, the two production scenarios are chosen as mere illustrative examples.

4 Detection of Seasonal Patterns

Following the clear seasonal pattern in production (Figure 10), this chapter will assess any evidence for, and the quality of, a seasonal pattern in seismicity. In Chapter 5 we will make us of Machine Learning methods to assess the modelling implication of any seasonality. In this Chapter, however, we will make use of more traditional hypothesis testing techniques to directly quantify the evidence for seasonality.

4.1 Data Pre-Processing

Before testing for seasonality, we perform three pre-processing steps, one step is applied in all experimental setups while two are optional: (i) detrending (applied in all setups); (ii) aftershock handling (optional); (iii) pressure delay correction (optional). The only cases where the detrending is not applied if for the GAM hypothesis test which is applied to the data without detrending. Each of the subsections below discusses one of the pre-processing steps in more detail. Please note that these pre-processing steps are only required by our hypothesis testing methods and are not used in Chapter 5 or Chapter 6.

4.1.1 Detrending

As can be seen in e.g. Figure 11, the rate of earthquake occurrences has increased over time. This upward trend is an important issue as some of the methods we propose for seasonality detection (such as hypothesis testing on grouped months) can be biased by such a trend and return incorrect results. As such, we detrend the data. There are many possible ways to detrend the data. When selecting a method, we are mindful that any seasonal structure in the time series must be preserved. We therefore estimate the trend 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. Figure 11 shows the results of this detrending where both original time-series and the detrended one are plotted.

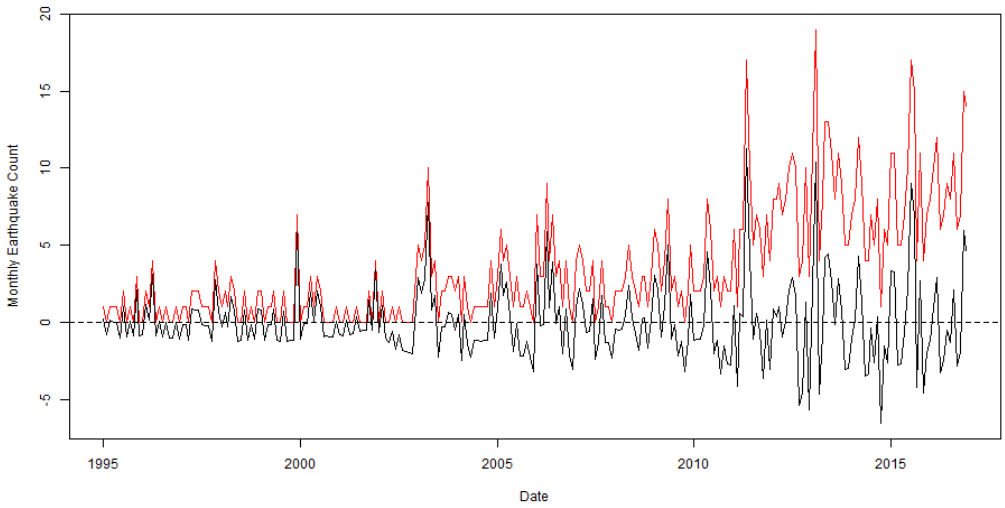

Figure 11: Monthly earthquake count for earthquakes with $M \geq 0$ within GFO from 1996-2016. Red: original count; Black: detrended with Moving Average. As the monthly count can be less than the moving average (minimum is zero), negative counts are observed for the detrended data.

4.1.2 Aftershock handling

Aftershocks are events which occur as a result of earlier events and are therefore not independent. Detecting and modelling aftershocks is a long-standing problem in geophysics. A detailed overview of aftershock removal methods is outside the scope of this study and we identify possible aftershocks via a pragmatic time-space window-based approach discussed in (Gardner & Knopoff, 1974). An alternative approach is via the ETAS model employed in e.g. (Bourne & Oates, 2015) and (Bourne & Oates, 2015). This is a probabilistic approach which assigns each event a probability of being an aftershock. Such an approach has proven successful when modelling events but is less suitable for our pre-processing as it does not give a definite separation between main and aftershock events. The method we use has three parameters: the time window T, the geospatial window D and the minimum main quake magnitude, M_{main}. Let x_0 and x denote respectively the epicenter of a main earthquake and a subsequent earthquake, t_0 and t respectively the time of occurrence of the main earthquake and a subsequent earthquake and M_0 and M the magnitude of the main earthquake and a subsequent earthquake, then a subsequent earthquake is an aftershock of this main quake if:

$$t_0 < t \leq t_0 + T$$
$$d(x, x_0) \leq D$$
$$M < M_0$$
$$M_0 \geq M_{main}$$

where $d(x, x_0)$ denotes the Euclidean distance between two points x and x_0. The values which we also use for our analysis are $T = 5$ days, $D = 5$ km and $M_{main} = 2$. In other words an aftershock should be after a main shock of at least 2 on the scale of Richter in a radius of 5 km around its epicentre and no later than 5 days after the main shock. All events fitting this definition are removed from the catalogue while all others are left unaltered. We note that the choice of parameters, and indeed our choice of aftershock definition, is a limitation as the results may be sensitive to our choices. For example, (Bierman, Paleja, & Jones, 2015) use the stricter definition of 3 days and 2.5km. Given that there is no universally accepted definition of an aftershock event we have chosen a method which is easily interpretable and parameter choices which reasonably reflect physics.

4.1.3 Pressure Delay Correction

Under the hypothesis that any seasonal pattern observed in the earthquake catalogue is caused by seasonal fluctuations in the production it is reasonable to ask if the seasonal pattern is constant across the entire field. We expect there to be a delay between a change in the production rate and a change in the rate of reservoir pressure depletion, as the impact of a rate change in a porous reservoir rock is a diffuse process. This delay will depend on the geology of the reservoir, i.e. permeability anisotropy, but will roughly be larger for locations which are further from production sites. We can account for this by modifying the dates of earthquake events by subtracting the delay corresponding to the location of the event.

The size of the delay is calculated as the time-delay corresponding to the maximum cross-correlation between the pressure at a location and the pressure at a reference location. The reference location is chosen to coincide with the production with the largest seasonal variation in production rate (Bourne S. J., 2018). Figure 45 shows maps of the measured correlations and phase delays for the Groningen field. Note that for some areas of the field the correlation is too weak and so the phase delay cannot be measured. We choose to remove these events when we apply pressure delay correction. This will have the effect of reducing the sample size however we expect

these events to have a very weak or no seasonal pattern which may obscure the pattern that is potentially present in other events.

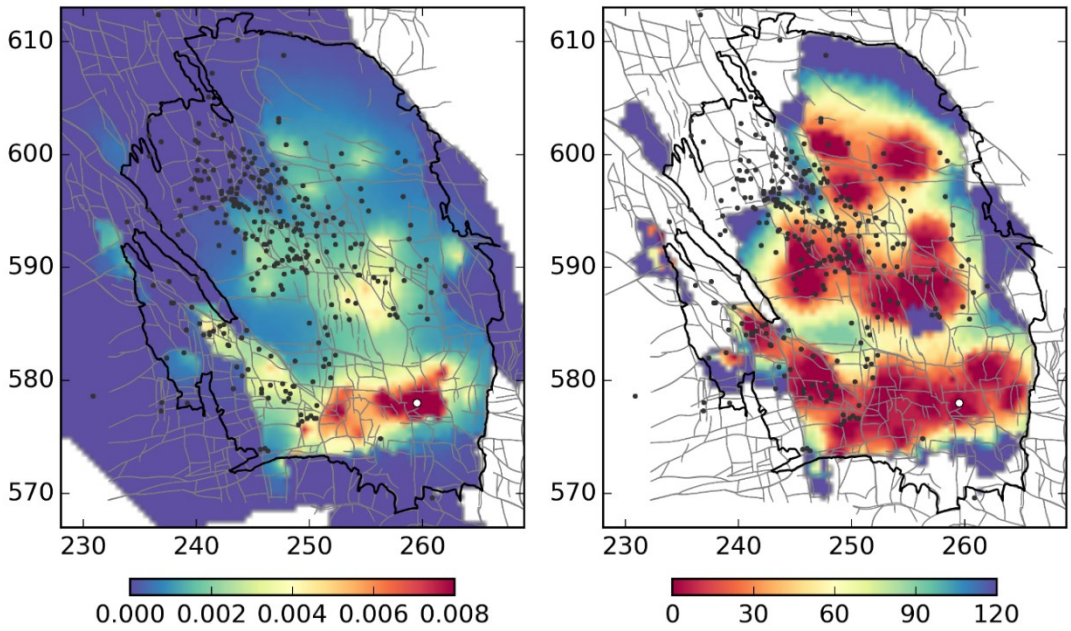

Figure 12: Reservoir Pressure Correlations (left) and Phase Delays (right, in days)

4.2 Seasonality test methodologies

In contrast to the common approach in the literature towards seasonality, where typically a single method is used, we use a consensus based setup based on the following four methods:

- Spectral analysis using Discrete Fourier Transform (DFT);
- Seasonal model fitting using Generalized Additive Models (GAM);
- Season-groups Parametric Hypothesis Testing (PHT);
- Season-groups Nonparametric Hypothesis Testing (NHT).

In our view, these four methods target three different angles to tackle the same problem. More specifically, DFT looks at signal decomposition in the Fourier domain, GAM explores how well our data fits into models with periodic components of interest, and hypothesis testing aims at discovering the differences among data groups (months or groups of months) 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 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 in a general terms they can be expressed as,

- H_0: The rate of seismic event occurrence does not have a repeating pattern of within year variation.
- H_1: The rate of seismic event occurrence has some repeating pattern of within year variation.

Each of the methods is described below.

4.2.1 Spectral analysis using DFT and Schuster Spectrum test.

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:

$$\text{DFT:} \quad X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-i2\pi kn/N}$$

$$\text{Inverse DFT:} \quad x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k \cdot e^{i2\pi kn/N}$$

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 13 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, Willsky, & Nawab, 1983) and (Kay, 1993) for more details.

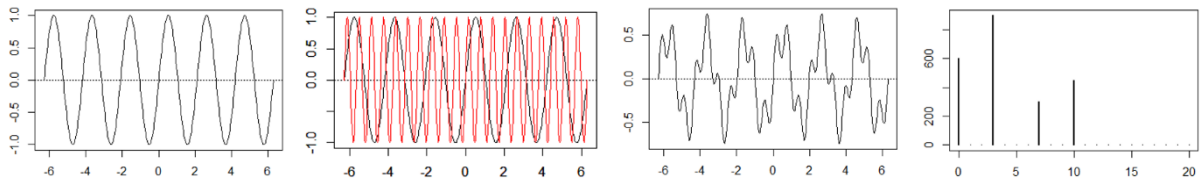

Figure 13: Visual illustration of DFT (right figure) of the time series in the middle-right figure. The time series is a superposition of a higher frequency signal (red, middle-left figure) and a lower frequency signal (left figure).

Our sampling time $T_s = 1$ 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 $T = 12$ months, i.e. yearly seasonal behaviour. We can then use Schuster's test, see for example (Ader & Avouac, 2013), to test for the presence of a significant peak in the periodogram. The null and alternative hypotheses for this test are as follows,

- H_0: The series does not contain a sinusoidal seasonal pattern with frequency 1 year^{-1}.
- H_1: The series has a sinusoidal seasonal pattern with frequency 1 year^{-1}.

Under the null hypothesis it can be shown that,

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

Here c is the expected value of the periodogram, this can be estimated as the sample mean of S_k. 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 assumes that the seismic event rate count, after detrending, can be represented in the Fourier domain as the sum or complex exponentials with independent normally distributed random amplitudes.

4.2.2 Model fitting using Generalized Additive Models (GAMs)

An alternative way to test for the presence of seasonality is by model selection. This is the process of evaluating the fit of models which contain seasonal terms and those which do not. This concept is explored in a more general context in Section 5 however in this section we will concentrate on a simple class of Generalised Additive Model (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. The spline chosen is cyclical to ensure that there is smoothness between all adjacent months including December and January. 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. As described in Section 2.1, this model form is also used in (Bierman, Paleja, & Jones, 2016). Again following (Bierman, Paleja, & Jones, 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 fit this model using 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: $s(i) = 0 \; \forall \; i \in \{1,12\}$.
- H_1: $\exists \; i \; \in \{1,12\}: s(i) \neq 0$.

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. We do this using the p-value estimated by the fitting method and a significant level of 5%.

4.2.3 Comparing Monthly Seismicity

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.

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 (Miller, 1981).

We have chosen to apply the test to individual months, $M = 12$, 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 all cases we apply two different hypothesis tests, a parametric and non-parametric test. For the parametric test we use 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: $C_i = C_j \; \forall \; i, j \in \{1,12\}, i \neq j$.
- H_1: $\exists i, j \in \{1,12\}, i \neq j: C_i \neq C_j$.

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 also use the non-parametric Kruskal-Wallis hypothesis test followed by Dunn's test with Bonferroni p-value adjustment (Lehmann & Romano, 2005), (Miller, 1981). 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 distributional differences between multiple groups. If we define the distribution function of the detrended count in the ith month as $F_i(x)$ then the null and alternative hypotheses for this test are as follows:

- H_0: $F_i(x) = F_j(x) \forall i, j \in \{1,12\}, i \neq j$.
- H_1: $\exists \mathrm{i}, \mathrm{j} \in \{1,12\} \; \mathrm{i} \neq \mathrm{j}: F_i(x) \neq F_j(x)$.

In R the standard implementations of ANOVA with Tukey's HSD and Kruskal-Wallis with Dunn's test were used.

4.3 Experimental Results & Interpretation

In this section we present the results of applying each of the four tests detailed in Section 4.2 to the factorial runs detailed in Table 1. We start by presenting a consensus based overview of the results for all runs and then follow up with a more detailed look at some of the individual runs. The consensus based results are shown in Figure 14 and Figure 15. In Figure 14 the number of positive test results, at the 5% significance level, is summed. The figure shows that the evidence for seasonality strongly depends on the experimental setup chosen. The choice of minimum magnitude M_{min} seems to be amongst the most important factors influencing seasonality detection. When taking M_{min} above the concordance magnitude of completeness $M_c^{cor} = 1.5$, the evidence for seasonality is absent or borderline. When the magnitude threshold is lowered to include an increasingly large range below M_c^{cor}, the evidence for seasonality increases. There is also an effect of epoch on the results: in particular the latest epoch (2010-2016) doesn't show any signs of seasonality, earlier epochs do in various degrees. Aftershock removal leads to overall higher seasonality detection. The effect of delay correction is more subtle: without aftershocks removed it seems to modestly increase seasonality detection, but with aftershocks removed the effect is inversed. Figure 15 shows the minimum p-value for the four hypothesis tests multiplied by 4. This is equivalent to the smallest p-value after Bonferroni correction, (Lehmann & Romano, 2005). Figure 15 shows some of the same trends as the previous plot but gives an overall more nuanced view. We can however see that some of the effects are marginal with only a small shift in p-value causing a large change in the number of positive test results. This again shows how sensitive our conclusions are to the choice of experimental setup. We note that in Figure 15 we only correct for the 4 different hypothesis tests and not the 80 different experimental setups. The 80 different setups arise from different choices of minimum magnitude, time period and pre-processing. Out

of these choices only a few will a priori be relevant to the reader. Additional correction should therefore be applied based on the number of parameter choices seen as relevant. Full Bonferroni correction, i.e. considering all parameter choices as relevant, leads to a significance threshold of $0.05/80 = 6.25\times10^{-4}$. This can be considered conservative as it does not account for the correlation between the tests and assumes all setups are seen as relevant.

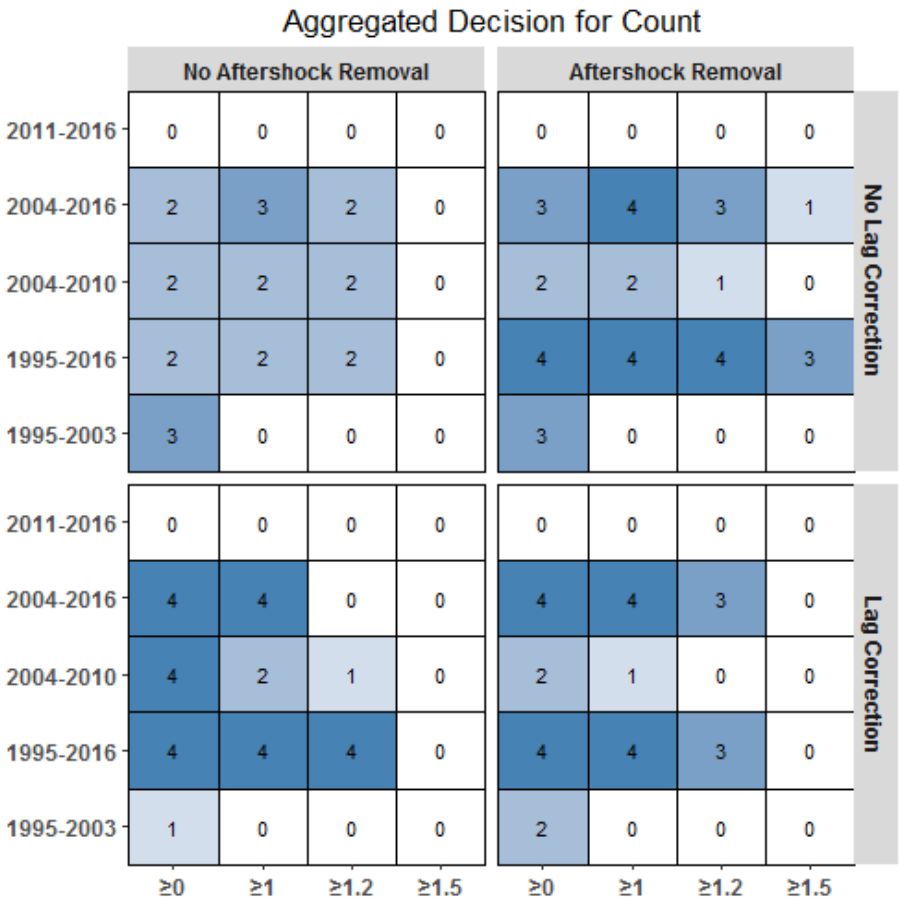

Figure 14: Aggregated Results of Hypothesis Tests. Each square represents one set of data filtering and pre-processing conditions. The numbers and square colouring indicate how many of the four tests rejected the null hypothesis of no seasonality at the 5% significance level.

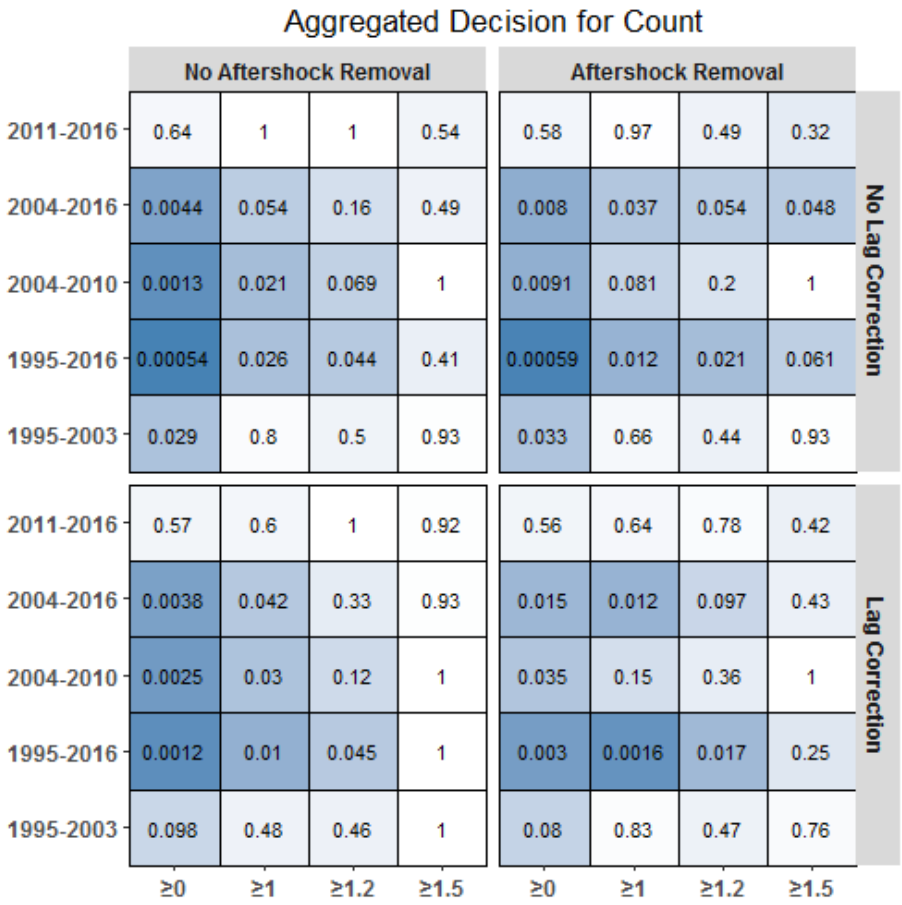

Figure 15: Aggregated Results of Hypothesis Tests. Each square represents one set of data filtering and pre-processing conditions. The numbers and square colouring show the minimum p-value of the 4 tests multiplied by 4 or equivalently the minimum Bonferroni adjusted p-value.

Figure 16 shows the p-values obtained from the DFT hypothesis testing. Looking at the results from the DFT we see that at the level of 0.05 the results are sensitive to the choice of epoch. There are no significant p-values for the period 2011 to 2016 and few significant results for the period 1995 to 2003. The magnitude range also plays an important role with no significant results for $M_{min} = M_c^{con} = 1.5$. Both of these effects may be in part due to the restrictions of either time or magnitude leading to a small number of events. There is no immediately obvious effect of either aftershock removal or pressure delay correction. Looking more closely at the p-values we see that the delay correction has the effect of increasing the p-values, thus showing that correcting for pressure delays in this way reduces the significance of any apparent seasonal patterns. This result is perhaps counter intuitive as we would expect the pressure delay correction to make the seasonal effect more pronounced. A possible explanation for this are that by removing events with unknown delays we have reduced the power of the test, or alternatively that our choice of pressure delay correction procedure is not correct.

We now consider the results from the GAM hypothesis test shown in Figure 17. The GAM hypothesis test gives similar results to the DFT with the choice of both epoch and magnitude range proving important. Indeed, the GAM test gives fewer significant results for higher magnitude ranges. These results seem to back up the findings of (Bierman, Paleja, & Jones, 2016), who concluded there was strong evidence for seasonality in small magnitude events but weaker evidence for events above the magnitude of completeness. For this method aftershock removal does affect

the results however the direction of the effect varies for different time periods and magnitude ranges, similarly delay correction can have an effect but the direction of this effect is not consistent.

Figure 18 and Figure 19 show the p-values obtained from the parametric and non-parametric hypothesis testing. In both methods we see that the choice to remove aftershocks plays a big role with most of the positive results being for runs where aftershocks have been removed. It is interesting to note that delay correction appears to reduce the effect of aftershock removal. Overall there are also far fewer positive results from these hypothesis tests than for either the DFT or GAM.

To put these results in the context of the question at hand, we note there are three key causes of measured seasonality: (i) seasonality in earthquake occurrence rates, (ii) seasonality in the smallest detectable event magnitude in the presence of seasonal noise variations and (iii) randomness in event occurrence. Our hypothesis testing methods are designed to exclude (iii) as a cause by calculating the probability of observing the measured seasonality under the null hypothesis of no seasonality, this is the p-value. An important way to distinguish between (i) and (ii) is limiting the earthquake catalogue to events equal or larger than the magnitude of completeness. The tests applied find very limited evidence for seasonality above the concordance magnitude of completeness M_c^{con}, hence there is at most borderline evidence for seasonality in earthquake occurrence rates. Lowering of the magnitude threshold below M_c^{con} increases the observation bias, the statistical power and the evidence for seasonality. Our analysis doesn't allow to say whether this evidence is due to true seasonal variations in the earthquake occurrence rates or observation bias due to including events below the magnitude of completeness. We note that the evidence for seasonality also depends on the other experimental factor values chosen. The sensitivity to the choice of setup combined with the simultaneous increase in observation bias and statistical power raises questions about the robustness of our findings. We therefore look more closely at some of the individual results.

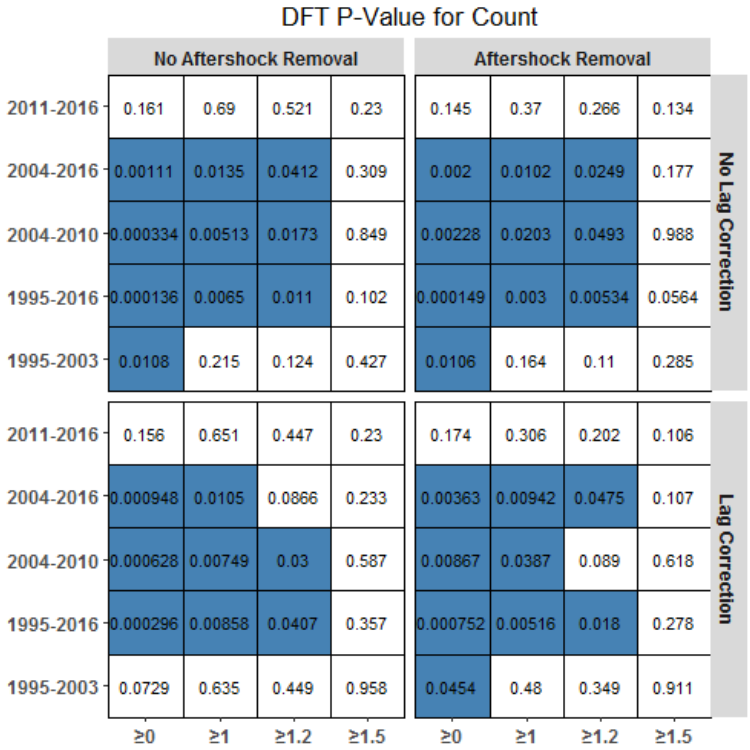

Figure 16: P-Values for DFT hypothesis test, each square represents one set of conditions. Blue squares indicate p-values ≤0.05.

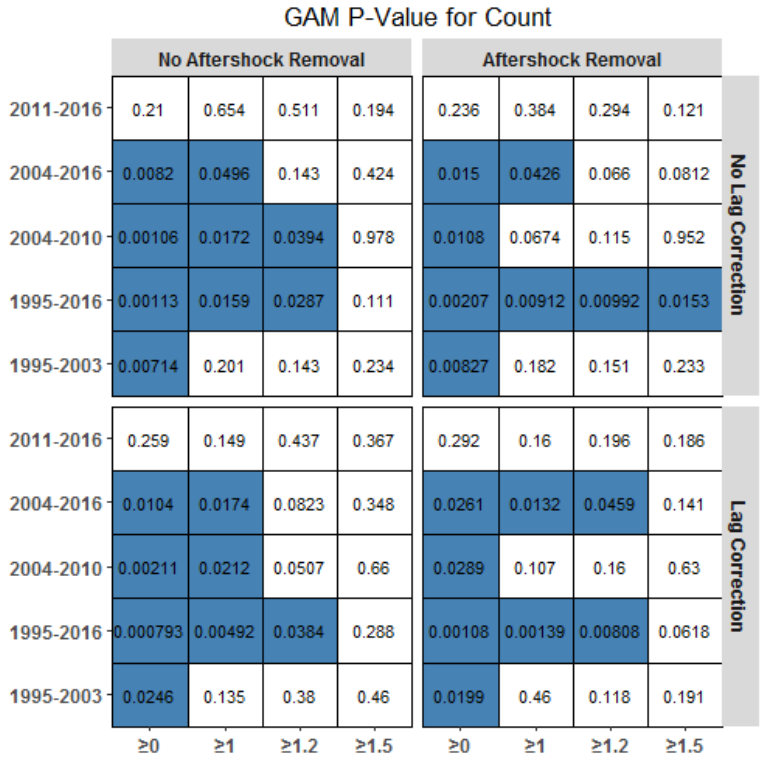

Figure 17: P-values GAM hypothesis test, each square represents one set of conditions. Blue squares indicate p-values ≤0.05.

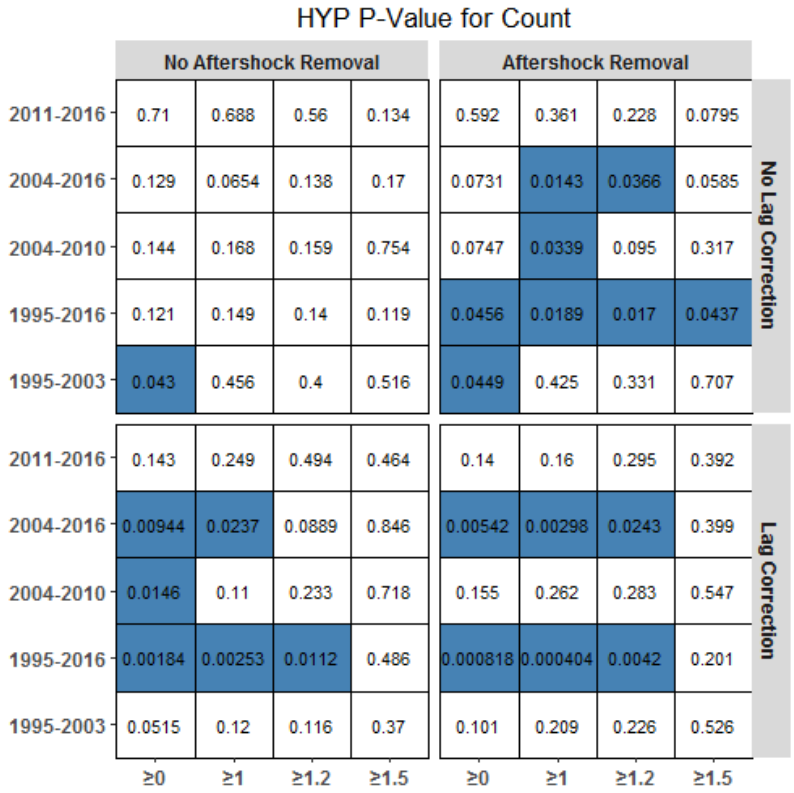

Figure 18: P-Value for the Parametric Hypothesis Test.

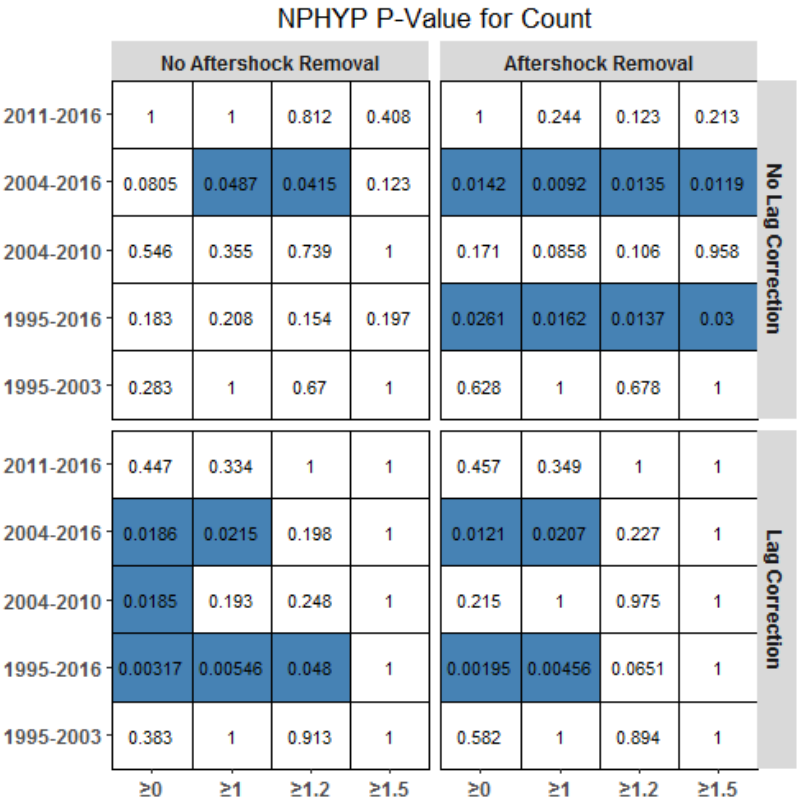

Figure 19: P-values for the Non-Parametric Hypothesis Tests.

4.3.1 Detailed Analysis 1

We now take a more detailed look at some of the runs. The first scenario is for the magnitude range ≥1.2 and the time period 2004-2016, without either aftershock removal or pressure delay correction. This scenario was chosen as for this time period and magnitude range the minimum magnitude can be determined with statistical reliability and the catalogue might be complete.

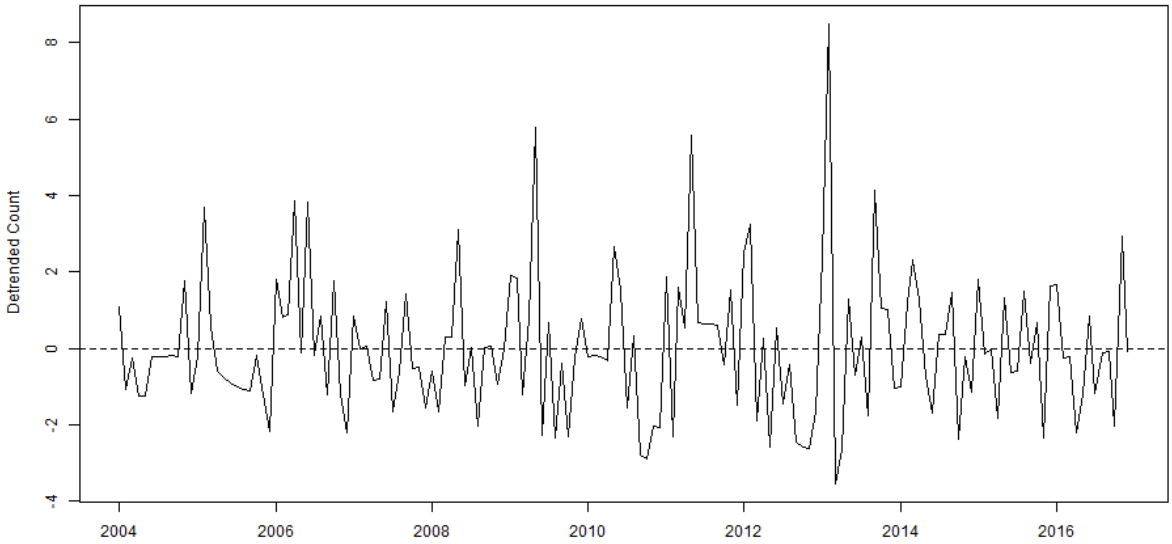

Figure 20: Detrended counts for the magnitude range ≥1.2 and the time period 2004-2016.

Figure 20 shows the detrended count data for this run. We note that the detrended series appears to be stationary with a constant variance but with some noticeable spikes.

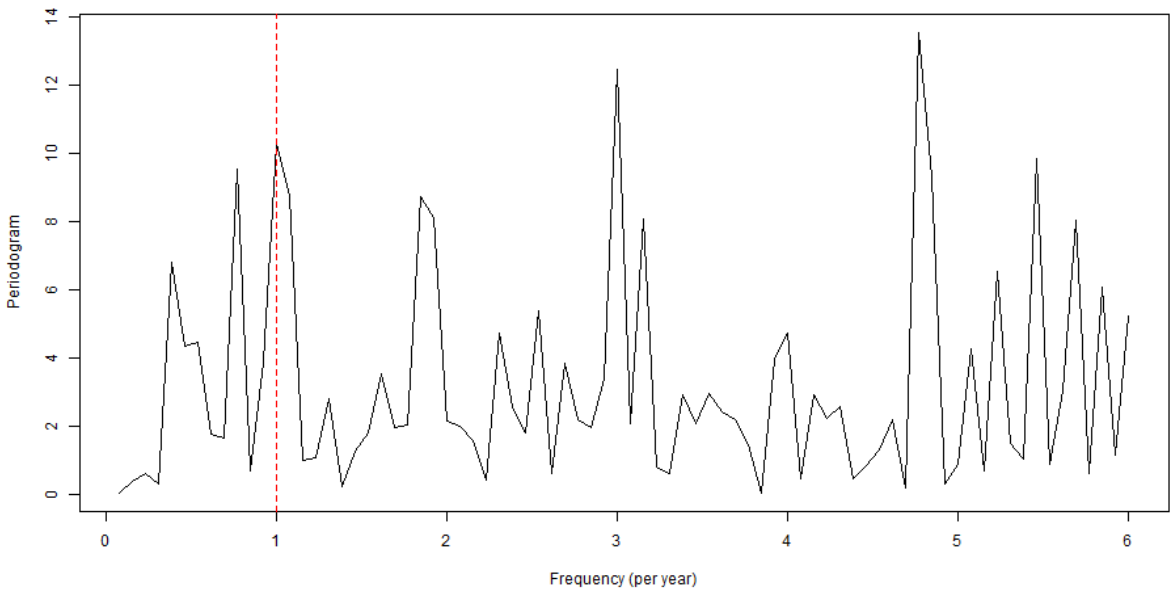

Figure 21: Fourier Periodogram for the detrended counts for the magnitude range ≥1.2 and time period 2004-2016, the vertical red line shows the frequency 1 year^{-1}.

Figure 21 shows the Fourier Periodogram for this series. The p-value in this case was 0.041 giving a positive result. This figure appears to back up the test result as there appears to be a peak in the spectrum at the frequency 1 year^{-1} alongside larger peaks at other frequencies, such as 3 year^{-1}. There is no obvious physical explanation for these higher frequency components. The additional spikes may be due to the non-normality of the series. Since we are considering count data, albeit

detrended counts, we do not expect the normality assumption to be entirely correct. This can be seen in the QQ plot in Figure 22. Alternatively, the extra spikes may be pointing to non-sinusoidal seasonal behaviour.

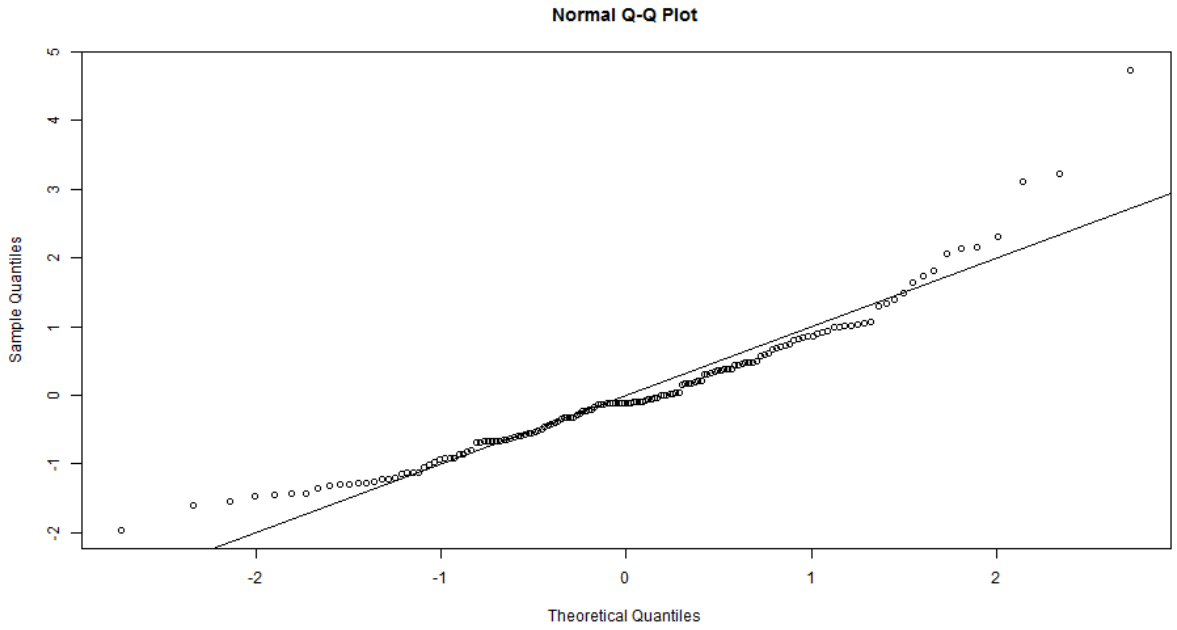

Figure 22: Normal QQ plot for detrended counts for the magnitude range ≥1.2 and time period 2004-2016.

The GAM test gives a p-value of 0.143 which is not significant. A plot of the fitted value from the two GAM models is shown in Figure 23. From this plot we see that the yearly mean model already captures a large amount of the variation and while the seasonal model does explain more of the variation it is a comparatively small improvement.

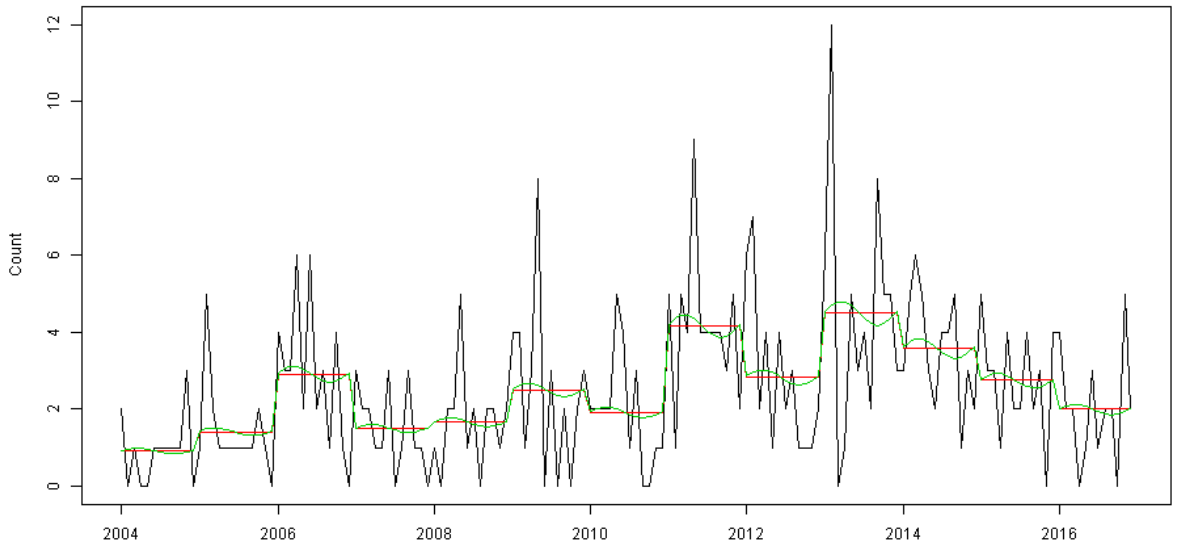

Figure 23: GAM fit, the black lines show the earthquake count, the red shows the fit to yearly means and the green is the model including seasonality.

Figure 24 shows the mean values of the detrended count for this run. The plot also shows 95% confidence intervals in the mean as calculated using Tukey's HSD method assuming independence between months. From this plot we can see the within yearly pattern for this data. There appears

to be a general downward trend in the number of events through the year. This does not agree with the expected sinusoidal pattern, although this may be obscured by the uncertainty in the data. Based on the hypothesis test the largest difference in means is between February and October, however this has not been judged significant with a p-value of 0.138. This Figure also gives some explanation to the additional peaks in the DFT spectrum as there appear to be local maxima in May and September.

Figure 25 shows box and whisker plots for the observations of the detrended count per month. From this plot it appears that the normality assumption is not met for all months. Some months, such as February, have outliers while others are skewed. The non-parametric hypothesis test found a significant difference between January and December with a p-value of 0.042. These months have differing means but are also both skewed in opposite directions. This explains why a rank based test gives a significant result where a parametric test did not.

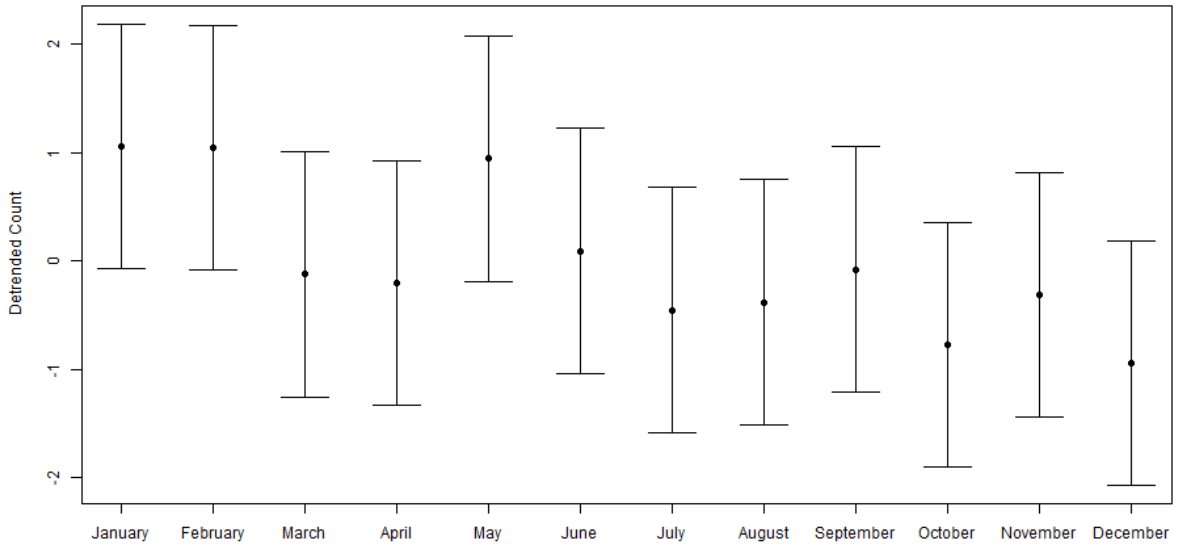

Figure 24: Means with 95% confidence intervals for monthly detrended counts, used for parametric month-by-month comparisons.

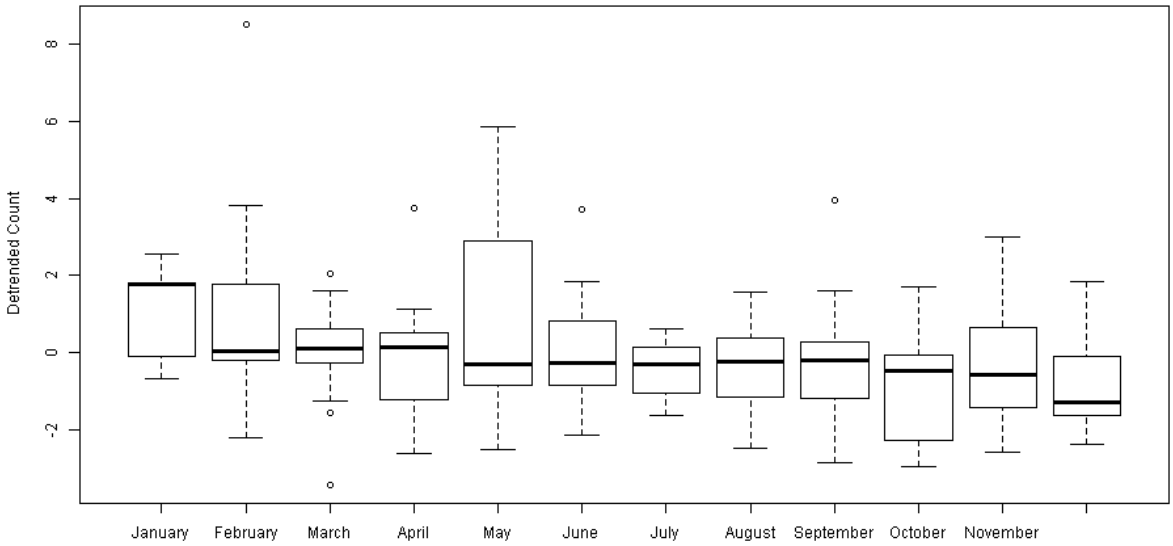

Figure 25: Box plots of Detrended Counts per month, used for non-parametric month-by-month comparisons.

4.3.2 Detailed Analysis 2

We now consider the runs for the time period 1995-2016 and magnitude range ≥1.5. This choice of magnitude of completeness ensures that there is no observation bias in the data caused by small magnitude earthquakes not being recorded. Specifically, we will consider the run with aftershock removal for which three of the four testing methods give a positive result.

Figure 26 shows the detrended time series of counts for this example, without aftershock removal in black and with aftershock removal in red. We note that based on the overall results the removal of aftershocks has the effect of increasing the significance of the findings from all the tests. Aftershocks are found throughout the time series although more frequently in the later half. It is not immediately obvious from this plot why their removal has caused the seasonality to become more pronounced.

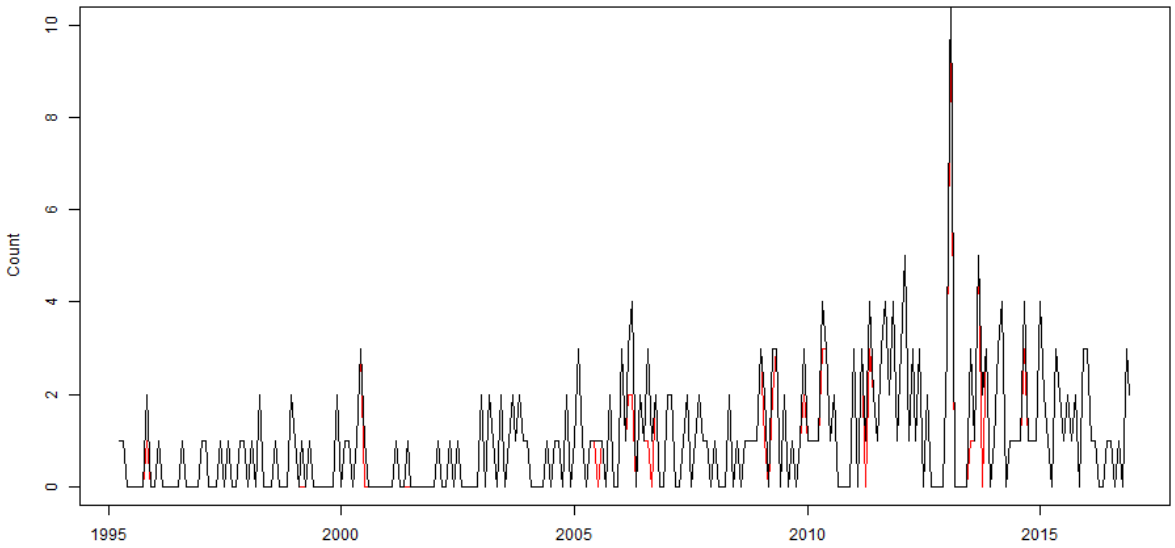

Figure 26: Earthquake Count with and without aftershock removal for the time period 1995-2016 and magnitude range ≥1.5, the black line shows the original and the red line shows the series with aftershock removal.

Figure 27 shows the Fourier Periodogram for this series. This peak is found not to be significant with a p-value of 0.056 and other frequency components which are more prominent in this series. This again could be an indication that non-normalities in the data are causing the test to become invalid. It may also be the case that the periodicity is not sinusoidal.

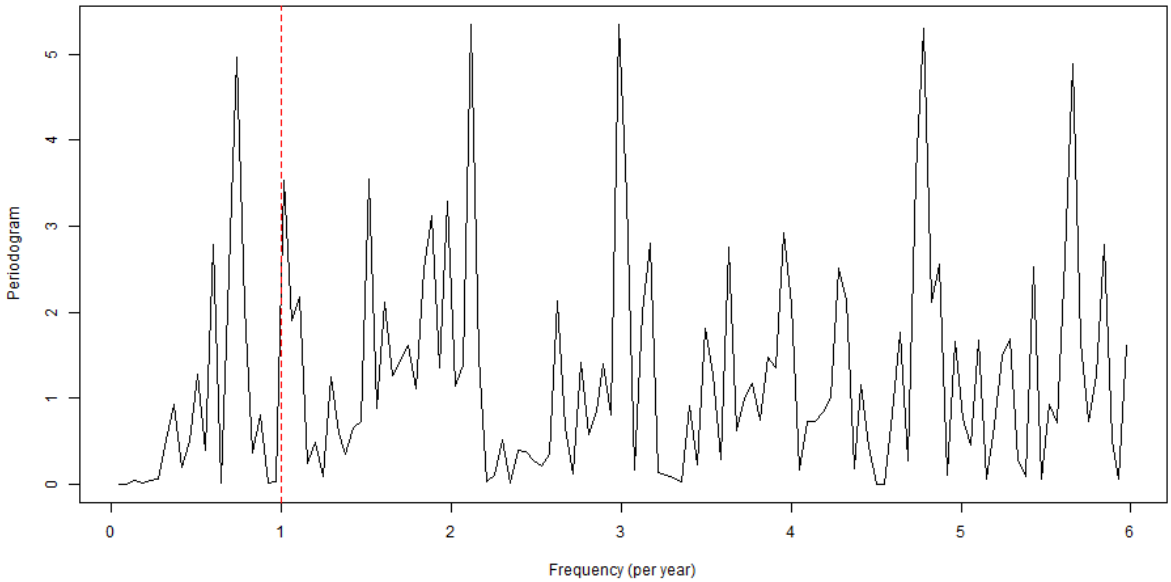

Figure 27: Fourier Periodogram for the Detrended Count with aftershocks removed, the vertical red line shows the frequency 1 year^{-1}.

Figure 28 shows the results of fitting the GAM to this data. It is clear from this plot that the method has found seasonality but not a sinusoidal type pattern, rather there is a sharp increase at the start of the year with the rest of the year being close to flat. This helps to explain the results of the DFT

test.

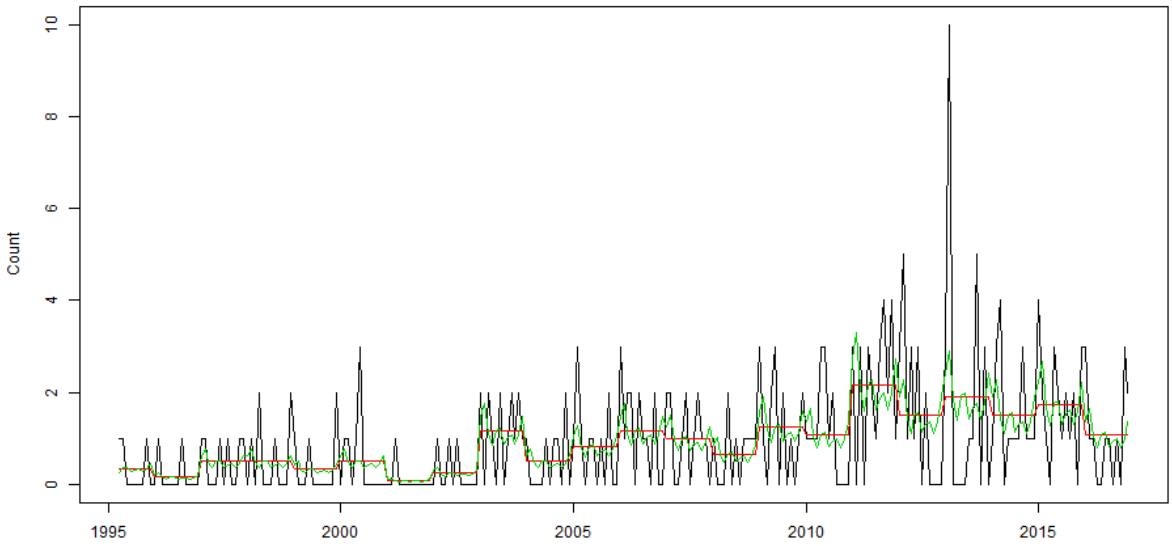

Figure 28: GAM fit, the black lines show the earthquake count, the red shows the fit to yearly means and the green is the model including seasonality.

Figure 29 shows the monthly mean and confidence intervals used in the parametric hypothesis test. This again backs up what has been found previously with January and February having the highest values. This plot also shows how the test output changes if we remove aftershocks. Looking at this we can see the reasons for the result becoming significant. Overall the confidence intervals are smaller due to a reduction in the variance of each month. There is also a slight drop in the mean for July, these two factors together reduce the p-value from 0.12 to 0.044.

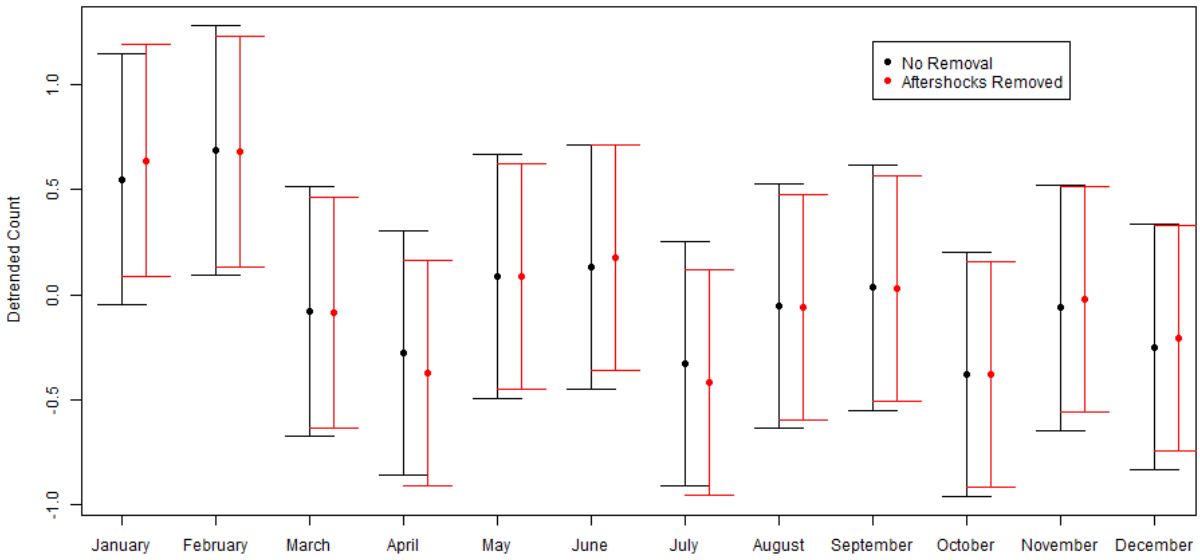

Figure 29: Mean and confidence intervals used for the parametric hypothesis test. The black lines are with no aftershock removal, the red lines are with aftershocks removed.

4.3.3 Detailed Analysis 3

For this analysis we will look at the effect of pressure delay correction. We select the run for the time period 1995 to 2016 and magnitude range ≥1.2 without aftershock removal. This run is selected because both hypothesis testing methods only give significant results if the pressure delay

correction is applied. The other two methods are significant in both cases. Figure 30 shows the detrended earthquake count both with and without the use of pressure delay correction. From this figure we see that many of the peaks in the series have simply been shifted in time. However, some of the largest peaks are significantly reduced.

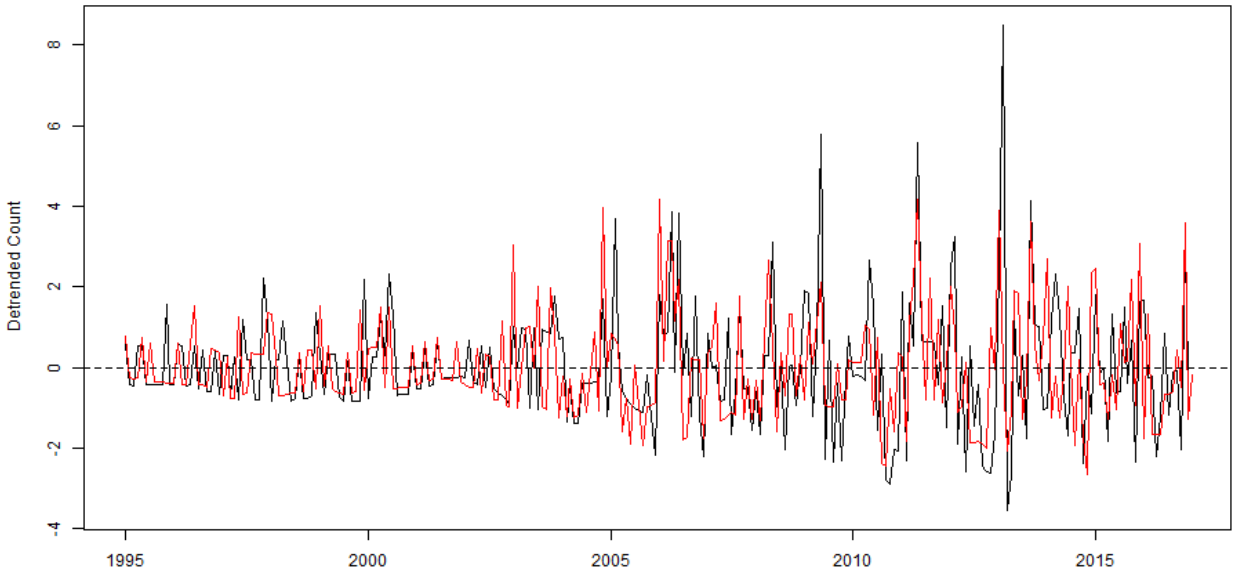

Figure 30: Detrended count without pressure delay correction (black) and with pressure delay correction (red).

Figure 31 shows the monthly mean of the detrended count with pressure delay correction and without. From this plot it is clear why the hypothesis test becomes more significant as it appears that events have been shifted from February into January. This has made the January count larger. The minimum has also shifted to August which is now significantly lower than January. This pattern is closer to what might be expected with a peak earthquake rate in the winter and a low in the summer. To fit this trend completely we would expect months such as February and March to have higher counts as opposed to only January.

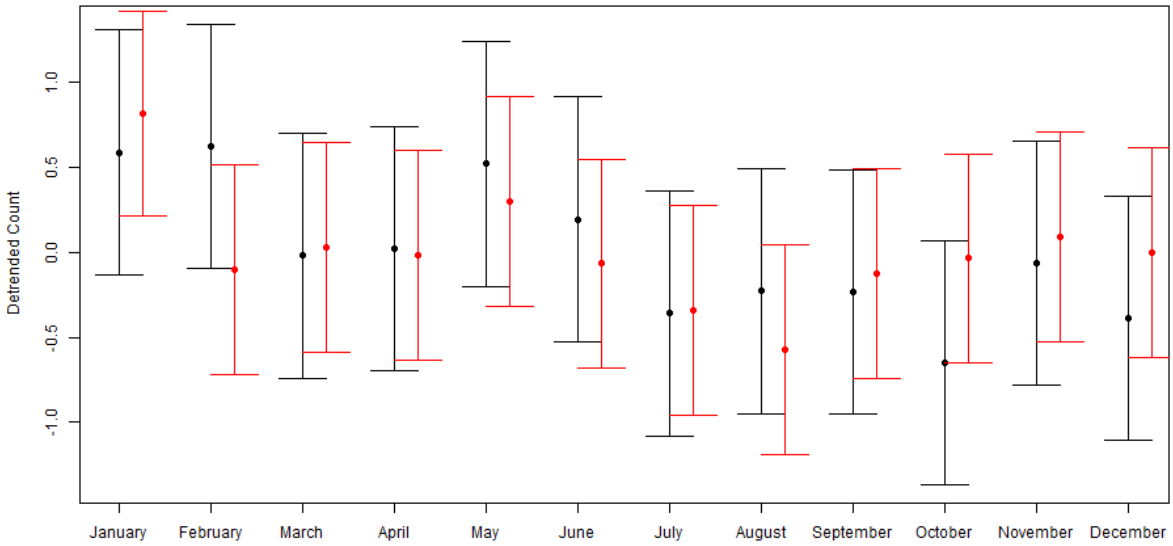

Figure 31: Monthly mean and 95% HSD confidence intervals. The black line shows the analysis without pressure delay correction and the red line is with pressure delay correction.

5 Comparison of Seasonal and Non - Seasonal Models

In Chapter 4 we focused our analysis on being able to detect any seasonal patterns within the earthquake catalogue. This analysis cannot by itself answer the question whether any seasonal pattern can be used to improve estimates of future seismicity. It could be the case that there are seasonal patterns in the earthquake rate but these are not detectable because they do not fit the assumptions of our tests, it is then possible that in a factorial approach of testing a range of nonparametric models some models will be able to detect and make use of this seasonality. In this Chapter we will perform a comparison between the predictive power of models which can make use of any seasonal information and those which cannot. The aim is to investigate whether we can distinguish forecast performance between models with and without a seasonal component. The question can be broken down in two sub questions:

1) Is it possible to build models which perform better than a baseline without including seasonal information?
2) Do models, which are able to outperform the baseline, perform significantly better if they can make use of seasonal information?

The first question addresses whether it is possible to build a viable model using input data which has had the seasonal pattern removed. In (Limbeck, et al., 2018) we found models which beat a baseline model in a paired hypothesis test. If we are not able to do the same for models which do not include seasonal information, then it is evidence that our models are making use of the seasonal patterns in the input data and these are giving improvements to the predictive power. The second question addresses if seasonality has given an improvement to the predictive power. This question is of most interest if we have found non-seasonal models which perform better than the baseline. An important element of this question is that we only consider models which are able to beat a baseline, this is because poorly performing models may be improved by including seasonality but this is only relevant if the improved model can be considered as a good model. In this case we cannot entirely discount non-seasonal models but we may still find that including seasonality does give a significant increase in accuracy.

5.1 Removing Seasonal Signals

To compare seasonal and non-seasonal models we need to be sure that the non-seasonal models are restricted in such a way that they do not contain any terms which are linked to the seasonality of the input data. For simple models, such as the GAM considered in Section 4.2.2, it is possible to ensure this by restricting the terms present in the model. However, for the machine learning methodology used in (Limbeck, et al., 2018), which including Random Forests, Support Vector Machines and more, such an approach is not always possible. This is especially true since the methodology aims to minimise the parametric assumptions made. We therefore opt to remove the seasonality in the models by first removing any seasonal patterns from the input data. In doing so we ensure that the models cannot be basing their forecasts on any seasonal patterns in the data. We apply a spline-based smoothing to any seismicity event rate forecast input data which may have a seasonal pattern. One drawback of this procedure is that it removes all sub year frequencies from the data. This includes the persistent yearly cycles we are interested in as well as any higher frequency cycles and any transient seasonal cycles. In the event of a significant difference between seasonal and non-seasonal model performance we would need to consider this question. Our smoothing procedure is as follows,

- Subsample the data to one observation per year.
- Interpolate the missing data points using a cubic spline based smoother.

Figure 32 shows an example of this procedure applied to the reservoir pressure data. It is very clear to see that original data has a seasonal pattern. The red line shows the same data but with the spline smoothing method applied to each reservoir location. As we can see this line does not appear to have any repeating seasonal pattern. We should note that there is no definitive proof that this method removes all seasonal information from all inputs but we believe it is sufficiently stringent to give meaningful results. We also note that by smoothing the inputs in this way there is the possibility of information leakage whereby a method is given information for future time periods when making predictions. This is only significant if future production is closely linked to the seismicity in the past year. We again believe this effect is not large enough to change our conclusions in a meaningful way.

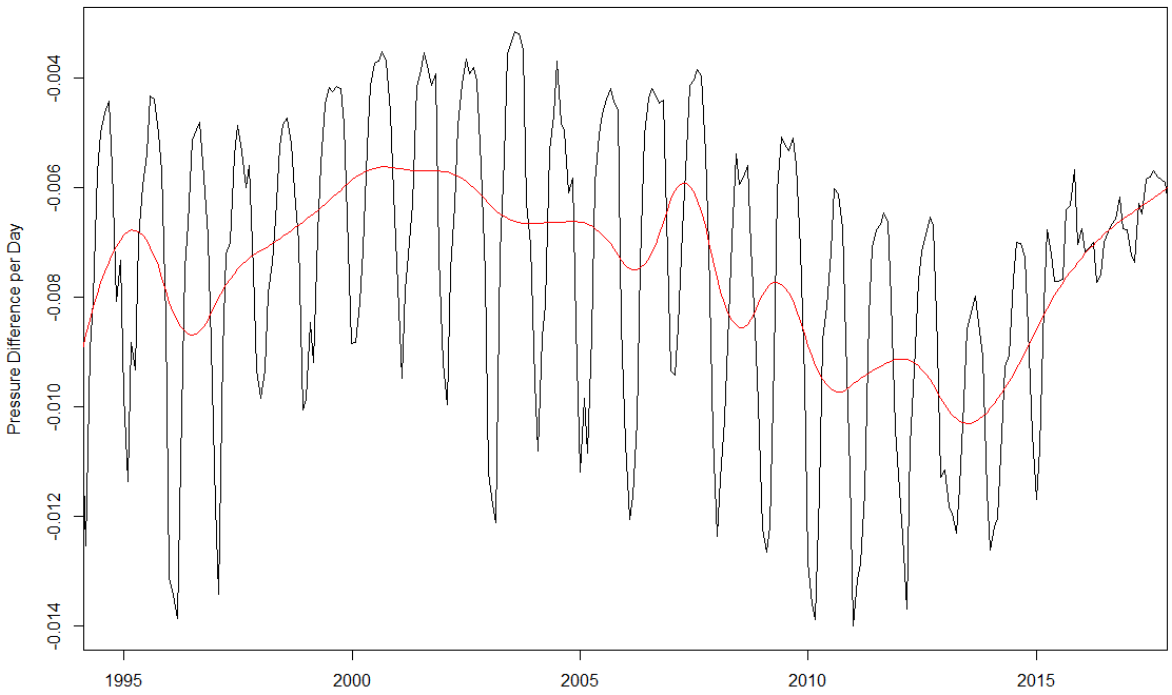

Figure 32: Plot of differenced average reservoir pressures. The black line shows the original data and the red line shows the spline smoothed data.

5.2 Seasonal and Non-Seasonal Results

In this section we present the results of comparing models built using both the seasonal and non-seasonal input data. We first compare the output in terms of the ability of models to beat a baseline. In this case, the baseline model we choose is a simple regression model between the event rate and the moving average of the change in reservoir pressure, which was reported as the best performing extrapolating baseline in (Limbeck, et al., 2018) and coined the "depletion moving average" there. Figure 33 shows a box and whisker plot summarising the results of performing an unpaired hypothesis test comparing the accuracy of each model with that of the baseline. We use the Mean Absolute Deviation and a one-sided Mann–Whitney test to determine if each model performs significantly better than the baseline. The null and alternative hypotheses for this test when comparing a model, $\mathcal{M}_1$, and the baseline model, $\mathcal{M}_0$, are as follows:

- H_0: $\mathrm{MAE}(\mathcal{M}_1) = MAE(\mathcal{M}_0)$.
- H_1: $MAE(\mathcal{M}_1) < MAE(\mathcal{M}_0)$.

As we can see there are no models, either seasonal or non-seasonal, which significantly beat the baseline in this test. This is perhaps unsurprising as this was also found in (Limbeck, et al., 2018).

We can conclude from this figure that there is no consistent shift in performance, either positively or negatively, if we remove seasonality from the data. There is perhaps a general drop in performance for the GLM models and an increase for KNN and KSVM models.

We now move to a paired hypothesis test. This test can be used as we are comparing paired predictions of the same time points. The test can be more powerful as it can detect more subtle differences between models, this does come with the possibility of a higher Type I error rate. We assume that this risk is acceptable as our main interest is in comparing the results of seasonal and non-seasonal models rather than looking at the absolute performance of any one model. Figure 34 shows a box and whisker plot of the results of the paired hypothesis tests using a Wilcoxon signed rank test. As we can see from this plot there are models which are significantly better than the baseline. This is true for both the seasonal and non-seasonal models. Already these results show that it is possible to build a model based on non-seasonal input data which passes the same significance tests we used to assess the seasonal models.

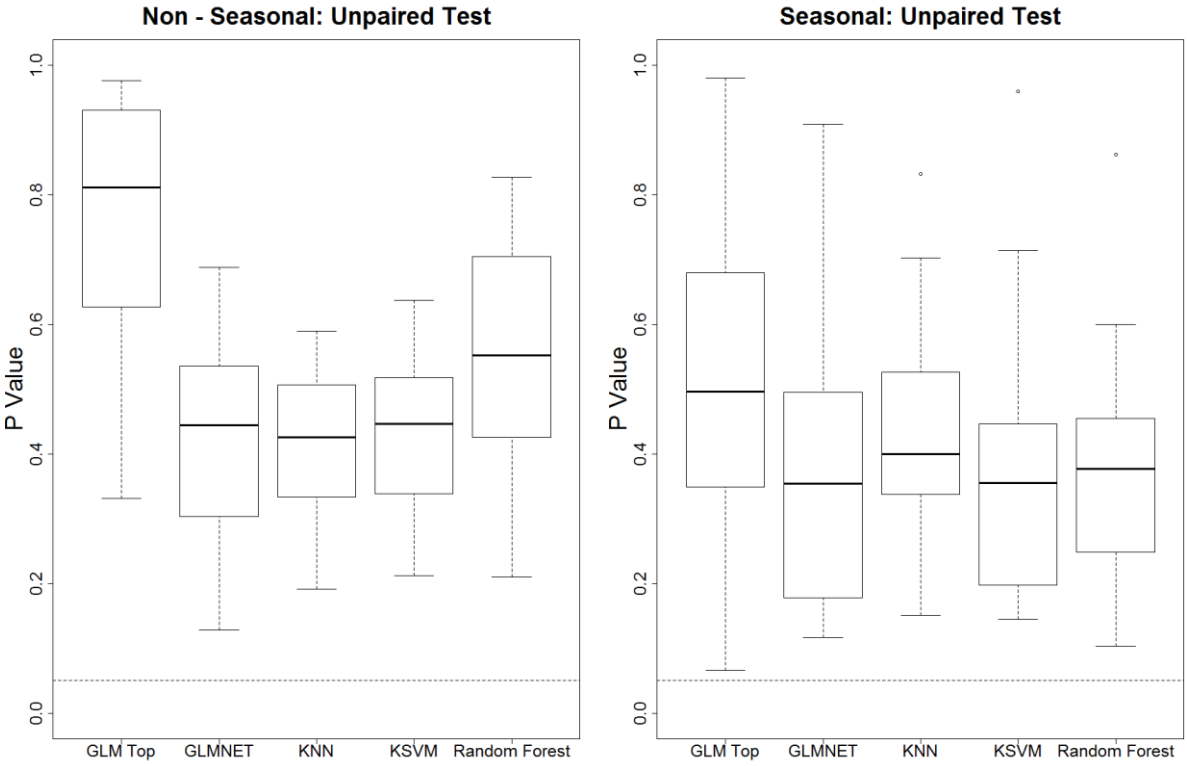

Figure 33: Results of Unpaired Hypothesis test comparing each model with the baseline, the dashed line in the plot shows the 5% significance level.

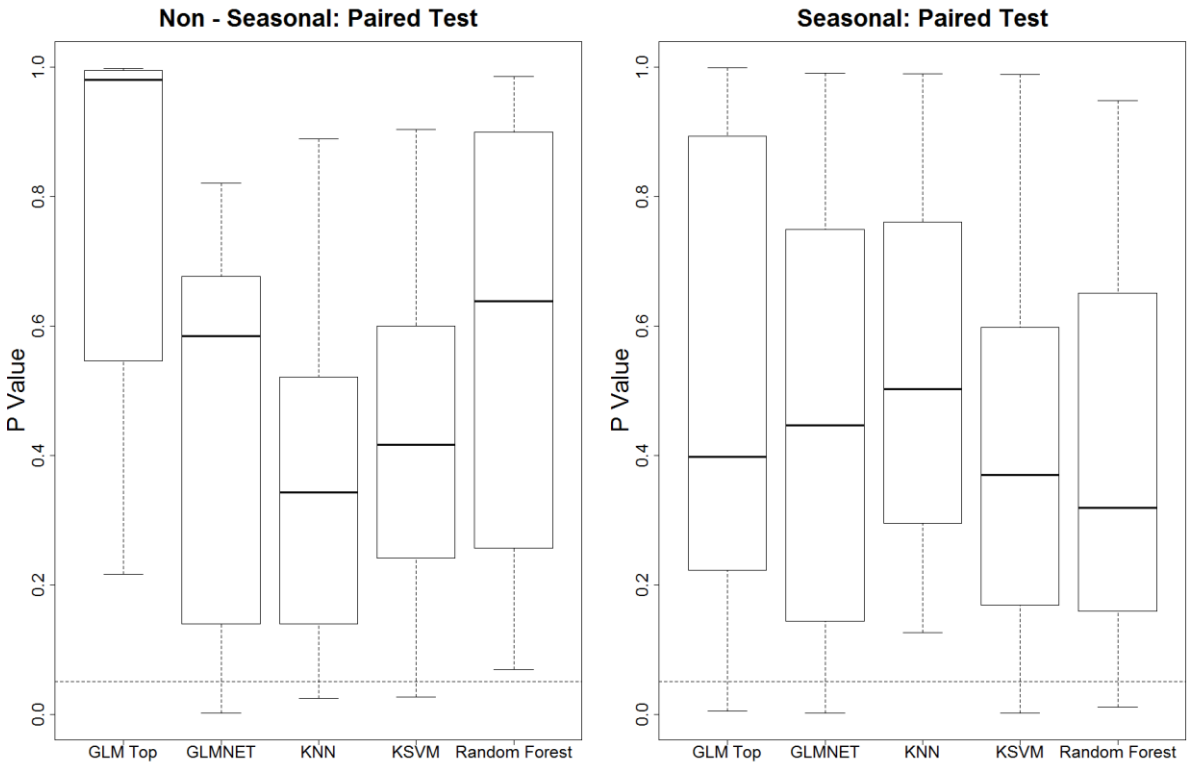

Figure 34: Results of Paired Hypothesis test comparing each model with the baseline, the dashed line in the plot shows the 5% significance level.

We have seen that non-seasonal models can pass a significance test and so cannot be discounted. Our next step is to do a direct comparison between the seasonal and non-seasonal models. The aim is to test if the seasonal models significantly outperform the equivalent non-seasonal models. In principle we can do this comparison between all possible pairs of models though this will give a prohibitively large set of comparisons. We therefore restrict ourselves to like-for-like comparisons, i.e. we will compare models which are of the same model type and are built on the same input data with the exception that one model uses the original data and the other model the smoothed data. We also exclude models which do not perform well enough to beat the baseline model in the paired hypothesis test. This gave a total of 13 comparisons where at least one of the two models being compared beats the baseline in the paired hypothesis test.

Figure 35 shows the results of this comparison for models which consider earthquakes with a magnitude greater than or equal to 1.2. Figure 36 shows the same information for models with a magnitude range of 1.5 and greater. Looking at these two figures we see that there are three cases where the seasonal model performs significantly better than the equivalent non-seasonal model. In general, we can see that there is a preference for models which include Seasonality to perform better. There are some differences between model types, but given the small number of models which perform well we cannot conclude if this result is due to chance or some difference in the character of the modelling techniques. Based on these results we can conclude that for certain specific models there appears to be a significant advantage to including seasonal information in the earthquake modelling. For the majority of comparisons made there was no significant differences and in some cases the non-seasonal model performed better. We therefore cannot conclude that in general including seasonal information gives a performance improvement.

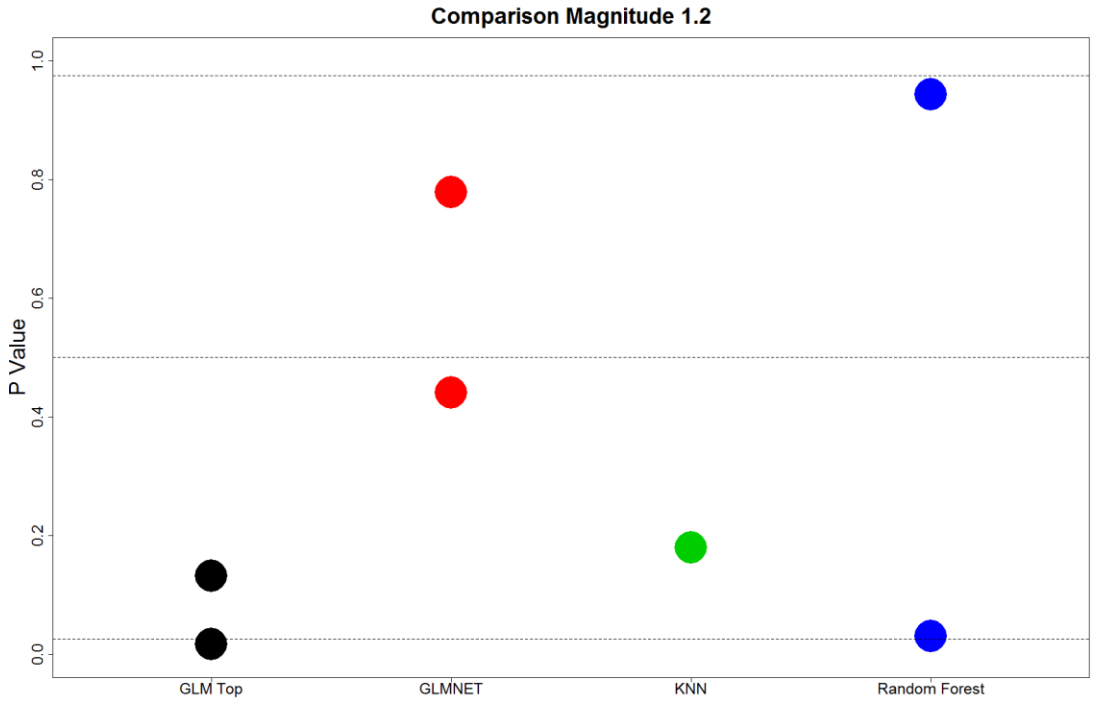

Figure 35: Comparison between Seasonal and Non-Seasonal Models for Magnitude $M \geq 1.2$, The dashed lines show the significance levels. Points above the upper line indicate that the Non-seasonal model performs better at the 5% significance level, points below the lower line indicate that the Seasonal model performs better at the 5% significance level. The middle line shows the boundary separating which model performs better.

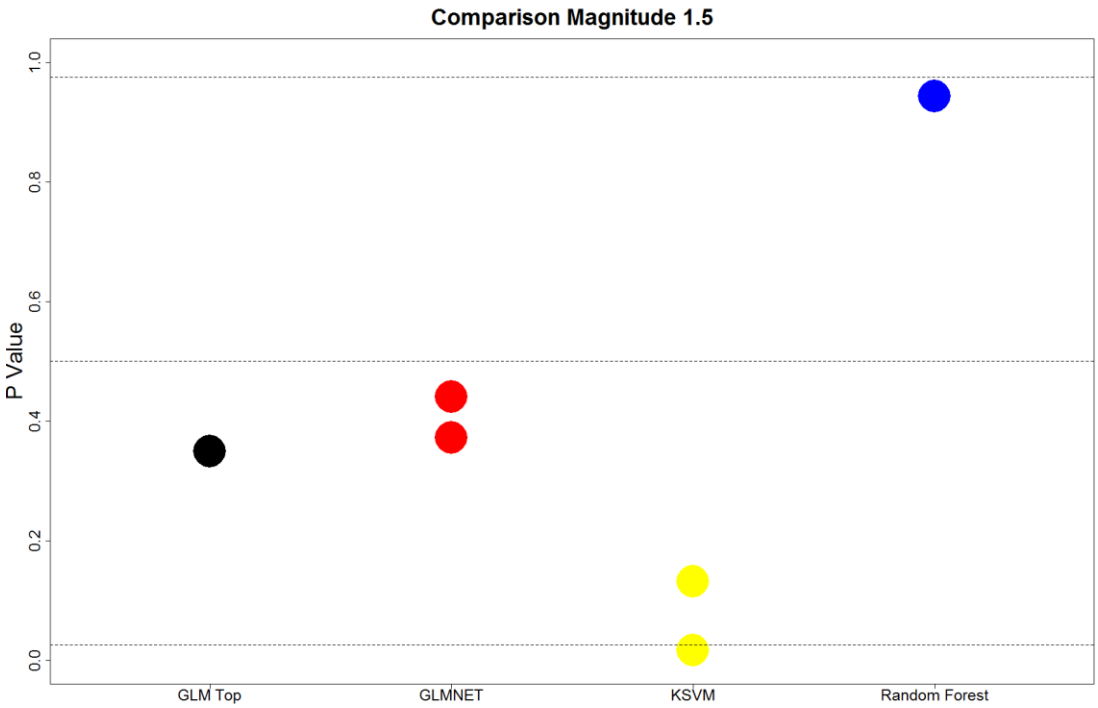

Figure 36: Comparison between Seasonal and Non-Seasonal Models for Magnitude $M \geq 1.5$, the dashed lines show the significance levels. Points above the upper line indicate that the Non-seasonal model performs better at the 5% significance level, points below the lower line indicate that the Seasonal model performs better at the 5% significance level. The middle line shows the boundary separating which model performs better.

6 Earthquake Rate Forecasts

As we have seen in previous Chapter there is limited evidence of the impact of any seasonal patterns on the accuracy of event rate forecast models. As a final step we investigate how the choice of a Seasonal or Non-Seasonal model could affect the trade-off between two production scenarios as discussed in Chapter 2: a "reduced volume" and a "reduced fluctuations" scenario. If seasonal patterns in production play a role in seismic event rates both of these scenarios could potentially lead to reduced seismicity as compared to the Baseline scenario. The expected production rates for these two scenarios as well as the Baseline scenario are shown in a larger time scale context in Figure 10, a zoom-in is shown in Figure 37.

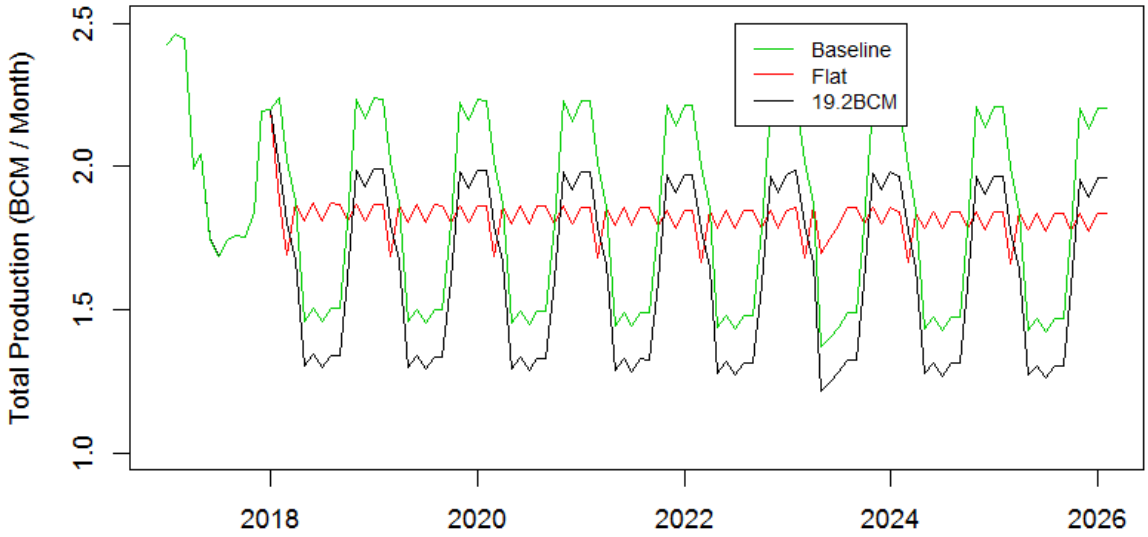

Figure 37: Monthly gas production rates for the 'Baseline', 'Flat' and '19.2BCM' Production Scenarios.

To prefer one production scenario over another we must be able to determine if one scenario is likely to lead to a lower rate of earthquake event rates than another. Here, the machine learning based seismicity event rate forecasting methodology of (Limbeck, et al., 2018) is used to forecast event rates for the two future production scenarios and the baseline scenario considered here. This Chapter considers a range of different extrapolating models from the beforementioned methodology and discuss the implications of each in terms of deciding between production scenarios. The models considered are shown in Table 5. These models were chosen as illustrative examples to show how the decision could vary between models which contain seasonal information and those which do not. We then selected two models, per magnitude range, which both beat the baseline and had the largest performance improvement over their equivalent seasonal or non-seasonal model.

Model	Minimum Magnitude	Includes Seasonality	Time Shifts
GLM Top	1.2	Yes	No
GLMNET	1.2	No	No
SVM	1.5	Yes	No
GLMNET	1.5	No	No

Table 5: Models used for future forecasts.

Figure 38 shows the future forecasts for the two models which have a minimum magnitude of 1.2. The forecast confidence intervals are based on the observed predictive accuracies of the model found from the test set over the required time horizon. There is very little difference in the earthquake rate forecasts for the two models. This is true for both the seasonal and non-seasonal model. The non-seasonal model forecasts are much smoother, this is likely in part due to the lack of a seasonal component but also could be due to the differences in model types. The forecasted event rates for both the reduced volume and flat scenarios are nearly indistinguishable – hence it is impossible to say with any degree of statistical significance which scenario would result in the lowest seismicity.

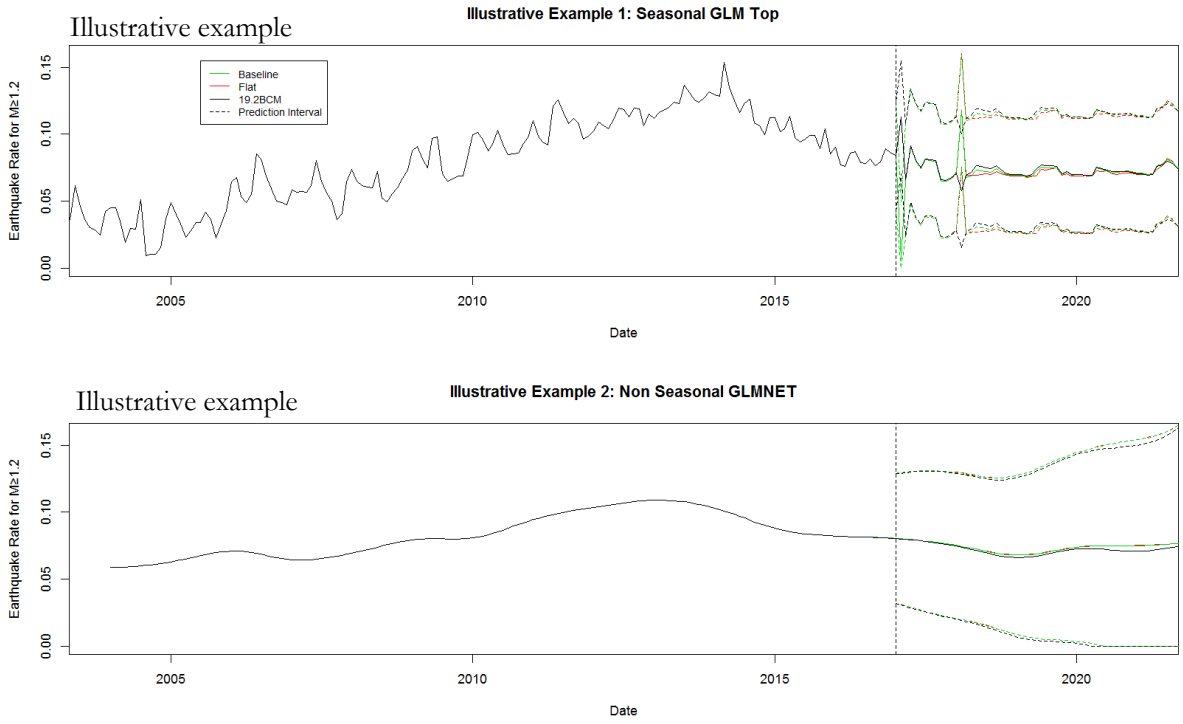

Figure 38: Scenario comparison for magnitudes $M \geq 1.2$. The vertical line shows the boundary between the training and testing period and the future forecasts. To the right of this line we can see the model output for both the 'Flat Production' Scenario (red), the '19.2BCM' scenario (black) and the baseline scenario (green). The solid lines show the expected values for the model with the dashed lines showing the 90% forecast interval.

Figure 39 shows the model forecasts for the two models based on a minimum magnitude of 1.5. Again, we see for these models there is very little difference in the forecasts for the two scenarios in terms of their forecast confidence intervals.

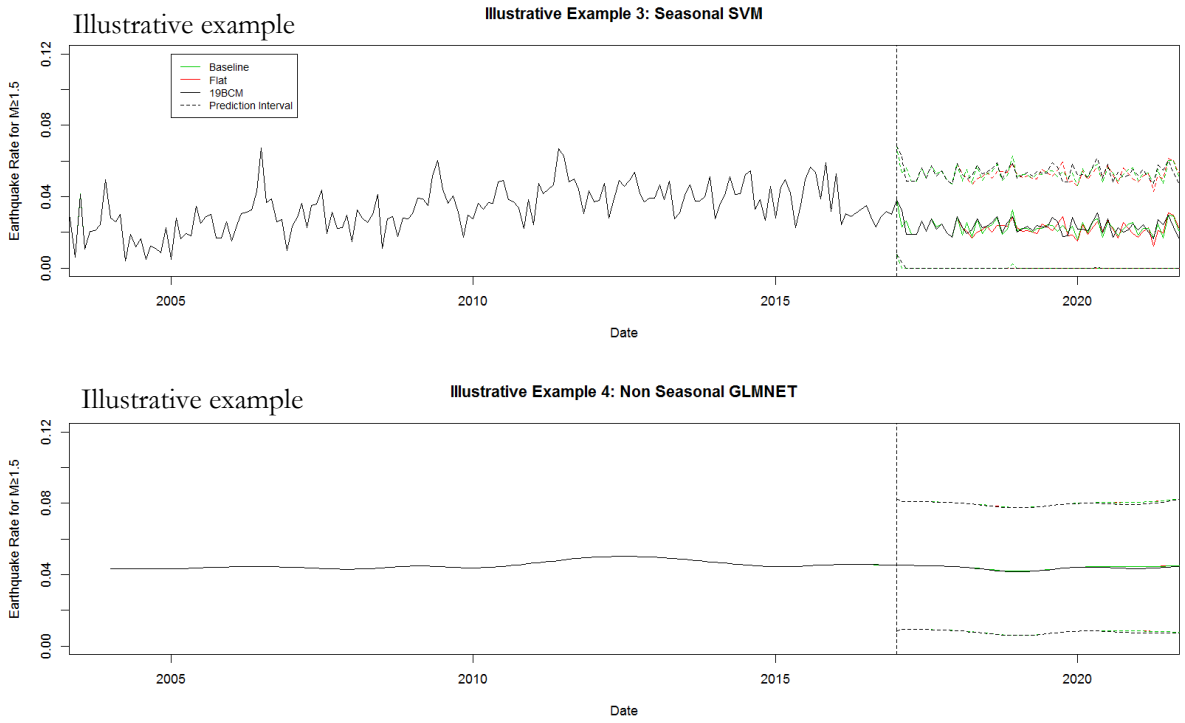

Figure 39: Scenario comparison for magnitudes $M \geq 1.5$. The vertical line shows the boundary between the training and testing period and the future forecasts. To the right of this line we can see the model output for both the 'Flat Production' Scenario (red), the '19.2BCM' scenario (black) and the baseline scenario (green). The solid lines show the expected values for the model with the dashed lines showing the 90% forecast interval.

7 Conclusions and Discussion

7.1 Evidence for seasonality in Earthquake Counts

In Chapter 4 we looked at the evidence for measured seasonality in the rate of earthquake occurrence. The key causes of any measured seasonality are (i) seasonality in earthquake occurrence rates, (ii) seasonality in the smallest detectable event magnitude in the presence of seasonal noise variations and (iii) randomness in event occurrence. Our hypothesis testing methods are designed to exclude (iii) as a cause by calculating the probability of observing the measured seasonality under the null hypothesis of no seasonality, this is the p-value. An important way to distinguish between (i) and (ii) is limiting the earthquake catalogue to events equal to or larger than the magnitude of completeness. We analysed evidence for measured seasonality via a factorial approach with factors including different minimum magnitudes, epochs, methods of aftershock removal and correction for pressure delays. Our results, as seen in the figures in Section 4.3, show there is at most borderline evidence for seasonality above the concordance magnitude of completeness M_c^{con}, hence there is at most borderline evidence for seasonality in earthquake occurrence rates. Lowering of the magnitude threshold below M_c^{con} increases the observation bias, the statistical power and the evidence for seasonality. Our analysis doesn't allow to say whether this evidence is due to true seasonal variations in the earthquake occurrence rates or observation bias due to including events below the magnitude of completeness. No seasonality was found for the latest epoch considered (2010-2016), which has on average the lowest magnitude of completeness but also has relatively low statistical power. For epochs containing earlier periods, the evidence for measured seasonality strongly depends on the experimental setup chosen. It seems that aftershock removal and to a lesser extend pressure delay correction improves the detection of seasonality, both effects together seem to partially negate each other though. For several representative cases we also looked in detail at the test results. We found that while there is some evidence for seasonality due to the tests giving positive results the further investigation found that the seasonal patterns did not necessarily follow the expected pattern and some of the test assumptions may have been violated. This can be seen in Sections 4.3.1, 4.3.2 and 4.3.3.

Lowering the magnitude range to below the consensus magnitude of completeness increases observation bias and statistical power at the same time. As such, an important avenue for future work is to estimate the seasonality detection threshold for our testing methods. The seasonality detection threshold is the minimum size of seasonal variation needed for us to conclusively reject the null hypothesis. This threshold can be estimated using a simulation study where we simulate earthquake catalogues with different levels of seasonal variation and apply our hypothesis tests. Such a study can give powerful insights both statistically (by enabling differentiation between the confounding effects of observation bias and statistical power) and in terms of limits on plausible physical earthquake generating mechanisms. Care must be taken to ensure that the model used for simulation closely matches the true earthquake generation process. Since the details of this process are not fully known it may be necessary to also consider a range of plausible simulation models. We would also need to consider different forms of seasonal patterns, as this is also not known in all detail, and this will likely also affect the detection thresholds of the different hypothesis tests. This point is particularly important as it also affects how we define the size of the seasonal effect which is currently not well defined. Failure to carefully consider these points would cause any results to only be valid for the specific modelling choices and so may not apply to the real data.

Another future step would be to test the sensitivity of our tests to the choice of aftershock removal method. It would be possible to test a range of different removal methods and different parameter choices for each method and look at the sensitivity of our test outcomes to these choices.

7.2 Evidence for Improvements in Model Accuracy

Chapter 5 compared the accuracies of models built on seasonal and non-seasonal input data. The aim was to investigate if the effect of any seasonality was pronounced enough to give a significant increase in the accuracy of models. We removed the seasonal information in the model input data by smoothing. We found that both models built on the original non-smoothed and on the smoothed data were able to beat a baseline model in a paired hypothesis test, indicating that non-seasonal models cannot be immediately discounted. In some cases we were able to detect a difference between the accuracies of the seasonal and non-seasonal models. This was not true for all model comparisons and indeed in some cases we found that smoothing the input data resulted in an increase in the predictive accuracy of the models. We also note that if we take into account the multiple comparisons being made the differences are not seen as significant. These results can be found in the figure in Section 5.2. We therefore cannot conclude that in general models which make use of seasonal patterns perform better than those that do not.

The reason for this result may be a simple case of needing more data. It is possible that given more data, a possible seasonal pattern could be leveraged by the models to improve forecast performance. We also do not explicitly account for any effects of pressure delays, as discussed in Section 4.1.3, which may overshadow the effects of seasonality.

7.3 The effects of Seasonality on Future Earthquake Forecasts

In Chapter 6 we analysed which of two production scenarios, one with reduced volume or one with reduced fluctuations, would result in a lower seismicity event rate compared to a baseline scenario. We found that the event rate forecasts for the two production scenarios were not statistically distinguishable. This is perhaps unsurprising in light of the results for the previous two questions.

Next to the points mentioned above, on several aspects of our approach further discussions are insightful, we address these per chapter:

Chapter 3: Data Sources:

- **Target definition**: this study investigates event rates defined as the number of earthquakes per unit of time above a minimum magnitude. Different definitions of event rates could be used, e.g. the number of earthquakes per unit of production or unit of pore pressure decrease instead of time. Alternatively, quantities like the earthquake energy released (which combines counts and magnitudes) could be investigated.

Chapter 4: Detection of Seasonal Patterns:

- **Detrending Methodology**: in this study we used a kernel smoother to detrend the seismicity event rates prior to testing, there are many other choices for detrending methods or time windows. This will affect the results to some extent however any well-chosen method should not alter the seasonal pattern being studied.
- **Consensus Aggregation**: when aggregating the results of the different test methods we chose to either count the number of positive results or take the smallest p-value after Bonferroni correction. More advanced ways of aggregation could be considered, e.g. ways which take into account the different powers and error rates of the various tests or consider the similarity of tests leading to correlation in their results.

- **Event Binning:** the test were applied to earthquake count data aggregated into monthly bins. We could have considered other time periods for the bins, in particular for the hypothesis tests on monthly means. We did not find an optimum way to choose the bin size which balanced small bins giving few events per bin and many zeros and large bins obscuring seasonal effects.

Chapter 5: Comparison of Seasonal and Non - Seasonal Models:

- **Removing Seasonality:** when distinguishing between seasonal and non-seasonal models we found the best way to make the comparison was to remove seasonality from the input data rather than alter the models themselves. In a similar way to the detrending there are many choices of smoothing method which we could have used. It is also not certain that our method ensures the models do not have access to seasonal information or future information.
- **Choice of Baseline Model:** all of our models were compared to a baseline to test their performance. We again could have made a different choice of baseline model which would have altered the test results.

Chapter 6: Earthquake Rate Forecasts:

- **Choice of Forecasting Models:** the models, whose future forecast were considered, were chosen as they were able to extrapolate for input data outside the range of the training data. We could extend this to non-extrapolating models if we can either show that the future input data is within the range already seen or that the models still have good forecast performance over larger forecast horizons.
- **Scenarios Considered:** We only looked at two possible future production scenarios, in principle there are many other choices for a production strategy. We could consider a more extreme range or do a designed experiment to test the strength of different factors.

Acknowledgements

The authors are indebted to Taco den Bezemer (NAM) and Jan van Elk (NAM) for their strong and continued support for this study – without their enthusiasm and trust this study would not be here.

We are indebted to the lead reviewers for insightful discussions throughout this study, much of which has been incorporated in this study. In alphabetic order:

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

This study has been reviewed within a larger audience at successive stages of its maturity. We owe a big "Thank You" to the extended review team for their constructive feedback. In alphabetic order:

- Dr Peter van den Bogert (Shell P&T);
- Dr Xander Campman (Shell P&T);
- Dr Pandu Devarakota (Shell P&T);
- Dr Munish Goyal (IBM Services)
- Dr Kees Hindriks (Shell P&T)
- MSc Stephen Lord (IBM Services);
- Dr Roger Yuan (Shell P&T);
- Dr Rick Wentinck (Shell P&T);
- Dr Mo Zhang (IBM Services).

As can be read in chapter 3, this study builds on the data provided by specialists. We would like to thank (alphabetically):

- Hermann Baehr (NAM), for providing subsidence data;
- Stijn Bierman (Shell P&T), for sharing his subsidence and compaction interpolations;
- Leendert Geurtsen (NAM), Per Valvatne (NAM) and Assaf Mar-Or (NAM) for providing both the dynamic reservoir data from MoReS and the production forecasts;
- Gerard Joosten (Shell P&T) for his support in exporting the HCT, HCM and related properties from MoReS;
- Richard Vietje (NAM), for providing historical production data;
- Clements Visser (NAM), for providing the fault data and estimates for the Groningen regions;
- Onno van der Wal (NAM), for providing compaction and subsidence data;
- Alan Wood (Shell P&T), for sharing a recent version of the Petrel model.

Finally, we would like to thank Robin Bakker (Shell SIEP), Harry van der Burg (Shell SITI), Maarten Veldhuizen (NAM), 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.* Shell Global Solutions International.

Bierman, S., Paleja, R., & Jones, M. (2015). *Statistical methodology for investigating seasonal variation in rates of earthquake occurrence in the Groningen field.* 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.* Shell Global Solutions International.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., . . . Jones, Z. (2016). mlr: Machine Learning in R. *Journal of Machine Learning Research*, 1-5.

Bourne, S. J. (2018, May 17). Personal Communication.

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, 122*, 10299-10320.

Bourne, S., & Oates, S. (2015). *An activity rate model of induced seismicity within the Groningen Field (part 1).* NAM.

Bourne, S., & Oates, S. (2015). *An activity rate model of induced seismicity within the Groningen Field (Part 2).* NAM.

Breiman, L. (2001). Random Forests. *Machine Learning*, 5-32.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science, 16*(3), 199-231.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 273-297.

Dempsey, D., & Suckale, J. (2017). Physics-based forecasting of induced seismicity at Groningen gas field, the Netherlands. *Geophysical Research Letters*, 7773-7782. doi:https://doi.org/10.1002/2017GL073878

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

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., Ruigrok, E., & Spetzler, J. (2017). Development of seismicity and probabilistic hazard assessment for the Groningen gas field. *Netherlands Journal of Geosciences*, s235-s245. doi:10.1017/njg.2017.20

Friedman, J., Hastie, T., & Tibshiranie, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 1-22.

Gardner, J. K., & Knopoff, L. (1974). Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian? *Bulletin of Seismological Society of America, 64*(5), 1363-1367.

Hastie, T., & Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall.

Hu, L.-Y., Huang, M.-W., Ke, S.-W., & Tsai, C.-F. (2016). The distance function effect on k-nearest neighbor classification for medical datasets. *Springerplus*, 1304.

Jordan, M., & Mitchell, T. (2015). Machine learning: trends, perspectives, and prospects. *Science*, 255-260.

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, Mar). PSHA for seismicity induced by gas extraction in the Groningen Field. *Presentation given at the 2015 Schatzalp Workshop on Induced Seismicity*. http://www.seismo.ethz.ch/en/static/schatzalp/2015/Kraaijpoel.pdf.

Lehmann, E. L., & Romano, J. P. (2005). *Testing Statistical Hypotheses* (3rd ed.). Springer.

Limbeck, J., Lanz, F., Barbaro, E., Bisdom, K., Park, T., Harris, C., . . . Nevenzeel, K. (2018). Machine Learning based induced seismicity event rate time series forecasts within the Groningen Field. *NAM*.

Mignan, A., & Woessner, J. (2012). *Estimating the magnitude of completeness for earthquake catalogs.* Community Online Resource for Statistical Seismicity Analysis. doi:10.5078/corssa-00180805

Miller, R. G. (1981). *Simultaneous Statistical Interference* (2nd ed.). Springer Verlag.

Ministry of Economic Affairs and Climate. (2018, Mar 29). *Kabinet: einde aan gaswinning in Groningen [Cabinet: end of gas production in Groningen]*. Retrieved from Rijksoverheid Nieuws [State News]: https://www.rijksoverheid.nl/actueel/nieuws/2018/03/29/kabinet-einde-aan-gaswinning-in-groningen

NAM. (2016). *Study and Data Acquisition Plan Induced Seismicity in Groningen, Update Post-Winningsplan 2016*. NAM.

NAM. (2016). *Winningsplan Groningen Gasveld 2016*. NAM.

NAM. (2017). *Groningen Measurement and Control Protocol (Translated into English from the Original Dutch document).* Technical Report.

Nelder, J., & Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society*, 370-384.

Nepveu, M., Van Thienen-Visser, K., & Sijacic, D. (2016). Statistics of seismic events at the Groningen field. *Bull Earthquake Eng, 14*, 3343-3362. doi:10.1007/s10518-016-0007-4

Oppenheim, A., Willsky, A., & Nawab, H. (1983). *Signals and Systems* (2nd ed.). Pearson Education Limited.

Paleja, R., & Bierman, S. (2016). *Measuring changes in earthquake occurrence rates in Groningen - update October 2016*. Shell Global Solutions International.

Pijpers, F. P. (2016). *A phenomenological relationship between reservoir pressure and tremor rates in Groningen.* CBS [Statistics Netherlands].

Pijpers, F. P. (2017). *Interim report: correlations between reservoir pressure and earthquake rate.* CBS [Statistics Netherlands].

Post, R. A. (2017). *Statistical inference for induced seismicity in the Groningen gas field.* Eindhoven University of Technology.

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

Sornette, D., & Sornette, A. (1999). General theory of the modified Gutenberg-Richter law for large seismic moments. *Bulletin of the Seismological Society of America, 89*(4), 1121-1130.

Spetzler, J., & Dost, B. (2017). Hypocentre estimation of induced earthquakes in Groningen. *Geophysical Journal International*, 453-465. doi:10.1093/gji/ggx020

Staatstoezicht op de Mijnen. (2016). *Advies Winningsplan Groningen 2016 [Advice Production Plan Groningen 2016]*. SodM. Retrieved from

https://www.sodm.nl/documenten/publicaties/2016/06/21/advies-sodm-winningsplan-groningen-2016

TNO, Geology Service Netherlands. (n.d.). *Groningen Gasfield*. Retrieved 12 19, 2017, from NLOG: http://www.nlog.nl/en/groningen-gasfield

Van Thienen-Visser, K., Fokker, P., Nepveu, M., Sijacic, D., Hettelaar, J., & Van Kempen, B. (2015). *Recent developments on the seismicity of the Groningen field in 2015*. TNO.

Van Thienen-Visser, K., Sijacic, D., Van Wees, J.-D., Kraaijpoel, D., & Roholl, J. (2016). *Groningen field 2013 to present Gas production and induced seismicity*. TNO.

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

Vlek, C. (2018). Induced Earthquakes from Long-Term Gas Extraction in Groningen, the Netherlands: Statistical Analysis and Prognosis for Acceptable-Risk Regulation. *Risk Analysis*. doi:https://doi.org/10.1111/risa.12967

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* , 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)*, 73(1):3-36.

Bibliographic information

Title	Seasonality analysis for induced seismicity event rate time series within the Groningen Field
Author(s)	T. Park (GSNL-PTX/D/S) H. Jamali-Rad (GSNL-PTX/S/IA) W. Oosterbosch (IBM Services) J. Limbeck (GSNL-PTX/D/S) F. Lanz (IBM Services) C. Harris (SUKEP-UPO/W/T) E. Barbaro (IBM Services) K. Bisdom (GSNL-PTX/S/RM) K. Nevenzeel (IBM Services)
Keywords	Induced seismicity, earthquakes, activity rate, seasonality, production strategies, analytics, data science, machine learning, statistics, NAM, Groningen
Date of Issue	August 2018
Period of Work	May 2017 – July 2018
US Export Control	Non US - Non Controlled
Issuing Company	Nederlandse Aardolie Maatschappij B.V. Upstream International Schepersmaat 2 9405 TA Assen The Netherlands

The copyright of this document is vested in Nederlandse Aardolie Maatschappij, B.V., Assen, The Netherlands. All rights reserved. Neither the whole nor any part of this document may be reproduced, stored in any retrieval system or transmitted in any form or by any means (electronic, mechanical, reprographic, recording or otherwise) without the prior written consent of the copyright owner.