Proceedings of the 126th European Study Group
Mathematics with Industry

SWI 2017

Amsterdam, January 23 – 27, 2017

Editors:
Daan Crommelin
Stella Kapodistria
Guus Regts
Chris Stolk
Peter van de Ven

Contents

Preface

These are the scientific proceedings of the 126th Study Group Mathematics with Industry (Studiegroup Wiskunde met de Industrie or SWI 2017), held at the Amsterdam Science Park campus, January 23-27, 2017. The SWI 2017 was co-organized by the Korteweg-de Vries Institute for Mathematics of the University of Amsterdam and Centrum Wiskunde & Informatica (CWI), the Dutch national research institute for mathematics and computer science in Amsterdam.

The proceedings are provided in two different formats. In this volume, the participants of SWI 2017 have provided their account of the week's developments, presented in English and aimed at a scientific audience. Each of the six groups has prepared a contribution that presents the problem they worked on, the approaches they attempted or used, and the results that they obtained.

Accompanying the present volume are the popular proceedings, written by science journalist Julia Cramer. These provide an account of the work meant for a general audience, written in Dutch.

The organisers of SWI 2017
Daan Crommelin, Stella Kapodistria, Guus Regts, Chris Stolk, Peter van de Ven

The fair value of a mortgage

Anatoliy Babič Mihail Bazhba Felix Beckebanze
Marko Dimovski Aldina Correia Eliana Costa e Silva
Marko Dimovski Robert Fitzner Stella Kapodistria*
David Koops Willem Moerkens
Irving van Heuven van Staereling Xiaoming op de Hoek
Bart Sevenster Jok Tang

Abstract

We consider the problem proposed by the Bank at the SWI 2017 meeting. In particular, the following directions of investigation were proposed:

Question 1: How should the bank calculate the fair value for it's current portfolio?

Question 2: What are the main drivers for prepayments?

Question 3: What is the biggest concern for the bank with the present low/negative rates regarding mortgages? How should the bank deal with this?

Question 4: What can the bank do to mitigate the risk of prepayments?

Question 5: What is the fair value of the banks portfolio (Dataset 2B)?

In this report, we deal with the mathematically oriented questions and we are interested in the modelling of the prepayments and their prediction.
KEYWORDS: mortgage, prepayments.

1 Introduction

Mortgages are an important tool in making housing available for people who do not have sufficient savings but still want to buy a house. In the Dutch economy, mortgages play an important role in consumer expenditure. The total household debt in the Netherlands in 2011 was 117% of GDP according to CBS, see (2). Household debt (including debt of non-profit institutions serving households) are the loans on the liabilities side of the balance sheet, excluding mutual household debts. While the non-financial corporations accounted for 107% of the GDP, (2). The bulk of household debts consists of residential mortgages. The mortgage debt has grown substantially between 1995 and mid-2012. After a period of decline between 2012 and 1014, household debts started rising as of September 2014, in particular the level of residential

*s.kapodistria@tue.nl

mortgage debt. The latter increased from 649 billion euros at the end of September 2014 to 669 billion euros at the end of June 2017. In the same period, non-mortgage debt rose from 88.5 to 91 billion euros. Despite increasing debt levels, household debt as a percentage of GDP declined in Q2 2017, because the increase in GDP was stronger than the increase in debt levels. At the end of June 2017, the household debt ratio had declined to 106.1 percent from 106.7 percent in March; the ratio has been falling since Q4 2012. In Q2 2017, the non-financial sector debt ratio declined from 114.1 to 112.7 percent. Total private sector debt amounted to 218.8 percent as a result, the lowest level since 2008.

Mortgages are for the bank an investment, as the bank loans the money in the form of a mortgage to the customer with a given interest rate. Although mortgages are a relatively secure investment opportunity due to their structured payment scheme and the interest rate which is fixed for a long time, mortgages do possess some risks for the issuer of the mortgage, typically the bank, which need to be taken into account when assessing the value of a mortgage. The most important risks related to mortgages are the risk of defaulting, which is the risk of customers not paying back their mortgage, and the interest rate risk, which is the risk caused by the uncertainty of future interest rates. These risks have been extensively researched both in practice and in the academic literature. A relatively less known risk for the issuer is that of the prepayment risk, which is the risk created by customers paying back (partially or in full) their mortgages earlier than the date stated in their contract. Prepayments can be done after the customers sell the house or refinance the mortgages, after which the mortgage ceases to exist. Refinancing can be lucrative if interest rates are low. It is also possible that customers pay back only a small percentage of the mortgage debt in order to reduce the interest that they have to pay, while staying under the extra instalments threshold so as to avoid penalties associated with prepayments.

A bank needs to meet its obligations: similarly to individuals, a bank needs to meet all its anticipated expenses, which in this case are the funding of loans, making payments on debt, etc. These payments need to be done by the bank using liquid assets, e.g., money. Ideally, a bank should maintain a level of liquidity that also allows it to meet any unexpected expenses without having to liquidate other assets. The bigger the cushion of liquid assets relative to anticipated liabilities, the greater the bank's liquidity. To this end, the bank makes a financial plan based on payments it will receive and the payments it needs to make. Thus, if many people simultaneously prepay their mortgage, the bank is thrown off this financial balance. Furthermore, because of the high costs of acquiring a mortgage, mortgages typically need to remain on the books for several years in order to be profitable.

A very recent trend noticed in the financial markets is that of the reduction of the average tenure of a residential mortgage (i.e., duration of a mortgage at one issuer), due mainly to consumers' increased willingness to switch lenders for a better deal. This is a worrisome trend for lenders, as they are exposed due to lack of liquidity and

due to the unprofitability of mortgages for the banks when they are prepaid at very short tenure.

In general, the following four drivers are shown to have a significant impact on prepayment behaviour, see (5):

- Age of the fixed rate loan – Typically, prepayment occurs after the end of the first fixed rate period.

- House price inflation – When house price inflation is high, the number of home moves increases. Increased activity in the housing market results in increased prepayment.

- Interest differential – This measures the tangible saving that a borrower could make by switching to another fixed rate or variable rate mortgage (typically by negotiating the mortgage anew with another bank - refinancing).

- Prepayment charges – These charges create a cost to prepayment that acts as a disincentive to prepay. We observed that charges over a certain level appeared to discourage prepayment significantly.

The goal of this report is to propose some models for the calculation of the fair value of a mortgage i.e. the value of a mortgage when taking prepayment risk into consideration.

Section 2 introduces the notation used in this analysis, Section 3 describes some common mortgage types and their characteristics. In Section 5, the risks for a bank are described. Section 6 introduces the way prepayments can be modelled and Section 7 introduces some models to incorporate prepayments in the valuation of a mortgage portfolio.

2 Prerequisites

Throughout this study, we consider mortgages with a contractual duration of M months, an initial coupon c (interest rate, expressed in % per month), and an initial principal P_0. For simplicity, we assume that the maximum duration for a mortgage is 30 years, i.e., $M \leq 360$. Furthermore, without loss of generality, we assume that the interest rate, c, is fixed throughout the entire contractual period, except if stated otherwise. We express the contractual payment (cashflow) by the borrower to the bank in month t by x_t. The amount of the monthly payment will be calculated in accordance to the type of mortgage.

3 Types of mortgages

Bullet mortgage

Bullet mortgages do not require the pay-off of the initial principal, P_0, throughout the contractual period. The monthly payment, x_t, thus only consists of the interest payment, i.e. $x_t = cP_0$ for all $t = 1, 2, \ldots, M-1$. The only exception is the last payment, $t = M$, in which the initial principal needs to be repaid, hence $x_M = P_0(1+c)$, see figure 1.

Linear mortgage

In a linear mortgage, the borrower repays the initial mortgage loan by a fixed amount every month. On top of this the borrower pays interest, but the interest payments will reduce over time since the borrower is gradually redeeming the initial loan. Since the mortgage amount will actually decrease, so will the interest payments. Say that the initial principle is P_0. Then, each month a payment of P_0/M is performed, plus the interest. So the monthly payment, x_t, is thus the sum of P_0/M plus the interest, cP_t, with $P_t = \frac{M+1-t}{M} P_0$. The monthly interest decreases every month, since the principle, P_t, decreases over time.

Straight line or level-pay or annuity mortgage

The characteristic of a straight line mortgage is that the monthly payment by the borrower is constant, x_* (assuming that there are no fluctuations in the mortgage interest rates). This means that initially the borrower pays a lot of interest, while the pay-off of the initial principle, P_0, is relatively small. This reverses towards the end of the mortgage term, when a smaller fraction of the monthly payment consists of interest payment. The constant monthly payment, x_*, can be calculated as follows

$$P_0 = \left(\sum_{t=1}^{M} \frac{x_*}{(1+c)^t} \right)^{-1} \Rightarrow x_* = \frac{cP_0(1+c)^M}{(1+c)^M - 1}. \tag{1}$$

The situation in the Netherlands

The Dutch housing market is shaped by four dominant forces, which stem from former political choices, see (3): (i) income tax deductibility of mortgage interest; (ii) a rental market in which not-for-profit social housing institutions have a combined market share of 84 percent; (iii) a scheme involving rent control and strong tenant protection; and (iv) a restrictive regulatory ('zoning') regime for the development and construction of new homes. The Dutch housing stock consists of 7.1 million houses, 56 percent of which are in the owner-occupied segment. This rate is below average in the eurozone. The low share of the owner-occupied segment in itself acts as

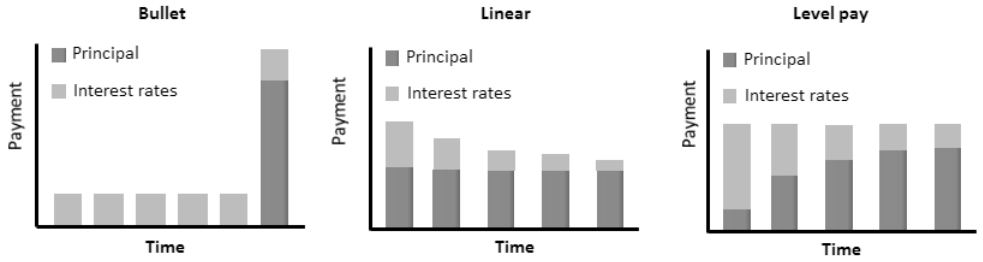


Figure 1: Cashflow (monthly payments) depiction for the various types of mortgages: Bullet, Linear and Level Pay mortgage

risk filter, since access to ownership is restricted to households with a good risk profile.

The Netherlands scores high in terms of mortgage debt. In fact, with a mortgage debt stock equalling 108% of the gross domestic product in 2012, see Figure 2. On the basis of this high debt burden, risks to the Dutch mortgage market are perceived to be elevated.

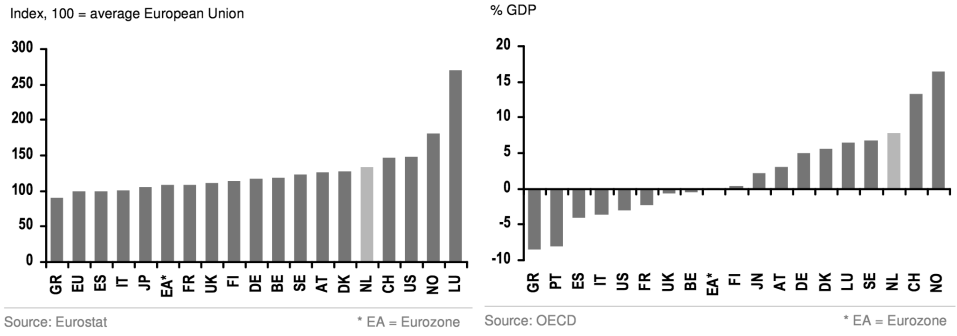


Figure 2: GDP per capita (2010), on the left, and current account balances (2011), on the right

In (1), the following conclusions were drawn for the Dutch mortgage market:

- The strong asset base of Dutch households and the full tax deductibility of interest payments are the primary reasons for the relatively high mortgage debt levels in the Netherlands.
- From 2013 onwards, it is expected that tax deductibility will only apply to amortising mortgage loans. Existing mortgage borrowers will be excluded from this change and continue to benefit from the existing tax regime. First-time

buyers will be hit by this change, which translates into higher net mortgage servicing costs.

- The Dutch housing market is unlikely to recover in the short run. Cyclical headwinds to the economy, a very low level of consumer confidence and structural changes to the housing and mortgage market are making people reluctant to buy a house. House prices are likely to decline further in the short run.

- Foreclosures rates are very low, especially in international comparisons. Although the modest recession is likely to result in an increase in the foreclosure rate, the resilient structure of the economy and the mortgage market will prevent a sharp increase in mortgage defaults.

Risks for the Dutch mortgage market

Till 2012, the main risks for the issuer of a mortgage loan were late payments and ultimately foreclosure. Late payments are generally managed well. Virtually all mortgage payments are automatically debited from current accounts. Payment failures are quickly discovered and notices are sent out usually within days. Statistics from various rating agencies show that mortgage arrears are very low in the Netherlands. By international comparison, both late payment and foreclosure rates are among the lowest in Europe.

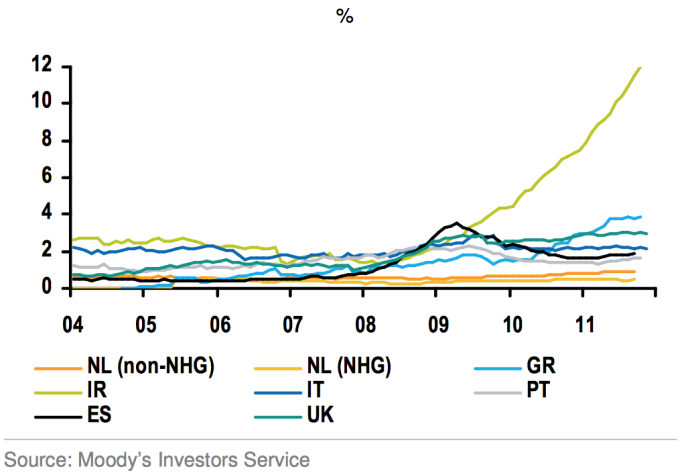


Figure 3: RMBS Prime 60+ days delinquency ratet

A very recent factor of risk is related to the housing price, (3). For about thirty percent of Dutch mortgages, the size of the mortgage exceeds the value of the un-

derlying property (4)[1]. These households suffer from negative home equity, their mortgages are 'underwater'. (4) shows that the underwater problem affects mainly younger households (20-40 years).

Key figures for the Dutch mortgage market

According to (3), the tax deductibility of mortgage interest has greatly influenced the Dutch mortgage market. It has encouraged 'interest only' mortgages, leading to high portfolio LTVs, and caused a large difference between LTIs based on gross and net income. Yet defaults and losses have remained very low, even in the recent crisis. The mortgage portfolio of lenders in the Dutch market consists of 3.5 million households: 83 percent of the 4.3 million Dutch homeowners carry a mortgage debt, with their property as collateral. In 2013, the total mortgage debt amounted to EUR 637 billion. The value of the housing stock amounts to EUR 1.07 billion.

4 Mortgage value

The objective of this study is to model and predict the *net present value* of a mortgage portfolio. The net present value of a mortgage portfolio is the sum of all future cash flows, $x_{i,t}$, of mortgage i at time t, discounted appropriately. In order to be able to generate income, the interest rate at which the bank borrows money, say r_t (expressed in % per month and called yield), must be smaller than the coupon rate, c_t, of the mortgages.

Let c_t denote the cashflow at time t, then its net present value at time 0 is given by $\frac{c_t}{(1+r_{\text{eff}})^t}$, assuming that r_{eff} is the effective interest rate (corresponding here to a monthly period). If the interest rate changes over time, then the net present value can be calculated by discounting every month and using the appropriate interest rate as they appear in the yield curve, see Figure 4.

We assume that the future cashflows from a borrower, c_t, are always larger or equal to the contractual payments, x_t. The exact amount of the additional payment, $d_t = c_t - x_t$, which is voluntary, depends on various parameters, which are discussed in more detail in §6. The net present value for a mortgage at time $t = 0$, (PV), with future cash flows c_t for $t = 1, .., M$ is given by

$$PV = \sum_{t=1}^{M} \frac{c_t}{(1 + r_{\text{eff}})^t} \tag{2}$$

[1]This figure is based on an approximated correction for mortgage-related savings (cf. Chapter 3). Without this correction, and including non-bank loans, the percentage of underwater loans amounts to 41 percent (2).

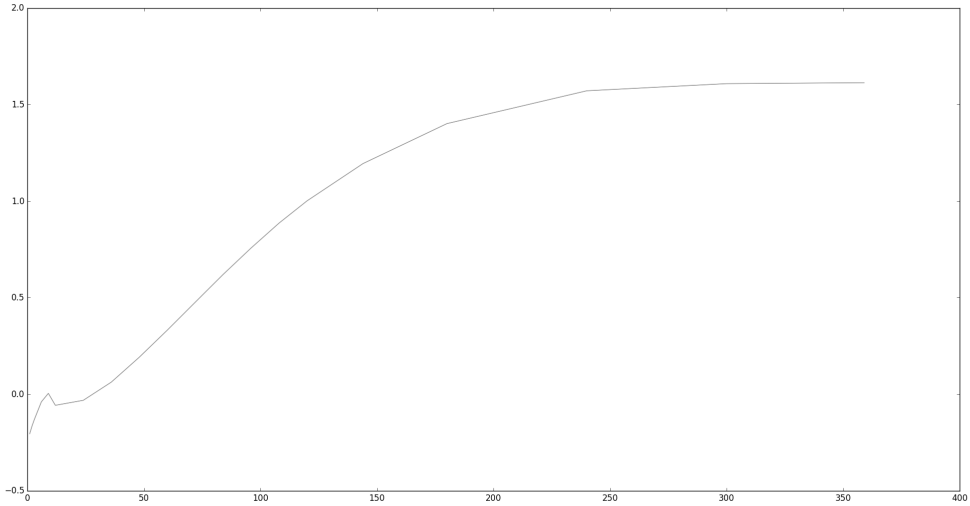


Figure 4: *Interpolated yield curve, 01-01-2016. Yield versus number of months (0 to 360).*

for a constant yield r_{eff}. If $r = r(t)$ varies over time, then

$$PV = \sum_{t=1}^{M} c_t \prod_{t'=1}^{t} \frac{1}{1 + r(t')}.$$

5 Risks of a mortgage portfolio

The net present value of a mortgage portfolio is the value of each discounted future cash flow of the mortgages. However, there are several risks that make valuation of the portfolio complicated; some of these risks are listed below.

- **Default risk** Default risk is the risk of customers not paying back (part of) their mortgages. This can be the result of people losing their jobs or having other problems with income which means they do not have enough money left to pay off their mortgages. Another reason is people dying before the end of the mortgage term.
- **Interest rate risk** Interest rate risk is the risk that banks have because of changing market interest rates and with it changing income and expenses. The interest that the banks receive on the loans and mortgages they have and the interest they have to pay on the savings accounts of their customers depend heavily on the market interest rate. A change in the market interest rate therefore influences the ratio of income and expenses.

- **Prepayment risk** Related to interest rate risk is the prepayment risk. Prepayment risk is the risk that customers pay back their loans or mortgages earlier than their contract says they have to, which means that the bank misses out on interest they would have received if no prepayment had been done. The prepayment risk depends on the interest rate since lower interest rates can make it more profitable for customers to pay back part of their mortgages, since the money they put on a bank account does not yield enough interest any more.

- **Reputational risk** Reputational risk is the risk that a bank has based on its reputation as a reliable business partner. If the bank has a bad reputation, customers may not want to use any of the financial products that the bank has to offer and instead go to one of its competitors.

- **Operational risk** Operational risk is the risk the bank has in its operations. According to Solvency 2 operations risk is "the risk of a change in value caused by the fact that actual losses, incurred for inadequate or failed internal processes, people and systems, or from external events (including legal risk), differ from the expected losses".

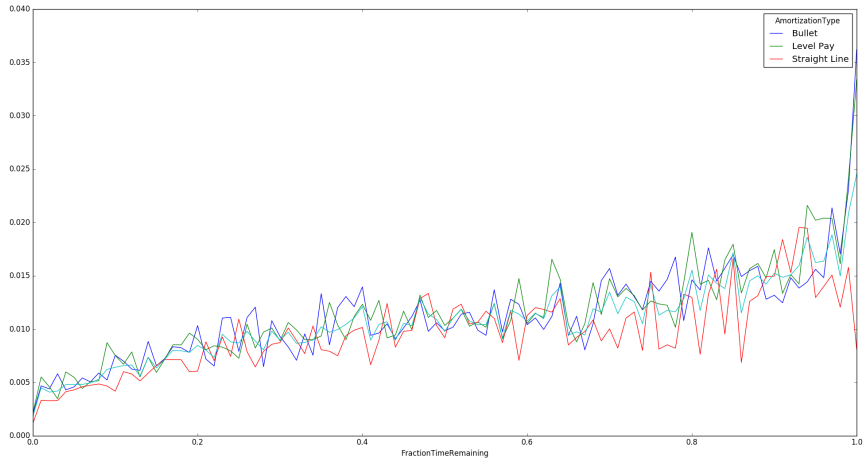


Figure 5: *Average fraction of outstanding principal that is prepaid, versus the fraction of time that is remaining until maturity, for each different type of mortgage (bullet, level pay, straight line). The light blue plot shows the average over all of these types.*

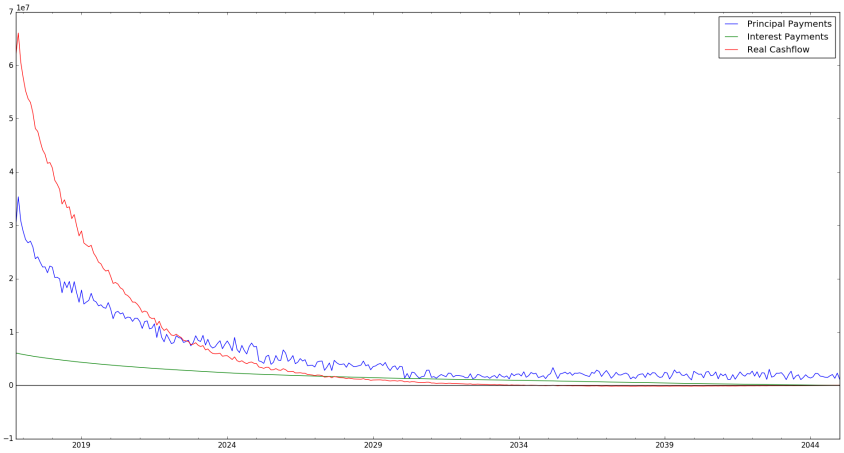


Figure 6: *The red curve is the (expected real (non-discounted) cash flow, based on prepayment model. The blue curve are the remaining principals according to the contractual payments. Green represents the interest payments.*

6 Prepayment modelling

In this section we zoom in on the prepayment risk, which is a significant risk in a mortgage portfolio. In order to manage this risk, it is important to have insight into the prepayment behaviour of clients. Therefore, we used the provided data[2] to find the relation between the duration (defined as the time in months since the start of the amortisation), and the average prepayment rate (defined as the prepayment made as a fraction of the outstanding principal, averaged over all contracts). This relation is represented in Figure 5. We see that in the beginning of the period, people typically prepay a smaller amount of their outstanding principal. We can fit a curve through this (for each type of mortgage, or just one curve if the differences are negligible). This curve can be used to generate the future expected cashflows, which consists of the contractual cashflows, in addition to the expected prepayments. The result of this is depicted in Figure 6. As a final step, we have to discount these cashflows by making use of the yield curve in order to obtain the present net value. The yield curve that we are provided with for the present time, does not contain a value for each month, but there are gaps. Therefore we interpolate the yield curve, after which we can use

[2]Due to confidentiality reasons and in order to guarantee the anonymity of the Bank that provided us with the data set, we cannot describe the data, but can only present some results obtained from the data set. We would also like to note that these results might be specific to the dataset we analysed and might not be replicable for other data sets, as the data set provided to us was synthetic.

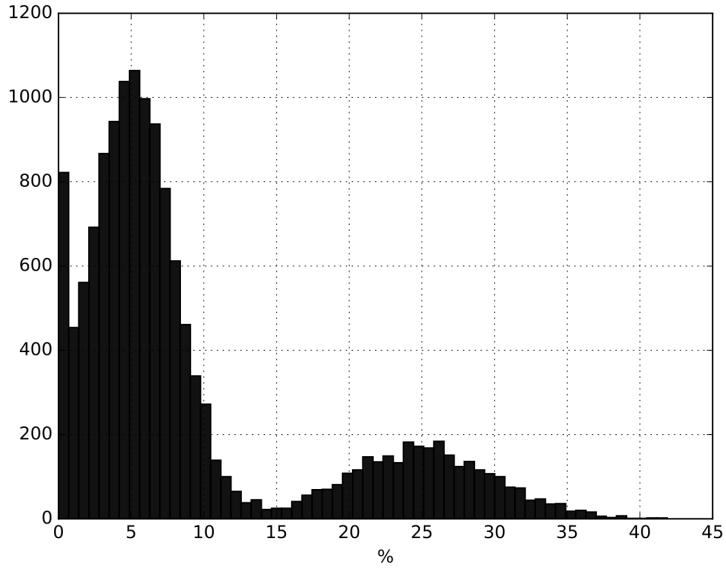


Figure 7: *Each bar represents the count of the number of people who prepay the corresponding percentage of their outstanding principal.*

it for discounting the future cash flows. The interpolated yield curve is depicted in Figure 4.

7 Improvements of the model

We could make improvements on the model by not only distinguishing between the different type of mortgages, but by also considering different types of prepayments (i.e., curtailment, relocation, refinancing). Within curtailment, we can make a further distinction, which is suggested by Figure 7. The low number of people who make prepayments at around 15% of the outstanding principal can be explained by the fact that there is a penalty inflicted on people who pay off more than 15% of their outstanding principal. Therefore, people who would like to make a curtailment, either avoid it by paying less than 15%, and if they really want to make a larger curtailment they make it worthwhile by substantially overshooting the 15% barrier.

8 Relocation distribution – Fitting Approach

This study focuses on the relocation cases for the mortgage portfolio of the bank, with interest time (maturity) equal to 360 months. By visual inspection, we consider the Gamma distribution as the best candidate to fit the data of the variable *ElapsedTime*,

see Figure 8. This variable represents the number of months elapsed at the time a client decides to relocate this mortgage. Other distributions were tested, such as the log-normal and beta, but the gamma distribution had the best fit.

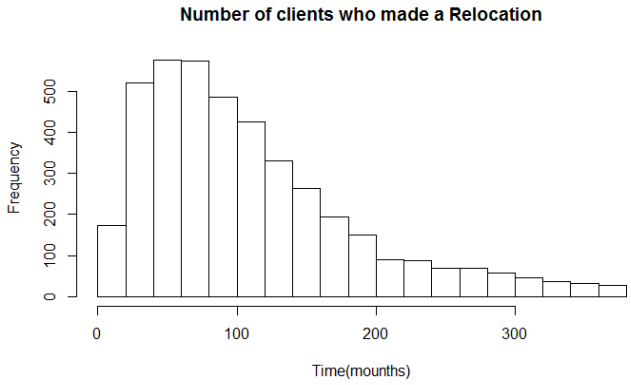


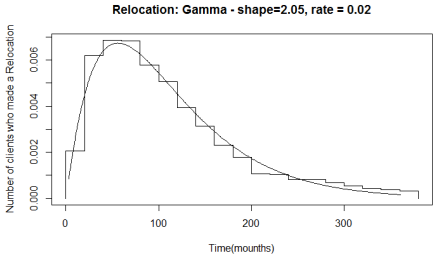
Figure 8: Histogram of the variable "Number of Months" accounting for the months that have elapsed since the beginning of the mortgage

Let $\mathbf{Y}$ be a random variable with the Gamma distribution with parameters ν and ν/μ, symbolically $\mathbf{Y} \sim Ga(\nu, \nu/\mu)$, then the density function of this random variable is written as

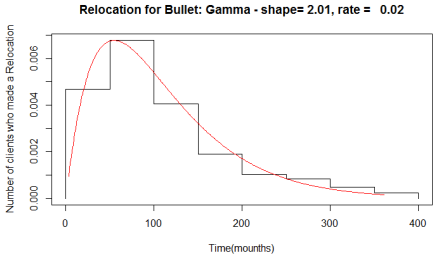
$$\begin{aligned} f(y|\nu,\mu) &= \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1}\exp\left(-\frac{\nu}{\mu}y\right) \\ &= \exp\left\{\nu(\theta y + \log(-\theta)) + (\nu-1)\log(y) - \log(\Gamma(\nu)) + \nu\log(\nu)\right\}, \end{aligned}$$

with $y > 0$, $\theta = -1/\mu$ and Γ the gamma function.

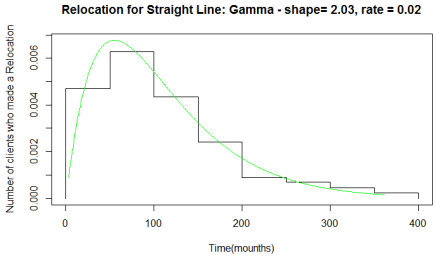
We can divide the data in the mortgage portfolio in 3 groups distinguishing for the three types of mortgages: Bullet (1401 clients), Level Pay (1404 clients) and Straight Line (1400 clients). In the first fit, we consider all three types of mortgages, obtaining the parameter estimates $\gamma = 2.0492072693$ and $\gamma/\mu = 0.0186694832$, see Figure 9a. If we consider just one type of mortgage, e.g. Bullet, we can repeat the fitting and obtain very similar results. More concretely, we obtain the following parameter estimates $\gamma = 2.0104627616$ and $\gamma/\mu = 0.0185100704$, see Figure 9b. If we consider the Straight Line, we obtain the following parameter estimates $\gamma = 2.0288771200$ and $\gamma/\mu = 0.0186201469$, see Figure 9c, and for Straight Line, we have the following estimates $\gamma = 2.1129085041$ and $\gamma/\mu = 0.0189148191$, see Figure 9d.



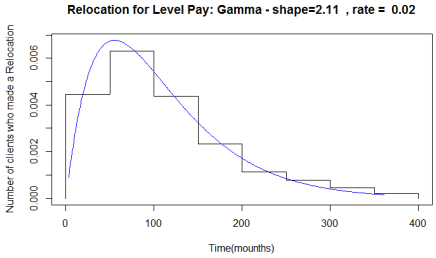
(a) $\gamma = 2.0492072693$ and rate $\gamma/\mu = 0.0186694832$



(b) Gamma $\gamma = 2.0104627616$ and rate $\gamma/\mu = 0.0185100704$ shaped histogram and fit curve - Bullet Amortisation Type



(c) Gamma $\gamma = 2.0288771200$ and rate $\gamma/\mu = 0.0186201469$ shaped histogram and fit curve - Straight Line Amortisation Type



(d) Gamma $\gamma = 2.1129085041$ and rate $\gamma/\mu = 0.0189148191$ shaped histogram and fit curve - Level Pay Amortisation Type

Figure 9: Fitting the data to the Gamma distribution with parameters ν and ν/μ

Using the well known Kolmogorov-Smirnov test (KS), we tested the similarities between the various estimates for the parameters comparing the case that all the data is aggregated versus the cases that we distinguish for each type of mortgage. In all cases, a significance level $>> 5\%$ was obtained, therefore we can consider that there exist no significant differences between the 4 distributions. Thus, the distribution of the months since the beginning of the mortgage to relocation does not depend on the type of amortisation and the same probability distribution can be used for all the amortisation types.

9 Relocation distribution – Survival Approach

In this section, we investigate the distribution of the relocation lifetime using notions from survival analysis. More concretely, we use the Kaplan-Meier (KM) estimator, which is a non-parametric statistic for the estimation of the survival function from

the relocation amortisation lifetime data. In the present analysis, considering as zero the instant in time at which the mortgage is contracted, the KM estimator gives, at each month after the begging of the mortgage, an estimation of the probability that a client has not relocated. The obtained KM estimator is illustrated in Figure 10.

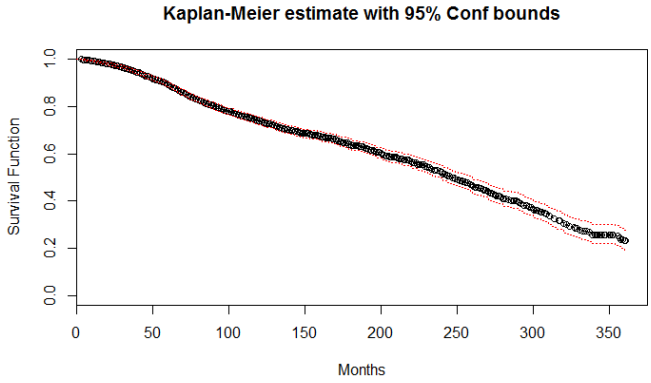


Figure 10: Kaplan - Meier estimator

References

[1] Ruben van Leeuwen and Philip Bokeloh (2012). *Mortgage Market in the Netherlands*, ABN AMRO Bank N.V., The Netherlands.

[2] CBSlink - Financial balance sheets and transactions by sectors; National Accounts http://CBSlink.cbs.nl/ (link).

[3] Dutch Banking Association (2014). *The Dutch Mortgage Market*, Nederlandse Vereniging van Banken, The Netherlands.

[4] Dutch Central Bank (2014). *Overview of Financial Stability*, Amsterdam: DNB.

[5] Simon Perry, Stuart Robinson, and John Rowland (2001). A study of mortgage prepayment risk. *Housing Finance International*, 16(2), 36–51.

Optimal dike heights around the IJsselmeer

Aida Abiad * Sander Gribling† Domenico Lahaye‡
Matthias Mnich§ Guus Regts¶ Lluis Vena‖

Abstract

In this note we show that the polytope associated with the IP model introduced by Zwaneveld and Verweij (6) is not integer. We also prove that, for a fixed number of dike segments, the problem can be solved in polynomial time. Similarly, we show that for a fixed number of allowed barrier heights, the problem can be solved in polynomial time.

1 Introduction

Protection against increasing sea levels is an important issue around the world. Optimal dike heights are of crucial importance to the Netherlands as almost 60% of its surface is under threat of flooding from sea, lakes, or rivers. This area is protected by more than 3500 kilometers of dunes and dikes. These dunes and dikes require substantial yearly investments of more than 1 billion euro (5).

Recently, Zwaneveld and Verweij (6) gave an integer programming model for a cost-benefit analysis to determine optimal dike heights that allows input-parameters for flood probabilities, damage costs and investment costs for dike heightening. The model by Zwaneveld and Verweij (6) is an improvement of the model proposed by Brekelmans et al. (1), who presented a dedicated approach with no optimality guarantee, and which was in turn an improvement of the original model by van Dantzig (4) from 1956. The latter was introduced after a devastating flood in the Netherlands in 1953.

*Department of Quantitative Economics, Maastricht University, Maastricht, The Netherlands (A.AbiadMonge@maastrichtuniversity.nl).

†CWI, Amsterdam, The Netherlands (gribling@cwi.nl)

‡Delft Institute for Applied Mathematics, The Netherlands (d.j.p.lahaye@tudelft.nl)

§Department of Quantitative Economics, Maastricht University, Maastricht, The Netherlands (mmnich@uni-bonn.de).

¶Korteweg de Vries Institute for Mathematics, University of Amsterdam, The Netherlands (guusregts@gmail.com)

‖Korteweg de Vries Institute for Mathematics, University of Amsterdam, The Netherlands (lluis.vena@gmail.com)

Our work is based on the IP model presented in a recent manuscript by Zwaneveld and Verweij (6), where the authors study the problem of economical optimal flood prevention in a situation in which multiple barriers dams and dikes protect the hinterland to both sea level rise as well as peak river discharges. Current optimal flood prevention methods (Kind (3), Brekelmans et al. (1)) only consider single dike ring areas with no interdependency between dikes. Zwaneveld and Verweij (6) present a model for a cost-benefit analysis to determine optimal dike heights with multiple interdependencies between dikes and barrier dams, and they also show that it can be solved quickly to proven optimality. The model was presented at the Study group Mathematics and Industry (SWI), taking place in Amsterdam in the last week of January 2017. It was our task at SWI 2017 to give a better understanding of the mathematical complexity of the model proposed by Zwaneveld and Verweij. The present report summarizes our approaches and results that were obtained during the week that SWI took place and the weeks after it.

We will follow the notation used in Zwaneveld and Verweij (6). Before going into the details of the problem, let us introduce some important terminology and the geographical configuration of the dikes in the Netherlands. A *dike segment* is a part of a dike that is protecting a region. It is possible that several segments protect the same area and in that case they are called a *dike ring*. In the Netherlands, dike ring areas and smaller dikes lie beneath the *Afsluitdijk*, sometimes denoted by the *barrier dam*, which is the most outer dike located in the north. The Afsluitdijk separates the North Sea and the IJsselmeer, an artificial lake.

In this paper we show that the polytope associated to the IP model introduced by Zwaneveld and Verweij (6) is not integer. Moreover, we present some sufficient conditions that allows the linear relaxation of the integer programming to avoid these non-integral points. We also prove that, for a fixed number of dike segments, the problem can be solved in polynomial time. Similarly, we show that for a fixed number of allowed barrier heights, the problem can be solved in polynomial time. This paper is organized as follows. In Section 2 we introduce the IP model that forms the subject of our investigations. In Section 3 we discuss integrality of the polytope. In Section 4 we propose an alternative approach to solve the problem by means of dynamic programming. Finally, in Section 5 we present a natural abstract version of the dike height problem, which allows for several variations and open problems.

2 IP Model formulation

In this section we present the model formulated in (6).

Throughout we use the following notation:

- D is the set of dike segments.
- H_D is the set of possible heights for a dike segment. For ease of notation, we do not let H_D depend on the dike segment, i.e., all dike segments have the same set of possible heights. We denote the height of a previous year by h_1, and that

of the current year by h_2. Likewise, H_B is the set of possible heights for the barrier dam and we denote the height of the barrier in the previous year by h_1^B, and that of the current year by h_2^B.

- T is the set of time periods at which changes to a dike segment can be made (e.g., one can assume that changes are scheduled per year), for simplicity we assume (with abuse of notation) $T = \{0, 1, \ldots, T\}$.

The decision variables are:

- $CY(t, d, h_1, h_2) \in \{0, 1\}$. The variable being one meaning that dike ring d is updated in time period t from height h_1 up to height h_2. If $h_1 = h_2$ then this dike ring segment is not strengthened in period t and remains at its previous height. This decision variable is used for tracking investment (and maintenance) costs.

- $DY(t, d, h_2, h_2^B) \in \{0, 1\}$. It is one if at the end of period t the barrier dam has height h_2^b, and dike segment d is of height h_2. This variable is used to connect investments in dike segments (and the barrier dam) to expected damages. Another way to view it is that this variable linearizes the 0-1 variable $\left(\sum_{h_1} CY(t, d, h_1, h_2)\right)\left(\sum_{h_1^B} B(t, h_1^B, h_2^B)\right)$.

- $B(t, h_1^B, h_2^B) \in \{0, 1\}$. It is one if the barrier dam (i.e., the Afsluitdijk) is updated in time period t from height h_1^B up to h_2^B. If $h_1^B = h_2^B$ then the barrier dam is not strengthened in period t and remains at its previous height. This decision variable is used for bookkeeping investment (and maintenance) costs, flood probabilities and related expected damage costs of the barrier dam.

The input parameters are:

- $Dcost(t, d, h_1, h_2)$ = costs for investment and maintenance, if dike ring d is strengthened in time period t from h_1 to h_2. If $h_1 = h_2$, the dike ring segment is not strengthened and these costs only represent maintenance costs.

- $Dexpdam(t, d, h_2, h_2^B)$ = expected damage, i.e.,

$$Dexpdam(t, d, h_2, h_2^B) = prob(t, d, h_2, h_2^B) \times damage(t, d, h_2, h_2^B),$$

  where $prob(t, d, h_2, h_2^B)$ and $damage(t, d, h_2, h_2^B)$ are respectively the probability of failure and the expected damage cost (the latter given that there is a flooding) in period t given the height of the segment h_2 and the height of the barrier h_2^B. Note that it is assumed that both the probability of failure and the expected damage upon failure of dike segment d only depend on the height of segment d and that of the barrier dam.

- $Bcost(t, d, h_1^B, h_2^B)$ = costs for investment and maintenance, if the barrier dam is strengthened in time period t from h_1^B to h_2^B. If $h_1^B = h_2^B$, the barrier dam is not strengthened and these costs only represent maintenance costs.

- $Bexpdam(t, h_2^B)$ = expected damage of a flooding of the barrier dam, i.e. $prob(t, h_2^B) \times damage(t, h_2^B)$, here $prob(t, h_2^B)$ and $damage(t, h_2^B)$ are respectively the probability of failure and the expected damage cost (the latter given that there is a flooding), in period t given the height of the barrier h_2^B.

All input parameters are calculated in net present value of a certain year (i.e. 2015, which is the starting year for our calculations) and represent price levels in a certain year.

The IP model is:

minimize

$$\sum_{t\in T}\sum_{d\in D}\sum_{h_1\in H_D}\sum_{h_2\geq h_1} Dcost(t,d,h_1,h_2)\cdot CY(t,d,h_1,h_2) + \tag{1}$$

$$\sum_{t\in T}\sum_{d\in D}\sum_{h_2\in H_D}\sum_{h_2^B} Dexpdam(t,d,h_2,h_2^B)\cdot DY(t,d,h_2,h_2^B) + \tag{2}$$

$$\sum_{t\in T}\sum_{h_1^B\in H^B}\sum_{h_2^B\geq h_1^B}\Big(Bcost(t,h_1^B,h_2^B)+Bexpdam(t,h_2^B)\Big)\cdot B(t,h_1^B,h_2^B) \tag{3}$$

subject to

$$CY(0,d,0,0)=1, CY(0,d,h_1,h_2)=0 \quad \forall d\in D, h_1,h_2\in H_D, h_2\geq h_1 \wedge h_2>0 \tag{4}$$

$$\sum_{h_1\leq h_2} CY(t-1,d,h_1,h_2) = \sum_{h_3\geq h_2} CY(t,d,h_2,h_3) \quad \forall t\in T_{>0}, d\in D, h_2\in H_D \tag{5}$$

$$\sum_{h_1\leq h_2} CY(t,d,h_1,h_2) = \sum_{h_2^B} DY(t,d,h_2,h_2^B) \quad \forall t\in T, d\in D, h_2\in H_D \tag{6}$$

$$B(0,0,0)=1, B(0,h_1^B,h_2^B)=0 \quad \forall h_1^B,h_2^B\in H_B, h_2^B\geq h_1^B\wedge h_2^B>0 \tag{7}$$

$$\sum_{h_1^B\leq h_2^B} B(t-1,h_1^B,h_2^B) = \sum_{h_3^B\geq h_2^B} B(t,h_2^B,h_3^B) \quad \forall t\in T\backslash\{0\}, d\in D, h_2^B\in H_B \tag{8}$$

$$\sum_{h_1^B\leq h_2^B} B(t,h_1^B,h_2^B) = \sum_{h_2} DY(t,d,h_2,h_2^B) \quad \forall t\in T, d\in D, h_2^B\in H_B \tag{9}$$

$$CY(t,d,h_1,h_2)\in\{0,1\} \quad \forall t\in T, d\in D, h_1\in H_D, h_2\geq h_1\in H_D \tag{10}$$

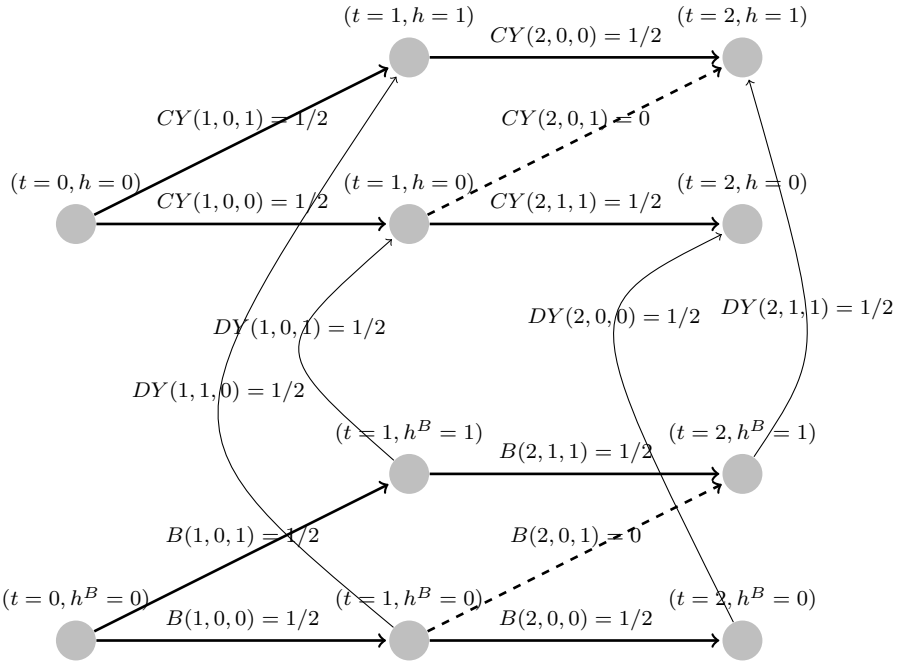
$$DY(t,d,h_2,h_2^B)\in\{0,1\} \quad \forall t\in T, d\in D, h_2\in H_D, h_2^B\in H_B \tag{11}$$

$$B(t,h_1^B,h_2^B)\in\{0,1\} \quad \forall t\in T, d\in D, h_2^B\geq h_1^B\in H_B \tag{12}$$

3 On the integrality of the polytope

In this section we show that, in general, there are vertices of the polytope defined by the linear relaxation of the constraints (when the integer values are considered to be in the interval $[0,1]$ instead of $\{0,1\}$), that have non integral coordinates.

Figure 1: Example of non-integer point.



The example involves the following sets indexing the variables.

- $T = \{0, 1, 2\}$
- one segment. Hence, we remove the dike index from all related variables.
- $H = \{0, 1\}$, $H_B = \{0, 1\}$

The point P, candidate to be a vertex of the polytope of the linear relaxation, has the following non-zero values:

- $CY(t, h_1, h_2)$: $CY(0,0,0) = 1$, $CY(1,0,1) = 1/2$, $CY(1,0,0) = 1/2$, $CY(2,1,1) = 1/2$, $CY(2,0,0) = 1/2$.
- $B(t, h_1, h_2)$: $B(0,0,0) = 1$, $B(1,0,1) = 1/2$, $B(1,0,0) = 1/2$, $B(2,1,1) = 1/2$, $B(2,0,0) = 1/2$.
- $DY(t, h_2, h_2^B)$: $DY(0,0,0) = 1$, $DY(1,0,1) = 1/2$, $DY(1,1,0) = 1/2$, $DY(2,1,1) = 1/2$, $DY(2,0,0) = 1/2$.

The example is summarized in Figure 3 where each arrow corresponds to one of the decision variables.

One can check that the example is a feasible solution (a point in the polytope). Indeed, the flow conditions are verified, as well as the equations linking the dummy variables DY and the CY's and B's (Equations (6) and (14)).

To argue that the point P is indeed a vertex of the polytope, we show that, for every line with non-zero direction vector $v = (x_0, \ldots, x_{14})$, and for every $\epsilon > 0$, either $P + \epsilon v$ or $P - \epsilon v$ is outside the polytope. Every coordinate x_i of v corresponds, uniquely, to a variable $B(\cdot)$, $CY(\cdot)$, or $DY(\cdot)$.

First observe that if x_i is the coordinate related to a variable that is either 0 or 1 in P, then $x_i = 0$, as otherwise, for any ϵ, either $P + \epsilon v$ or $P - \epsilon v$ would be outside of the polytope. Hence, the only x_i that may be non-zero, are those for which the coordinate i in P is in the open interval $(0, 1)$.

In our example, every equation involves at most 2 variables on each side of the equality, one of them being either 0 or 1. Hence the implications written below are forced by the previous observation. Assume, for instance, that the coefficient x_i corresponding to $B(2, 1, 1)$ in v is negative.

- Then, by the flow constraints (Equation (8)), the coefficient of $B(1, 0, 1)$ is negative.
- Then, by the flow constraints, the coefficient of $B(1, 0, 0)$ is positive.
- Then, by the flow constraints, the coefficient of $B(2, 0, 0)$ is positive.

Now, using the equations that link the variables B and DY, we obtain that the the coefficient of $DY(2, 1, 1)$ is positive, which implies that

- the coefficient of $CY(2, 1, 1)$ in v is positive.
- Then, by the flow constraints, the coefficient of $CY(1, 0, 1)$ is positive.
- Then, by the flow constraints, the coefficient of $CY(1, 0, 0)$ is negative.
- Then, by the flow constraints, the coefficient of $CY(2, 0, 0)$ is negative.

Observe now that this implies that the coefficient of $DY(2, 0, 0)$ has to be negative. However, let us now look at the coefficients of $DY(1, 0, 1)$ and the one corresponding to $DY(1, 1, 0)$.

If we use the links between the variables DY and B, the coefficients corresponding to the variables $DY(1, 0, 1)$ and $DY(1, 1, 0)$ in v have to be negative and positive respectively. However, if we look at the equations linking the variables DY and CY, the signs of the coefficients should have the opposite sign. Thus, these coefficients should be zero, implying that all the other coefficients have to be 0, which shows that no non-zero vector v exists.

The first coefficient involved in the argument was the one involving the variable $B(2, 1, 1)$. Since the implications described here involve all the non-zero variables of the point, and the implications are reversible, the result now follows.

3.1 Avoiding the non-integral points

We present here a sufficient condition on the objective function (1)–(3), that guarantees that either the linear relaxation of the integer program finds an integral point as a solution, or that there is an integral point in the optimal face and a procedure to find it.

Proposition 1. *Assume that, for every $h_2 \leq h_2'$ and $h_2^B \leq h_2'^B$ the objective function satisfies:*

$$D_{expdam}(t,i,h_2',h_2^B) + D_{expdam}(t,i,h_2,h_2'^B) \geq D_{expdam}(t,i,h_2,h_2^B) + D_{expdam}(t,i,h_2',h_2'^B) \tag{13}$$

and that, if $h_1 \leq h_1'$ and $h_2 \leq h_2'$, then, for every t,

$$B_{cost}(t,h_1,h_2') + B_{cost}(t,h_1',h_2) \geq B_{cost}(t,h_1,h_2) + B_{cost}(t,h_1',h_2') \tag{14}$$

and, for every t and d,

$$D_{cost}(t,d,h_1,h_2') + D_{cost}(t,d,h_1',h_2) \geq D_{cost}(t,d,h_1,h_2) + D_{cost}(t,d,h_1',h_2') \,. \tag{15}$$

Then, there is an optimal solution of the linear relaxation of the IP model in Section 2 with integer coordinates.

Proof of Proposition 1. The problem from Section 2 can be thought of as several intertwined min-cost flow problems (see Section 5), one for each dyke, and one for the barrier.

Let x_0 be a solution point given by the linear relaxation, and assume it is non-integral. Using the monotone relations (14) and (15), the paths of the non-zero flows that x_0 defines for each of the dykes and the barrier can be assumed to be completely ordered (as otherwise, the flow values on the edges might be modified while maintaining the value of the in flow and out flow at each vertex while not increasing the objective function). So, we obtain a layered flow, where no two flow-paths strictly cross between two layers of vertices corresponding to two different consecutive times. In particular, for each of the dykes d, we can talk about a top path U_d (the height profile being always larger or equal than all the other height profiles), and a bottom path L_d, whose heights are smaller or equal than all the other height profiles. There is also a top U_B and bottom L_B paths for the flow of the barrier.

Observe that, as x_0 is non-integral, at least one of the variables DY is non-integral (either not equal to zero or not equal to one). Let $DY_{\min}$ be the minimal distance of the non-integral variables to either 0 or 1.

Using (13) as a guideline repeatedly, we modify the variables DY from x_0 to create a new feasible solution x_1 in which the variables $DY(t,i,h_2,h_2^B)$ are "untangled". In particular, we can assume that

$$\begin{aligned} &DY_{x_1}(t,i,h_2(U_i),h_2^B(U_B)) = \\ &\qquad = \min\left\{ \sum_{h_2} DY_{x_0}(t,i,h_2,h_2^B(U_B)), \sum_{h_2^B} DY_{x_0}(t,i,h_2(U_i),h_2^B) \right\} \end{aligned}$$

and that

$$DY_{x_1}(t,i,h_2(L_i),h_2^B(L_B)) = \\ = \min\left\{\sum_{h_2} DY_{x_0}(t,i,h_2,h_2^B(L_B)), \sum_{h_2^B} DY_{x_0}(t,i,h_2(L_i),h_2^B)\right\}$$

by reassigning some mass of the variables DY that are crossed. The remaining variables of x_0 are kept equal in x_1. The reassignment is done in a way to preserve the flow constraints, so x_1 remains feasible. By (13), x_1 has the same objective value as x_0, since x_0 is optimal.

Let $F_{\min}$ be the minimal difference to 0 or 1 of the flow through each L_d, U_d for every dyke d and L_B or U_B, which can be assumed to be the minimal value of

$$\min_{t,i}\left\{DY_{x_1}(t,i,h_2(U_i),h_2^B(U_B)), DY_{x_1}(t,i,h_2(L_i),h_2^B(L_B))\right\}$$

We note that x_1 is not a vertex of the polytope. Indeed, for any dyke d, we can pair up $L_d \leftrightarrow L_B$ and $U_d \leftrightarrow U_B$. Using (14) and (15), this pairing is well defined and consistent. In particular, we can redirect an ϵ flow ($0 < \epsilon \leq F_{\min}$) from each of the L_d to U_d and from L_B to U_B, or viceversa (the redirection of the flow should be done on each of the paths simultaneously, either from upper to lower paths, or from lower to upper ones). Since there exists a d (or B) for which the paths L_d and U_d differ, this flow-redirection by ϵ gives a different point on the polytope of feasible points and shows that x_1 is not a vertex of the polytope.

Furthermore, for every $\epsilon > 0$, the mentioned flow redirection should give the same value of the objective function (since otherwise x_0 would not have been an optimal solution). Hence we can choose to redirect the flow at our convenience; we redirect it so that the edge whose flow-value is $F_{\min}$ becomes either 0 or 1 (depending on whether its value is closer to 0 or to 1, if $F_{\min} = 1/2$, we arbitrarily redirect the flow either way). In particular, we have obtain a new solution x_2 where the number of edges with non-integral flow has been reduced, at least, by one. This procedure can be iterated until no non-integral flows are found. Therefore, an integral vertex of the polytope in the optimal face of the linear relaxation of the integer program is found. □

4 Alternative approaches

A feasible solution to the integer program presented in Section 2 can be interpreted as a choice of height $h^d(t)$ for each dike segment at each time period t, and a height $h^b(t)$ of the barrier dam. Abstractly, the cost of these height series can be written as a sum of cost terms which depend only on the 'upgrade' done in period t to segment d (i.e., a heightening of the dike, or merely the maintenance cost), we denote this by $\mathsf{cost}^d(h^d(t-1),h^d(t),t)$ for segment d, and by $\mathsf{cost}^b(h^b(t-1),h^b(t),t)$ for the barrier. Finally, there is also an expected damage cost for upgrading the dike and barrier to heights $h^d(t)$ and $h^b(t)$ in period t, denoted by $\mathsf{dam}^{d,b}(h^b(t),h^d(t),t)$. The problem

modeled in Section 2 can thus be written in the following way:

$$\begin{aligned}
&\mathsf{d}-\mathsf{opt} = \\
\min\Big\{ &\sum_{t\in[T]} \mathsf{cost}^b(h^b(t-1),h^b(t),t) + \sum_{d\in D} \mathsf{cost}^d(h^d(t-1),h^d(t),t) + \mathsf{dam}^{d,b}(h^b(t),h^d(t),t) \\
&\quad \text{s.t.} \quad h^d(t) \in H_D, h^b(t) \in H_B \text{ for } d \in D, t \in T \\
&\qquad\quad h^d(t) \geq h^d(t-1) \text{ for } d \in D, t \in T \\
&\qquad\quad h^b(t) \geq h^b(t-1) \text{ for } t \in T \Big\}
\end{aligned}$$

The linear relaxation of the integer programming model presented in Section 2 can be solved in time polynomial in $|D|, |T|, |H_D|$, and $|H_B|$. However, in general there is no guarantee that the returned solution is integral, see Section 3. In the next two sections we describe two different approaches to solving this problem. Both approaches have the benefit of solving the integer problem exactly. However, this comes at a cost: both approaches give a polynomial time algorithm only if one of the parameters is regarded as a constant. The first approach is to solve the integer program by ways of a dynamic program. The second approach comes down to enumerating all possible height profiles of the barrier dam, and for each profile solving shortest path problems on small graphs.

4.1 Dynamic programming

There are two key observations to be made. First, the second part of the objective function decomposes naturally into a sum of $|D|$ terms, each of which depends only on the barrier height and one segment. Secondly, for each time period the cost only depends on the dike/barrier heights at times $t-1$ and t. Together this allows us to solve the problem using a dynamic program. The recursion will be on the time period. We maintain the following table: $\mathsf{opt}(h^b, \mathbf{h^s}, t)$ for all $t \in T, h^b \in H_B, \mathbf{h^d} \in (H_D)^D$. The interpretation is as follows, $\mathsf{opt}(h^b, \mathbf{h^d}, t)$ is equal to the minimum cost made, up to time t, if the barrier and segments are of height h^b and $\mathbf{h^d}$ at time period t respectively. We can compute the entries of this table as follows:

$$\begin{aligned}
\mathsf{opt}(h^b, \mathbf{h^d}, t) = \min\Big\{ &\mathsf{opt}(h^b - i^b, \mathbf{h^d} - \mathbf{i^d}, t-1) + \mathsf{cost}^b(h^b - i^b, h^b, t) + \\
&\mathsf{cost}(\mathbf{h^d} - \mathbf{i^d}, \mathbf{h^d}, t) + \mathsf{dam}(h^b, \mathbf{h^d}, t) : \\
&h^b - i^b \in H_B, \mathbf{h^d} - \mathbf{i^d} \in (H_D)^{|D|} \Big\}
\end{aligned}$$

It follows that each entry of the table can be computed in time $\mathcal{O}(|H_B||H_D|^{|D|})$. Hence, all entries of the table can be filled in time $\mathcal{O}\big((|H_B||H_D|^{|D|})^2 \cdot |T|\big)$. Using the interpretation of $\mathsf{opt}(h^b, \mathbf{h^d}, t)$ it follows that

$$\mathsf{d}-\mathsf{opt} = \min_{h^b \in H_B, \mathbf{h^d} \in (H_D)^{|D|}} \mathsf{opt}(h^b, \mathbf{h^d}, T)$$

This shows the following result:

Theorem 4.1. *One can determine* $\mathsf{d}-\mathsf{opt}$ *in time* $\mathcal{O}\big((|H_B||H_D|^{|D|})^2 \cdot |T|\big)$.

4.2 Shortest paths

In the previous section we have seen an algorithm for computing the optimal dike/barrier height profiles which has polynomial runtime for a fixed number of dike segments, in this section we present a different algorithm, based on shortest paths, that runs in polynomial time when the number of possible barrier heights is fixed. We present an algorithm that computes $\mathsf{d}-\mathsf{opt}$ in time

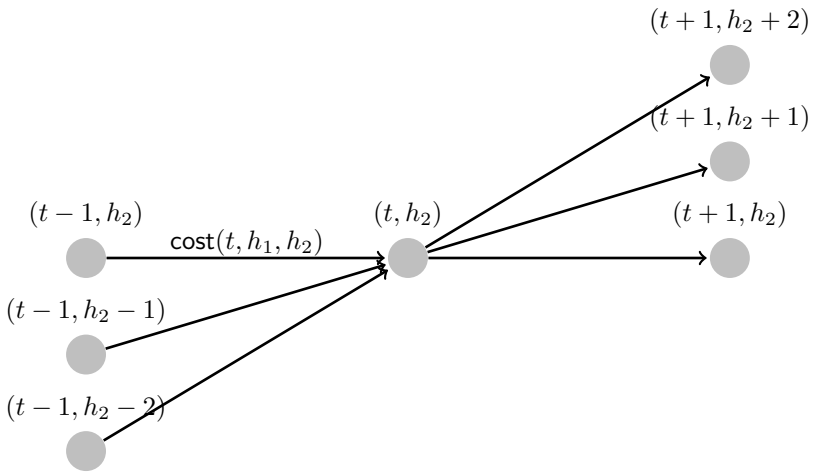
$$\mathcal{O}\left(\overbrace{|D|}^{\#\text{ segments}} \cdot \underbrace{(T\cdot|H_D|)^2}_{\text{Complexity shortest path}} \cdot \overbrace{T^{|H_B|}}^{\#\text{ barrier height profiles}}\right).$$

To illustrate the basic idea we first discuss the algorithm for the setting of one dike segment and no barrier, we then add a barrier dam and from that the generalization to multiple dike segments and barriers easily follows.

4.2.1 One dike segment, no barrier

First consider the situation with only one dike segment and no barrier. In this case the problem of minimizing the cost at time period T becomes equivalent to finding a shortest p-q path in the following graph. The source $p = (0, 0)$ is the initial height of the dike at time 0. Then, for each time $t \in [T]$ and each possible height of the dike h, we define a node (t, h). Finally we define a sink node q. The edges are defined as follows. We first add an edge between $(0, 0)$ and $(1, h)$ for each $h \in H_D$, with weight $\mathsf{cost}(0, h, 1)$, similarly for each $t \in [T]$ and height pair $h_1 \leq h_2$ there is an edge from $(t-1, h_1)$ to (t, h_2) with weight $\mathsf{cost}(h_1, h_2, t)$ equal to the financial cost associated to the decision of raising the dike segment from height h_1 to h_2 in time period t. Notice that since there is no barrier, we can assume that the expected damage cost $\mathsf{dam}(t, h)$ are incorporated in $\mathsf{cost}(h_1, h_2, t)$. Finally, the nodes (T, h) are all connected to the sink q. In the figure below the incoming and outgoing arcs of a node (t, h_2) are sketched for some $0 < t < T$ and $h_2 \in H_D$. One observes that, indeed, the shortest p-q path corresponds to the best strategy of heightening this dike segment.

Recall, the shortest p-q path in a graph $G = (V, E)$ with nonnegative edge weights can be found in time $\mathcal{O}(|V|^2)$ using Dijkstra's algorithm.



4.2.2 One dike segment, a barrier

We now consider the case of a single dike segment and a barrier. The observation we need to make is that the total financial cost incurred by upgrading the dike segment from height h_1 to height h_2 in time period t no longer only depend on the dike segment, they also depend on the height of the barrier at time point t. This means that we cannot solve a shortest path problem for the barrier and dike segment separately: the costs on the dike segment graph depend on the path chosen in the barrier graph.

The key idea is that if we fix the height of the barrier at each time t, then we reduce to the previous setting where all the costs are known. Hence, the optimization problem $\mathsf{d-opt}$ can be solved by minimizing over the possible height profiles $h^b(t)$ of the barrier over time, the minimum cost of a p-q path in the network defined in the previous section (using the costs associated to $h^b(t)$) plus the cost of implementing height profile $h^b(t)$. The outer minimization over the possible height profiles $h^b(t)$ is performed by enumeration, which takes time roughly $T^{|H_B|}$. This means that the optimal investment strategy for both the dike segment and barrier can be found in time

$$\mathcal{O}\left((T \cdot |H_D|)^2 \cdot \binom{T}{|H_B|}\right) = \mathcal{O}\left((T \cdot |H_D|)^2 \cdot T^{|H_B|}\right).$$

4.2.3 Multiple dike segments and a barrier

The approach of the previous section easily generalizes to the setting of multiple dike segments and a barrier. Once a height profile $h^b(t)$ of the barrier dike is fixed, the optimal height profiles of each of the different dike segments can be computed independently. Hence the problem of finding the optimal investment strategy for multiple dike segments and a barrier can be solved in time

$$\mathcal{O}\left(|D| \cdot (T \cdot |H_D|)^2 \cdot T^{|H_B|}\right).$$

This approach generalizes to the setting of multiple barriers and dike segments (where the costs of a dike segment at time t may depend on the height of several barriers). The complexity will be of the form

$$\mathcal{O}\left(|D| \cdot (T \cdot |H_D|)^2 \cdot T^{|H_B||B|}\right),$$

where $|B|$ is the number of barriers. One should note that the above approach assumes the same discretization in time of the barrier and dike segments. It seems reasonable to assume a coarser discretization for the barrier of say T_B steps, this would reduce the above-mentioned formula to

$$\mathcal{O}\left(|D| \cdot (T \cdot |H_D|)^2 \cdot (T_B)^{|H_B||B|}\right).$$

5 An abstraction of the problem

In this section we present a natural abstract version of the dike height problem, which allows for several variations and questions, which we believe have not been considered in the literature before. We believe that studying these variations may shed more light on the complexity of the dike height problem.

In the dike height problem we essentially have two directed graphs where each path in one of the two graphs (the one modeling the height of the barrier dam) influences the cost of arcs in the other graph. It is not difficult to show that if we were to allow any kind of influence of the path in the one graph on the cost of arcs in the other graph, the problem would automatically become NP-hard. Indeed, one can easily show that in this case the problem contains the problem of finding two vertex disjoint paths in a directed graph, which is NP-complete (2).

For this reason, we consider the following restricted problem.

Definition 5.1. For $k \in \mathbb{N}$, a *k-layered* graph is a directed graph $D = (V, A)$ such that V is partitioned into *layers* $V = V_0 \cup V_1 \cup \ldots \cup V_k \cup V_{k+1}$ such that each $a \in A$ is from V_i to V_{i+1} for some $i = 0, \ldots, k$ and where V_0 and V_{k+1} both consist of a single vertex and where $|V_1| = |V_2| = \cdots = |V_k|$. We denote the arcs between V_i and V_{i+1} by $A[V_i, V_{i+1}]$ and we refer to $|V_1|$ as the *partition size*.

Definition 5.2 (Minimum intertwined-cost path)**.**
Input: two k-layered graphs $G_1 = (V_1, A_1), G_2 = (V_2, A_2)$, with partitions $V_i = V_1^{(i)} \cup \ldots \cup V_k^{(i)}$, respectively, cost functions $c_1 : A_1 \to \mathbb{R}_{\geq 0}$, $c_2 : A_2 \to \mathbb{R}_{\geq 0}$ and for each $i = 1, \ldots, k$ a map $m_i : V_i^{(2)} \times A[V_{i-1}, V_i] \to \mathbb{R}_{\geq 0}$.

Given a path $P_2 = (a_1, v_1, a_2, v_2, \ldots, a_k, v_k, a_{k+1})$ from $V_0^{(2)}$ to $V_{k+1}^{(2)}$ and a path $P_1 = (a'_1, \ldots, a'_{k+1})$ from $V_0^{(1)}$ to $V_{k+1}^{(1)}$, we define the *cost* of the pair (P_1, P_2) as

$$\text{cost}(P_1, P_2) = \sum_{i=1}^{k+1} (c_1(a_i) + c_2(a'_i)) + \sum_{i=1}^{k+1} m_i(v_i, a_i).$$

Output: the minimum cost of a pair of paths (P_1, P_2) over all pairs and a pair of paths (P_1^*, P_2^*) attaining this minimum.

In the Minimum intertwined-cost problem, the dependence of $\text{cost}(P_1, P_2)$ on the path P_2 is linear in the edges of P_2. It is not difficult to see that the dike height problem in Section 4.2.2 can be modeled as a special case of the Minimum intertwined-cost path problem, where for both graphs the arcs between V_i and V_{i+1} are somewhat restricted. More precisely, if we identify each $V_i^{(2)}$ $(i = 1, \ldots, k)$ with $H_B =: \{h_1, \ldots, h_t\}$ then the only arcs that are present are of the from (h_i, h_j) with $h_i \leq h_j$. This particular fact allowed us in Section 4.2.2 to give an algorithm for the problem, which runs in polynomial time if we consider the size of the sets in the partition of the vertices of the second graph as a constant. Clearly if the bipartite graphs between $V_i^{(2)}$ and $V_{i+1}^{(2)}$ are complete, then this dynamic programming approach will not work. It would be interesting to find out if some other approach may yield an efficient algorithm.

We end this section with some concrete questions.

Question 1. *Is the Minimum intertwined-cost path problem NP-hard?*

If this question has a positive answer, then it makes sense to consider the following questions.

Question 2. *Under which conditions on the bipartite graphs $G_j[V_i^{(j)}, V_{i+1}^{(j)}]$, ($j = 1, 2$, $i = 1, \ldots, k$) is there a polynomial time algorithm for the Minimum intertwined-cost path problem?*

Question 3. *Suppose the partition size of G_2 is constant. Under which conditions on the bipartite graphs $G_j[V_i^{(j)}, V_{i+1}^{(j)}]$ ($j = 1, 2$, $i = 0, \ldots, k$) is there a polynomial time algorithm for the Minimum intertwined-cost path problem?*

Acknowledgements

We thank Kees Roos for helpful discussions and for presenting the ideas of (1). We moreover thank Peter Zwaneveld from CPB and André Woning from Rijkswaterstaat for useful background information and suggesting the problem to us as well as Gerard Verweij from CPB for giving us information about how the LP model is solved.

References

[1] R. Brekelmans, D. den Hertog, K. Roos and C. Eijgenraam, Safe dike heights at minimal costs: the nonhomogeneous case, *Oper. Res.* **60(6)** (2012), 1342–1355.

[2] S. Fortune, J. Hopcroft and J. Wyllie, The directed subgraph homeomorphism problem, *Theoretical Computer Science* **10** (1980), 111–121.

[3] J. Kind, Maatschappelijke kosten-batenanalyse Waterveiligheid 21e eeuw (MKBA WV21), Deltares Report, Delft, The Netherlands, 2011.

[4] D. van Dantzig, Economic decision problems for flood prevention, *Econometrica: Journal of the Econometric Society* (1956), 276–287.

[5] P.J. Zwaneveld and G. Verweij, Safe Dike Heights at Minimal Costs: An Integer Programming Approach, *CPB Discussion Paper* **277** (2014).

[6] P.J. Zwaneveld and G. Verweij, Economic Decision Problems in Multi-Level Flood Prevention, manuscript (2017), private communication.

Equalizing the Cost of Health Insurance

Casper Beentjes Alessandro Di Bucchianico
Christian Hamster Ajinkya Kadu Irene Man
Keith Myerscough Marta Regis Omar Richardson

Abstract

The Dutch government compensates health insurance companies when insuring individuals who are estimated to have high health care costs. This is necessary to avoid insurers not offering services to certain groups or not providing them with a high quality of service. It is, however, unknown to what extent the differences in health care expenses by different groups of people are truly due to a poorer or better health status. We explore several statistical approaches that facilitate explaining the cause of these differences.

Keywords: health insurance, risk equalisation, model selection, predictive model, explanatory model, model selection, lasso, elastic net, ridge regression, clustering.

1 Introduction

Health care costs in the Netherlands are paid for by private insurance companies, who receive their funds from two different sources. In the Netherlands each adult chooses an insurance company and pays a fixed premium per month. Insurers are free to set their premium, but it has to be the same for all insured adults. The first source of income for insurance companies is this monthly premium paid by all their customers.

The second source is a subsidy from the government. This subsidy is different for different insured individuals, based on a number of indicators that estimate the general health of the insured. The goal of this differentiation is to equalize the risk carried by insurers when insuring different people. Without such equalization, it would be profitable for insurers to target certain groups they estimate will generate larger health care costs and an incentive for insurers to offer good care and services to those in need would be lacking. This report focuses on the second source of funding, the risk equalization fund.

Determining the correct amount of funding for each insured individual is a challenging task and involves political considerations. This task is very important since health care costs in the Netherlands are increasing significantly due to an ageing population, which is a threat to the affordability of the national health care system. The particular problem we intend to address is that the current approach uses past health

care expenses as the basis for the estimation of required health care costs. These past expenses, however, are not necessarily a good indicator of the truly required health care costs. Several factors may lead to inflated expenses, such as, but not limited to: a propensity to 'consume' health care if it is readily available, the deliberate exaggeration of diagnoses by care givers to increase turnover and profit and inefficiencies in the execution of certain treatments. On the other hand, the real expenses do not see where care was required, but not consumed due to financial incentives, such as the legally imposed deductible of several hundred euros or the loss of income for self-employed people. These two deficiencies both have grave consequences for the functioning of the health care market. The first error, overestimating the truly required costs, removes the incentive for health insurers to put pressure on care providers to make their business more cost efficient. Furthermore, as there is a fixed total budget, an overestimated budget for one group directly harms another. The cases where the insured persons would benefit from more care but are for some reason unable to obtain this are now largely ignored by the system. This could be detrimental to their long term health, and is certainly an ethically questionable situation.

2 The current model

In this section we will outline the current procedure used by the government to determine the funding for health insurance companies. We will particularly focus on how the funding for risk equalization is computed. The risk equalization model is calculated for the total funding and then adjusted for a set premium by VWS. After the real premium collection by insurers, an insurance company either gets money from the risk equalization fund or contributes to the risk equalization fund based on the outcome of the risk equalization model. The basis of the risk equalization is a linear regression model that aims to predict the health care costs for each individual based on a number of personal variables that are deemed a good indicator of their general 'health status'. We will detail which variables are used — and to some extent why — in Section 2.1. Due to the time required for processing all health care providers' accounts of realized costs, there is a three year lag in the cost prediction. This implies the risk equalization funds are determined for 2017 using a regression model based on costs from 2014 and the characteristics (FKG/DKG) from 2013.

2.1 Model parameters

The variables used for risk equalization are intended to be variables that indicate how healthy a person is likely to be. Originally, this included only the age and gender of individuals, but the number of variables included in the model has vastly increased since then.

We have summarized the used variables in Table 1 and noted a few remarkable properties of the data below. The categories `s_fkg`, `s_dkg` and `s_hkg` all indicate specific use of products and are therefore strongly linked to specific health issues.

Table 1: A list of model parameters used in the government's risk equalization scheme

Variable prefix	Explanation
`normbedrag_somatisch`	Gross compensation of health insurer for somatic costs in the postcode based on risk equalization model
`s_iedereen`	Total number of people in the postcode, given in 'insured years' to account for people who are insured for the full year
`s_kost`	Total costs for various types of somatic care
`s_totale_kosten`	Grand total of somatic costs
`s_lgnw`	Age and gender categories
`s_fkg`	Pharmaceutical cost groups
`s_ape`	Postcode 'region'
`s_dkg`	Diagnosis cluster groups
`s_mhk`	History of medical expenses (top percentiles over past two or three years)
`s_hkg`	Medical devices-based cost group
`s_avi`	Source of income
`s_ses`	Social economic status
`s_FGG`	Physiotherapy use
`s_VGG`	Nursing and caregiving (at home) costs
`s_GGG`	Geriatric rehabilitation care
`s_gsm`	Comorbidity

The (ten) postcode 'regions' indicated by the `s_ape` variables will be discussed in greater detail in Section 2.2. Some of the categories (e.g. `s_avi` and `s_ses` or `s_fkg` and `sdkg`) are strongly correlated, resulting in a strong multicollinearity, which we will discuss in Section 4.3. It should also be noted that the category `s_avi` contains a wealth of different types of income, such as benefit schemes and regular employment, but also contains elements for students and the self-employed.

2.2 Linear regression procedure

The regression procedure used by the government consists of three steps. The first step is a linear regression that fits the grand total cost (variable `s_totale_kosten`) based on all the predictive variables (variables `s_lngw`–`s_gsm`) except for the postcode regions (`s_ape`). This first regression is performed using two constraints: (i) those coefficients associated to age and gender must result in a total that matches the true total when ignoring the other variables, and (ii) the other groups of coefficients must all result in a zero sum. These constraints result in an easier interpretation of regression parameters and facilitate the comparison of the parameters obtained from different models or different years, but in our view do not affect the result of

the regression. An exeception are pharmaceutical costs, since each individual can have more than one `s_fkg`, while for example the `s_dkg` of each individual can be categorized in one group only.

After the first step, the model is further refined by looking at 'regional' variations. The residuals from the first step are aggregated to a postcode level, and the postcodes are then clustered into ten 'regions' based on their aggregated residuals. This clustering is performed by considering ten deciles of these residuals, and thus the 'regions' do not necessarily have any geographical cohesion. A final regression is then performed using the postcode region (`s_ape`) for each individual, again constrained to a zero-sum correction.

3 Our goal

We would like to immediately point out an important distinction between the *prediction* and the *explanation* of health care costs. The current model is intended for prediction to equalize the risk for healthcare insurers. The question(s) posed relate to both explanation and prediction, to understand how the various parameters in play affect the realized healthcare costs and to what extent these are down to the actual health status of individuals. It is known in statistics that models that perform well for prediction, may perform poorly for explanation. We combine these questions in a single research question:

> What are appropriate ways to find and explain geographical differences in healthcare costs?

The answer to this question will be given by a number of different data analysis tools. We exemplify – and to some extent justify – these methods by discussing the results of their application to the aggregated data set that has been made available to us.

4 Results

The proposed models can be divided into three categories. First, we study ordinary linear regression models similar to the model used by the government, but we focus on the selection of the most significant variables. Secondly, we look at more sophisticated linear regression models in order to obtain models that are good for either explanation or prediction. Finally, we perform clustering of postcodes (or other aggregation levels) to investigate analogies and differences. This approach is significantly different from the other two and is aimed specifically at explanation rather than prediction.

We would like to stress that we are aiming solely at providing *tools* that may be used in assessing the proper choice of model and variables. Any such choice impacts the funding of health insurers and consequently their behaviour on both the healthcare market (buying healthcare from providers) and the consumer market (selling insurance to individuals). The choices made should be justifiable to the public or at least to their elected representatives. It is our contention that *after* using tools that might

seem opaque to assess the behaviour of certain models, it is possible to obtain better models that balance transparency on the one side with predictor accuracy on the other. Note, however, that an accurate predictor of realized costs might in fact be predicting something different to the necessary costs of health care.

Beside model fitting, finding which elements in a data set are outliers with respect to a given regression provides useful information on the quality of the regression. If outliers share traits that are not captured by the regression model, it may be useful to include a quantifier for these traits in the model. The outliers will also clearly demonstrate the regional differences, potentially providing an even stronger motivation for our current exercise. We discuss the results of outlier detection (aggregated at the level municipalities for privacy reasons) in Section 4.2.

We also perform recursive variable selection to get an understanding of which variables in the model are most relevant. Subsequently ranking the different models based on how well they model the variation in the data relative to the required number of variables — quantified by Mallows' C_p, see Mallows (1973) — provides valuable insight into the optimal number of variables to include. Particularly, this will indicate if the current model is over- or underfit. This is important because overfitting may lead to spurious explanations of differences. Section 4.3 contains the results of this part of the analysis.

To avoid issues arising due to the collinearity between many of the predictor variables (in particular, the danger of important variables being erroneously declared non-significant since significance is divided over several similar predictor variables), we consider adding regularization to the regression. Such regularization can facilitate the interpretation of different resulting parameters for the regression by removing some ambiguity from the system. It does, however, introduce some new (meta)parameters that need to be computed a priori. We discuss some preliminary results from this approach in Section 4.4.

The final approach we elaborate upon does not involve a regression technique, but instead attempts to find clusters of similar postcodes. Within these clusters, it may be easier to identify individuals, postcodes or groups at a different level of aggregation that are remarkably different from others in the cluster. A clustering based on data at the postcode level, the finest at our disposal, is presented in Section 4.5.

Before detailing the results of these approaches, we briefly outline some of the work that was done in preparing the data.

4.1 Data and preparation

Currently the government collects a large number of 'health status' variables on an individual person basis as input for their risk equalization calculations. These indicator variables are, amongst others, information on location of residence, age, gender, social economic status, source of income, healthcare costs in the previous three years and the morbidity of the individual. The morbidity is split in somatic morbidity and mental morbidity. For somatic morbidity the government includes 30 (classes of) diseases each with its own list of specific medicine use and medical treatments (Zorginstituut

Nederland, 2017). The use of these determines whether an individual is registered for that specific disease in the government database. For mental morbidities a similar approach is used, although with fewer diseases. The resulting dataset then contains a total of 225 variables per individual of which 26 are mental health care costs specific and 17 are used solely in the model for regional variation at the postcode level.

The original dataset held by the government is too confidential to work with as it contains information on the health of individual citizens. Therefore, we only had access to a data set aggregated at the four digit postcode level [1]. This results in 3838 postcodes with the combined 225 health status variables of the people living in those postcodes together with an extra variable depicting the total number of people in that postcode.

Due to this aggregation a few peculiarities creep into the dataset. Firstly the dataset contains several postcodes that consist solely or partially of PO boxes[2]. These do in fact not correspond to a physical location in the Netherlands where people are registered to live. Multiple scenarios exist why people can be registered under a PO box, such as when someone does not have a fixed address or lives abroad. In that case the health insurers will often register the costs of the insured person on the postcode of the insurer, which can be a PO box. For the purpose of explaining geographical differences in the Netherlands we exclude these particular set of postcodes as any geographical information on the people in these groups is lacking[3]. However, when making predictions for the health care costs for the next year the individuals in these postcodes have to be included, since they do after all contribute to the total costs and need to be included in the risk equalization calculations. This reiterates our earlier point that there should be a difference between the explanation and prediction approaches.

Aside from the PO box issue we now have the issue that postcodes can widely vary in the number of residents registered. As a result we observe a vast range of different scales of many of the indicator variables. We therefore normalise all the variables to the number of registered residents. The indicator variables then represent the average values for a registered insured person at the various postcodes.

[1] Dutch postcodes follow a four digits plus two letter format, e.g. 1234 AB. The aggregation puts together all the residents of 1234 AB and 1234 CD into the same category 1234.

[2] In major cities like Amsterdam and Rotterdam such postcodes are the ones ending at 00 or 01, but this varies from city to city. E.g., PO boxes in Nijmegen are postcodes ending on 00, 01, 03, 04 and 31 (see e.g. `http://postcodebijadres.nl/postbus+Nijmegen`.

[3] There remain postcodes that are partially PO boxes and partially residential addresses and we make the decision to omit these from the data used for the explanation as well whenever we can locate these postcodes.

4.2 Outlier identification

Summary

Goal Identify regions with exceptional costs or characteristics. These identified regions can then be investigated more thoroughly to determine if the risk adjustment model is appropriate and sufficient for these more extraordinary groups of insured. In addition exclusion of extreme values can improve the estimation of the expected costs of a group of insured.
Method Studentized residual analysis after linear regression on a regional basis for the costs and age/gender.
Main result There are a few regions with exceptional costs and characteristics. For example, Urk, Lelystad, Pekela, Koggenland, Weesp and Oegstgeest have an extraodanary age/gender-profile. Vlist, Onderbanken, Pekela, Vlieland, Menseradiel, Son en Breugel and Oud-Beijerland have exceptional costs. Pekela seems extreme with low age and high costs.
Recommendation Further research of characteristics (next to age or gender) of inhabitants in the extraordinary regions is recommended. This can lead to new characteristics that can be included in the risk equalization model. Compare the current risk equalization model with a model where extreme values are excluded to get a feeling how strongly the average results are biased by the outliers.

As mentioned in Section 2, the current model used has three steps, of which the first is a straightforward linear regression and the second and third steps aim to correct for regional variation. In the interest of simplicity, we will only consider the first step of the procedure. Due to restrictions on what we are allowed to publish, the data is first aggregated to the level of municipalities before studying outliers. We perform two linear regressions using a different subset of the available variables. The first uses only the age and gender distribution of each municipality and the second uses all predictive variables that are used in the government model.

Figure 1 displays the regression fit and the studentized residuals of that fit for each municipality in the data set. The left panel shows the results using only age and gender in the model, the right uses all predictive variables. The studentized residuals represent a rescaling of the residuals (i.e, the differences between the observations and the fitted values from the model) such that it is comparable to a standard normal distribution under the assumption that the error of the linear regression is truly Gaussian. In this way the residuals are scale free, i.e. their values are independent of the unit used for the response variable, so that it is possible to have a universal threshold to detect outliers. Red lines in Figure 1 indicate a threshold for outliers chosen at 2.5 times the standard deviation from the zero mean. The factor 2.5 is a rule-of-thumb to decide on suspect observations, based on the approximate standard normal distribution of the scaled residuals.

When using only age and gender in the model, there are three outliers either side of the 2.5 standard deviation threshold. This is not surprising in itself, but the very low fit for the municipality of Urk is exemplary of a broader trend that 'cheaper' municipalities are underestimated. It should be noted that Urk is a fairly unique location; it is a former island that still retains a somewhat isolated character. When using all predictive variables, there is a marked skew in the outliers with many more

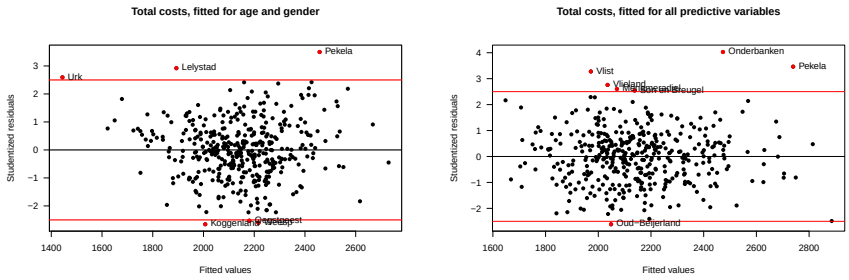


Figure 1: The studentized residuals after linear regression on a municipal level, based on only age and gender (left) or on all predictive variables (right). On the x-axis are the predicted values of the models. These studentized residuals put deviations from predictions on a universal scale. The usual threshold is 2.5. The municipalities that are highlighted in red have total costs that are not predicted well by the respective models

municipalities having a severe, positive residual. As a final note, we wish to point out the municipality of Pekela retains a strong, positive residual when using all variables.

4.3 Variable selection

Summary

Goal In the current risk equalization model there is a risk of over-fitting due to the large number of included variables. There are also some problems with multicollinearity that make results difficult to interpret. In this part of the study we identify the variables with the highest impact on the results and which variables can be left out with little impact on the result.
Method We use a stepwise forward model selection and a stepwise backward model selection procedure to find the variables with highest and lowest impact.
Main result The forward and backward selection procedure both lead to the conclusion that variables need to be excluded in order to avoid overfitting. The two methods have partially overlapping results concerning variables that can be excluded. However, it is difficult to determine which variables should be left out due to multicollinearity.
Recommendation Use this method to determine which variables potentially can be deleted. For the variables with multicollinearity problems, determine politically which variables should be left out. Estimate the regression model without these variables to determine the impact on the regular descriptives of the model. Use a measure for multicollinearity in addition to the current descriptives to judge the performance and validity of the model.

There are several different ways to study which variables are important so that they should be included in a regression model. Note that this is not the same as ranking the variables in a model according to their importance (see Grömping (2007) for an excellent discussion of relative importance in regression analysis). In this section, we

focus on forward and backward regression, which are elementary, heuristic ways for iterative model selection.

Stepwise *forward* model selection starts with a (trivial) linear regression model with no variables and then at each the variable that results in the lowest R^2 error term is added to the regression. As such, an increasingly complex regression is constructed. *Backward* model selection starts using *all* model variables and then iteratively removes those variables that have the smallest impact on the error. Both methods result in a hierarchy of models and a list of variables for each model. The models are then ranked by a score that balances the complexity of the model with the accuracy of the fit. This is intended to counter the overfitting that would occur if only the R^2 error is used as a norm – in that case the more variables the better, which may lead to overfitting. By studying which variables are used most, or at least used by the best models, provides useful information on which (categories of) variables have the most predictive power.

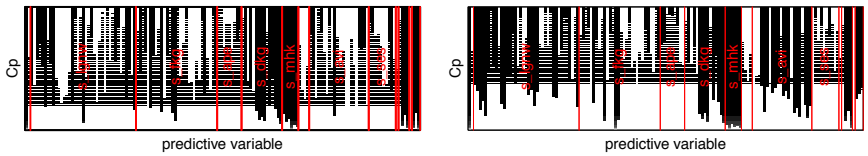


Figure 2: Ranking of models based on Mallows' C_p. Each row of either chart represents a linear regression using different variables, indicated by the dark cells in that row. The left-hand panel uses forward selection, the right hand panel uses backward selection. Note that the vertical axes are ranked, but correspond to slightly different Mallows' C_p values (the lower the better), both ranging from 136 up to roughly 4900.

Figure 2 demonstrates the results of forward (left panel) and backward (right) model selection. The models are ordered vertically by Mallows' C_p (see Mallows (1973) and Gilmour (1996), an alternative to R^2 that penalizes for having too many variables), a lower score indicates a better model. Note that there is no linear scale on the axis. It is well-known in the statistical literature that R^2 is not a model selection criterion (see e.g. Kvålseth (1985)). Dark pixels indicate variables included in the model, the darkness of the colour scales linearly with Mallows' C_p. The individual variables have not been labelled on the abscissa, since this would be too dense to read (but they can be easily read off a tabular output). Instead the different categories have been indicated.

Variables that are included in the top rows in the diagrams provide the most useful information in predicting health care costs. We observe in both forward and backward selection that the best scoring models are those with roughly half the number of

variables included. Models that include more variables rank low, indicating that using all these variables is overfitting the data. This conclusion is particularly apparent from the forward selection, but also holds in the backward case.

The difference in the variables selected by the forward and backward procedures is minimal. We therefore first focus on aspects that are visible in both, a few differences will be pointed out later. The most striking feature is the ubiquitous inclusion of the history of health care expenditure (`s_mhk`) the list of variables. These categories reflect top quantile health care use in previous years, and probably is particularly useful in predicting the high costs for chronic patients. Similarly, a history of high costs for nursing and caregiving is also a strong indicator of realized health costs. Comorbidity (`s_gsm`, far right) is also included in all models with a high rank. However, this is likely to cause multicollinearity in models that also include `s_fkg`, `s_mhk` and `s_hkg`. The *diagnostic* cluster groups are mostly good indicators, with the notable exception of cluster groups 1, 3, 5 and 10. As we have no information on the meaning of these clusters, we can draw no further conclusions from this.

The forward selection selects substantially fewer variables from the age and gender (`s_lgnw` and 'source of income' (`s_avi`) categories. It appears that this is compensated for by the inclusion of more socio-economic status (`s_ses`) variables. This is possible in part due to the multicollinearity embedded into the variables by making an explicit division by age in the variables from the socio-economic status and source of income categories. Besides this deliberate multicollinearity, we suspect there is a strong correlation between source of income and socio-economic status, leading to ambiguity in the choice of variables.

While this variable selection procedure is of limited sophistication, it does point to two suggestions. First, the current model appears to be overfitting the data, as suggested by the improved Mallows' C_p for models with fewer variables. Second, the multicollinearity embedded (in part deliberately) into the model stands in the way of interpreting the individual significance of certain variables.

4.4 Advanced regression techniques

Summary

Goal Solving the problem of multicollinearity with other regression techniques.
Method Alternative regression techniques: LASSO regression, Ridge regression and Elastic Net.
Main result The Elastic Net outperforms both the Ridge regression and LASSO regression by achieving a much lower Mean Squared Error, while controlling the number of variables.
Recommendation Explore the impact of the new methods on the regression model (on individual level). Compare the current descriptives and compare the coefficients of the regression. Describe these models for non-mathematicians in order to let them comprehend these models and interpret the results.

The standard linear regression approach fails to find good explanatory models when the data contains strong multicollinearity. It is evident from Section 4.3 that the dataset has many collinear variables, so that one may miss significant explanatory variables. To avoid this issue, we propose the elastic-net approach from Zou and Hastie

(2005). This approach can be thought of as a combination of the established ridge (Hoerl and Kennard (1970)) and the modern LASSO (Least Absolute Shrinkage and Selection Operator, see Tibshirani (1996)) regression techniques. These approaches have in common that they add L_1 regularizations to the L_2 (= least squares) criterion in the regression procedure, and thereby handle the multicollinearity in the data. We refer to Hesterberg et al. (2008) for an accessible review where these methods are put in perspective, while extensive treatments can be found in the monographs Efron and Hastie (2016) and Hastie et al. (2015). However, ridge regression has the disadvantage that it does not lead a parsimonious model (it does shrink unimportant variables, but they do not shrink to zero). The LASSO does shrink unimportant variables to zero, but the LASSO selects at most n variables, where n is the number of observations. It also tends to select only one variable from a group of correlated variables, ignoring the others. To overcome these limitations, the elastic net adds a quadratic part to the penalty ($\|\beta\|_2^2$), where $\|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2}$ denotes the 2-norm of a vector $\mathbf{x}$. Mathematically, the regression problem is now written as

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \|\mathbf{Y} - X\beta\|_2^2 + \lambda \left((1-\alpha)\|\beta\|_2^2 + \alpha\|\beta\|_1 \right) \right\}, \tag{1}$$

where $\mathbf{Y}$ are the actual costs, β is a vector of weights for variables, X represents the values of these variables for various postcodes, and $\alpha \in [0,1]$. The $\|\mathbf{x}\|_1 = \sum_i |x_i|$ denotes the 1-norm. In this way the elastic net combines the advantages of these methods while minimising their disadvantages. Sometimes a factor $1/2$ is put in front of the $\|\beta\|_2$ term for mathematical convenience. Note that the elastic net includes ridge regression and the LASSO as special cases through the choices $\alpha = 0$ and $\alpha = 1$, respectively.

The main drawback of the elastic net is that it requires to select appropriate values for the regularisation parameter λ and the elastic net parameterα. In principle, one can use cross validation techniques or a grid search to find optimal values for these parameters. For correct values of λ and α, we get a vector $\widehat{\beta}$, which highlights the most important variables in the regression.

To test the elastic net approach, we fit our model on two thirds of the data and test on the remaining data. We consider 11 different values for α from 0 to 1. In Formula (1), $\mathbf{Y}$ represents the total Somatic costs, while X represents the various variables corresponding to somatic costs except the postcode clusters, that is, all variables from Table 1 except `s_ape`.

Table 2 presented the mean squared error (MSE) for a few different tested values of the parameter α. The extreme values $\alpha = 0$ and $\alpha = 1$ correspond to ridge regression and LASSO, respectively. The minimal value is found for the parameter $\alpha = 0.7$

To present a little more insight into the behaviour of the regression techniques corresponding to different values of α we illustrate the regression result with Figures 3–5. Figure 3 shows, for the ridge regression ($\alpha = 0$), the coefficients $\widehat{\beta}$ against λ on the left and the mean squared error (MSE), measured as $\|Y - X\beta\|_2^2$, on the right-hand side. Similarly, Figures 3 and 4 show the variation for elastic net (with the optimal $\alpha = 0.7$, see Table 2) and LASSO ($\alpha = 1$) respectively.

α	MSE
0 (Ridge)	42960.8
0.2	41876.1
0.5	39314.6
0.7	**37012.1**
0.9	40353.4
1.0 (LASSO)	44236.1

Table 2: Mean-squared error for various values of the elastic net parameter α (on test data). Smaller values indicate better fit.

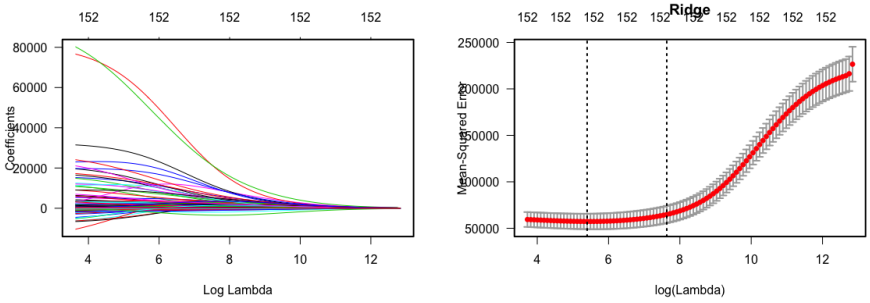


Figure 3: Ridge Regression ($\alpha = 0$). The number of coefficients (left) and the respective mean square error (right) variation with regularization parameter λ. This picture helps to find the right balance to a small number of coefficients (parsimony) while at the same time controlling the mean squared error (measure for model fit).

From each of the left hand-panels, we observe that only relatively few variables contribute strongly to the linear model. In the case of the LASSO, this is explained by multicollinearity. In the right-hand panels, vertical dashed lines indicate the region in which the MSE attains its minimal value.

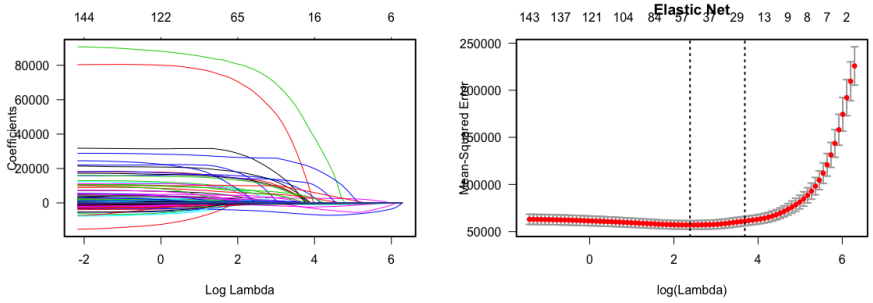


Figure 4: Elastic Net (for $\alpha = 0.7$). The coefficients (left) and the respective mean square error (right) variation with regularization parameter λ. This picture helps to find the right balance to a small number of coefficients (parsimony) while at the same time controlling the mean squared error (measure for model fit).

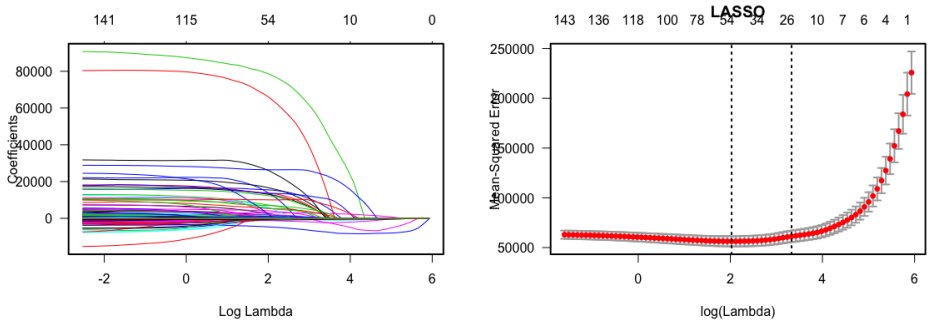


Figure 5: Lasso Regression ($\alpha = 1$). The coefficients (left) and the respective mean square error (right) variation with regularization parameter λ. This picture helps to find the right balance to a small number of coefficients (parsimony) while at the same time controlling the mean squared error (measure for model fit).

4.5 Clustering approaches

Summary

Goal Cluster postcode entries based on similarity of features, which constitute a proxy for the health profile. Within a cluster of postcode entries with similar feature values, a postcode entry with exceptional costs might indicate inefficiency.
Method We used a k-mean clustering method using a standard distance metric (Euclidean distance) on a transformed (e.g. splitting age and gender and collapsing age groups within AVI) features.
Main result There is a clear distinction between urban regions and suburbs, when clustered based on age as well as based on AVI. For the age clustering, the suburban postcode entries more frequently have low total health costs.
Recommendation This method can be extended by clustering based on different or larger sets of features. Moreover, we suggest to use this method as a new approach to discover regions with inefficient care, which is to search for outliers with high health care costs within clusters with the similar feature values.

In this section we explore the possibilities and benefits of cluster analysis on the provided dataset. We aim to use these techniques to find a natural structure present in the postcode entries (*instances*) based on the similarity in the explanatory variables (*features*). Our objective is twofold: we hope to gain intuition of the meaning of different combinations of explanatory variables, and we may find unknown patterns and relations unexplained by the regression models currently in use.

There exist many clustering techniques, but almost all of them follow similar steps:

1. Define a distance metric on the set of instances.
2. Formulate a decision rule that determines whether an instance belongs to a cluster.
3. Iteratively separate *or* group instances until every one is classified.

Among commonly techniques are hierarchical clustering, distribution-based clustering, and centroid-based clustering.

To provide some context: hierarchical clustering algorithms often evaluate the 'distance' between two observations, i.e. some quantitative notion representation of the difference between their properties. Nearby observations are linked to form a cluster, and nearby clusters are merged to larger clusters. This is a convenient strategy for exploratory clustering approaches, but expensive to apply to large data sets and difficult to interpret for high-dimensional data.

Distribution-based clustering algorithms try to find a set of clusters by choosing from a family of distributions that matches the observations.

While this collection of algorithms generally has no difficulty clustering all observations, these methods are prone to overfitting data. In addition, for many observation properties (especially in our dataset) it is difficult to find an underlying family of distributions.

Centroid-based clustering is based on the assumption that each cluster has a central observation: a centroid. Observations are classified based on their distance to the

nearest centroid. The benefits of this type of approach is that the created clusters have an intuitive and well-defined meaning: they can be interpreted by their centroid observations. The downside is that a priori the number of clusters must be known, and that the method allows only for the creation of clusters with a specific shape: convex clusters.

In our analysis, we choose a centroid-based clustering algorithm: k-means clustering. This choice is based on the fact that the data is high-dimensional and that before we start our analysis, we lack knowledge of how the data is structured or how the clustering could be interpreted. It is worth the effort to investigate if other algorithms are more suitable.

4.5.1 Distance metric

In our implementation, we used the function `kmean` of the statistical software `R`, which performed the k-mean clustering using standard distance metrics such as the Euclidean distance. However, Euclidean distance has no valid meaning for the variables in their current state. Since the spread in money-related variables is much higher than in age-related variables, a transformation is required to ensure that the Euclidean distance is normalized. Furthermore, from an information entropy perspective, some variables in the data may be redundant. These are also transformed to a more compact form. In the following section, we explain in what way the data is non-uniform and redundant, and how we transform it.

4.5.2 Feature transformation

The variables as given in the data are of different types. Some count the number of individual belonging to a category of a binary attribute (e.g. the number of individual using certain drug) or to a category of a categorical attribute (e.g. the number of individuals with source of income high). As, for each attribute, each category is represented by a variable, there are as many variables as there are categories. In fact, some variables represent the count of a category from a certain age group, which is a finer resolution. As a result, an attribute that has been broken down in many categories and in age groups comprises many variables.

The exact distribution in age group for each category seems redundant, therefore we apply the following transformation:

- We collapse the age-gender categories into gender and add a feature with the gender-ratio per instance.
- We collapse the age-'source of income' categories into age.
- We convert the binary variables to ratios.

Although we have reduced the number of variables of some attributes, some attributes still have many variables. To uniformly distribute the contribution of each

attribute to the distance, we weight each with the reciprocal of the number of categories such that the total weight of all categories sums to one.

Finally, we weigh each of the variables with the reciprocal of its observed variance. This follows from our assumption to let all features be equally important.

Our transformation creates a uniform information representation, and combine some of them into new features. Although it loses some accuracy in the exact age distribution of some attributes, this greatly reduces the dimensionality of the data, which is a welcome benefit for the quality of the cluster analysis.

It should be noted that we do not have to include all features when we compute a cluster. Different subsets of features may yield different clusterings. For this reason, our scripts allow for an arbitrary subset of features.

4.5.3 Finding the optimal cluster

The k-means clustering algorithm is a probabilistic algorithm that finds clusters in the following way. First, it randomly chooses k centres in the feature space. Each of these centres represents an initial cluster. Then, for each cluster it repeatedly executes the following steps.

1. Find the closest instance and add it to the cluster.
2. Compute the new centre of all instances in the cluster.

The clustering is finished when all the instances belong to one of the k clusters. The quality of a set of clusters is determined by computing the sum of squares (CSS) of each of the instances with the centre of its cluster. The lower the CSS, the better the clustering. This procedure is repeated an arbitrary number of times, each time choosing randomly new initial locations, and finally choosing the clustering with the minimal CSS. The k-means clustering algorithm can find clusterings for any number of clusters. A heuristic way to determine the optimal number of clusters is to find the elbow in the plot of the CSS against the number of clusters.

4.5.4 Results

In our exploratory analysis, we used two different groups of features to generate two clustering. The first clustering uses the age categories as features and the second uses the source of income (`s_avi`). The elbow method determines the optimal number of clusters for both clusterings to be 3. A detail of the geographic distribution of the clusters using age categories is depicted in Figure 6. In both maps, clusters clearly distinguish between urban (Amsterdam, Utrecht, Almere, Amersfoort) and suburban areas. We now discuss the findings for the two clusters separately.
Figure 7 shows the characteristic profiles for the clusters based on age. The left-hand panel shows the distribution of people over the age groups per cluster, the right-hand panel shows the distribution of health care costs. We see that clusters 1 (black) and 3 (red) have relatively young age profiles, of which cluster 1 has the most centralized distribution of health care costs. One would expect young clusters (i) to be associated

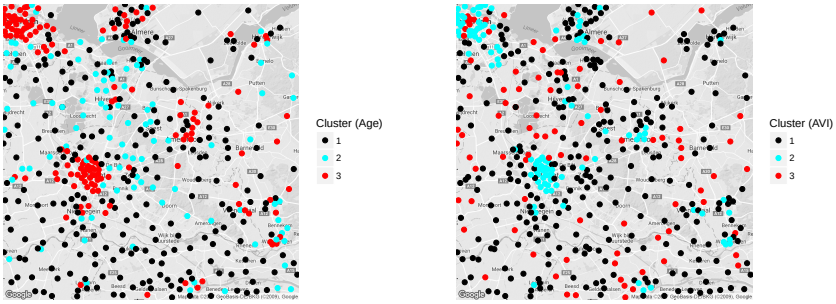


Figure 6: Detail of maps showing the result of clustering postcodes based on age (left) and based on AVI categories (right).

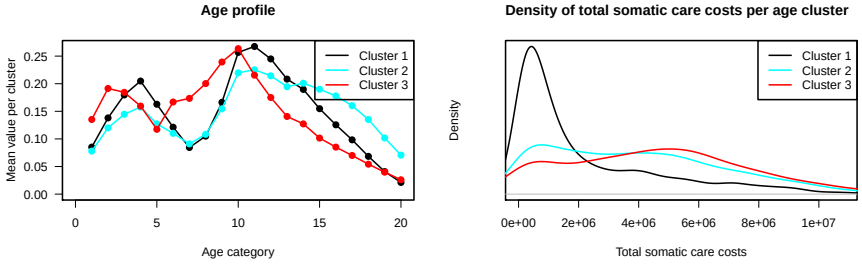


Figure 7: Cluster profiles when clustering based on age. Left: distribution of people over different age groups per cluster. Right: distribution of health care costs per cluster.

with better health and therefore low health care costs and (ii) to cluster around the urban area, which are somewhat confirmed by in our results.

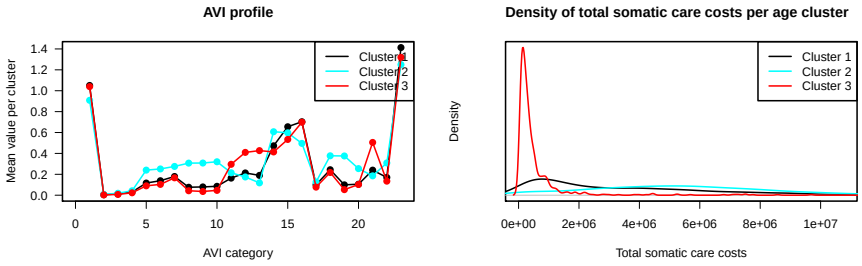


Figure 8: Cluster profiles when clustering based on source of income (`s_avi`). Left: distribution of people over different source of income categories. Right: distribution of health care costs per cluster.

Figure 8 shows the characteristic profiles for the clusters based on source of income (`s_avi`). The left-hand panel shows the distribution of people over the source of income categories per cluster, the right hand panel shows the distribution of health care costs. Cluster 1 (black) has the lowest distribution of health care costs. It would be interesting to see whether it agrees with the actual meaning of the avi-categories.

Due to time constraints, we were only able to analyse clustering based on age and avi-categories. For further analysis, it would be interesting to explore the clustering based on different sets of features. For instance, a larger set of features may better describe the health profile of regions.

Moreover, we suggest a new approach to discover municipalities with inefficient care, which is to search for outliers with high health care costs within clusters. Conceptionally, this approach is similar to outlier identification as discussed in Section 4.2, since it also searches for outliers after adjustment for explanatory variables.

5 Conclusions

The Dutch government compensates health insurance companies when insuring individuals who are estimated to require more or more expensive health care. This is necessary to avoid insurers avoiding certain groups or not providing them with a high quality of service. To estimate the required costs, the government uses a number of personal characteristics that are deemed good indicators of the general health status. The basis for this estimator is a linear regression model that fits the real health care costs based on the chosen parameters. It is, however, unclear whether the realized costs are due to a difference in health status, or due to other reasons that affect the health care expenses made. The model employed for the Dutch government is made for prediction rather than explanation.

We used several different techniques that investigate these differences, providing a first step towards understanding if these differences are preferably compensated for or not. The conclusions are somewhat diverse — in part due to how the research was carried out. Studying the outliers in the data using a linear regression model revealed no surprising results. We did, however, find that certain variables in the model are of much greater importance than others. Comparing Mallow's C_p for models using different subsets of the variables suggests the current model might suffer from overfitting. This method can help in simplifying the risk equalization model together with the clustering approach. Work needs to be done on the impact of these methods on the usual criteria for evaluating the risk equalization model. Elastic net regression provides for regularization of the model parameters (and thus avoids overfitting), at the cost of introducing a single metaparameters. With this regularization in place, it is easier to explain the impact of various parameters on quality of the regression. By applying clustering techniques we exposed a remarkable difference in the health expenditure between clusters based on only a few of the prognostic variables. Differences *within* these clusters may provide valuable information on possible causes for health care expenditure differences.

The techniques presented in this work all contribute to a greater understanding of the factors influencing the health care costs. One main conclusion is that we can leave out variables with no or limited loss of the quality of the model. The methods we used can help in selecting variables that contribute and variables that have no contribution. It can also help in selecting and combining variables in the current model in order to be capable of interpreting the results of the regressions and solve the issues with multicollinearity in the model.

6 Recommendations

Besides the methods presented above, we also have a number of suggestions for techniques that may improve the prediction, or improve the understanding of the factors at play. In particular we briefly discuss two approaches and their merits.

Since the data show significant differences across The Netherlands such as dependencies on the region, on the municipality, on the geographical location, and on the presence of academic hospitals just to name a few, it is strongly advisable to take into account this heterogeneity when fitting a unique model on all the available data. An option is the inclusion of random effects in a simple linear regression model, the so-called linear mixed model of Laird and Ware (1982). The approach is similar to linear regression, but now part of the observed effect is supposed to be due to some random effects. Consequently, the estimate is a sum of two terms: the product of a design matrix (i.e. a matrix of the covariates) with a vector of coefficients, the fixed effects, and the product of another design matrix (containing the same or other covariates) with a vector of random coefficients drawn from a particular distribution. Including these random effects shrinks the estimates of the first deterministic part towards the mean. Since the second term is a random sample drawn from a wider

population, this allows taking into account the effects of some variables that are only partially observed, such as the overall health status. Furthermore, this method is advantageous in terms of estimation. In fact, fitting a traditional regression model including a fixed effect for each unit can become cumbersome when the chosen unit is small, and thus the number of units and corresponding coefficients is large. The bigger advantage of random effects model with respect to fixed effects model is the reduction in the number of parameters. That is, if in a regression model we include a variable for each region (province, zipcode or any other cluster) and we have n regions, then we will have to estimate n coefficients (one for each region). Instead, if we include the random effect for that same variable (region, province, zipcode or any other cluster), then the number of parameters to be estimated reduces to 2 (mean and variance) in case of assumption of normality for the random effects. I.e. the n coefficients (one for each region) are samples from a normal distribution with mean μ and variance $sigma^2$.

Introducing the random effects reduces to one the number of parameters to be estimated to account for the heterogeneity among units, namely the variance of the random effects. For further details and implementation, see for example Verbeke and Molenberghs (2009), Verbeke et al. (2010), and Fitzmaurice et al. (2008).

Another approach is given by the mixture model, which fit different distributions (or the same distribution with different parameters) for each region (province, zipcode or any other cluster). If all the data can be modelled by the same distribution, then there is no need for mixture model - the data is homogeneous. On the other hand, if there is heterogeneity in the data, it can be captured by this flexible model fitting different distributions of the data on different regions.

This method can be used not only to fit a completely new model, but also to check whether the distribution of the data of a certain supposed subpopulation is indeed different from the others. Since it is possible to include and merge almost any desired distribution, it results in an extremely flexible and thus powerful tool in capturing and explaining heterogeneity.

Many statistical softwares have built-in procedures to fit both linear mixed models and mixture models, see R, SAS, STATA, SPSS and Matlab among others. As highlighted earlier, heterogeneity is visible at different levels, thus when fitting these models one might want to explore the effect size of various subpopulations at different scales, such as postcode, municipality, province, and so on.

References

B. Efron and T. Hastie. *Computer Age Statistical Inference.* Cambridge University Press, 2016.

G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs. *Longitudinal Data Analysis.* CRC Press, 2008.

S. Gilmour. The interpretation of mallows's C_p-statistic. *The Statistician*, 45(1): 49–56, 1996.

U. Grömping. Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2):139–147, 2007.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity*. CRC Press, Boca Raton, Florida, 2015.

T. Hesterberg, N. Choi, L. Meier, and C. Fraley. Least angle and ℓ_1 penalized regression: A review. *Statistical Surveys*, 2:61–93, 2008.

A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

T. Kvålseth. Cautionary note about R^2. *The American Statistician*, 39(4):279–285, 1985.

N. Laird and J. Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

C. Mallows. Some comments on C_p. *Technometrics*, 15(4):661–675, 1973.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc. Series B (Methodological)*, pages 267–288, 1996.

G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer Science & Business Media, 2009.

G. Verbeke, G. Molenberghs, and D. Rizopoulos. Random effects models for longitudinal data. In *Longitudinal Research with Latent Variables*, pages 37–96. Springer, 2010.

Zorginstituut Nederland. Zvw 2017, 2017. URL `https://www.zorginstituutnederland.nl/financiering/risicoverevening-zvw/zvw-2017`.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

Quiescent Periods during Helicopter Landings on Ships

Krzysztof Bisewski, Bart M. de Leeuw, Bart Kamphorst, Hans Kraaijevanger, Ivan Kryven, Julia Kuhn, Alberto Montefusco*, Michael Muskulus, Tommaso Nesti, Yuliia Orlova, Mark Peletier

Abstract

The problem of helicopter landing on ships has been recently studied by MARIN (MAritime Research Institute Netherlands) with the purpose of helping the naval crew, and in particular the HLO (Helicopter Landing Officer), to take decisions in a fast and reliable way. The basic issue consisted in the prediction of time intervals, called *quiescent periods* (QPs), where the ship motion is sufficiently moderate for the helicopter to be able to land in safe conditions. The ingredients at our disposal were a set of wave data that were simulated by MARIN with their proprietary software FREDYN. Our first goal, then, was to study the statistics of QPs and to identify patterns. The second objective was to use the same data to make predictions on the basis of a few deterministic and stochastic models. The results show that these models are indeed able to capture several features of the waves, such as repetitions of special patterns and memory effects, and surely deserve further investigation and extension. The last approach was purely analytical: first we focus on the question whether a given sum of n harmonics will have QPs or not. After analyzing the cases $n = 1, 2, 3$ in full detail we present a general criterion for the existence of QPs for the case of arbitrary n. We also give estimates for the frequency and probability of QPs in a signal composed of many random harmonics.

KEYWORDS: helicopter landing on ships, quiescent period, sea waves, narrow-banded signal, random harmonics, level crossing, stationary process, autoregressive model, logistic regression, change-point detection, forecasting, Markov chain

*Corresponding author. E-mail address: alberto.montefusco@mat.ethz.ch

1 Introduction

> "–Well, you must understand, signore,
> that the scirocco blows for three days if it starts on Tuesday.
> Nine days if it starts on Friday.
> But if it hasn't blown itself out by the tenth day,
> then it goes on for 21 days."
>
> *from L. Visconti's screen adaptation of* Death in Venice
> *by* Thomas Mann

Marine operations, both civilian and military, often require a helicopter to land on a ship or other vessel. Safely landing a helicopter requires the landing pad to be approximately stationary for a period of twenty or thirty seconds. Often, such *quiescent periods* (QP) alternate with periods of stronger ship motion, in which landing is impossible. In such cases a *Helicopter Landing Officer* (HLO) on the ship is responsible for guiding in the helicopter and coordinating its descent.

The landing operation consists of two phases. In the first phase, the HLO assesses the general state of the sea at that moment. This is done on the bridge or inside a cabin, and in this phase the HLO observes the sea and has access to a variety of instruments. When the HLO decides that the frequency of quiescent periods is sufficiently high, he signals the helicopter to approach the ship and to start hovering above the landing pad, and takes position outside, next to the landing pad, in view of the helicopter.

In this second phase the HLO maintains eye contact and radio contact with the helicopter pilot, and observes the ship motion through his legs and eyes. When the HLO believes that a quiescent period is imminent, he signals the pilot to land on the pad. During this operation the pilot has no view of the deck, and is completely dependent on the HLO for guidance.

MARIN (MAritime Research Institute Netherlands) is a Dutch organization with the broad goal of studying operations and decommissioning of ships and offshore platforms, bulk and surface hydrodynamics, as well as nautical training and regulations. Currently, they have an open project on helicopter landing on ships, with which they decided to participate in the SWI 2017. The problem posed by MARIN consists of two questions, each related to one of the two phases described above.

First, MARIN is interested in the distribution of quiescent periods in ship motion, given a certain sea state. This would help the HLO to judge whether the ship motion allows for the helicopter landing to take place in the following minutes with a reasonable accuracy.

Secondly, to make the final phase both more efficient and safe, MARIN would like to give the HLO a further instrument to predict the initiation of quiescent periods with a very short advance, in the order of few seconds. This is why, in our work, we developed tools for predictions, given a history of signals of ship motion.

This report is about the properties of certain *signals*. We will be considering two types of signals:

- *Synthetic* signals, created by adding harmonics (sines and cosines, or complex versions of these) with varying frequencies and amplitudes;

- *Data* signals, given to us by MARIN, which describe the movement of the ship in response to certain "sea states".

In reality MARIN generated the data signals by feeding certain well-chosen synthetic signals as "wave input" to a ship simulator called FREDYN, which outputs the movement of a specific ship in response to these waves. For the purposes of this report, however, we consider these data as "externally given".

The data are time series of the motion of a ship under a predefined wave spectrum. As a ship, for our purposes, may be considered as rigid body, what really matters for us is the set of the six coordinates that fully characterize the motion. In marine jargon, these coordinates assume specific names, which are shown in Figure 1.

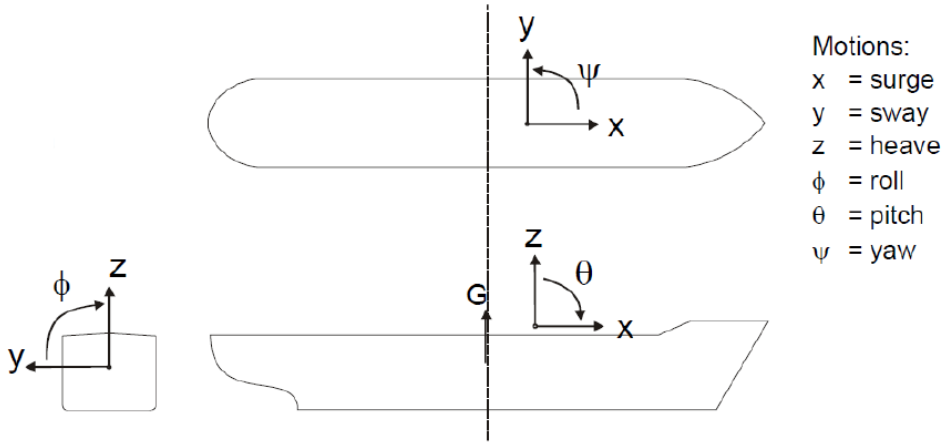


Figure 1: Nomenclature for the ship motion in the marine jargon.

In Sec. 3, we will give more details on this set of data: how it has been generated, how we have used it and what we can say about it. Before doing that, in Sec. 2, we will give a short review of the basic theory of signals that is needed in this report, and in Sec. 4 we will study synthetic signals from an analytical viewpoint. In Sec. 5, we will model the data signals by means of various techniques, with the common aim of predicting quiescent periods.

2 Some signal theory

2.1 Signals

For us, signals are functions defined on $\mathbb{R}$ (as for synthetic signals) or on a discrete set (as for the data), with values that are real or complex. Given a signal f on $\mathbb{R}$, the Fourier transform $\mathcal{F}(f)$ or $\hat{f}$ is the complex-valued function of frequency ω given by

$$\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(t) e^{-i\omega t} \, \mathrm{d}t.$$

As it stands, this integral is only defined if $f \in L^1(\mathbb{R})$; however, a natural extension exists Stein and Weiss (1971) to the set of all tempered distributions $\mathcal{S}'(\mathbb{R})$, by exploiting Parseval's theorem

$$\int_{\mathbb{R}} \hat{f}(\omega)\hat{g}(\omega) \, \mathrm{d}\omega = \int_{\mathbb{R}} f(t)g(t) \, \mathrm{d}t. \tag{1}$$

We will use this extension without mentioning it.

In the discrete case, the signal is only sampled at a finite number of points in time $x_0, x_1, \ldots, x_{n-1}$. Usually these points are multiples of a sampling interval Δ, i.e., the t-th sample x_t is observed at time $t\Delta$. The discrete Fourier transform is then

$$\hat{x}(\nu) = \frac{1}{n} \sum_{t=0}^{n-1} x_t e^{-2\pi i \nu t}. \tag{2}$$

Similar to the continuous Fourier transform, the harmonic functions implicit in Eq. 2 are orthogonal when the frequencies are restricted to the set of *Fourier frequencies*, $\nu_j = j/n$,

$$\sum_{t=0}^{n-1} e^{2\pi i \nu_j t} e^{-2\pi i \nu_k t} = \begin{cases} n & \text{if } j \equiv k \pmod{n}, \\ 0 & \text{otherwise,} \end{cases}$$

and this guarantees the existence of the inverse transform,

$$x_t = \sum_j \hat{x}(\nu_j) e^{2\pi i \nu_j t}. \tag{3}$$

The discretization leads to two phenomena: frequencies higher than the *Nyquist frequency* $1/(2\Delta)$ have an alias in the interval $0 \le \nu \le 1/(2\Delta)$, i.e., appear as an artificial contribution to one of these frequencies. A second undesirable phenomenon is *leakage*, i.e., the appearance of a contribution in the transform at a frequency ν because of the presence of a signal at a different frequency ν_0. This happens (only) if the frequency ν_0 is not a Fourier frequency. More details about this and other practical aspects of Fourier analysis can be found in Bloomfield (2000).

2.2 The harmonics

The *harmonic functions* are an important set of examples. If $f(t) = \cos \omega_0 t$, then $\hat{f}(\omega) = \sqrt{\pi/2}(\delta_{\omega_0} + \delta_{-\omega_0})(\omega)$, where δ_{ω_0} is the Dirac delta function at ω_0; if $f(t) = \sin \omega_0 t$, then $\hat{f}(\omega) = -i\sqrt{\pi/2}(\delta_{\omega_0} - \delta_{-\omega_0})(\omega)$; and if $f(t) = e^{i\omega_0 t}$, then $\hat{f}(\omega) = \sqrt{2\pi}\delta_{\omega_0}(\omega)$. These examples illustrate the general fact that the function f is real-valued if and only if $\hat{f}$ is conjugated-even, i.e. $\hat{f}(\omega) = \overline{\hat{f}(-\omega)}$; similarly, f is purely imaginary iff $\hat{f}$ is conjugated-odd.

Consider the function $f(t) = ae^{i\omega_0 t}$, where $\omega_0 \in \mathbb{R}$ and $a \in \mathbb{C}$. The number ω_0 is called the *angular frequency* and is expressed in radians per second. It can be written as

$$\omega_0 = 2\pi\nu_0, \tag{4}$$

where ν_0 is the *ordinary frequency* expressed in hertz. The word "frequency" can refer to both the angular frequency ω_0 or the ordinary frequency ν_0, depending on the context. The complex number a is called the *complex amplitude*, and contains both the usual amplitude information and information on the phase, since (writing $a = \alpha e^{i\varphi}$, for $\alpha, \varphi \in \mathbb{R}$),

$$ae^{i\omega_0 t} = \alpha e^{i(\omega_0 t + \varphi)} = \alpha\Big[\cos(\omega_0 t + \varphi) + i\sin(\omega_0 t + \varphi)\Big].$$

2.3 Energy spectra and sea states

The *energy spectrum* of a signal f is the real-valued function $\omega \mapsto |\hat{f}(\omega)|^2$. If 'energy' of a function $f \in L^2$ is defined as the L^2-norm $\int |f|^2$, then the value $|\hat{f}(\omega)|^2$ represents the energy of the Fourier component of f with frequency ω, since from (1) we have

$$\int_{\mathbb{R}} |f(t)|^2 \,\mathrm{d}t = \int_{\mathbb{R}} |\hat{f}(\omega)|^2 \,\mathrm{d}\omega.$$

An important type of signal is related to the *sea state*, which is a description of the waves at a certain moment. For our purposes, a sea state is defined by an energy spectrum of the waves, as a function of a two-dimensional frequency (ω_1, ω_2), although in the rest of this report we will mostly disregard the two-dimensionality and consider functions of one variable only: the sea state then describes the energy spectrum of a function f of one variable, which describes the waves. In this interpretation f can be interpreted either as giving the wave height at a fixed point in space as a function of time t, or as giving the wave height at a fixed moment in time as a function of a spatial variable x. We will usually consider the former. (Again there is a difficulty here: we want to consider "waves" as elements of $L^\infty(\mathbb{R})$, as in the case of the harmonics, but such waves have infinite spectrum, since $|\delta_\omega|^2$ cannot be defined as a distribution. For these cases the concept of energy spectrum can be made meaningful by considering large intervals and taking a limit under appropriate rescaling. We omit the details.)

The energy spectrum of a function f alone does not uniquely characterize the function f, since it does not contain any phase information. In addition, for simulation

purposes the spectrum needs to be discretized. This leads to constructing sample functions f, which are assumed to be representative of the waves, of the form

$$f(t) = \sum_{j=1}^{n} a_j e^{i\omega_j t}, \quad \text{or the real part of this } f,$$

where the a_j and ω_j are chosen randomly from the energy spectrum, in such a way as to make $|\hat{f}|^2$ approximately equal to the assumed spectrum. It is natural in such a setup to choose the distribution of $\arg a_j$, i.e. of the phases, to be uniform on $[0, 2\pi)$, reflecting the fact that the energy spectrum contains no information about the phases.

2.4 Narrow-bandedness and its consequences

We observed that the data provided to us by MARIN is *narrow-banded*: the frequencies present in the signal are concentrated in a fairly narrow interval (see Figure 2). This results in a signal with a fairly recognizable period, and an amplitude that varies on a larger scale.

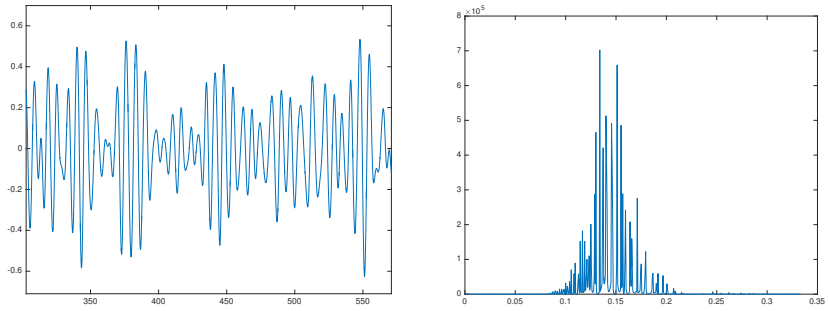


Figure 2: A representative section from *heave* data from MARIN (see the next Section for details). Left is the signal as function of time, right is the spectrum.

Because of this narrow-bandedness the time course resembles an amplitude modulation of a fixed-frequency oscillation, and in the rest of this report we use this way of viewing the signals. This has a number of consequences:

1. The essential information in the data is already encoded in the local maxima and minima; in the data processing that we do, we thus first extract the local maxima and minima, and use the sequence of those data points.

2. For the analysis, one would like to concentrate on the properties of the "envelope" that appears "obvious" to the human eye, since quiescent periods of more than a fraction of the period of the underlying oscillation are one-to-one related to periods in which the amplitude of this envelope is small.

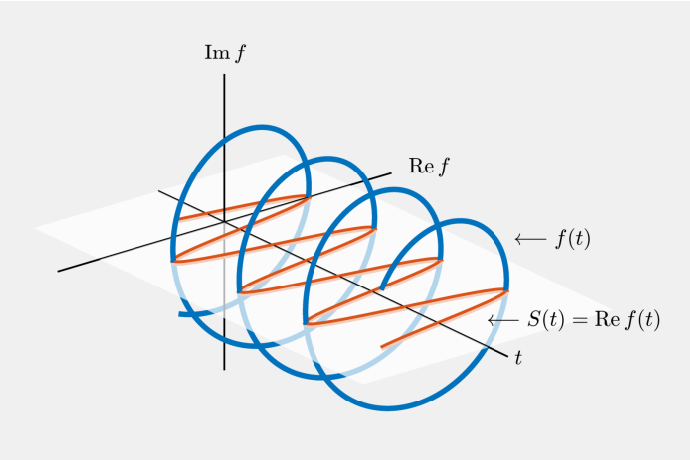


Figure 3: Graphical representation of the *analytic signal*: the red curve is the real signal and the blue complex curve is the corresponding analytic signal.

We now explore this second aspect more in detail. A real-valued signal has a spectrum that is symmetric with respect to frequency 0; "narrow-banded" for a real-valued signal means that the spectrum is concentrated around ω_0 and $-\omega_0$ for some $\omega_0 \neq 0$.

From any complex signal $f(t)$ one can easily construct a real-valued signal $S(t)$ by taking its real part,

$$S(t) = \mathrm{Re}\, f(t). \tag{5}$$

The inverse operation is not unique, however, since there obviously exist many complex-valued signals with the same real part. We can use this freedom of constructing a well-chosen complex counterpart of a given real-valued signal to make the spectrum appear only at positive frequencies. Given a real-valued signal S, its associated *analytic signal* f is defined by concentrating all of the Fourier transform on the *positive* frequencies, i.e. we set

$$\hat{f}(\omega) := \begin{cases} 0 & \omega < 0 \\ \hat{S}(0) & \omega = 0 \\ 2\hat{S}(\omega) & \omega > 0. \end{cases}$$

After transforming $\hat{f}$ back to f, the function f is now complex-valued, and can be interpreted as a an "interpolated" version of the function S, in the sense that (5) holds; and it is an interpolated version that "only rotates in one direction" in the complex plane, as shown in Figure 3. The function f can also be represented as

$$f(t) = S(t) + iH[S](t), \tag{6}$$

where $H[S]$ is the Hilbert transform of S.

The analytic signal now gives us an opportunity to make the concept of "envelope" precise. In general, from each complex-valued function $t \mapsto f(t)$ one can define the real-valued *instantaneous amplitude* and *instantaneous phase* by writing

$$f(t) = A(t)e^{i\phi(t)}, \qquad \text{for some } A(t), \phi(t) \in \mathbb{R}. \tag{7}$$

If the function f is continuous, then A and ϕ can also be taken continuous, and A and ϕ are unique up to adding multiples of 2π to the phase.

The property that f is narrow-banded corresponds to the fact that $\phi'(t)$ is close to ω_0. If f is narrow-banded, then A varies slowly (we illustrate this in Section 4.2), and as a result we can use the function A as a working concept for the intuitive idea of the "envelope".

3 Data signals and their quiescent periods

In the last decades several programs have been developed to study the motion of ships under the forcing of sea waves. MARIN uses its own software, denominated FREDYN, which studies the dynamic behavior of a steered ship subjected to waves and wind. A description can be found in the website MARIN. As the software is a proprietary one, MARIN provided us with several sets of data, varying for time length, direction and spectrum of the waves.

The input of the program was a train of waves given by randomly sampling a well-defined spectrum, typical of the North Sea. The output that was relevant for us consisted of six time series of the six coordinates of ship motion, sampled at regular time intervals.

In our analyses, we mostly focused on the *heave* coordinate *at the landing pad*, since – together with the roll – it is the most important variable for helicopter landing. Although operative conditions for helicopter landing on ships are not well defined by any regulation, there exist such rules for landing on offshore platforms. According to the latter, MARIN suggested the following requirements for a quiescent period:

- peak-to-trough amplitude of *heave* $< 3\,\mathrm{m}$;
- single *roll* amplitude $< 3°$;
- time duration of at least $30\,\mathrm{s}$.

These represent rather strict requirements, which might be relaxed, and are surely too stringent for navy operations.

The first question that MARIN asked us concerns the distribution of quiescent periods. In the present section, we will address this problem by looking at the data signals that we received from MARIN. Some of the data sets were not representative enough either in time duration, or wave spectra didn't include non-quiescent periods. Thus, we considered only a few representative data sets, collected in Table 1.

Alias	U	μ	H_s	T_p	T	Motion sensor
D1	10 kn	180°	3 m	8 s	18000 s	HELI
D2	10 kn	180°	3 m	8 s	7200 s	HELI
D3	10 kn	180°	3 m	8 s	1800 s	HELI (wave spreading)
D4	10 kn	180°	5 m	8 s	1800 s	HELI

Table 1: Data sets generated by the computer program FREDYN. The meaning of the simulation parameters is as follows: U - ship speed, μ - wave direction, H_s - significant wave height, T_p - peak wave period, T - simulation time.

3.1 Distribution of Quiescent Periods

In this section, we will describe the procedure of data pre-processing and the idea of finding QPs in the considered system. According to MARIN's definition of QPs explained above, only several data sets were suitable for this analysis as for some data sets the system never went out of the quiescent state.

First of all, roll and heave are chosen as the most representative coordinates. Due to the definition of the QPs, only extrema of the signals of these two coordinates are taken into account as points lying between extrema don't contribute to the analysis. For purposes of convenience, we suggest to work with absolute values of signals. In this case, the *single amplitude* is the height of the peak; the *peak-to-trough amplitude* is the sum of heights of two neighboring peaks.

In Figure **??** one can see the absolute values of the signals for the roll and heave coordinates from the data set D4 of Table 1. It appeared that in all data sets the signal for the roll coordinate was not exceeding the threshold of 3°. Thus we agreed with MARIN to lower the threshold for single roll amplitude from 3° to 2° in order to illustrate the whole QP search procedure. Green asterisks denote those peaks that do not fall into the definition of the QP for the considered coordinate. Thus, the QPs are those areas, which lie between green asterisks. In the plot we illustrate QPs with an indicator function, which takes the value 1 if extrema are in a QP, and 0 otherwise:

$$\mathbb{1}_{QP} = \begin{cases} 1, & x \in \text{QP} \\ 0, & \text{otherwise.} \end{cases}$$

As we have separate QPs for roll and heave, we can determine QPs for the whole system. For this purpose, we take an intersection of these areas for both signals. According to the definition, we consider only those periods that last longer than 30 seconds.

Further, we would like to look at the distribution of the duration of QPs. From Figure 5 we can see that the statistics of QPs is not good enough. Thus, we apply the same search procedure on the data set from the longer simulation of 18000 seconds (D1). On the upper plot in Figure 5 we can see how often QPs with different duration

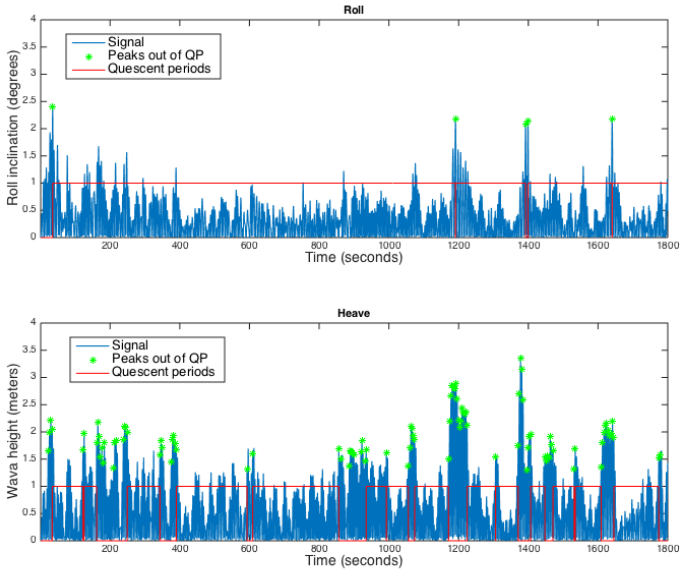


Figure 4: Example of quiescent periods found in a raw data set from the simulation of 1800 seconds for roll (up) and heave (down) signals. QPs are illustrated with an indicator function (red line), which takes the value 1 if absolute values of extrema are in a QP, and 0 otherwise. Green asterisks indicate extrema which do not fall into the definition of s QP for roll and heave respectively.

appear in the system. The lower plot in Figure 5 corresponds to the distribution of the time intervals when the system is not in a QP.

From the plot in Figure 5 we notice that the distribution of the time intervals for QPs reminds of the shape of the probability density function of the exponential distribution, in which case one could model the occurrences of random events as a Poisson process. This observation may be verified by statistical hypothesis testing, which has not been done in the current work. Furthermore, the histogram of the durations of QPs captures the information about the sea state in a specific time interval. Thus, it could help the HLO to judge the behavior of the sea and estimate how many QPs one might expect in the current situation.

3.1.1 Summary

The aim of this section was to examine the data signals generated by FREDYN from a descriptive standpoint and gain an idea about the nature of the occurrences of QPs in waves. Upon analyzing the histogram of the durations of QPs, one may assume that the data follows an exponential distribution. However, to conclude this, we

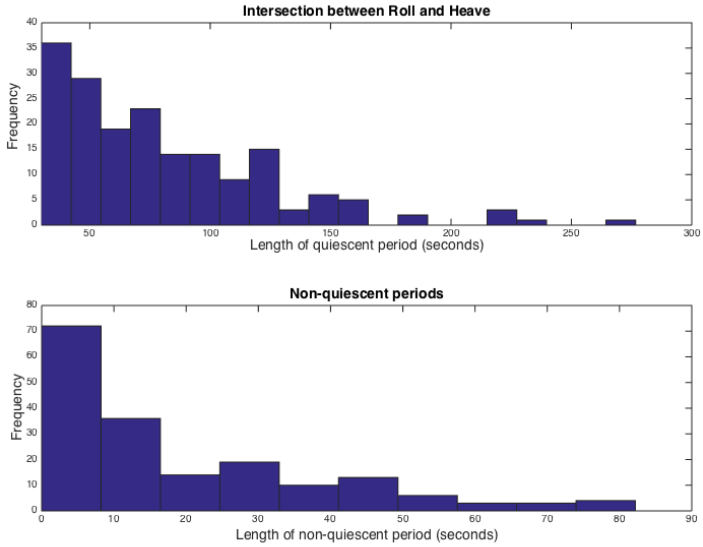


Figure 5: Distribution of the duration of quiescent (up) and non-quiescent (down) periods in a data set D1. The upper plot shows the number of occurrences of QPs with different time duration from the considered simulation. The lower plot depicts how many non-QPs fall in the bins for different duration in the same simulation.

would need to analyze longer simulations with more variations in the wave profile and perform a statistical hypothesis test. If the test confirms the exponential distribution, one might consider to model the occurrences of QPs according to a Poisson process.

3.2 Qualitative patterns

An interesting way of studying qualitative patterns in the signal related to QPs is the use of *event-related analysis*. After QPs have been defined and identified in the signal, one cuts the time series into short segments around the beginning of each QP and aligns these periods such that the QPs start at the same relative time (or lag). An example is shown in Figure **??** for the extrema of the heave signal in the data set D2. In fact, only the absolute values of the extrema were used in this analysis, as otherwise QPs starting with negative or with positive extrema would be mixed and the relevant information would be averaged out. The start of the QP, i.e., the event used for the alignment of the signals, is marked with a vertical red line. The condition imposed by the event is that the first extremum before the event has to lie above the threshold (marked by the horizontal red line), and the first extremum inside the event has to lie below it.

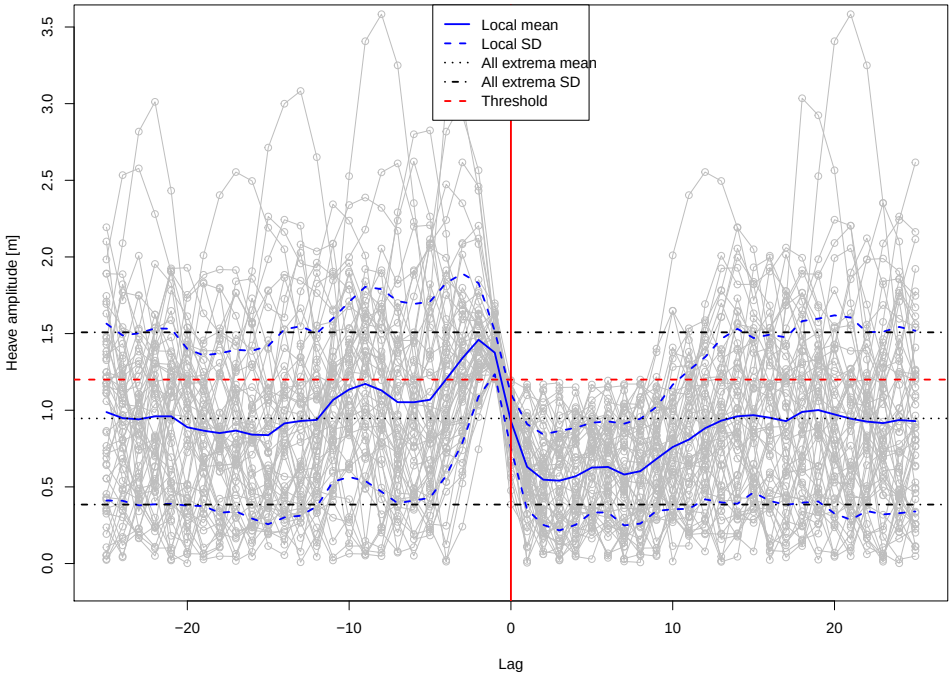


Figure 6: Transition to quiescent period in the data set D2. Shown are subsequent absolute values of extrema of the heave signal, conditional on the event that a quiescent period starts (marked by red vertical line). For simplicity, here the quiescent period has been defined to be at least 30 seconds of heave signal below a threshold value of 1.2 m. The mean and standard deviation of the individual time traces are indicated (blue curves), as well as the overall mean and standard deviation of the (absolute values of) extrema (black lines).

What is somewhat unexpected, and therefore interesting, is that the extrema seem to have been higher than average already for about 5 waves (equal to 10 extrema) before the event, on average. The length of this period corresponds to the average length of the QP in this case, which is also about 5 waves – although this might be a coincidence. After the QP, the statistical properties of the extrema quickly approach the overall distribution indicated in the figure (i.e., the blue curves approach the black lines), within about 4 extrema.

This and related figures (e.g. for different conditions imposed on the extrema) can provide important hints for what patterns are present in the signals and how to

exploit these. One example of a more quantitative analysis of these patterns will be given in Sec. 5.4.2.

4 Distribution of QPs by analytic estimates

The motion of the ship is the net result of the mechanics of the ship and the forces exerted on the ship by the waves. Exactly characterizing the forces on the ship that result from the waves is non-trivial, and beyond our scope. Instead of focusing on the ship, we have therefore focused on the waves.

More precisely, we have addressed the question

> Given a signal on $\mathbb{R}$ with specified spectrum and random amplitudes and phases, what is the distribution of quiescent periods?

Again, this requires specification, since a typical spectrum has a full support. Instead we consider signals with discretized spectra, of the form

$$f(t) = \sum_{j=1}^{n} a_j e^{i\omega_j t}, \tag{8}$$

for some finite n, where a_j are complex amplitudes chosen such that the spectrum of f resembles a given spectrum, and such that the phases are uncorrelated. As discussed in Section 2.4 this complex signal f can be 1-to-1 related to a real signal S, which is simply obtained from f by taking its real part (see (5)),

$$S(t) = \sum_{j=1}^{n} \alpha_j \cos(\omega_j t + \phi_j), \tag{9}$$

where $\alpha_j = |a_j|$ and $\phi_j = \arg a_j$. We emphasize that for any real-valued signal S of the form (9) its associated complex signal f is uniquely defined and should be seen as its analytic representation (see Section 2.4).

In the software FREDYN the ship model is driven by one or more of such signals, representing wave trains from different directions. In this case $n \approx 100$, but we will also address the small-n case; it turns out that interesting insight can be gained from $n = 2$ and $n = 3$, for instance.

4.1 Definition of quiescent periods

In the context of a general signal of the form (8), describing the behaviour of waves, it does not make much sense to consider a quiescent period as defined by an absolute criterion. Instead we consider quiescent periods as defined by a relative criterion, characterized by two parameters and a choice of norm:

Definition 4.1. Let $\tau > 0$ and $\theta > 0$ be given. Given a signal of the form (8) a quiescent period is defined by the property

$$\|f\|_{[t,t+\tau]} \leq \theta \,\text{`average'}\big(\|f\|_{[t',t'+\tau]}\big). \tag{10}$$

Here $\|f\|_{[t,t+\tau]}$ can be any norm of f that is calculated over the time section $[t, t+\tau]$; we will consider two different norms below. The parameter θ is a threshold: a quiescent period is a period in which the norm of f over that period is less than θ times the average value of the norm. The 'average' can be interpreted in two ways – either the average over times t', or the expectation of the randomly chosen coefficients. We will use both below.

4.2 The narrow bandwidth assumption

It is unclear to us how to characterize the rate of occurrence of quiescent periods in a completely arbitrary signal. In order to make the question more amenable to analysis we concentrate in all of Section 4 on the case of *narrow bandwidth*, as discussed in Section 2.4: we assume that there exists a *reference frequency* $\omega > 0$ and a *bandwidth* $\varepsilon \geq 0$ such that

$$|\omega_j - \omega| \leq \varepsilon \ll \omega \quad \text{for all } j = 1, 2, \ldots, n. \tag{11}$$

We refer to Figure 7 for a graphical illustration of this assumption.

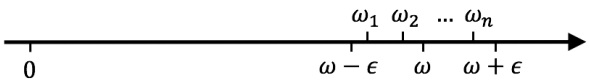


Figure 7: All angular frequencies ω_j are ε-close to the reference frequency ω.

Using the narrow-bandedness assumption, we rewrite the complex signal $f(t)$ defined in (8) as

$$f(t) = e^{i\omega t} f_0(t) \tag{12}$$

so that the function f_0 can be written in terms of the real-valued amplitudes $\alpha_j > 0$ and phases ϕ_j as

$$f_0(t) = \sum_{j=1}^{n} \alpha_j e^{i[(\omega_j - \omega)t + \phi_j]}. \tag{13}$$

Since $f_0(t)$ only differs by a factor $e^{i\omega t}$ from $f(t)$, its polar form

$$f_0(t) = A(t) e^{i\phi_0(t)} \tag{14}$$

has the same instantaneous amplitude $A(t)$ as $f(t)$, whereas the instantaneous phases $\phi(t)$ and $\phi_0(t)$ are related by

$$\phi(t) = \omega t + \phi_0(t). \tag{15}$$

This implies that the corresponding real signal $S(t) = \operatorname{Re} f(t)$ can be written as

$$S(t) = A(t) \cos \phi(t) = A(t) \cos(\omega t + \phi_0(t)). \tag{16}$$

If the bandwidth ε is small, the value $f_0(t)$ moves slowly through the complex plane since it follows from (11) and (13) that its velocity is bounded by

$$|f_0'(t)| \leq \varepsilon \sum_{j=1}^{n} \alpha_j.$$

This allows us to bound the time derivatives of both the instantaneous amplitude $A = |f_0|$ and the reduced phase ϕ_0. Differentiating (14) we find

$$f_0'(t) = A'(t)e^{i\phi_0(t)} + i\phi_0(t)A(t)e^{i\phi_0(t)},$$

so that

$$A'(t) + i\phi_0'(t)A(t) = f_0'(t)e^{-i\phi_0(t)}.$$

Splitting the left-hand side into real and imaginary parts, we find that the instantaneous amplitude $A(t) = |f_0(t)|$ is slowly changing,

$$|A'(t)| \leq |f_0'(t)| \leq \varepsilon \sum_{j=1}^{n} \alpha_j,$$

and also that the phase rate $\phi_0'(t)$ of $f_0(t)$ is small,

$$|\phi_0'(t)| \leq \frac{|f_0'(t)|}{A(t)} \leq \varepsilon \frac{\sum_{j=1}^{n} \alpha_j}{A(t)},$$

provided $f_0(t)$ stays away from the origin. In that case it follows from (15) that the phase rate $\phi'(t)$ of $f(t)$ is approximately equal to the reference frequency ω,

$$\phi'(t) = \omega + \phi_0'(t) \approx \omega. \tag{17}$$

We conclude that the real signal $S(t) = \operatorname{Re} f(t)$ can be written in the form (16), where the instantaneous amplitude $A(t)$ is the modulus of the slowly varying complex-valued function $f_0(t)$ defined in (13), and the instantaneous (angular) frequency $\omega(t) = \phi'(t)$ is approximately equal to the reference frequency ω (see (17)).

This remark allows us to refocus our attention. The *reference time period* associated with the reference frequency ω is given by

$$T = \frac{2\pi}{\omega}. \tag{18}$$

In practice, the minimal length τ of a quiescent period is significantly longer than T. This implies that the real-valued signal S can only be small over a time τ if the amplitude A also is small over that period (i.e., the smallness can not come from the cosine in (16); it has to come from A). Therefore, in our quest for suitable quiescent periods we can limit ourselves to time intervals where the instantaneous amplitude $A(t)$ is small; or equivalently, we can focus on f_0 instead of f. Our aim therefore becomes

Find periods (or characterize the probability of periods) such that the modulated signal f_0 is small over a period τ.

In the following we will first take a "deterministic" approach, which is followed by a "stochastic" approach. In the "deterministic" approach, we derive criteria for the existence of quiescent periods for arbitrary real signals S of the form (9) (and their complex counterpart f defined in (8)). In dedicated subsections we first consider the cases $n = 1$, $n = 2$ and $n = 3$ in detail before we analyze the case of arbitrary n. After completing the "deterministic" case we turn our attention to the stochastic case, where the complex amplitudes a_j of the complex signal f in (8) are stochastic variables. In that case we will study quiescent periods of randomly sampled signals.

4.3 The deterministic case for $n = 1$

If $n = 1$, the real signal $S(t)$ defined in (9) consists of a single cosine,

$$S(t) = \alpha_1 \cos(\omega_1 t + \phi_1), \qquad \alpha_1 > 0,\ \phi_1 \in \mathbb{R}. \tag{19}$$

In this case the bandwidth is equal to $\varepsilon = 0$ and the reference frequency is equal to $\omega = \omega_1$. Quiescent periods longer than the reference value T defined in (18) only occur if α_1 is small enough, and in that case the quiescent period lasts forever.

For completeness we note that the associated complex signal $f(t)$ defined in (8) has instantaneous amplitude $A(t) \equiv \alpha_1$ and instantaneous phase $\phi(t) \equiv \omega_1 t + \phi_1$, showing that $f(t)$ moves on a circle with radius α_1 centered around the origin with uniform angular velocity ω_1. In contrast, the complex signal $f_0(t)$ defined in (13) is constant, and corresponds to a fixed point in the complex plane. In Figure 8 we have displayed the signal $S(t)$ and the (constant) instantaneous amplitude $A(t)$ of its associated complex signal for $n = 1$, $\alpha_1 = 1$, $\omega_1 = 1$, $\phi_1 = 1$.

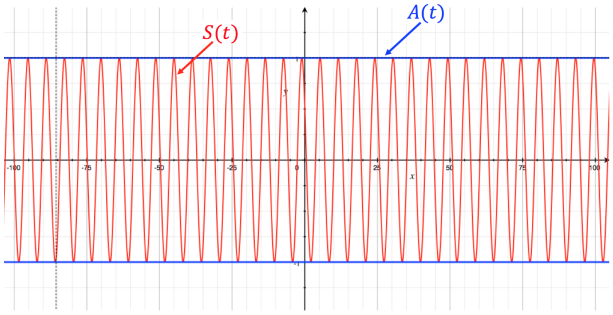


Figure 8: For $n = 1$, $\alpha_1 = 1$, $\omega_1 = 1$, $\phi_1 = 1$ we have displayed the real signal $S(t)$ and the instantaneous amplitude $A(t)$ of its associated complex signal for $t \in [-100, 100]$.

4.4 The deterministic case for $n = 2$

If $n = 2$ we assume without loss of generality that $\omega_1 < \omega_2$. We set $\omega = \omega_1$ so that the bandwidth equals $\varepsilon = \omega_2 - \omega_1$. The complex signal $f_0(t)$ defined in (13) is given by

$$f_0(t) = \alpha_1 e^{i\phi_1} + \alpha_2 e^{i(\varepsilon t + \phi_2)}, \qquad \alpha_1, \alpha_2 > 0,\ \phi_1, \phi_2 \in \mathbb{R}. \tag{20}$$

Clearly $f_0(t)$ moves on a circle with center at $\alpha_1 e^{i\phi_1}$ and radius α_2 with a relatively low constant velocity given by

$$|f_0'(t)| = \alpha_2 \varepsilon. \tag{21}$$

For the corresponding instantaneous amplitude $A(t) = |f_0(t)|$ we find

$$\begin{aligned} A(t) &= |\alpha_1 e^{i\phi_1} + \alpha_2 e^{i(\varepsilon t + \phi_2)}| = |\alpha_1 + \alpha_2 e^{i(\varepsilon t + \Delta\phi)}| \\ &= \sqrt{\alpha_1^2 + \alpha_2^2 + 2\alpha_1\alpha_2 \cos(\varepsilon t + \Delta\phi)}, \end{aligned}$$

where

$$\Delta\phi = \phi_2 - \phi_1.$$

Clearly, A is a periodic function (with period $2\pi\varepsilon^{-1}$) that varies between its minimum $|\alpha_1 - \alpha_2|$ and its maximum $\alpha_1 + \alpha_2$. Quiescent periods only occur if this minimum is small enough. This is the case if α_1 is sufficiently close to α_2. In Figure 9 we have displayed such an example with $\alpha_1 \approx \alpha_2$, $\varepsilon = 0.11$ and $\Delta\phi = -1$.

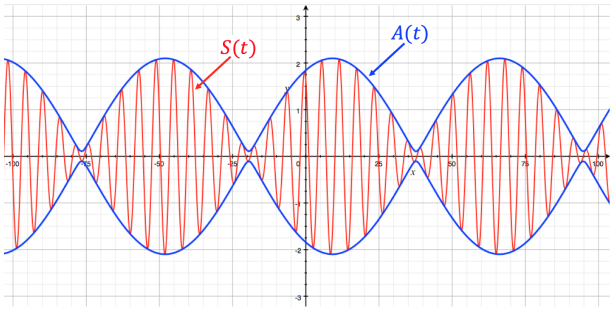


Figure 9: A typical example of a real signal $S(t)$ and the instantaneous amplitude $A(t)$ of its associated complex signal for $t \in [-100, 100]$. Here we have chosen $n = 2$, $\alpha_1 = 1$, $\alpha_2 = 1.1$, $\omega_1 = 1$, $\omega_2 = 1.11$, $\phi_1 = 1$, $\phi_2 = 0$.

4.5 The deterministic case for $n = 3$

If $n = 3$ we assume without loss of generality that $\omega_1 < \omega_2 < \omega_3$. We define $\varepsilon_1 = \omega_2 - \omega_1$ and $\varepsilon_3 = \omega_3 - \omega_2$ (see Figure 10).
Setting $\omega = \omega_2$ the complex signal $f_0(t)$ defined in (13) is given by

$$f_0(t) = \alpha_1 e^{i(-\varepsilon_1 t + \phi_1)} + \alpha_2 e^{i\phi_2} + \alpha_3 e^{i(\varepsilon_3 t + \phi_3)}, \qquad \alpha_1, \alpha_2, \alpha_3 > 0,\ \phi_1, \phi_2, \phi_3 \in \mathbb{R}. \tag{22}$$

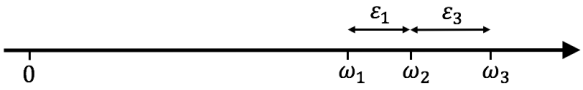


Figure 10: For $j = 1, 3$ the distance $|\omega_j - \omega_2|$ is denoted by ε_j.

This shows that the trajectory of $f_0(t)$ is the result of the superposition of two circular motions with relatively low angular velocities ($-\varepsilon_1$ and ε_3). In Figure 11 we have displayed two such trajectories.

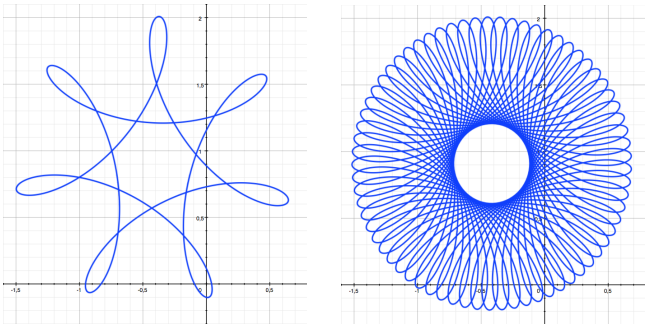


Figure 11: The trajectory of the complex signal $f_0(t)$. In the left figure we have chosen $n = 3$, $\alpha_1 = 0.4$, $\alpha_2 = 1$, $\alpha_3 = 0.7$, $\omega_1 = 0.95$, $\omega_2 = 1$, $\omega_3 = 1.02$, $\phi_1 = 1$, $\phi_2 = 2$, $\phi_3 = 3$. In the right figure we have only slightly changed ω_1 from 0.95 into 0.951.

In general we can distinguish the following two cases:

- **The "rational" case:** the ratio $\varepsilon_3/\varepsilon_1$ is a rational number
- **The "irrational" case:** the ratio $\varepsilon_3/\varepsilon_1$ is irrational

Both cases displayed in Figure 11 are 'rational' since the ratios $\varepsilon_3/\varepsilon_1$ are $2/5$ and $20/49$, respectively. In the general 'rational' case there exist two positive integers k and ℓ such that

$$\frac{\varepsilon_3}{\varepsilon_1} = \frac{\ell}{k}, \tag{23}$$

where we may assume, without loss of generality, that k and ℓ are relatively prime. One easily verifies that in this case the complex signal $f_0(t)$ has a periodic orbit with period

$$\Delta t = 2\pi k \varepsilon_1^{-1} = 2\pi \ell \varepsilon_3^{-1}. \tag{24}$$

For the two cases displayed in Figure 11 the periods are $\Delta t = 200\pi$ and $\Delta t = 2000\pi$, respectively. For the graphs of the corresponding real signals $S(t)$ we refer to Figures 12 and 13. In Figure 12 (which corresponds to the left trajectory in Figure 11) we

see that the amplitude $A(t)$ has indeed a period $\Delta t = 200\pi$ and that each period has exactly one quiescent period. In Figure 13 (which corresponds to the right trajectory in Figure 11) we have limited the time window to $[-300, 2700]$, which is less than half the period $\Delta t = 2000\pi$ of the amplitude $A(t)$. Comparing the latter figure to Figure 12, we see that both graphs are very similar for times in the interval $[-300, 500]$, including the two quiescent periods marked with a black arrow. This is not surprising since the only difference between both cases is a slightly different value of ω_1. For later times, the difference between both graphs becomes more pronounced, which also illustrates the fact that the period of the amplitude $A(t)$ in the second graph is 10 times as large as the amplitude of $A(t)$ in the first graph.

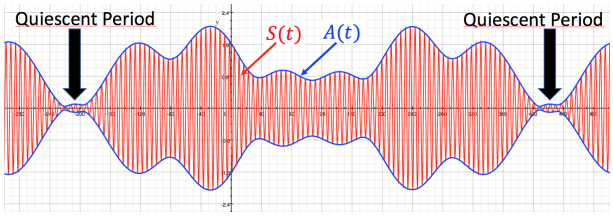


Figure 12: For the case displayed in Figure 11 on the left, this is the graph of the real signal $S(t)$ (in red) and the instantaneous amplitude $A(t) = |f_0(t)|$ of its associated complex signal (in blue) for $t \in [-300, 500]$. The period of the amplitude function A is 200π, which is exactly the distance between two quiescent periods.

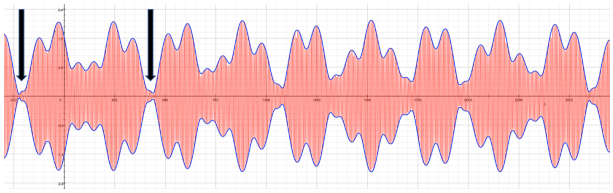


Figure 13: For the case displayed in Figure 11 on the right, this is the graph of the real signal $S(t)$ (in red) and the instantaneous amplitude $A(t) = |f_0(t)|$ of its associated complex signal (in blue) for $t \in [-300, 2700]$.

We finally discuss the "irrational" case, where the number $\varepsilon_3/\varepsilon_1$ is not a rational number. In this case, as opposed to the rational case, the trajectory of the complex signal $f_0(t)$ is not periodic. One easily verifies directly from its definition in (22) that the trajectory of $f_0(t)$ is contained in the complex annulus Ω given by

$$\Omega = \{z \in \mathbb{C} : \ |\alpha_1 - \alpha_3| \leq |z - \alpha_2 e^{i\phi_2}| \leq \alpha_1 + \alpha_3\}. \tag{25}$$

With some imagination, such an annulus can already be recognized in Figure 11 on the right. Indeed, if we change the value $\omega_1 = 0.951$ in the right example into an arbitrary

irrational number close to 0.951, the corresponding trajectory would "densely" fill the complete annulus Ω. This means that for each $z \in \Omega$, $T > 0$, $\varepsilon > 0$ there exists a time $t > T$ with $|f_0(t) - z| < \varepsilon$.

The existence of quiescent periods for the irrational case depends on the proximity of the origin to the annulus Ω. If the distance $d(0, \Omega)$ is small (which is the case, for example, if $0 \in \Omega$), there will exist infinitely many quiescent periods, but the spacing of these periods will be chaotic (as opposed to the regular spacing in the rational case). It easily follows from the definition of the annulus in (25) that the proximity criterion for the existence of quiescent periods is given by

$$|\alpha_1 - \alpha_3| \ \leq \ \alpha_2 \ \leq \ \alpha_1 + \alpha_3, \tag{26}$$

which is equivalent to the more symmetric condition that each of the numbers α_1, α_2, α_3 is less than or equal to the sum of the other two. The latter condition can further be rewritten into the single condition

$$\max(\alpha_1, \alpha_2, \alpha_3) \ \lesssim \ \frac{1}{2}(\alpha_1 + \alpha_2 + \alpha_3). \tag{27}$$

4.6 The deterministic case for arbitrary n

We consider an arbitrary real signal S of the form (9) with angular frequencies $\omega_j > 0$, real amplitudes $\alpha_j > 0$ and phase shifts $\phi_j \in \mathbb{R}$. By renumbering we can assume without loss of generality that

$$\alpha_1 \geq \alpha_2 \geq ... \geq \alpha_n. \tag{28}$$

Setting $\omega = \omega_1$ and $\varepsilon_j = \omega_j - \omega_1$ $(j = 2, 3, ..., n)$, the complex signal $f_0(t)$ defined in (13) is given by

$$f_0(t) = \alpha_1 e^{i\phi_1} + \sum_{j=2}^{n} \alpha_j e^{i(\varepsilon_j t + \phi_j)}. \tag{29}$$

We make again a distinction between the "rational" and "irrational" case. In the rational case the numbers $\varepsilon_2, \varepsilon_3, ..., \varepsilon_n$ are rationally dependent, which means that there exist integers $k_2, k_3, ..., k_n$, not all zero, such that

$$k_2\varepsilon_2 + k_3\varepsilon_3 + ... + k_n\varepsilon_n = 0. \tag{30}$$

In the irrational case the numbers $\varepsilon_2, \varepsilon_3, ..., \varepsilon_n$ are rationally independent, which means that the only way for (30) to hold is that all integers $k_2, k_3, ..., k_n$ are zero.

We first deal with the **irrational case**. In that case it follows from Kronecker's theorem (Hardy and Wright, 1979, Theorem 444) that the trajectory of $f_0(t)$ is contained in and densely fills the set

$$\Omega = \{\alpha_1 e^{i\phi_1} + \sum_{j=2}^{n} \alpha_j z_j : \ z_j \in \mathbb{C}, |z_j| = 1\}. \tag{31}$$

One easily verifies (with induction) that the set Ω is a (closed) annulus in the complex plane with center $z = \alpha_1 e^{i\phi_1}$ and (external/internal) radii given by

$$r_{\text{ext}} = \alpha_2 + \alpha_3 + ... + \alpha_n$$
$$r_{\text{int}} = \max(0, \alpha_2 - \alpha_3 - ... - \alpha_n).$$

It follows that the distance from the origin to the set Ω is equal to

$$d(0, \Omega) = \max(0, \alpha_1 - \alpha_2 - ... - \alpha_n). \tag{32}$$

Since quiescent periods are periods in which $A(t) = |f_0(t)| \approx 0$, there exist quiescent periods if and only if $d(0, \Omega) \approx 0$, which is equivalent to the condition

$$\alpha_1 - \alpha_2 - ... - \alpha_n \lesssim 0.$$

Hence we have shown that in the irrational case there exist quiescent periods if and only if

$$\max(\alpha_1, \alpha_2, ..., \alpha_n) \lesssim \frac{1}{2}(\alpha_1 + \alpha_2 + ... + \alpha_n). \tag{33}$$

In the **rational case** (which should be seen as exceptional) the situation is slightly different. In that case the trajectory of $f_0(t)$ is still contained in the set Ω, but it does not densely fill that set. Hence condition (33) is necessary but not sufficient for the existence of quiescent periods.

4.7 Random sampling of signals for arbitrary n

We now turn to the case of an arbitrary number n of harmonics, still under the narrow-bandwidth assumption. The case of arbitrary n arises when representing a "general" signal with a certain given spectrum. In practice, e.g. for the simulation tool FREDYN, frequencies ω_j and complex amplitudes a_j are drawn randomly from a distribution modeled on the spectrum. Since the spectrum does not contain information about the phases, the phases are chosen following a uniform distribution.

We mimic this situation as follows. First we assume that a set of frequencies $\omega_j \in \mathbb{R}$, $j = 1, \ldots, n$ are *given*, once and for all. Next we assume that $a_1, \ldots, a_n$ are independent, centered, complex Gaussian random variables, i.e. $a_j \sim \mathcal{CN}(0, \sigma_j I_2)$, for some $\mathcal{C} > 0$ and $\sigma_j > 0$, where I_2 is the two-dimensional identity matrix. We then let f_0 be given by

$$f_0(t) = \sum_{j=1}^{n} a_j e^{i(\omega_j - \omega)t}. \tag{34}$$

The above-mentioned slower time scale of $f_0(t)$ corresponds to the fact that $|\omega_j - \omega| \leq \varepsilon \ll \omega$.

By choosing the coefficients to be random variables in $\mathbb{C}$, the functions f and f_0 become random variables in $L^\infty(\mathbb{R}; \mathbb{C})$; the assumption that the coefficients are *normal*

makes the functions f and f_0 *Gaussian processes.*[1] Because time translation corresponds to multiplying the coefficients by unit-length complex numbers, and because the coefficients are normally distributed with mean zero and isotropic covariance, the process is *stationary.*

4.8 The level-crossing approach for arbitrary n

The study of extremes of a stochastic process has been a topic of great interest in engineering. For stationary processes, the main tool has been Rice's formula for the expected number of level crossings Rice (1944) and its generalizations. The most recent account of this theory has been given by Lindgren (2013). For a Gaussian stationary process X_t with zero mean, as we're considering here, the number of upcrossings of the level $u > 0$ per unit time is given by

$$\mu^+(u) = \frac{1}{2\pi}\sqrt{\frac{\lambda_2}{\lambda_0}}e^{-u^2/(2\lambda_0)},$$

where $\lambda_k = \int_{-\infty}^{\infty} |\omega|^k \mathrm{S}(\omega)\,\mathrm{d}k$ are the spectral moments of X_t; here, if we choose $X = f$ as in (8), then we have

$$\lambda_k = \sum_{j=1}^{n} |a_j|^2 |\omega_j|^k.$$

The up-crossings of the mean level define the *mean period* $T_2 = 1/\mu^+(0) = 2\pi\sqrt{\lambda_0/\lambda_2}$.

Using this approach, Cramér and Leadbetter (1967) have studied the following problem: A process X_t is said to *fade* below a level u if the envelope R_t of X_t has a downcrossing of the level u. The *length of the fade* is the time between a downcrossing and the next upcrossing of the level u by R_t. This corresponds closely to our notion of a quiescent period (for a single variable, e.g. the heave signal).

Let us quote Lindgren here (Lindgren, 2013, p.261): *'One of the most intriguing problems in stationary process theory is that of the distribution of the length of excursions above a critical fixed level. Even for Gaussian processes, no explicit solution is known, except in a few cases.'* However, Lindgren then goes on to present *'a method to numerically calculate the* exact *distributions of excursion length'*, based on the evaluation of an infinite dimensional expectation for the so-called Slepian model. Unfortunately this is beyond the scope of this report, but could be very useful for the first problem posed by MARIN. Some of the numerical calculations are available in the WAFO Matlab toolbox The WAFO group (2011).

Generalizing the analysis to vector processes, Lindgren even mentions the phenomenon of *the seventh wave*, i.e. "the observation that waves on a shore or on the ocean seem to have a typical regularity of one big wave followed by six smaller ones" (Lindgren, 2013, p.271). The expected number of u-upcrossings of the envelope $R(t)$

[1] A Gaussian process is a stochastic process whose finite marginals are distributed according to multivariate normal distributions.

per unit time interval is given by

$$\mu_R^+(u) = \sqrt{\frac{\lambda_2(1-\rho^2)}{2\pi\lambda_0}}\frac{u}{\sqrt{\lambda_0}}e^{-u^2/(2\lambda_0)},$$
$$= \sqrt{\frac{\lambda_0\lambda_2-\lambda_1^2}{2\pi\lambda_0^3}}ue^{-u^2/(2\lambda_0)},$$

where $\rho^2 = \lambda_1^2/(\lambda_0\lambda_2)$ is the squared correlation between Hilbert transform and derivative of the process. The inverse of this corresponds to the result given by Cramér and Leadbetter (1967) for the mean length of a fade. And the average number of envelope u-upcrossings per mean period is

$$T_2\mu_R^+(u) = \sqrt{2\pi(1-\rho^2)}\frac{u}{\sqrt{\lambda_0}}e^{-u^2/(2\lambda_0)},$$

and this corresponds to the inverse of the average number of waves per envelope upcrossing.

4.9 Alternative estimates for arbitrary n

In this report we also derive a different type of estimate. As remarked above, the modulus $|f(t)|$ equals the modulus $|f_0(t)|$ for all t. We exploit this by choosing the norm $\|f\|_{[t,t+\tau]}$ in Definition 4.1 to be the sup-norm of f on $[t,t+\tau]$, i.e. $\|f\|_{L^\infty(t,t+\tau)} := \sup_{s\in[t,t+\tau]}|f(s)|$. Then, it follows that $\|f\|_{L^\infty(t,t+\tau)} = \|f_0\|_{L^\infty(t,t+\tau)}$; also, since the process is stationary, the distribution of $\|f\|_{L^\infty(t,t+\tau)}$ is independent of t, so that $\mathbb{E}(\|f\|_{L^\infty(t,t+\tau)})$ is independent of t. In this context we interpret the "average" mentioned in Definition 4.1 as this expectation.

Then the probability of a quiescent period of length τ at time t equals

$$\mathbb{P}\Big(\|f\|_{L^\infty(t,t+\tau)} \le \theta\mathbb{E}(\|f\|_{L^\infty(t,t+\tau)})\Big) = \mathbb{P}\Big(\|f_0\|_{L^\infty(t,t+\tau)} < \theta\mathbb{E}(\|f_0\|_{L^\infty(t,t+\tau)})\Big), \tag{35}$$

and as we already mentioned this probability is independent of t.

As it is difficult to analyse $\|f_0\|_\infty$ directly, we first focus on the L_2-norm $\|f_0\|^2_{L^2(t,t+\tau)} := \int_t^{t+\tau}|f_0(t')|^2\,dt' \le \tau\|f_0\|^2_{L^\infty(t,t+\tau)}$. We prove the following theorem:

Theorem 4.2 (Estimate of the distribution of the L^2-norm)**.** *Let f_0 be the Gaussian process that we construct above, and assume that $\varepsilon \ll 2\pi/\tau$. Then*

$$\mathbb{P}\Big(\|f\|^2_{L^2(t,t+\tau)} \le \theta^2\mathbb{E}(\|f\|^2_{L^2(t,t+\tau)})\Big) \approx 1-e^{-\theta^2}. \tag{36}$$

Note that the condition $\varepsilon \ll 2\pi/\tau$ is stronger than the earlier narrow-bandedness assumption $\varepsilon \ll \omega = 2\pi/T$, whenever $\tau > T$ (see the discussion on page 65). Under this assumption, over an interval $(t,t+\tau)$, the signal looks like a single harmonic (whose amplitude and phase can be viewed both as random for fixed t, or alternatively as t-dependent for each realization).

Proof. The L^2-norm of f_0 is readily computed. We have

$$\|f_0\|^2_{L^2(0,\tau)} = \int_0^\tau f_0(s)\overline{f_0(s)}\,\mathrm{d}s = \sum_{j,k=1,\dots,n} a_j\overline{a_k}\int_0^\tau e^{i(\omega_j-\omega_k)s}\,\mathrm{d}s = \sum_{j,k=1,\dots,n} A_{jk}\,a_j\overline{a_k},$$

where

$$A_{jk} = \int_0^\tau e^{i(\omega_j-\omega_k)s}\,\mathrm{d}s = \begin{cases} \frac{1}{i(\omega_j-\omega_k)}\left(e^{i(\omega_j-\omega_k)\tau}-1\right) & j\neq k \\ \tau & j=k. \end{cases}$$

Since $\varepsilon \ll \omega$, we replace A_{jk} by its limit τ, i.e. $A_{jk}=\tau$ for all j,k. Then

$$\|f_0\|_{L^2(0,\tau)} = \tau\left|\sum_{j=1}^n a_j\right|^2.$$

Next we determine the distribution of $|\sum_{j=1}^n a_j|^2$. Note that the a_j's are assumed to be independent and centered complex Gaussian variables with variance matrices $\sigma_j^2 I_2$, and therefore we have:

$$\sum_{j=1}^n a_j \sim \mathcal{CN}(0,\sigma^2 I_2),$$

where $\sigma^2 = \Sigma_{j=1}^n \sigma_j^2$. It follows that:

$$\left|\sum_{j=1}^n a_j\right|^2 \sim \sigma^2(Z_1^2+Z_2^2) \sim 2\sigma^2 Z,$$

where Z_1, Z_2 are independent, standard normal random variables and Z follows a standard exponential distribution (the sum of the squares of two independent standard normal random variables is exponentially distributed with mean 2). In other words, the squared norm $\|f_0\|^2_{L^2(t,t+\tau)}$ follows an exponential distribution with parameter $2\tau\sigma^2$.

Therefore, using the formula for the exponential cumulative distribution function we obtain:

$$\begin{aligned}\mathbb{P}\Big(\|f\|^2_{L^2(0,\tau)} < \theta^2\mathbb{E}\|f\|^2_{L^2(0,\tau)}\Big) &= \mathbb{P}\Big(\|f_0\|^2_{L^2(0,\tau)} < \theta^2\mathbb{E}\|f_0\|^2_{L^2(0,\tau)}\Big) \\ &\approx \mathbb{P}\left(\tau\Big\|\sum_{j=1}^n a_j\Big\|_2 < 2\theta^2\tau\sigma^2\right) = 1-e^{-\theta^2}.\end{aligned}$$

□

If we prefer to have an estimate of the norm $\|f\|_{L^\infty(t,t+\tau)} = \|f_0\|_{L^\infty(t,t+\tau)}$, then we can use the Gagliardo-Nirenberg interpolation inequality (Nirenberg, 2011) to derive this from the previous estimate. This inequality gives an estimate of the supremum norm in terms of the L^2-norms of f_0 and f_0'. Although there are various versions in the literature, we prove our own because it gives us control over the constants:

Lemma 4.3. *For any $f \in C^1([0,\tau];\mathbb{C})$,*

$$\frac{1}{\tau}\|f\|^2_{L^2(0,\tau)} \le \|f\|^2_{L^\infty(0,\tau)} \le \frac{2}{\tau}\|f\|^2_{L^2(0,\tau)} + \tau\|f'\|^2_{L^2(0,\tau)}. \tag{37}$$

Proof. The first inequality is immediate. For the second, we write for any $s,t \in [0,\tau]$

$$|f(t)|^2 = |f(s)|^2 + 2\,\mathrm{Re}\int_s^t f(\sigma)f'(\sigma)\,d\sigma \le |f(s)|^2 + \frac{1}{\tau}\|f\|^2_{L^2(0,\tau)} + \tau\|f'\|^2_{L^2(0,\tau)}.$$

Integrating left and right over $s \in [0,\tau]$, and taking the supremum over $t \in [0,\tau]$, we find

$$\tau\|f\|^2_{L^\infty(0,\tau)} \le \int_0^\tau |f(s)|^2\,ds + \|f\|^2_{L^2(0,\tau)} + \tau^2\|f'\|^2_{L^2(0,\tau)} = 2\|f\|^2_{L^2(0,\tau)} + \tau^2\|f'\|^2_{L^2(0,\tau)}.$$

This proves the result. □

From this inequality we deduce the following theorem.

Theorem 4.4 (Estimates of the distribution of the infinity-norm)**.** *Assume the same conditions as Theorem 4.2. Setting $\tilde\theta^2 := \theta^2\tau^{-1}\mathbb{E}(\|f\|^2_{L^2(t,t+\tau)})$, we have*

$$\mathbb{P}\Big(\|f_0\|^2_{L^\infty(t,t+\tau)} \le \tilde\theta^2\Big) \lesssim 1 - e^{-\theta^2}$$
$$\mathbb{P}\Big(\|f_0\|^2_{L^\infty(t,t+\tau)} \le \tilde\theta^2\Big) \gtrsim 1 - e^{-\theta^2/2}.$$

Note the scaling of $\tilde\theta$: since $\|\cdot\|_2^2$ scales as τ, and $\|\cdot\|_\infty$ scales as 1, we rescale the L^2-norm by τ in the definition of $\tilde\theta$ in order to make $\tilde\theta$ τ-invariant.

Proof. Above we already calculated that

$$\|f_0\|_2^2 = \sum_{j,k=1,\dots,n} A_{jk}\,a_j\overline{a_k}.$$

Similarly, we see that

$$\|f_0'\|_2^2 = \sum_{j,k=1,\dots,n} a_j\overline{a_k}i(\omega_j-\omega)\overline{i(\omega_k-\omega)}\int_0^\tau e^{i(\omega_j-\omega_k)s}\,\mathrm{d}s = \sum_{j,k=1,\dots,n} \widetilde{A}_{jk}\,a_j\overline{a_k},$$

where

$$\widetilde{A}_{jk} := (\omega_j-\omega)(\omega_k-\omega)A_{jk}.$$

Therefore, using Lemma 4.3,

$$\sum_{j,k=1,\dots,n} A_{jk}\,a_j\overline{a_k} \le \tau\|f_0\|^2_{L^\infty(0,\tau)} \le \sum_{j,k=1,\dots,n} B_{jk}\,a_j\overline{a_k},$$

where $B = 2A + \tau^2 \widetilde{A}$, i.e. $B_{jk} = (2 + \tau^2(\omega_j - \omega)(\omega_k - \omega))A_{jk}$.

As before we use the narrow-bandedness assumption that $\varepsilon \ll 2\pi/\tau$, which implies that $A_{jk} \approx \tau$ and $B_{jk} \approx 2A_{jk} \approx 2\tau$; then the inequalities above reduce to

$$\left|\sum_{j=1}^{n} a_j\right|^2 \leq \|f_0\|^2_{L^\infty(0,\tau)} \leq 2\left|\sum_{j=1}^{n} a_j\right|^2.$$

In the proof of Theorem 4.2 we already observed that $|\sum_{j=1}^{n} a_j|^2$ is exponentially distributed with parameter $2\sigma^2$; therefore

$$\mathbb{P}\Big(\|f_0\|^2_{L^\infty(t,t+\tau)} \leq \tilde{\theta}^2\Big) \leq \mathbb{P}\left(\left|\sum_{j=1}^{n} a_j\right|^2 \leq \tilde{\theta}^2\right) \approx 1 - e^{-\tilde{\theta}^2/2\sigma^2},$$

and

$$\mathbb{P}\Big(\|f_0\|^2_{L^\infty(t,t+\tau)} \leq \tilde{\theta}^2\Big) \geq \mathbb{P}\left(2\left|\sum_{j=1}^{n} a_j\right|^2 \leq \tilde{\theta}^2\right) \approx 1 - e^{-\tilde{\theta}^2/4\sigma^2}.$$

The assertion of the theorem follows from remarking that $\tilde{\theta}^2 = \theta^2 \tau^{-1} \mathbb{E}\|f_0\|^2_{L^2(0,\tau)} = \theta^2\, 2\sigma^2$. □

4.10 Discussion

The various results mentioned above all give partial characterizations of the probability of the appearance of quiescent periods in a narrow-banded signal.

For the "deterministic" case, the small-n results illustrate how quiescent periods may or may not recur in deterministically chosen sums of harmonics, and show how a precise characterization quickly becomes complex as the number n of harmonics increases. For the generic "irrational" case, however, we were able to derive a general necessary and sufficient condition for the existence of QPs.

For the "stochastic" case with arbitrary n, by choosing random coefficients, with uniformly distributed phases and normal amplitudes, we can leverage the property that the signal is a Gaussian process to characterize rates of upcrossings; possibly the Slepian-model can lead to a more precise characterization of the distribution of quiescent periods.

We also derived some estimates of our own for the probability distribution of quiescent periods defined by the L^2 and the L^∞ norm, under the assumption of strong narrow-bandedness. Although each of these various results covers only part of the picture, together they do give some insight into the occurrence of quiescent periods in sums of harmonics.

5 Deterministic and stochastic models for prediction of QPs

As a second main step, MARIN would like to help the HLO to predict quiescent periods with high confidence. We pursued several approaches to this problem, by both deterministic and stochastic models, with different levels of success. The underlying hope is that the signal in a finite time interval contains enough information to allow for forecasts in the very near future. This means that certain patterns are repeating in the ship motion.

5.1 Fourier continuation of the signal

The ship motion is assumed to be a second-order stationary stochastic process X_t that can be described by a continuous spectrum $S(f)$. In fact, as the sea surface elevation can be considered a Gaussian process, and the ship dynamics can be assumed to be linear, the resulting ship motion response is also a Gaussian process. In simulations, e.g. the ones performed by MARIN, realizations of this process are generated in the form of time series that share the same second-order statistical properties. The most common method is superposition of a large number of frequency components with randomized phases

$$f(t) = \sum_k \sqrt{2S(\omega_k)\Delta\omega_k}\cos(\omega_k t + \delta_k), \tag{38}$$

where δ_k are drawn from the uniform distribution on the interval $[0, 2\pi]$. This method can be readily extended to the multivariate setting Shinozuka and Jan (1972). Mathematically, there is thus a difference between the simulated signals and ship motions that are measured in reality.

Nevertheless, in both cases the underlying structure of the signals suggests that Fourier analysis might be a useful tool to understand – and possibly predict – the signals in question. Naively, one would suspect that if one estimated the Fourier decomposition of the signal, i.e. the frequencies, amplitudes and phase angles, one could simply continue the signal and predict its future evolution. For example, in Eq. 38 the randomness appears only in the phases. Each realization of this process, however, is a deterministic function. Of course, for real-world data the situation is more unclear, but let us focus on the simpler case of simulated data for now.

The main difficulty in practice is that such an analysis is based only on a finite time series $(x_0, x_1, \ldots, x_{n-1})$, whereas the underlying signal is defined on all of $\mathbb{R}$. The discrete Fourier transform can be used to estimate the frequency components of the signal, but it is essentially a Fourier series. Since Fourier series of a non-periodic function are really the Fourier series of the periodic extension of the function, this assumes that the past history of the ship motion $(x_0, \ldots, x_{n-1})$ is repeated periodically. In other words, prediction based on Fourier continuation of the signal consists of trivially repeating the signal from the start of the analysis period. This is illustrated in

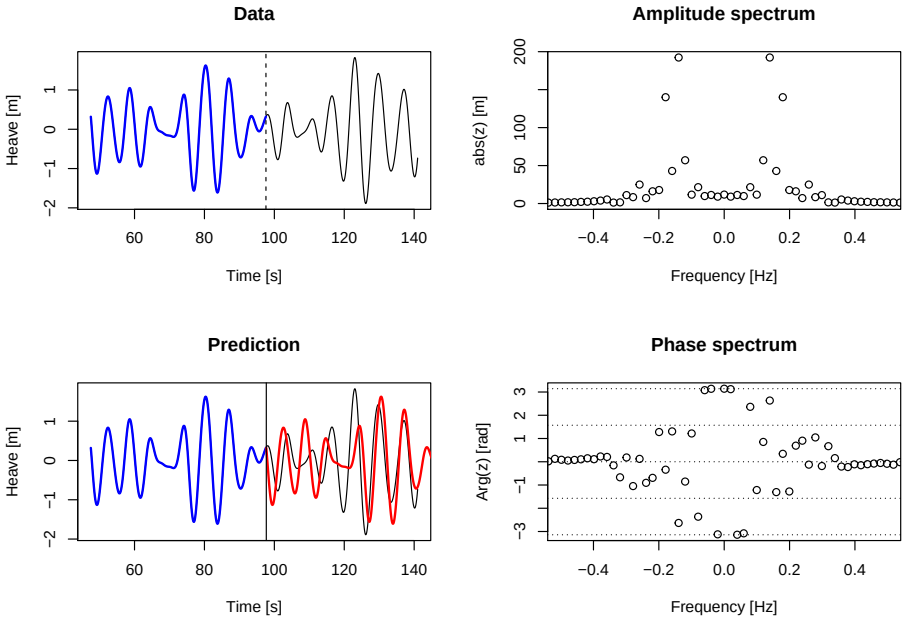


Figure 14: Discrete Fourier transform of ship heave signal. Prediction is based on periodic continuation of the signal (see text).

Figure 14. The continuation therefore depends on the length of the past history that is used. It is not clear how this can lead to a usable predictor of future ship motions. One might average over different lengths of the past of the signal, but the resulting variance in the prediction is too large to be useful.

5.2 Prediction in stationary processes

Prediction in stationary processes has been studied already by Kolmogorov. A very accessible introduction is given by Fristedt et al. (2007). An extensive treatment was given in Yaglom (1962), and the following is simply an application of his approach. Let us consider here the *extrapolation problem* for a stationary random sequence $(x_i)_{t\in\mathbb{Z}}$, with the *mean square extrapolation error* as error criterion. This is the problem of minimizing

$$\sigma^2_{m,n} = \mathbb{E}\left[|x_{t+m} - g(x_{-1}, x_{-2}, \dots x_{-n})|^2\right] \tag{39}$$

over all extrapolation functions g. We restrict ourselves here to the class of linear extrapolation functions

$$g(x_{-1}, x_{-2}, \dots, x_{-n}) = \alpha_1 x_{-1} + \alpha_2 x_{-2} + \cdots + \alpha_n x_{-n}.$$

If the sequence x_i is a Gaussian process (which can be assumed here) this is no restriction: it can be shown that in this case the best linear extrapolation formula coincides with the best possible extrapolation formula (Yaglom, 1962, ch.20).

Let us assume that we know the the correlation function

$$C(j,i) = \mathbb{E}[x_j x_i]$$

of the sequence x_i. Because of stationarity, this does not depend on time i, but only on the *lag* $k = j - i$, such that

$$C(k) = \mathbb{E}[x_{i+k} x_i]$$

for any $i \in \mathbb{Z}$.

The normal equations corresponding to the minimization problem in Eq. 39 are

$$\left.\frac{\partial \sigma^2_{m,n}}{\partial \alpha_k}\right|_{\alpha_1=a_1,\ldots,\alpha_n=a_n} = -C(m+k) + \sum_{i=1}^{n} a_i C(k-i) = 0 \quad (k = 1, 2, \ldots, n). \quad (40)$$

This is simply a linear system of n equations in n unknowns, which under the conditions assumed here always has a unique solution. The best linear extrapolation formula is then

$$\hat{x}_{t+m} = a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_n x_{t-n}$$

and the corresponding mean square error is given by

$$\sigma^2_{m,n} = C(0) - \sum_{k=1}^{n} \alpha_k C(m+k).$$

Yaglom remarks that this approach is impractical since the solution of Eq. 40 is tedious for $n > 10$ and continues to develop a spectral theory of the solution, applicable whenever the correlation function or spectral density is a known rational function, as well as the theory for the case of continuous time. However, this was written at a time when mainframe computers had only 10 KB of memory. For our purposes, the above approach seems the most direct and useful.

Testing this approach with the time series consisting of the extrema of the heave signal, we start by looking at the autocorrelation function in Figure 15. The alternating nature of the extrema process hides the relevant information, and it becomes more natural to consider the absolute extrema. It can be seen that after about 4 values the absolute extrema are not correlated anymore, within the estimated statistical uncertainty. Note that we removed the mean of the signals before the analysis, so subsequent results are for zero-mean processes.

Setting up the linear prediction for the absolute extrema process is straightforward. Figure 16 shows the one-step ahead prediction (top) and the three-step ahead prediction (bottom). As expected for this linear method, the n-step ahead prediction approaches the mean value (of zero) for increasing n, and the prediction becomes less reliable.

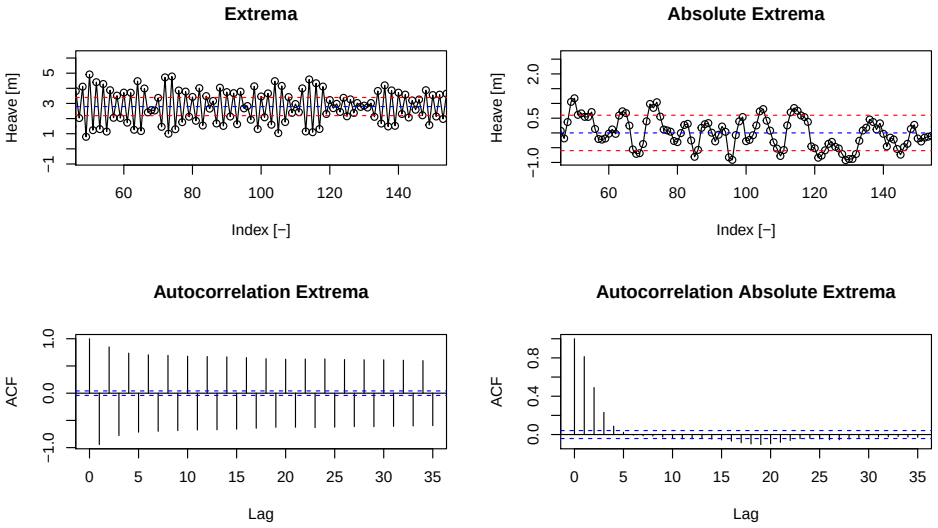


Figure 15: Autocorrelation of extrema process and absolute extrema process.

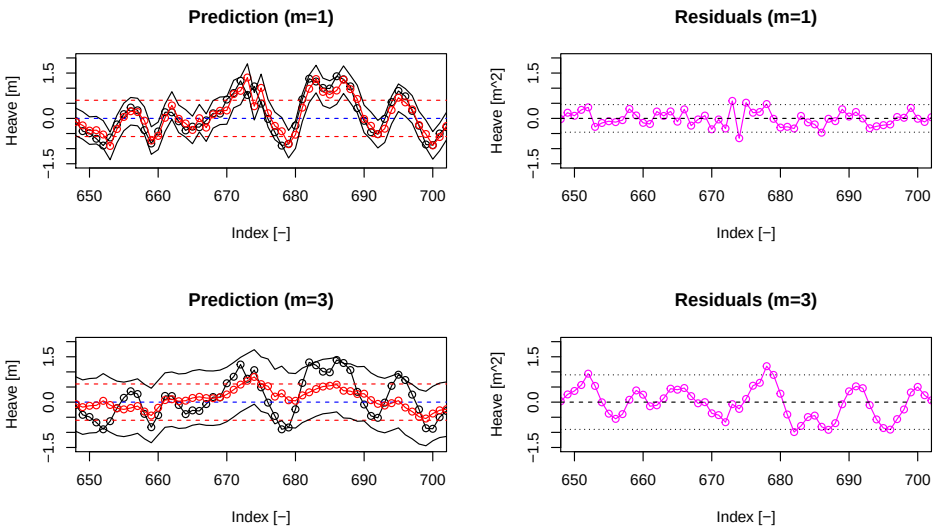


Figure 16: Linear prediction of absolute extrema process. An example for two different values of the step ahead m are shown. In both cases a long history ($n = 400$) was used. Root mean square error estimates are shown in addition.

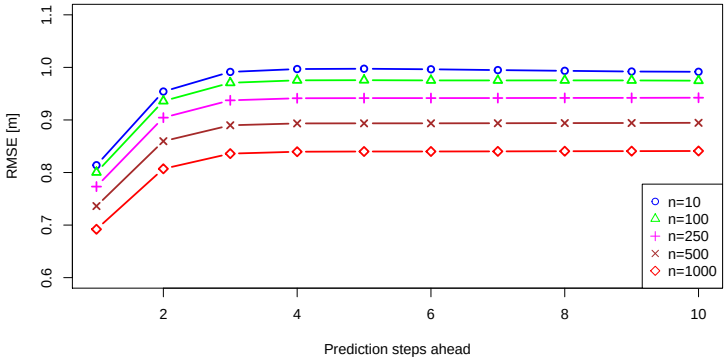


Figure 17: Accuracy of linear prediction for different lengths n of past history. Root mean square error (RMSE) against number of prediction steps.

Figure 17 gives an indication of the accuracy of prediction that can be achieved with this method, in terms of root mean square error between prediction and known signal. It can again be seen that prediction beyond 4 steps ahead (amounting to about two waves) is not really possible.

5.3 Statistical modeling of ship movements

In this section we consider the approach of fitting a stochastic model to the data on ship movements that can be used to predict or test for the occurrence of a quiescent period. In practice, the HLO appears to base his decision as to when to call the helicopter in for a landing attempt on the ship movements that occurred during the recent past. This suggests that it ought to be possible to predict the occurrence of a quiescent period based on past observations. In this section we concentrate on linear models, and outline some preliminary ideas on which models may be useful. We focus on the modeling of the wave envelope by considering observations of the ship motion of the recent past.

In Section 5.3.1 we consider *autoregressive moving average* (ARMA) models, which are commonly used to model economic time series but have widespread applications in other areas (Brockwell and Davis, 2009). We also provide a preliminary example in which we fit an ARMA model to the sequence of extrema of the heave data set provided by MARIN. In Section 5.3.3 we explain how one can use sequential hypothesis testing as an aid to decide whether or not a quiescent period has commenced, given a fitted ARMA model. In Section 5.3.2 we propose a variant of a *logistic regression* model as a possible improvement to the ARMA model for the problem at hand. We provide a brief summary in Section 5.3.4.

5.3.1 Autoregressive Moving-Average model

The basic ARMA model is defined as follows. We assume time is slotted into time epochs of equal length that we index by $t \in \mathbb{N}$. Let $(Z_t) \in \mathbb{R}^d$ denote a sequence of Gaussian independent and identically distributed (i.i.d.) random variables with zero mean and variance σ^2. Such a sequence is often referred to as *white noise process*. Suppose the data sequence of interest is a realization of a stochastic process $(X_t) \in \mathbb{R}^d$. Then the process is referred to as ARMA(p,q) process if it satisfies the recursion

$$X_t = c + \sum_{i=1}^{p} A_i X_{t-i} + \sum_{j=0}^{q} B_j Z_{t-j}, \tag{41}$$

where $c \in \mathbb{R}$, and A_i and B_j are coefficient matrices of suitable dimensions. For background on ARMA modeling see, for example, Brockwell and Davis (2009).

It is a virtue of the ARMA model that forecasting based on this model is particularly easy. Given the observations up to time t, we can predict the next vector of data points by

$$\hat{X}_t = c + \sum_{i=1}^{p} A_i x_{t-i} + \sum_{j=1}^{q} B_j z_{t-j},$$

where we have replaced Z_t by its expected value zero. Thus, the model can be used to predict the magnitude of the ship movements in the near future.

We now provide a small example where we fitted a univariate ARMA model to the series of heave data. We focussed on this data series because the magnitude of consecutive heave movements seems to be particularly important for the decision of the HLO to initiate a landing attempt. We expect, however, that the predictive capability of the model can be improved by including other relevant time series.

First, we recall that the relevant information for predicting a quiescent period is included in the envelope. We therefore extract the sequence of local extrema of the heave data series. We then take the absolute value of the extrema and center the resulting time series by subtracting the mean value: indeed, the amplitude is what affects the helicopter landing.

To estimate the model parameters, we used the package "forecast" in the statistical computing language R. We fitted the model to a training set of 200 data points, resulting in an ARMA(2,0). With this model specification and the estimated coefficients, we ran diagnostic tests on the residuals to verify that the latter are Gaussian white noise. We then used the model to predict the subsequent 10 data points, see Figure 18.

We remark that the accuracy of the prediction did not improve with a larger training set; seemingly, the series can be modeled as ARMA only locally. Further testing with multivariate ARMA is needed to optimize the data to be included in the model: we included only the extrema of the heave data series, but other data such as roll and pitch motion may be significant as well. It is also possible to attempt to model the amplitude of the wave heaves rather than the absolute value of each extreme point

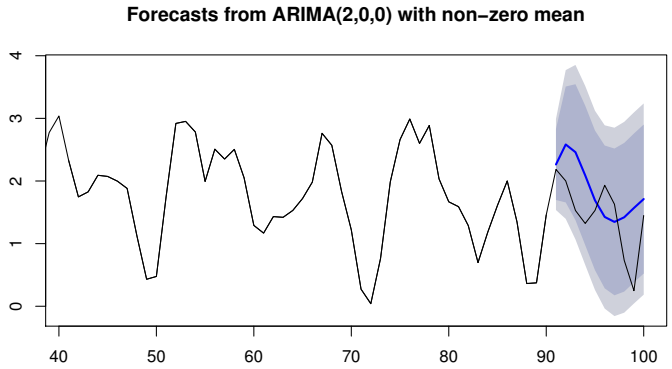


Figure 18: Example of a forecast for heave extrema based on an ARMA(2,0). The shaded area indicates the 90% (dark grey) and 95% (light grey) confidence intervals.

as we did in this preliminary experiment. It may also be that predicting the actual value of the time series based on simple linear models is not possible with sufficient accuracy. In the next section therefore we suggest a logistic regression model that can be used to decide whether or not a quiescent period has commenced or is about to commence.

5.3.2 Logistic regression

In the preliminary experiment we presented in Section 5.3.1, the forecasts we obtained with the ARMA model corresponded to rather large confidence intervals. We suggested a number of steps that may help to remedy this issue. Note, however, that our objective is to decide whether or not to expect a quiescent period; predicting the actual value of the time series is not strictly necessary for this purpose. As an alternative to the ARMA model we therefore propose the following *logistic regression* model.

Let Y denote a discrete random variable taking values in $\{0, 1\}$, where the realization 1 indicates that the current time period is quiescent. Let π denote the probability that $Y = 1$. We now seek to explain the realization of π by current and past observations. For example, consider

$$\log\left(\frac{\pi}{1-\pi}\right) = \sum_{j=1}^{p}\sum_{i=0}^{k}\beta_{i,j}X_{i,t-j}, \tag{42}$$

where $\beta_{i,j}$ denote the coefficients. Here, $X_{i,t-j}$ denotes the random variable corresponding to an observation obtained at time $t-j$ of a particular type of ship movement labelled by i.

A simple logistic regression model can be estimated using maximum likelihood techniques that are readily available in any standard statistical programming language; see (Kabacoff, 2015, Section 13.2) for an example using R. Note, however, that the sequence $(X_{i,t-j})_j$ is not independent; therefore care has to be taken that the correlation between included variables is not too strong. If variables are nearly perfectly correlated, the matrix of coefficients is nearly singular, which can lead to problems with standard estimation procedures (this is known as *multicollinearity*).

In order to estimate a model for explaining Y, we need to label each data point of the training set by 1 or 0 depending on whether or not it lies within a quiescent period. We remark that for a period to qualify as quiescent, it must be of sufficient length, say $T = 5$. Thus, if we collect measurements every Δ time units, where $\Delta < T$, then, we must observe nearly perfect positive correlation between values of Y_t: If $Y_t = 1$ then we must have that neighbouring points also have realization 1. Furthermore, Y_t is not necessarily measurable at time t: We only know whether or not we should label Y_t as quiescent after we observed a period of length T, during which the waves were quiet. Suppose, for example, that $\Delta = 1/T$ and consider the first observation we collect (at time Δ, that is), which we denote by Y_Δ. Then we need to observe $T - 1$ more data points before we can determine whether Y_Δ is part of a quiescent period. Therefore, we cannot use Y_Δ to predict the value of $Y_{2\Delta}$, say. This explains why we did not include past observations Y_t as explanatory variables on the right-hand side of Eq. 42.

A possible alternative is to group data into sliding windows such that for each new observation arriving the oldest observation is discarded. If the size of the windows coincides with the minimum length of a quiescent period, then windows are not perfectly correlated, and we can determine whether or not the previous window was quiescent, namely, if all observations in the previous window corresponded to a quiescent period. This alternative framework leads to a model of the form

$$\log\left(\frac{\pi}{1-\pi}\right) = \sum_{j=1}^{p}\sum_{i=0}^{k}\beta_{i,j}X_{i,t-j} + \sum_{k=1}^{q}\gamma_k Y_{t-k}^{w}, \tag{43}$$

where Y_t^w denotes the random variable describing whether or not the collection of data points belonging to the window that ends at time t.

To gain more certainty as to whether or not a quiescent period has commenced, the predicted future values of the relevant time series may be supplemented by the outcome of a statistical hypothesis test. We briefly discuss such a procedure in the next section for the ARMA model example.

5.3.3 Change point detection

In this section we explain how change point detection procedures can be applied to test a stationary ARMA time series for a change in the mean value. Specifically, we focus on the popular CUSUM method that was originally suggested by Page (1954). Similar procedures have been considered in Basseville and Nikiforov (1993); Chen and Gupta (2012); Robbins et al. (2011).

First, note from Eq. 41 that setting the initial white noise terms equal to zero, the sequence of residuals can be extracted from the sequence of observations as

$$\hat{Z}_t = x_t - c - \sum_{i=1}^{p} A_i X_{t-i} - \sum_{j=1}^{q} \hat{Z}_{t-j}.$$

A shift in the mean value of size μ in the sequence of observations therefore results in a shift in the mean value of the *innovations* $\hat{Z}_t$. Apart from the jump in the mean value, this sequence is i.i.d. Gaussian, so that we can focus on the easier problem of testing a sequence of independent Gaussian random variables. For further details on this and a comparison to the approach of testing the sequence of observations directly, see Basseville and Nikiforov (1993); Kuhn et al. (2014); Robbins et al. (2011).

The CUSUM method is essentially a sequential application of a log-likelihood ratio test. Consider testing the data in sliding windows of fixed size n. We wish to test whether at any time within the window the mean value of the sequence (Z_t) has changed from θ_0 to θ_1, say. Denote the hypothesis that such a change in mean has occurred at time k by $H_1(k)$. Thus, under $H_1(k)$ we have $\mathbb{E}[Z_t] = \theta_0$ for $t < k$ and $\mathbb{E}[Z_t] = \theta_1$ otherwise. Instead, under the null hypothesis H_0 we have $\mathbb{E}[Z_t] = \theta_0$ for all $t \in \{1, \ldots, n\}$.

Denoting by p_θ a normal density with mean θ, the log-likelihood ratio test statistic for testing the first window is

$$S_k := \sum_{t=k}^{n} Y_t := \sum_{t=k}^{n} \log \left(\frac{p_{\theta_1}(\hat{Z}_t)}{p_{\theta_0}(\hat{Z}_t)} \right)$$

(note that $Y_t = 0$ for $t < k$ since for such t the distribution of Z_t is equal under H_0 and $H_1(k)$). Obviously, the ratio of likelihoods $p_{\theta_1}(\hat{Z}_t)/p_{\theta_1}(\hat{Z}_t)$ is large if $p_{\theta_1}(\hat{Z}_t) > p_{\theta_1}(\hat{Z}_t)$, that is, if it is more likely to observe $\hat{Z}_t$ assuming that $H_1(k)$ is true. We would thus decide in favor of $H_1(k)$ if the test statistic S_k is large in some sense.

In order to decide whether a change point has occurred at some point k within the current window, we therefore need to check whether there is a $k \in \{1, \ldots, n\}$ such that S_k exceeds a certain critical value, b, say. As a result, the statistic for the *composite* test (that is, H_0 versus $\bigcup_{i=1}^{k} H_1(k)$) is

$$t_m := \max_{k \in \{m-n+1, \ldots, m\}} S_k(m), \tag{44}$$

where m is the label of the current window, and $S_k(m)$ denotes the test statistic corresponding to the innovations in the m-th window. Then, for a given threshold $b > 0$, the CUSUM method raises an alarm (indicating that a change has occurred) at time t_a, with

$$t_a := \inf \left[m : t_m \geq b \right]. \tag{45}$$

The name of the test is explained by noting that the test statistic t_m can be rewritten in terms of the cumulative sums $T_k := \sum_{t=1}^{k} \log p_{\theta_1}(\hat{Z}_t)/p_{\theta_0}(\hat{Z}_t)$ as follows,

$$t_m = T_m - \min_{k \in \{m-n+2, \ldots, m\}} T_{k-1}.$$

This is convenient with respect to computational efficiency since T_m equals $T_{m-1} + Y_m$, and computing $\min_{k\in\{m-n+2,\dots,m\}} T_{k-1}$ only involves comparing the minimum computed at time $m-1$ with T_{m-1}. The choice of the threshold b can be based on simulation, or using approximations to the false alarm probability (see Kuhn et al. (2016)). For an example with multivariate data sequences see Kuhn et al. (2014).

5.3.4 Summary

The methods suggested in this section require more extensive testing. In particular, for the ARMA modelling approach other variables should be included besides the extrema of the heave movements. The logistic regression approach may be more suitable given that the objective is to discern between quiescent and non-quiescent periods, and should also be investigated based on numerical experiments. As suggested, one may use a change-point-detection procedure as a further indicator as to whether a quiescent period has commenced. Assuming that the HLO is risk averse, we would recommend that a quiescent period is then only announced if both the test and the predicted values indicate that such a period has started.

5.4 Short-term forecasting

In this section we will investigate a possibility of short-term forecasts of quiescent periods by solely analysing the ship motion data. That is to say we regard the motion data as a discrete-time stochastic process with memory. In this process, the states at time points $t_i,\ i \geq 0$ are correlated with the previous states at $t_{i-1}, t_{i-2}, \dots, t_{i-k},\ 0 < k < i$. Since the original motion data is not supplied in a form of discrete states but as samples of a continuous-time function, one needs to convert the sampled signal into a discrete time series first. All in all, three questions crystallise as central to this analysis:
1) How to define patterns in data?
2) What correlation between the patterns is observable?
3) How good are the forecasts that can be made on the basis of observed patterns?
Let $f(t) \in C^2[0,\infty)$ represent one component of the measured signal. Without loss of generality we assume the signal $f(t)$ has zero mean value, $\int_0^\infty f(t)\,\mathrm{d}t = 0$. Furthermore, for the sake of simplicity we restrict our attention to local extrema of $f(t)$, that are in view of the smoothness class isolated points,

$$F = \left\{ f(t) : \frac{\mathrm{d}}{\mathrm{d}\xi}|f(\xi)|_{\xi=t} = 0 \text{ and } \frac{\mathrm{d}^2}{\mathrm{d}^2\xi}|f(\xi)|_{\xi=t} < 0 \right\}.$$

Occurrence times t naturally induce a strict order on F which allows us to speak of a sequence $F_i,\ i = 1, 2, \dots$ In this way, each peak is characterised by a couple $(F_i, T_i) \in (0,\infty)^2$, and the whole signal by a sequence of peaks: $S = ((F_1, T_1), (F_2, T_2), \dots)$, where F_i denotes the peak height and $T_i = \frac{t_{i+1} - t_{i-1}}{2}$ the duration. Furthermore, a configuration for d consecutive peaks, that is a d-tuple $s = ((F_1, t_1), \dots, (F_d, t_d))$, is a

point in $\Omega = (0, \infty)^{2d}$. We will now consider the probability space $(\Omega, \mathcal{F}, \mu_F),\ \mathcal{F} = 2^{\Omega}$, containing the d-tuples as outcomes. For given $p \in \mathcal{F}$, the probability measure $\mu_F\, p$ tells us how often the elements of p occur in the signal,

$$\mu_F\, p := \lim_{n\to\infty} \frac{1}{n-d} \sum_{i=d}^{n} \mathbf{1}_p(S_{i-d:i}),$$

where $S_{i-d:i}$ denotes a fragment of the signal S, and $\mathbf{1}_p$ is the indicator function for event p. Some events from $\mathcal{F}$ can be represented as a union tensors products. Let,

$$P^d = \left\{ \bigcup_{i=1}^{m} p_i : p_i \in \bigotimes_{j=1}^{d} [a_j, b_j], 0 < a_j < b_j \right\} \subset \mathcal{F}.$$

We refer to events $p = p_0 \times p_1,\ p_0 \in P^{d_1}, p_1 \in P^{d_2},\ d_1 + d_2 = d$ as patterns. For each pattern p there is a signal F such that $\mu_F\, p > 0$, which is not generally the case for events that are not patterns. For a given pattern p, we will now quantify its suitability for forecasting. Suppose one finds a d_0-tuple representing p_0 in the data. Is the expectation that a d_1-tuple from p_1 will follow immediately after a good forecast? Formally, the answer to this question unfolds into three distinct statistical estimates:
a) probability to find p_0, is simply given by $P_0(p) = \mu_F\, p_0$;
b) probability that p_0 is followed by p_1, $P_1(p) = \frac{\mu_F\, p}{\mu_F\, p_0}$;
c) probability that p_1 is preceded by p_0, $P_2(p) = \frac{\mu_F\, p}{\mu_F\, p_1}$.
The estimate P_0 tells us how often we can perform the forecast based on this pattern. The estimate P_1 tells us how reliable this forecast will be, and the estimate P_2 tells us what fraction of all p_1 in the signal is predictable via the pattern. For example, it may happen that p_0 is always followed by p_1 which makes this combination of patterns a reliable prediction ($P_1 = 1$). If in the same time, p_1 is preceded by many other patterns, then $(p_0 \times p_1)$ is reliable but not very efficient combination ($P_2 \approx 0$). Finally, if besides the above-stated, p_0 alone is not frequently observed then the prediction is reliable but practically useless, as one has to wait long, before the opportunity to assert a forecast comes ($P_0 \approx 0$). And so the problem of good forecasting given a sample of the signal shapes as a search for such $p \in \mathcal{F}$ that scores high on all three estimates P_0, P_1, P_2. Below, we will consider a few semi-heuristic choices on how such a search can be performed.

5.4.1 Markov model

Let $W_F = \{[b_{i-1}, b_i],\ i = 1, \ldots, n : b_i > b_{i-1},\ b_i \in (0, \infty)\}$, $W_T = (0, \infty)$ and $d = 2$. We search for patterns from $p \in (W_F \times W_T)^d \subset \mathcal{F}$. We are discretising the peak height into n bins and ignore the duration of the peaks completely.

This way, the prediction scheme with $d = 2$ becomes identical to a Markov chain. To do this, we classify the wave heights in a number of bins and then count how often transitions between bins occur. We can also include a finite history, by classifying wave heights of two successive extrema and counting transitions between pairs of

n	1	2	3	4	5
Bin	0-0.198	0.198-0.323	0.323-0.448	0.448-0.607	0.607-1.47
Number of extrema	1041	1042	1043	1042	1043

Table 2: Numbers of peaks in each bin from the chosen system of 5 bins

extrema or by counting for how many successive extrema the waves are above a certain threshold before a quiescent period is entered. An optimised combination of bin widths, number of bins and history depth will be needed for the best possible prediction, but a full exploration of all these algorithmic choices is beyond our scope here.

We consider the wave heights for a run of 5 hours. In these 5 hours there are 5212 extrema in the data, with the largest deviation from the mean equal to 1.47 meters. We choose to use 5 bins, with the limits on the bins such that each of the 5 intervals specified by the bins has equal numbers of extrema. This is summarized in Table 2. The slight variation in numbers of extrema is due to rounding on the bin widths. We now simply count the transitions between bins and use this to construct a matrix $\hat{M}_2$ that, at index (n, m), counts how often a wave of height n evolves into height m:

$$\hat{M}_2 = \begin{pmatrix} 620 & 294 & 104 & 24 & 0 \\ 291 & 353 & 252 & 122 & 24 \\ 99 & 266 & 344 & 261 & 73 \\ 26 & 105 & 273 & 399 & 238 \\ 5 & 24 & 70 & 236 & 708 \end{pmatrix}. \tag{46}$$

It is clear that there is some structure in the wave pattern, namely that waves of a certain height are likely to be followed by waves of comparable height.

The question whether it is sufficient to only consider a history depth of one extremum may be raised. This assumption underlying the analysis leading to $\hat{M}_2$ may simply be tested using the data. To do this we first normalise the columns of $\hat{M}_2$ to 1, which makes it into a probability transition matrix M_2. The normalisation is chosen such that if we are in a state and multiply it from the left with M_2, we always go to some other state and the total probability of being in any state is conserved. We can then compute M_2^2, which models the process of taking two steps with our Markov model M_2, and compare it to the transition matrix that skips over one extremum, M_2^s. Then, if the assumption that only the current state matters for forecasting holds, we should have that $M_2^2 = M_2^s$. These two matrices are shown in below:

$$M_2^2 = \begin{pmatrix} 0.44 & 0.29 & 0.17 & 0.08 & 0.02 \\ 0.29 & 0.27 & 0.22 & 0.16 & 0.07 \\ 0.17 & 0.22 & 0.25 & 0.23 & 0.13 \\ 0.08 & 0.15 & 0.23 & 0.28 & 0.26 \\ 0.02 & 0.06 & 0.13 & 0.26 & 0.52 \end{pmatrix}, \quad M_2^s = \begin{pmatrix} 0.31 & 0.29 & 0.22 & 0.13 & 0.05 \\ 0.30 & 0.25 & 0.18 & 0.17 & 0.11 \\ 0.20 & 0.21 & 0.22 & 0.22 & 0.15 \\ 0.14 & 0.15 & 0.22 & 0.25 & 0.24 \\ 0.05 & 0.10 & 0.16 & 0.23 & 0.46 \end{pmatrix}. \tag{47}$$

It can be seen from Eq. 47 that M_2^2 and M_2^s are not identical. The question is then if this is just because we do not have enough data, or because our modelling choice of having the bins in Table 2 and considering a state space of only the current extremum is not good enough. To test this properly, we need a way of comparing these matrices while taking into account that due to statistical fluctuations we expect the estimation of transition probabilities of rare events to be worse than the estimation for common events. Furthermore, we would like to be able to compare matrices of different sizes, because changing the number of bins or history depth changes the size of the state space and hence the dimensions of the matrices. Let $n = 5212$ be the number of extrema, $\hat{M}_2^s$ the unnormalised version of M_2^s, $\times$ the element-wise product of matrices, and $||.||_F$ the Frobenius norm and define

$$e(\hat{M}^s, M^s, M, n) := ||\hat{M}^s \times (M^s - M) \times (M^s - M)||_F/n. \tag{48}$$

Then $e(\hat{M}_2^s, M_2^s, M_2, 5212) = 0.0021$. To interpret this number we shall compare it to the Markov model for the state space with the same bins, but with a history of two extrema. The corresponding 25×25 transition probability matrices are not shown here, but inspection of their entries shows that after a sharp decline in extremum height the likelihood of multiple low extrema is highest. The estimation quality is given by $e(\hat{M}_4^s, M_4^s, M_4, 5212) = 0.00046$. We conclude that the data are better described by taking a longer history depth and that multiple low extrema are most likely if a sharp decline in extremum height is found.

5.4.2 Counting waves

An obvious way to account for longer history is to simply increase the pattern length d in the previous approach. Such decision will quickly lead us to a big number of patterns each with a very low frequency of occurrence and hence poorly represented in finite samples of the signal. We will instead construct a heuristic system of patterns that covers a big part of the whole configurational space and is a formalization of the already observed strategy described by the HLO: counting peaks.

A pattern for a single peak with a height below a quiescent threshold, b_q, is given by

$$p_q = (0, b_q] \times [0, \infty).$$

If a peak belongs to this pattern, its height $F_i \in (0, b_q]$ and the duration is arbitrary $T_i \in (0, \infty)$. In a similar fashion we define a pattern with non-zero number of peaks having all the heights below b_q and the total duration exceeding t_q.

$$p_Q = \bigcup_{k=1}^{\infty} \bigcup_{\sum q_i \geq t_q} \bigotimes_{i=1}^{k} (0, b_q] \times [q_i, \infty).$$

If – on a signal fragment S – $\mu_S\, p_Q > 0$, then $\mu_S\, p_q > 0$. Consider now a sequence of $k+2$ peaks that consists of: a peak below the quiescence threshold b_q, k peaks above

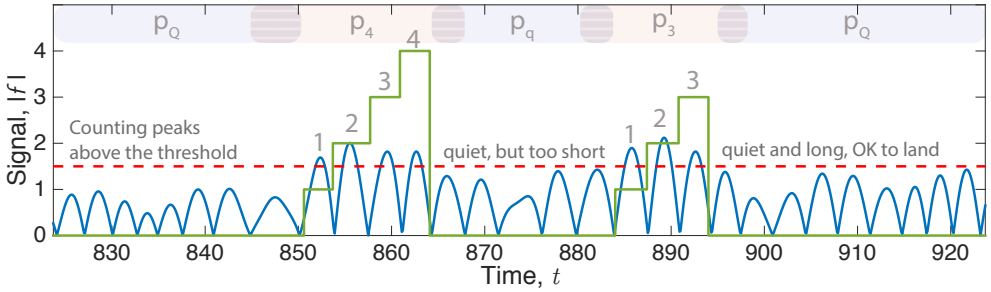


Figure 19: A sample of the signal from dataset D1 together with a matched system of patterns.

the threshold b_s, and again a peak below the threshold b_q. The corresponding pattern is given by

$$p_k = p_q \times ([b_s, \infty) \times (0, \infty))^k \times p_q.$$

Now the idea is to investigate the occurrence of patterns $p_k \times p_Q$ for $k = 1, \ldots$ This idea has a very simple practical interpretation.
Suppose one is counting all peaks above the threshold b_s. Every time a peak with amplitude below $b_q < b_s$ comes, one resets the counter to zero. We would like to know whether the count number at the resetting helps in predicting long quiescent periods.

An example of matching patterns from this system to the data is given in Figure 19. As before, we investigate the efficiency of the forecasting according to three measures: P_0, P_1, P_2. Figure 20 presents results for dataset D1 (see Table 1). The figure rates patterns p_k according to measure P_1 (top panel) and T/P_0 (bottom panel), where T is the average distance between peaks. There are a few empirical observations to make here. Firstly, not all patterns are equally good in the prediction. Secondly, the longer a pattern is, the less frequently it is represented in the signal. Thirdly, we see an artefact caused by the finite size of the signal sample: pattern p_9 predicts the quiescence period with probability one precisely because it occurred only once in the sample. On another hand, p_6 leads to very certain predictions, yet its average waiting time, approximately 30 min, is longer than practical limitations. In principle, one can combine p_6 with a pattern that occurs more frequently but has a lower prediction rate, say p_1, to compromise on predictability and reduce the waiting time. Additionally, the partition into patterns is based on parameters b_q, b_s, t_q. While b_q and t_q define the quiescent period and cannot be adjusted, b_s is a free parameter that may influence the quality of the predictions. This motivates the following optimization procedure.

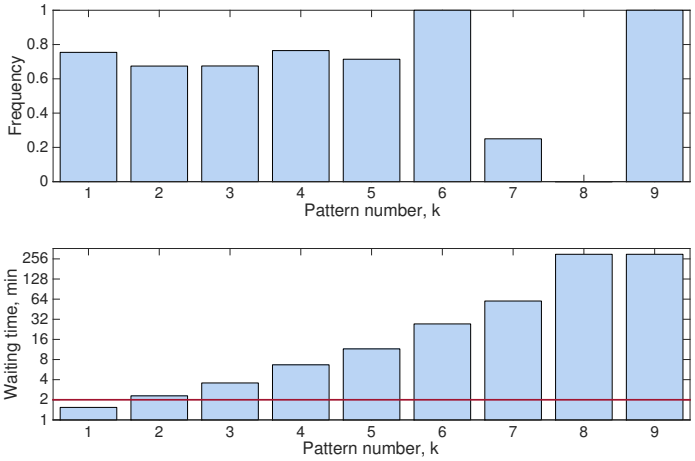


Figure 20: Efficiency of forecasts due to patterns p_k, as measured according to forecast certainty P_1 and waiting the time $1/P_0$. The values for method parameters are: $b_s = 1.72$.

Let $\omega_k = \frac{P_0(p_k)}{\sum\limits_{k=1}^{\infty} P_0(p_k)}$ are relative frequencies for pattern k, then the cost function

$$c(k_1, k_2, \dots) = \frac{\sum\limits_{i=1}^{\infty} \omega_{k_i} P_1(P_{k_i})}{\sum\limits_{i=1}^{\infty} \omega_{k_i}}$$

gives the average prediction rate for a union of patterns p_{k_i}, where k_i form a subset in $\mathbb{N}$. The task is to choose such a subset of indexes that the union of the corresponding patterns has best expected prediction rate. These requirements are crystallized as the following optimization problem,

$$\begin{aligned} c(k_1, k_2, \dots) &\to \min, \\ \{k_1, k_2 \dots\} &\subset \mathbb{N}, \\ \mathrm{w}_t\left(\bigotimes_i p_{k_i} \right) &\le w_{\max} \\ b_s &\in [b_q, \infty), \end{aligned}$$

where $\mathrm{w}_t(p)$ denotes the waiting time for a pattern p, and $w_{\max}$ is the upper constrain on the waiting time, in this report $w_{\max} = 2$ min unless stated otherwise.

Figure 21 features the prediction rates and waiting times for patterns after such an optimisation has been carried out for dataset D1 (see Table 1). The resulting optimal subset of indexes is $S_o = \{1, 2, 4, 5, 6, 7, 9, 10\}$ and the optimal value for

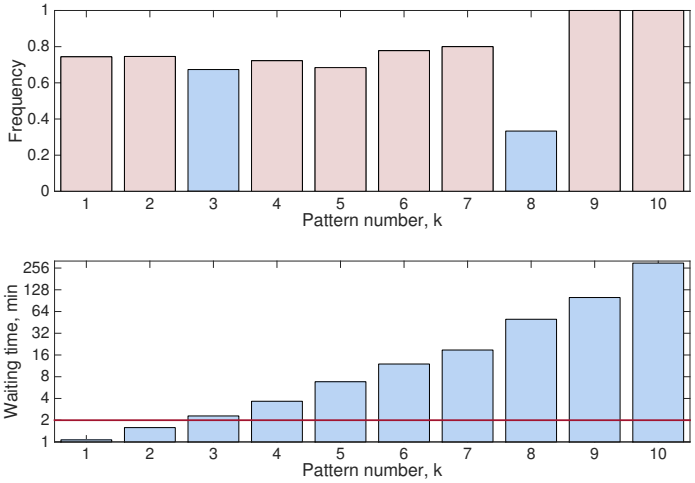


Figure 21: Efficiency of forecasts due to patterns p_k, as measured according to forecast certainty P_1 and waiting the time $1/P_0$. The values for method parameters are: $b_s = 1.545$.

$b_s = 1.5750$. The optimal set of parameters leads to the expected prediction rate 0.74; the occurrences of the combined prediction pattern $\bigotimes_{k \in S_o} p_k$ are separated by average waiting time of 1.85 min. In total, 78% of all quiescent periods are predictable via this combined pattern. This frequency of predictable events is limited by two factors: the choice for the pattern, which is in part heuristic and thus can be improved; the randomness of the signal that is a feature of data and cannot be manipulated.

All in all, we performed prediction tests/optimisation of the patterns on four datasets, shortly referred to as D1, D2, D3, D4, as shown in Table 1.

Table 3 provides the quality measures for all combination of optimisation/prediction. Data sets D1, D2 are two finite uncorrected samples produced for the same model parameters. One notices that the prediction quality changes little if we optimise on D1 and then predict on D2 or D3 (the first line of Table 3) as opposed to the scenario when we optimise and predict on the same dataset (the diagonal of Table 3). This may suggest that the optimised pattern grasps some universal property of the data. The situation changes when we analyse datasets with distinct simulation parameters, e.g. comparing dataset D1 to D4, that features larger wave height. In this case, when trained on D1, the prediction certainty on D4 is much smaller. When trained on D4 and then predicting on D1 the prediction certainty is relatively high but the waiting time is a magnitude larger. This scenario demonstrates that the optimised pattern does depend on the software parameters (that, in turn, mimic the sea state).

Optimisation on D4 results in no solution unless we increase the upper constraint on the waiting time. Such behaviour is connected to the fact that there are not many quiescent periods in this dataset.

Optimize \ Predict	D1	D2	D3	D4
D1	$\chi = 0.804$ $t = 1.33$ $f = 0.30$	$\chi = 0.68$ $t = 2.0$ $f = 0.79$	$\chi = 0.53$ $t = 3.0$ $f = 0.55$	$\chi = 0.12$ $t = 5.0$ $f = 0.75$
D2	$\chi = 0.73$ $t = 1.73$ $f = 0.83$	$\chi = 0.71$ $t = 1.87$ $f = 0.87$	$\chi = 0.58$ $t = 2.30$ $f = 0.72$	$\chi = 0.12$ $t = 7.5$ $f = 0.5$
D3	$\chi = 0.74$ $t = 2.0$ $f = 0.7$	$\chi = 0.71$ $t = 2.10$ $f = 0.78$	$\chi = 0.74$ $t = 1.87$ $f = 0.89$	$\chi = 0.09$ $t = 10$ $f = 0.37$
D4*	$\chi = 0.76$ $t = 17.60$ $f = 0.08$	$\chi = 0.76$ $t = 15.0$ $f = 0.1$	$\chi = 0$ $t = \text{n/a}$ $f = 0$	$\chi = 0.67$ $t = 7.5$ $f = 0.5$

Table 3: Prediction and pattern optimisation on various datasets. The prediction quality is measured by certainty χ, pattern waiting time t (min) and fraction of predictable events, f. *For optimisation on dataset D4 the upper constrain on average waiting time was relaxed to $w_{\max} = 8$ min.

5.5 Summary

Instead of processing the full data from the motion sensor, we narrowed our attention to the sequence of extrema values (the peaks). Patterns in such a sequence are defined as a subsequence of peaks with heights that fall within specific bounds. From this point of view a pattern is a manifold in the the peak configuration space. Given an observed sequence, the frequency of pattern occurrence can be computed as number of times such manifold was hit by samples from the data. Special interest present those patterns that combine non-quiescent period followed by a quiescent one.

Software-simulated data were analyzed for occurrence of patterns. Similar patterns were found in uncorrelated sample data that were produced with the same simulation parameters specifying the sea state. The patterns differ when different sea-state parameters are used. A somewhat naive choice for patterns as a tensor product allows one to assert predictions on quiescent period with 80% certainty and acceptable (from operation time point of view) frequency on some datasets. We expect that the certainty can be improved by a cleverer choice for pattern manifolds.

6 Conclusions

Given several simulations of ship motion, we tried to identify the distribution and initiation of quiescent periods (QPs) by various techniques with the common aim of pattern recognition. Moreover, within reasonable assumptions on the response of the ship to the forcing of the sea waves, we claimed that studying the more general problem of finding QPs in a sum of (deterministic or random) harmonics is relevant to make statements about the occurrence of QPs in ship motion.

The first thing we realized is that the essential information of the motion is contained in the extrema of the waves, and that this is encoded in the Hilbert transform of the signal. We then gave a statistical description of the distribution of QPs and a qualitative picture of the typical ship motion around a QP. While the former suggests modeling the occurrence of QPs by a Poisson process (even though this argument has still to be statistically tested), the latter information constitutes the first tool that we have for prediction of QPs.

Whenever ship motion is essentially coincident with the sea motion and its spectrum is narrow-banded, we gave analytical estimates of both probability and frequency of quiescent periods in a sum of deterministic and random harmonics. We reviewed the cases of one, two and three deterministic harmonics: the second one encodes the phenomenon of beating and is the prototype to have a first understanding and definition of a quiescent period; the third case already contains a lot of the features of the most general case, for which we derived a general criterion for the existence of QPs.

We then considered the case of arbitrarily many random harmonics. First, we applied existing methods to characterize rates of upcrossing of a fixed threshold. Next, we gave estimates for the distribution of QPs according to two different definitions of a QP and in terms of both the hight and the length of a QP.

The methods of fast prediction of quiescent periods are based on recognizing pattern in the time series via Fourier continuation and a few stochastic models for stationary processes. While the former, at this level of analysis, doesn't seem to be useful, the latter look promising. Indeed, we were able to identify several structural properties in the data.

The methods via the extrapolation problem perform well in the case of short-term prediction, but deteriorate when prediction is sought for longer futures. The autoregressive models are able to provide a reasonable forecast in some cases, but with rather scarce statistical confidence. A logistic regression was proposed, too, but it has still to be tested, together with a change-point-detection procedure. We remark that the simulations we have performed are limited to the data series of the heave coordinate. We feel that the inclusion of other variables may help the predictive power of such models.

The final approach described in this report is looking at the data from the standpoint of the theory of Markov processes. We were able to identify a few waves patterns, interpret the data as a random sequence of patterns, investigate the "memory content" of that stochastic process, and implement prediction. Some patterns gave rise to fairly good predictions, specifically when a series of particularly high waves are followed by a QP. We expect that this could be improved further by a better choice of the patterns themselves.

Acknowledgments

We are very thankful to MARIN, in particular in the persons of Ed van Daalen and Jos Koning for giving us such an interesting and stimulating challenge. We hope that this work will give MARIN sufficient motivation and inspiration to continue the project with new ideas and renovated enthusiasm.

References

M. Basseville and I. V. Nikiforov. *Detection of Abrupt Changes: Theory and Application.* Englewood Cliffs, Prentice Hall, N. J., 1993.

P. Bloomfield. *Fourier analysis of time series: An introduction.* Hoboken, John Wiley & Sons, 2000.

P. J. Brockwell and R. A. Davis. *Time series: theory and methods.* Springer Series in Statistics. Springer-Verlag New York, 2009.

J. Chen and A. Gupta. *Parametric Statistical Change Point Analysis: with Applications to Genetics, Medicine, and Finance.* Springer-Verlag Berlin, 2012.

H. Cramér and M. R. Leadbetter. *Stationary and related stochastic processes: Sample function properties and their applications.* Mineola, Dover Publications, 1967.

B. Fristedt, N. Jain, and N. Krylov. *Filtering and prediction: a primer*. Providence, American Mathematical Society, 2007.

G. H. Hardy and E. M. Wright. *An introduction to the theory of numbers*. Oxford University Press, fifth edition edition, 1979.

R. Kabacoff. *R in Action: Data Analysis and Graphics with R*. Manning Publications Co., Greenwich, CT, USA, 2015. ISBN 1617291382, 9781617291388.

J. Kuhn, W. Ellens, and M. Mandjes. Detecting changes in the scale of dependent Gaussian processes: A large deviations approach. In B. Sericola, M. Telek, and G. Horváth, editors, *Analytical and Stochastic Modeling Techniques and Applications*, Lecture Notes in Computer Science, pages 170–184. Springer International Publishing, 2014.

J. Kuhn, M. Mandjes, and T. Taimre. False alarm control for change point detection: Beyond ARL. *Submitted*, 2016.

G. Lindgren. *Stationary stochastic processes: Theory and applications*. Boca Raton, CRC Press, 2013.

MARIN. *Dynamic Stability Simulation*. URL `http://www.marin.nl/web/Facilities-Tools/Software/Dynamic-Stability-Simulation.htm`. [accessed on 31-03-2017].

L. Nirenberg. On elliptic partial differential equations. In *Il principio di minimo e sue applicazioni alle equazioni funzionali*, pages 1–48. Springer, 2011.

E. Page. Continuous inspection scheme. *Biometrika*, 41:100–115, 1954.

S. Rice. Mathematical analysis of random noise. *Bell System Technical Journal*, 23: 282–332, 1944.

M. Robbins, C. Gallagher, R. Lund, and A. Aue. Mean shift testing in correlated data. *Journal of Time Series Analysis*, 32:498–511, 2011.

M. Shinozuka and C.-M. Jan. Digital simulation of random processes and its applications. *Journal of Sound and Vibration*, 25:111–128, 1972.

E. M. Stein and G. L. Weiss. *Introduction to Fourier Analysis on Euclidean Spaces*. Mathematical Series. Princeton University Press, 1971. ISBN 9780691080789.

The WAFO group. *WAFO - A MATLAB Toolbox for Analysis of Random Waves and Loads; Tutorial for WAFO version 2.5*. Center for Mathematical Sciences, Lund University, 2011. URL `http://www.maths.lth.se/matstat/wafo/`.

A. Yaglom. *Vvedenie v teoriiu statsionarnykh sluchainykh funktsii. An introduction to the theory of stationary random functions*. Mineola, Dover Publications, 1962.

Modelling of fluid mixing and dynamics in curved pipelines

Thijs Bouwhuis, Daan Crommelin, Olfa Jaïbi, Vivi Rottschäfer, Ray Sheombarsing, Bas van 't Hof

Abstract

KEYWORDS: Navier Stokes, Modelling, Fluid Dynamics, Advection-diffusion equation, Multiphase flow

1 Introduction

During the Study Mathematics with Industry held in Amsterdam we worked on a challenge formulated by Shell about the mixing of fluids in curved pipelines. The question originates from a problem that can occur when transporting oil and gas through pipelines. This transportation of hydrocarbon fluids through pipelines in a safe and efficient way is a major challenge for the petrochemical industry. Especially in rough conditions like the ones that are present on the bottom of the ocean where temperatures typically lie around 4°C. Many oil and gas fields lie beneath inland waters and offshore areas around the world, and the exploration, drilling and development of oil and gas fields in these underwater locations is called subsea. When oil and gas flow out of a subsea well the fluids are transported through pipelines on the ocean floor to offshore production platforms. These pipelines, can stretch for many kilometres, forming a large infrastructure. Because the seabed is not perfectly flat, there are segments of pipeline which will not lie horizontal but under an angle or even vertical.

When oil and gas are produced from a well, it is usually a mixture of the two which is often co-flowing with water, sand particles and other contaminants. A phenomenon related to the presence of water that can cause a lot of problems is hydrate formation, typically gas hydrates. These hydrates are *solids* which are crystalline water-based: they consist of a gas molecule (e.g. methane, ethane, propane and carbon dioxide) which is trapped in a water cavity composed of hydrogen bonded water molecules. Macroscopically, hydrates form a slurry which is quite similar to wet snow. Single gas hydrates can cluster together and form structures. When these structures grow, they can form a hydrate plug that blocks the full cross sectional area of the pipe.

Hydrates only form under specific circumstances, namely at low temperatures and high pressure. These circumstances arise, for instance, when an oil and gas well (re-)starts production and the pipeline is filled with cold fluids, including water. To prevent hydrates from forming the pipeline is usually flushed with a hydrate inhibitor. Such a hydrate inhibitor chemically acts the same as the antifreeze fluid one uses in a car. A common hydrate inhibitor is methanol. In general, the aim is to use as little methanol as possible, since it is both an expensive and dangerous fluid. That is one of the reasons why Shell wants to be able to better predict how methanol will mix into a pipeline filled with water.

1.1 Problem description

For our study, we start with a pipeline filled with water. Then, from one entrance of the pipeline, methanol is flushed into it at a constant speed. The challenge that Shell posed was:

What is the concentration of methanol along the pipeline as a function of time and space, when looking at different geometries of the pipe such as the presence of curves and sections of the pipeline under an angle?

Determining this concentration is not straightforward since there are several effects that have to be taken into account. The first one is the difference in the densities: the density of methanol is approximately 800 kg/m^3, whereas that of water is approximately 1000 kg/m^3. Because of this density difference, the methanol tends to 'float' on the water. This results in different behaviour of the methanol in the water along the various sections of the pipeline. In downward sloped sections, the density difference will result in a stable front of methanol that moves down. In horizontal or upward sloped sections a layer of methanol will form and float on top of the water. When observing a cross section of the pipe, one can see a distinct region of a 'light' fluid on top of a 'heavy' fluid. This phenomenon is called stratification.

In addition, we have to take into account that water and methanol are miscible. This means that they are able to fully dissolve in one another. This in contrary to immiscible fluids (e.g. oil and water) for which there will always exist a distinct layer between the two fluids. There are some additional effects (e.g. viscosity differences, surface tension) which will play a role in reality, but these will not be accounted for in this study.

This report is structured as follows: First, a physical background in fluid dynamics is presented with the relevant equations and their derivation. In section 3 appropriate notations and conventions are introduced. The problem is then approached from two different angles: in section 4 a 3D tranformation of coordinates is studied, intended to focus on the mixing interface of the miscible fluids. In section 5 a 1D model reduction approach is proposed, in which the along-pipeline direction is the only remaining spatial coordinate in the resulting model. This 1D model is solved numerically, as dis-

cussed in section 6. Results from simulations with this numerical model are presented in section 7.

2 Navier-Stokes

In this section we provide a brief description of the Navier-Stokes equations. The contents of this section are not meant as a detailed exposition of the field but should rather be thought of as a simple and heuristic introduction to fluid dynamics. Furthermore, the idea's presented in this section are standard and no originality is claimed. The interested reader is referred to Chorin and Marsden (1979) for a more comprehensive introduction into the field of fluid dynamics.

Suppose $\Omega \subset \mathbb{R}^3$ is an open subset which contains a fluid with mass density $\rho(t,x)$, where $t \geq 0$ and $x \in \Omega$. Let $\mathbf{u}(t,x)$ denote the velocity of a fluid particle starting at $x \in \Omega$ at time t. In other words, the trajectory $t \mapsto \varphi(t,x)$ of a fluid particle starting at x satisfies the differential equation

$$\frac{d}{dt}\varphi\,(t,x) = \mathbf{u}\,(t,\varphi(t,x))\,.$$

The Navier-Stokes equations are based on two basic principles: conservation of mass and Newton's second law. In order for the computations in the following sections to be valid we shall henceforth assume that ρ, φ and $\mathbf{u}$ are sufficiently smooth.

2.1 Conservation of mass

In this section we derive an equation for the conservation of mass. To this end, suppose $B \subset \Omega$ is an open subset. Then the rate of change of mass of the fluid contained in B is given by

$$\frac{d}{dt}\int_B \rho \,\mathrm{d}V = \int_B \frac{\partial \rho}{\partial t}\,\mathrm{d}V.$$

We assume that the change of mass in B is only caused by fluid flowing in from $\Omega \setminus B$ or flowing out from B. In particular, the rate at which fluid comes in or escapes through ∂B is

$$-\int_{\partial B} \langle \rho\mathbf{u}, \mathbf{n}\rangle \,\mathrm{d}A = -\int_B \mathrm{div}\,(\rho\mathbf{u})\,\mathrm{d}V,$$

where $\langle\cdot,\cdot\rangle$ is the standard Euclidian product on $\mathbb{R}^3$ and $\mathbf{n}$ is the *outward* (unit) normal vectorfield on ∂B. Therefore, conservation of mass is equivalent to

$$\int_B \frac{\partial \rho}{\partial t}\,\mathrm{d}V = -\int_B \mathrm{div}\,(\rho\mathbf{u})\,\mathrm{d}V. \tag{1}$$

In turn, this implies that

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho \mathbf{u}) = 0, \tag{2}$$

since (2) holds for any open subset $B \subset \Omega$.

2.2 Newton's second law

In this section we use Newton's second law and the conservation of mass to derive an equation for the velocity field $\mathbf{u}$. The idea is straightforward: we simply compute the rate of change of momentum of the fluid, the net force acting on the fluid, and then use Newton's second law to relate the two.

Rate of change of momentum The acceleration of a fluid particle at $x \in \Omega$ at time t is given by

$$\frac{d^2}{dt^2}\varphi(t,x) = \frac{\partial \mathbf{u}}{\partial t}(t, \varphi(t,x)) + \mathbf{u} \cdot \nabla \mathbf{u}(t, \varphi(t,x)),$$

where

$$\mathbf{u} \cdot \nabla \mathbf{u} := \sum_{j=1}^{3} \frac{\partial \mathbf{u}}{\partial x_j} u_j.$$

Let $B \subset \Omega$ be an open subset as before and set $B_t := \varphi(t, B)$. Then the momentum of the fluid initially contained in B at time t is given by

$$\int_{B_t} \rho \mathbf{u} \, \mathrm{d}V = \int_B (\rho \mathbf{u}) \circ \varphi \cdot \det D_x \varphi \, \mathrm{d}V.$$

The conservation of mass (2) and a tedious (but straightforward) computation can now be used to show that the rate of change of momentum is given by

$$\frac{d}{dt} \int_{B_t} \rho \mathbf{u} \, \mathrm{d}V = \int_{B_t} \rho \left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} \right) \mathrm{d}V. \tag{3}$$

Forces acting on the fluid Next, we explain how to model the forces acting on the fluid. One can separate these forces into two categories:

(*i*) forces which act "directly" on the fluid particles in B_t,

(*ii*) forces which act on B_t through its boundary.

It is out of the scope of this text to give a detailed treatment of all the forces acting on the fluid. Instead, we have chosen to give two simple but representative examples of how to model forces of either type. We will use these examples to derive a simplified

equation for the rate of change of momentum. In the next section we will then proceed by stating the full Navier-Stokes equations with the understanding that the forces appearing in the equation are derived by using the principles presented in this section.

A simple example of a force of type (i) is gravity. Indeed, gravity is a force which acts "directly" on each fluid particle in Ω. In the easiest case, the force on B_t due to gravity is given by

$$\mathbf{F}_g = \int_{B_t} \rho g \, \mathrm{d}V,$$

where $g \approx 9.81 \text{ m/s}^2$ is the gravitational acceleration.

The general procedure for modeling forces of the second type is to derive an integral formulation of the force by using the Divergence Theorem. Let us, for example, consider the internal force $\mathbf{F}_p$ which corresponds to the fluid pressing on itself. One could attempt to model this force by assuming the existence of a function $p : [0, \infty) \times \Omega \to \mathbb{R}$, usually called the *pressure*, so that the force on ∂B_t due to the fluid outside of B_t is given by

$$\mathbf{F}_p = -\int_{\partial B_t} p\mathbf{n} \, \mathrm{d}A = -\int_{B_t} \nabla p \, \mathrm{d}V,$$

where $\mathbf{n}$ is the outward unit normal on B_t. We remark, however, that in reality there is also another non-tangential force acting on the boundary of B_t which contributes to the internal force and is related to the *viscosity* of the fluid.

If gravity and internal pressure are the only forces acting on the fluid, i.e. $\mathbf{F}_{net} = \mathbf{F}_g + \mathbf{F}_p$, then

$$\int_{B_t} \rho \left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} \right) \mathrm{d}V = \int_{B_t} (\rho g - \nabla p) \ \mathrm{d}V \tag{4}$$

by Newton's second law and (3). Hence

$$\rho \left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} \right) = \rho g - \nabla p, \tag{5}$$

since (4) holds for any open subset B. The latter equation is essentially an infinitesimal formulation of Newton's second law.

2.3 Navier-Stokes equations

In this section we combine the conservation of mass and Newton's second law to state the Navier-Stokes equations. We start with the simplified considerations from the previous section and explain why the resulting system is ill-posed. We then resolve this issue by introducing the notion of *incompressiblity*. Finally, we state the full set of Navier-Stokes equations for an incompressible fluid.

An ill-posed Navier-Stokes equation The equations for the conservation of mass and Newton's second law from the previous sections yield the following simplified system:

$$\begin{cases} \rho\left(\dfrac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u}\right) = \rho g - \nabla p, \\ \dfrac{\partial \rho}{\partial t} + \operatorname{div}(\rho\mathbf{u}) = 0. \end{cases} \tag{6}$$

The unknowns in (6) are the the mass density ρ, the internal pressure p and the three components of the velocity field $\mathbf{u}$. Note, however, that the system in (6) is underdetermined, since we have five unknowns but only four equations. A possible solution to this problem is to take the conservation of energy into account.

The total energy of the physical model consists of *kinetic* and *internal* energy. The kinetic energy of the fluid is simply the classical energy related to the motion of the fluid on a macroscopic level. More precisely, the kinetic energy of the fluid initially contained in B at time t is given by

$$E_{kin}(t, B) = \frac{1}{2}\int_{B_t} \rho \|\mathbf{u}\|^2 \,\mathrm{d}V,$$

where $\|\cdot\|$ denotes the Euclidian norm on $\mathbb{R}^3$.

The internal energy E_{in} is related to the potential energy and microscopic motion of the fluid molecules. A detailed treatment of the internal energy requires thermodynamical considerations and is out of the scope of this text. We remark, however, that it is possible to balance the number of equations and unknowns by adding a scalar equation based on the conservation of energy:

$$\frac{dE}{dt} = 0, \quad E := E_{kin} + E_{in}.$$

The incompressible Navier-Stokes equations Another strategy for balancing the number of equations and unknowns is to introduce a so-called equation of state, providing an algebraic relation between the pressure and the fluid properties, the density in this case. A simple approach is to assume that the fluid is *incompressible*, i.e., φ preserves volume. This is equivalent to requiring that $\operatorname{div}(\mathbf{u}) = 0$. It depends on the properties of the fluid whether this assumption is realistic or not. For water and methanol in liquid state, this is generally a suitable assumption.

If the velocity field is divergence free, then the equation for the conservation of mass (2) can be explicitly solved. To see this, suppose that $\operatorname{div}(\mathbf{u}) = 0$, then

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho\mathbf{u}) = \frac{\partial \rho}{\partial t} + \langle \nabla\rho, \mathbf{u}\rangle = 0,$$

where $\langle\cdot,\cdot\rangle$ denotes the Euclidian inner product on $\mathbb{R}^3$. Consequently,

$$\frac{d}{dt}\int_{B_t} \rho \,\mathrm{d}V = \int_{B_t}\left(\frac{\partial \rho}{\partial t} + \langle\nabla\rho, \mathbf{u}\rangle\right)\mathrm{d}V = 0,$$

by the same computation as in (3). In other words, if $\mathbf{u}$ is divergence free, then φ preserves mass (the converse holds as well), i.e.,

$$\int_B \rho(0,x)\ \mathrm{d}V = \int_{B_t} \rho(t,x)\ \mathrm{d}V = \int_B \rho\left(t,\varphi\left(t,x\right)\right) \det D_x\varphi(t,x)\ \mathrm{d}V,$$

for all $t \geq 0$. Therefore,

$$\rho\left(t,\varphi\left(t,x\right)\right) = \frac{\rho(0,x)}{\det D_x\varphi(t,x)} = \rho(0,x), \quad t \geq 0,\ x \in \Omega, \tag{7}$$

since B was arbitrary and $\det D_x\varphi(t,x) \equiv 1$ (because $\operatorname{div}(\mathbf{u}) = 0$). In particular, the mass density is independent of time along trajectories of the fluid.

We are now ready to state the *incompressible* Navier-Stokes equations:

$$\begin{cases} \rho\left(\dfrac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u}\right) = -\nabla p + \mu\Delta\mathbf{u} + \rho g, \\ \operatorname{div}(\mathbf{u}) = 0, \end{cases} \tag{8}$$

where $(t,x) \in [0,T] \times \Omega$ and

- $p : [0,\infty) \times \Omega \to \mathbb{R}$ is the internal pressure,
- μ is the *dynamic viscosity* of the fluid,
- $g \approx 9.81$ is the acceleration of gravity,
- $T > 0$ is a prescribed integration time.

The unknowns in (8) are the internal pressure p and the velocity field $\mathbf{u}$. The mass density ρ is explicitly given by the initial and boundary conditions, as can be inferred from (7). Therefore, the number of unknowns and equations in (8) is balanced. Finally, the system should be supplemented with an initial condition $\mathbf{u}_0 : \Omega \to \mathbb{R}^3$ and suitable boundary conditions. These are dictated by the physical model under consideration.

3 Notation and conventions

Here we introduce the coordinates/variables as seen in Figure 1. The pipeline is fully described using the following coordinates:

- s is tangential to the central line of the pipeline. It is oriented along the flow, which we chose to be from left to right (water flowing in from the left entrance)
- w is the vertical direction starting from the the central line. It is normal to the central line and the radial direction q but not to the mixing layer.

- q is normal to both s and w. It is pointing out of the paper in the sketch shown in Figure 1. We will ignore this coordinate in all our subsequent transformations since we assume that the liquid is homogeneously distributed along a vertical cross section (the mixing layer is horizontal).

- α denotes the angle that the central line makes with the horizontal. It is positive in case the pipeline is sloping downwards and negative in case the pipeline is sloping upwards (see sketch).

- c and A denote the concentration of methanol in the fluid and the area of the fluid (see Figure 1), respectively. Since we only have two components, the concentration of water $\tilde{c}$ satisfies $\tilde{c} = 1 - c$.

- The subscripts u and l denote the *upper* and *lower* regions, with respect to the vertical position of the fluids.

- D_w is the normal diffusion coefficient in the w-direction.

- ψ is a mixing term that will be used in the 1D model in section 5.

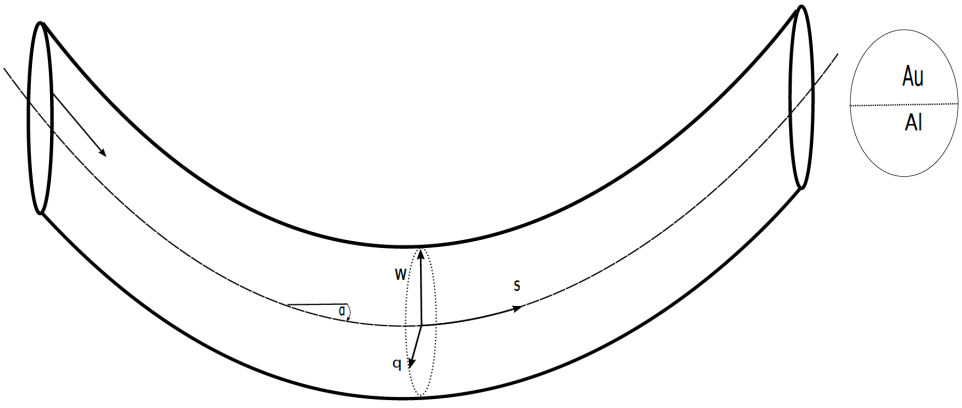


Figure 1: Sketch of the 3D pipeline and a cross-section of the pipeline in the vertical direction.

4 3D Approach: Coordinate transformation along the central line

4.1 3D co-moving frame

The non-trivial curvature of pipelines makes it difficult to model the flow of different fluids and the change in concentration. To account for this, we suggest to perform a coordinate transformation that allows us to focus on the specific needs: computing the concentration in the case of miscible fluids. In this section we discuss how such a transformation can be carried out. Although not fully complete at this point, the ideas presented in this section may provide a useful approach when worked out in more detail. We leave a more detailed exploration of these ideas for future study.

The performed coordinate transformation follows the fluid interface and allows for a stretching in the direction normal to the flow (so where the diffusion is highest between the two fluids), a method also known as asymptotics. We assume that the fluids are evenly distributed along a vertical cross-section, as depicted in Figure 1. Therefore, the y-direction can be omitted when it comes to the spatial distribution of the fluids. Therefore, our 3D model reduces to a 2D model, centred along the central line of the pipe s. The height of the interface between the fluids, can be parametrized as a (non-trivial) function of position and time. Define $h(s,t)$ as the height of the interface surface, oriented along the w-direction, which is defined to be normal to the interface surface. Then, for any time t, the interface at point s_0 has height $h(s_0,t)$.

For immiscible fluids, the concentration of methanol is represented by a Heaviside function, with changing point at $h(s,t)$:

$$c_0(s,t) = \begin{cases} 0 & \text{if} \quad w < h(s,t) \\ 1 & \text{if} \quad w > h(s,t) \end{cases} \tag{9}$$

Note that the immiscible solution has a discontinuous volume fraction c_0. The mass fraction can only be 1 or 0, because there is no mixing. The velocity and pressure fields are continuous, but there may be discontinuities in their derivatives.

Due to the discontinuity in the volume fractions, the advection-diffusion equations only hold in integrated form.

4.2 Immiscible and miscible solutions

We will try to find the solution of the miscible system

$$\rho\frac{\partial \vec{v}}{\partial t} + \rho\vec{v}^T\nabla\vec{v} = \mu\nabla^2\vec{v} - \nabla p + \rho\vec{g}, \tag{10}$$

$$\nabla\cdot\vec{v} = 0, \tag{11}$$

$$\frac{\partial c}{\partial t} + \nabla\cdot c\vec{v} = D\nabla^2 c. \tag{12}$$

where D is small. The first two equations are the incompressible Navier-Stokes equations (8) discussed before. The third equation is the advection-diffusion equations for

c, with the flow velocity $\vec{v}$ from (10)-(11).

Because the interface moves with the fluid, the time derivative of the water height is given by the *kinematic boundary condition* for the interface in two dimensions. :

$$\left(0, 0, \frac{\partial h}{\partial t}\right) \cdot \vec{n} = \vec{v} \cdot \vec{n} \quad \Leftrightarrow \quad \frac{\partial h}{\partial t} + \frac{\partial h}{\partial s} v_{0,s} = v_{0,w}. \tag{13}$$

A (non-unit) upward normal vector $\vec{m}$ to the interface is given by

$$\vec{m} := \left(-\frac{\partial h}{\partial s}, 1\right). \tag{14}$$

The unit upward normal vector $\vec{n}$ is found by scaling $\vec{m}$:

$$\vec{n} := \frac{\vec{m}}{|\vec{m}|}. \tag{15}$$

The directions parallel to the interface are called $\vec{a}$ and $\vec{b}$:

$$\vec{a} := \frac{(1, \frac{\partial h}{\partial s})}{|(1, \frac{\partial h}{\partial s})|} \quad , \quad \vec{b} := \vec{n} \times \vec{a}. \tag{16}$$

Introduce the coordinate transformation:

$$\tilde{s}(s, \chi, t) = s - \frac{\partial h(s,t)}{\partial s} \chi\delta \tag{17}$$

$$w(s, \chi, t) = h(s,t) + \chi\delta \tag{18}$$

where χ represents the stretching along the w-axis. Then the derivatives in terms of

the new coordinates become:

$$\frac{\partial c}{\partial s} = \frac{\partial \chi}{\partial \tilde{s}}\frac{\partial \tilde{s}}{\partial s}\frac{\partial c}{\partial \chi} + \frac{\partial \tilde{s}}{\partial s}\frac{\partial c}{\partial \tilde{s}} \tag{19}$$

$$= \left(-\frac{\partial h}{\partial s}\delta\right)^{-1}\left(1 - \frac{\partial^2 h}{\partial s^2}\chi\delta\right)\frac{\partial c}{\partial \chi} + \left(1 - \frac{\partial^2 h}{\partial s^2}\chi\delta\right)\frac{\partial c}{\partial \tilde{s}} \tag{20}$$

$$\frac{\partial c}{\partial w} = \frac{\partial \chi}{\partial w}\frac{\partial c}{\partial \chi} + \frac{\partial \tilde{s}}{\partial \chi}\frac{\partial \chi}{\partial w}\frac{\partial c}{\partial \tilde{s}} \tag{21}$$

$$= \delta^{-1}\frac{\partial c}{\partial \chi} - \frac{\partial h}{\partial s}\frac{\partial c}{\partial \tilde{s}} \tag{22}$$

$$\frac{\partial^2 c}{\partial s^2} = \frac{\partial^2 h}{\partial s^2}\delta\left(\frac{\partial h}{\partial s}\delta\right)^{-2}\left(1 - \frac{\partial^2 h}{\partial s^2}\chi\delta\right)\frac{\partial c}{\partial \chi} + \left(-\frac{\partial h}{\partial s}\delta\right)^{-1}\left(-\frac{\partial^3 h}{\partial s^3}\chi\delta\right)\frac{\partial c}{\partial \chi} \tag{23}$$

$$+ \left(\frac{\partial h}{\partial s}\delta\right)^{-2}\left(1 - \frac{\partial^2 h}{\partial s^2}\chi\delta\right)^2\frac{\partial^2 c}{\partial \chi^2} - \frac{\partial^3 h}{\partial s^3}\chi\delta\frac{\partial c}{\partial \tilde{s}} \tag{24}$$

$$+ 2\left(\frac{\partial h}{\partial s}\delta\right)^{-1}\left(1 - \frac{\partial^2 h}{\partial s^2}\chi\delta\right)\frac{\partial^2 c}{\partial \chi \partial \tilde{s}} + \left(1 - \frac{\partial^2 h}{\partial s^2}\chi\delta\right)\frac{\partial^2 c}{\partial \tilde{s}^2} \tag{25}$$

$$\frac{\partial^2 c}{\partial w^2} = \delta^{-2}\frac{\partial^2 c}{\partial \chi^2} + \left(\frac{\partial h}{\partial s}\right)^2\frac{\partial^2 c}{\partial \tilde{s}^2} - 2\delta^{-1}\frac{\partial h}{\partial s}\frac{\partial^2 c}{\partial \chi \partial \tilde{s}} \tag{26}$$

$$\tag{27}$$

Then the LHS of equation (12) for the concentration becomes:

$$\frac{\partial c}{\partial t} + \nabla \cdot c\vec{v} = \frac{\partial c(\tilde{s}, \tilde{w}, t)}{\partial t} + \nabla \cdot c(\tilde{s}, \tilde{w}, t)\vec{v} \tag{28}$$

$$= \frac{\partial c(\tilde{s}, \tilde{w}, t)}{\partial t} + \frac{\partial c(\tilde{s}, \tilde{w}, t)}{\partial s} \cdot v_s + \frac{\partial c(\tilde{s}, \tilde{w}, t)}{\partial w} \cdot v_w \tag{29}$$

$$= \frac{\partial c(\tilde{s}, \tilde{w}, t)}{\partial t} + \left(\left(-\frac{\partial h}{\partial s}\delta\right)^{-1}\left(1 - \frac{\partial^2 h}{\partial s^2}\chi\delta\right)\frac{\partial c}{\partial \chi}\right. \tag{30}$$

$$\left. + \left(1 - \frac{\partial^2 h}{\partial s^2}\chi\delta\right)\frac{\partial c}{\partial \tilde{s}}\right) \cdot v_s + \left(\delta^{-1}\frac{\partial c}{\partial \chi} - \frac{\partial h}{\partial s}\frac{\partial c}{\partial \tilde{s}}\right) \cdot v_w \tag{31}$$

Furthermore, the RHS of (12) becomes:

$$
\begin{aligned}
D_w \nabla^2 c &= D_w \left(\frac{\partial^2 c(\tilde{s},\tilde{w},t)}{\partial s^2} + \frac{\partial^2 c(\tilde{s},\tilde{w},t)}{\partial w^2} \right) && (32)\\
&= D_w \cdot \left(\frac{\partial^2 h}{\partial s^2}\delta \left(\frac{\partial h}{\partial s}\delta \right)^{-2} \left(1 - \frac{\partial^2 h}{\partial s^2}\chi\delta \right) \frac{\partial c}{\partial \chi} \right. && (33)\\
&+ \left(-\frac{\partial h}{\partial s}\delta \right)^{-1} \left(-\frac{\partial^3 h}{\partial s^3}\chi\delta \right) \frac{\partial c}{\partial \chi} && (34)\\
&+ \left(\frac{\partial h}{\partial s}\delta \right)^{-2} \left(1 - \frac{\partial^2 h}{\partial s^2}\chi\delta \right)^2 \frac{\partial^2 c}{\partial \chi^2} - \frac{\partial^3 h}{\partial s^3}\chi\delta \frac{\partial c}{\partial \tilde{s}} && (35)\\
&+ 2\left(\frac{\partial h}{\partial s}\delta \right)^{-1} \left(1 - \frac{\partial^2 h}{\partial s^2}\chi\delta \right) \frac{\partial^2 c}{\partial \chi \partial \tilde{s}} + \left(1 - \frac{\partial^2 h}{\partial s^2}\chi\delta \right) \frac{\partial^2 c}{\partial \tilde{s}^2} && (36)\\
&+ \left. \delta^{-2}\frac{\partial^2 c}{\partial \chi^2} + \left(\frac{\partial h}{\partial s} \right)^2 \frac{\partial^2 c}{\partial \tilde{s}^2} - 2\delta^{-1}\frac{\partial h}{\partial s}\frac{\partial^2 c}{\partial \chi \partial \tilde{s}} \right)
\end{aligned}
$$

These expressions can lead to the advection-diffusion equation in the new coordinates. Further work is needed to apply a similar approach, using the same coordinate stretching, to equations (10)-(11). As mentioned at the start of this section, such a further exploration is beyond the scope of this report, and is left for future study.

5 1D approach: Averaging over the concentrations

5.1 A two-layer model with one space dimension

In this section we discuss a simple model for mixing and diffusion of fluids in a pipeline. We consider a situation with two layers with different fluid mixtures, one above the other. This vertical stratification can be the result of e.g. density differences, with the heaviest mixture in the lower layer and the lightest in the upper layer. Furthermore, we assume that each layer contains a mixture of two fluids, methanol and water. We remark that two natural extensions of this simple set-up are (i) to model more than two layers in the vertical, or even consider a situation of continuous vertical stratification, and (ii) to let each mixture consist of more than two fluids. Clearly, the number of layers and the number of mixture components need not be the same.

The fluid mixtures in the upper and lower layers have different horizontal velocities. The time evolution of the fluid mixtures are governed by 1-dimensional advection-diffusion equations for the upper and lower layer separately. The spatial coordinate in these advection-diffusion equations is s, the coordinate along the central line of the pipe. The two layers exchange fluid at the layer interface, modelled with source/sink terms in the horizontal advection-diffusion equations. These source/sink terms are derived from a vertical diffusion equation. For simplicity, we ignore here the angle of the pipeline, and assume that the pipeline is oriented horizontally so that a vertical

cross-section forms a perfect circle in a plane orthogonal to the direction of s. The circle has the diameter of the pipe, denoted D, so we have $0 \leq w \leq D$ for the vertical coordinate w. We denote by h the height of the layer interface, i.e. the lower layer extends from $w = 0$ to $w = h$, and the upper layer from $w = h$ to $w = D$.

From here on, we use notations with subscripts u and l to denote quantities for the upper and lower layer, respectively. The cross-sectional area of the upper layer is A_u, and that of the lower layer is A_l. Clearly $A_u + A_l = A$ with A the total cross-sectional area $A = \pi R^2$ and $R = D/2$ the pipe radius. Given the layer interface height h, we have

$$A_u = R^2 \cos^{-1}((h-R)/R) + (R-h)\sqrt{2hR - h^2} \quad \text{and} \quad A_l = \pi R^2 - A_u \,. \tag{37}$$

5.2 Coupled advection-diffusion equations

We denote by c_u the volume fraction of methanol in the upper layer. By construction, the volume fraction of water in the upper layer, denoted by $\tilde{c}_u$, satisfies $\tilde{c}_u = 1 - c_u$. Likewise, the lower layer methanol and water fractions are denoted c_l and $\tilde{c}_l$, satisfying $c_l + \tilde{c}_l = 1$. Furthermore, let u_u and u_l be the fluid velocities (in the s-direction) in the upper and lower layer. We model the time evolution of the fractions $c_u(s,t)$ and $c_l(s,t)$ with advection-diffusion equations coupled by a source/sink term:

$$\partial_t(c_u A_u) + \partial_s(u_u c_u A_u) = \partial_s(D_u \partial_s(c_u A_u)) + \psi \tag{38a}$$

$$\partial_t(c_l A_l) + \partial_s(u_l c_l A_l) = \partial_s(D_l \partial_s(c_l A_l)) - \psi \tag{38b}$$

We denote partial derivatives with respect to s and t by ∂_s and ∂_t, respectively. The velocity fields $u_u(s,t)$ and $u_l(s,t)$ are given. We assume that the effective axial diffusion coefficients (D_u and D_l) are constant in s and t, and that they are identical in the upper and lower layer, i.e. $D_u = D_l$. Finally, the term ψ is a source/sink term that accounts for the exchange/mixing of fluids between the two layers. Below, we derive an expression for ψ based on a diffusion equation in the vertical direction.

As can be seen, the volume fractions c_u, c_l depend only on (s,t) in our model set-up here. Thus, these fractions are assumed constant over the upper ($w > h$) and lower ($w < h$) parts of the pipe cross-section. Any exchange of fluids between the layers, as modelled by ψ, is assumed to be mixed instantaneously within each layer in the directions perpendicular to s. This will guide the derivation of ψ.

5.3 Exchange between layers: a source/sink term from the heat equation

In our model set-up, there is no advection in the vertical direction, only diffusion. We start our derivation of ψ by considering the methanol volume fraction in a pipe cross-section (i.e., s is fixed) to be a function of both the vertical coordinate w and time t, so $c = c(w,t)$. The time evolution is governed by the diffusion equation

$$\partial_t c = \partial_w(D_w \partial_w c) \tag{39}$$

with diffusion coefficient D_w. Let the initial state of c be the piecewise-constant (in w) profile with c_u in the upper layer and c_l in the lower layer. Thus, $c(w,0) = c_l + H(w-h)(c_u - c_l)$ with h the interface height and $H(.)$ the Heaviside function. We assume $c_u > c_l$ so that the lower layer fluid mixture is heavier than the mixture in the upper layer (as water is heavier than methanol).

If D_w is independent of w, (39) reduces to the heat equation in 1-d. Below, we use a simple analytical solution for the heat equation on $\mathbb{R}$, although strictly speaking, the domain for our problem is finite since w is bounded by the pipe wall. A more refined treatment, beyond the scope of this report, would be to take account of this finite domain size (note that the curvature of the pipe wall makes the characterization of the finite domain complicated). We remark that our primary interest is in diffusion over small time intervals, so that most of the exchange is (very) close to the layer interface and effects of finite domain size may not have much impact.

Consider the following standard initial value problem for the heat equation on $\mathbb{R}$:

$$\partial_t v = \kappa\, \partial_x^2 v\,, \qquad x \in \mathbb{R}\,, \qquad v(x,0) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{40}$$

with diffusion constant $\kappa > 0$. The solution at time $t > 0$ is

$$v(x,t) = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left(\frac{x}{\sqrt{4\kappa t}}\right) \tag{41}$$

with erf(.) the error function Temme (1996). From the solution at t we can calculate the amount of exchange over the time interval $[0,t]$ across the interface at $x = 0$ in this standard problem:

$$\begin{aligned} \int_0^\infty [v(x,t) - v(x,0)]\,dx &= \lim_{x^*\to\infty} \frac{1}{2} \int_0^{x^*} \left[\mathrm{erf}\left(\frac{x}{\sqrt{4\kappa t}}\right) - 1\right] dx \\ &= \lim_{x^*\to\infty} \frac{1}{2} \left[\sqrt{\frac{4\kappa t}{\pi}} \left(e^{-(x^*)^2/(4\kappa t)} - 1\right) - x^* + x^* \mathrm{erf}\left(\frac{x^*}{\sqrt{4\kappa t}}\right)\right] \\ &= -\sqrt{\frac{\kappa t}{\pi}} \end{aligned} \tag{42}$$

where we have used that $\mathrm{erf}(x) \to 1 - \dfrac{\exp(-x^2)}{x\sqrt{\pi}}$ as $x \to +\infty$ Oldham et al. (2009).

Transforming the standard problem above to the diffusion equation (39) of interest to us, we obtain for the exchange over a time interval dt the following:

$$\int_h^\infty [c(w,dt) - c(w,0)]\,dw = -(c_u - c_l)\sqrt{\frac{D_w\,dt}{\pi}} \tag{43}$$

To obtain an expression for the source/sink term ψ from this, we must take into account that the vertical exchange takes place over the layer interface with length $2\sqrt{2hR - h^2}$, hence it should be proportional to this length.

Furthermore, an important assumption is that the amount of exchanged fluid is instantaneously mixed throughout the upper and lower parts of the pipe cross-section, with cross-sectional areas A_u and A_l, respectively. Thus, if we consider the upper layer at location s and time t, the change in c_u over a time interval dt due to fluid exchange between the layers is

$$\begin{aligned} c_u(s,t+dt) &= \frac{c_u(s,t)A_u(s,t) + \text{exchange}}{A_u(s,t)} \\ &= c_u(s,t) - \frac{[c_u(s,t) - c_l(s,t)]\sqrt{D_w\, dt/\pi}\; 2\sqrt{2h(s,t)R - h^2(s,t)}}{A_u(s,t)} \\ &= c_u(s,t) - \frac{[c_u(s,t) - c_l(s,t)]\, F(s,t)\, \sqrt{D_w\, dt}}{A_u(s,t)} \end{aligned} \tag{44}$$

where F is dependent on the interface height $h(s,t)$:

$$F(s,t) = 2\sqrt{\frac{2h(s,t)R - h^2(s,t)}{\pi}} \tag{45}$$

We note that A_u depends on (s,t) through $h(s,t)$, see (37). Also, we neglect the (presumably small) change in $h(s,t)$ (and thus A_u) over the time interval dt.

The advection-diffusion equations (38) describe the time evolution of $c_u A_u$ and $c_l A_l$ rather than c_u and c_l. As a result, the factor A_u in (44) drops out and we obtain for the source/sink term

$$\psi(s,t) = \lim_{dt \downarrow 0} \; -[c_u(s,t) - c_l(s,t)]\, F(s,t)\, \sqrt{\frac{D_w}{dt}} \tag{46}$$

Note that ψ diverges in the $dt \to 0$ limit, a consequence of our set-up with a sharp layer interface at which the mixture fractions are discontinuous. It implies that ψ should be interpreted in a weak or distributional sense. For numerical time integration with time step Δt we will use $\psi(s,t)\,\Delta t \approx -[c_u(s,t) - c_l(s,t)]\, F(s,t)\, \sqrt{D_w\, \Delta t}$.

We conclude this section with some remarks about the vertical diffusion coefficient D_w. Above, we assumed it to be independent of w to obtain an expression for ψ from the 1-dimensional heat equation. It would make sense to let D_w depend on the (local) shear, i.e. the horizontal velocity difference between the two layers, $|u_u(s,t) - u_l(s,t)|$. A large shear may generate small-scale turbulence at the layer interface, enhancing the effective vertical diffusivity. We leave further exploration of this issue for future study.

6 Implementation

In this section we provide a concise description of the numerical method employed to approximate solutions of the coupled 1-d advection-diffusion equations presented in the previous section. For notational convenience, we will replace the subscripts u and

l by the integers 1 and 2, respectively, and refer to the upper and lower region as the first and second region, respectively.

The system of equations to be solved is

$$\begin{cases} \dfrac{\partial}{\partial t}\left(A_i c_i\right) + \dfrac{\partial}{\partial s}\left(u_i A_i c_i\right) = \dfrac{\partial}{\partial s}\left(D_i \dfrac{\partial}{\partial s}\left(A_i c_i\right)\right) + (-1)^{i+1}\,\psi, \\ c_i\left(t,0\right) = \alpha_i, \quad \dfrac{\partial c_i}{\partial s}\left(t,L\right) = \beta_i, \\ c_i\left(0,s\right) = c_i^0(s), \end{cases} \tag{47}$$

for $1 \leq i \leq 2$, where

- $T > 0$ is the integration time,
- $L > 0$ is the length of the pipe,
- $u_i : [0,T] \times [0,L] \to \mathbb{R}$ is the prescribed speed of the mixture in the i-th region in the direction of the pipe,
- $D_i \in \mathbb{R}$ is the diffusion coefficient of methanol in the i-th region,
- $\psi : [0,T] \times [0,L] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ models the diffusion across the interface,
- $c_i^0 : [0,L] \to \mathbb{R}$ is the initial concentration of methanol,
- $\alpha_i, \beta_i \in \mathbb{R}_{\geq 0}$.

The Dirichlet-boundary conditions at $s = 0$ correspond to a constant stream of methanol being pumped into the pipe. The Neumann-boundary conditions at $s = L$ are used to model the outward flux of methanol at the end of the pipe.

The strategy is to first discretize (47) in space by using the *Finite Volume Method* (FVM). This results in a system of nonlinear ODEs. The solution of this ODE is then approximated by using the so-called θ-method. Both methods are discussed in more detail below.

6.1 Discretization in space

In this section we explain how to discretize (47) in space by using the FVM. The main idea of the FVM is to approximate the averages of $(c_i)_{i=1}^2$ instead of the pointwise values. To this end, partition $[0,L]$ into $N \in \mathbb{N}$ subdomains of equal size and let $\left\{ s_j := \delta_s \left(j - \frac{1}{2}\right) : 1 \leq j \leq N \right\}$ denote the midpoints of these subdomains, where $\delta_s = \frac{L}{N}$ (see Figure 2). The objective is to approximate the averages

$$\bar{c}_{i,j}(t) := \frac{1}{\delta_s} \int_{s_{j-\frac{1}{2}}}^{s_{j+\frac{1}{2}}} c_i(t,s)\ \mathrm{d}s, \quad 1 \leq i \leq 2,\ 1 \leq j \leq N,$$

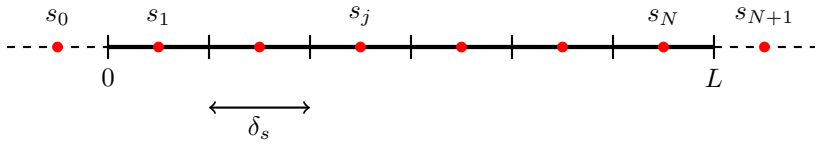


Figure 2: The interval $[0, L]$ is subdivided into N subdomains of size $\delta_s = \frac{L}{N}$. The boundaries of these subdomains are depicted as black vertical lines. The red dots correspond to the associated midpoints $(s_j)_{j=1}^N$. The additional grid-points $s_0 = -\frac{\delta_s}{2}$ and $s_{N+1} = L + \frac{\delta_s}{2}$ are needed to approximate $\dfrac{\partial c_i}{\partial s}(t, 0)$, and $\dfrac{\partial c_i}{\partial s}(t, L)$, respectively, with central differences.

on each subdomain at some prescribed points in time $\{0 = t_0 < t_1 \ldots < t_m = T\}$. Observe that if δ_s is sufficiently small and c_i is sufficiently regular, the averages $\bar{c}_{i,j}$ constitute accurate approximations of the point-values $(c_i(t, s_j))_{j=1}^N$.

Let $1 \leq i \leq 2$, $1 \leq j \leq N$ and take the average of (47) around s_j to obtain the following equation:

$$
\begin{aligned}
A_i \left[\frac{d\bar{c}_{i,j}}{dt} + \frac{(u_i c_i)\left(t, s_{j+\frac{1}{2}}\right) - (u_i c_i)\left(t, s_{j-\frac{1}{2}}\right)}{\delta_s} \right] \\
= \frac{A_i D_i}{\delta_s} \left[\frac{\partial c_i}{\partial s}\left(t, s_{j+\frac{1}{2}}\right) - \frac{\partial c_i}{\partial s}\left(t, s_{j-\frac{1}{2}}\right) \right] \\
+ \frac{(-1)^{i+1}}{\delta_s} \int_{s_{j-\frac{1}{2}}}^{s_{j+\frac{1}{2}}} \psi\left(t, s, c_1(t, s), c_2(t, s)\right) \, \mathrm{d}s. \qquad (48)
\end{aligned}
$$

We will now explain how to discretize the latter equation in space for fixed time t. In order for the following arguments to make sense, we will henceforth assume that δ_s is sufficiently small.

Discretization of the spatial derivatives To approximate the spatial derivatives in the righthand-side of (48) we would like to use the (second order) central difference approximation

$$
\frac{\partial c_i}{\partial s}\left(t, s_{j+\frac{1}{2}}\right) \approx \frac{c_i(t, s_{j+1}) - c_i(t, s_j)}{\delta_s} \approx \frac{\bar{c}_{i,j+1}(t) - \bar{c}_{i,j}(t)}{\delta_s} \qquad (49)
$$

for $0 \leq j \leq N$. However, the latter approximation only makes sense for $1 \leq j \leq N-1$, since $\bar{c}_{i,0}$ and $\bar{c}_{i,N+1}$ are undefined. In order to make sense of (49) for $j = 0$ and $j = N$ we formally introduce additional ghost nodes $s_0 = -\frac{\delta_s}{2}$ and $s_{N+1} = L + \frac{\delta_s}{2}$, see Figure 2.

The value of $\bar{c}_{i,0}$ is determined by taking an average over the two neighboring nodes and using the boundary condition at $s = 0$. In other words, since

$$\alpha_i = c_i(t, 0) \approx \frac{c_i(t, s_1) + c_i(t, s_0)}{2}$$

we set $\bar{c}_{i,0} := 2\alpha_i - \bar{c}_{i,1}$. Similarly, the value of $\bar{c}_{i,N+1}$ is determined by using the Neumann-boundary condition at $s = L$. That is, since

$$\beta_i = \frac{\partial c_i}{\partial s}(t, L) \approx \frac{c_i(t, s_{N+1}) - c_i(t, s_N)}{\delta_s},$$

we set $\bar{c}_{i,N+1} := \delta_s \beta_i + \bar{c}_{i,N}$. We can now use (49) to approximate the spatial derivatives for $0 \leq j \leq N$.

Approximation of the nonlinearity If the map $s \mapsto \psi(t, s, c_1(t, s), c_2(t, s))$ is sufficiently regular (at the very least L^1), then

$$\begin{aligned}
&\frac{1}{\delta_s} \int_{s_{j-\frac{1}{2}}}^{s_{j+\frac{1}{2}}} \psi(t, s, c_1(t, s), c_2(t, s)) \, \mathrm{d}s \\
&\quad \approx \psi(t, s_j, c_1(t, s_j), c_2(t, s_j)) \\
&\quad \approx \psi(t, s_j, \bar{c}_{1,j}(t), \bar{c}_{2,j}(t)),
\end{aligned} \tag{50}$$

for $0 \leq j \leq N$. Alternatively, one could use numerical quadrature to approximate the integral. The latter could potentially yield more accurate approximations provided $s \mapsto \psi(t, s, c_1(t, s), c_2(t, s))$ is sufficiently smooth.

Approximation of the advection term To approximate the advection term in (48) we simply approximate the average of c_i, as before, by using its values at the neighboring nodes:

$$\begin{aligned}
(u_i c_i)\left(t, s_{j+\frac{1}{2}}\right) &\approx u_i\left(t, s_{j+\frac{1}{2}}\right) \frac{c_i(t, s_{j+1}) + c_i(t, s_j)}{2} \\
&\approx u_i\left(t, s_{j+\frac{1}{2}}\right) \frac{\bar{c}_{i,j+1}(t) + \bar{c}_{i,j}(t)}{2},
\end{aligned} \tag{51}$$

for $0 \leq j \leq N$. Recall that we are assuming that u_i is *known*, in the sense that we can evaluate it at any $(t, s) \in [0, T] \times [0, L]$ on the computer.

6.2 Discretization in time

In this section we explain how the spatial discretizations from the previous section can be used to approximate solutions of (47). Substitution of (49), (50), and (51)

into (48) yields a system of nonlinear equations of the form

$$\begin{cases} \dfrac{d\bar{c}_i}{dt} = Q \cdot \bar{c}_i + (-1)^{i+1} \Psi\left(t, \bar{c}_1, \bar{c}_2\right), t \in [0,T], \\ \bar{c}_i\left(0\right) = \left[c_i^0\left(s_j\right)\right]_{j=0}^{N}, \end{cases} \qquad 1 \le i \le 2, \tag{52}$$

where $\bar{c}_i := \begin{bmatrix} \bar{c}_{i,0} & \dots & \bar{c}_{i,N} \end{bmatrix}^T$, Q is the $(N+1) \times (N+1)$ matrix which encodes the linear part of the equations (i.e. it is the discretization of the advection and diffusion term), and $\Psi : [0,L] \times \mathbb{R}^{N+1} \times \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ corresponds to the nonlinear part associated to (50).

For notational convenience, we introduce the map $F : [0,T] \times \mathbb{R}^{2(N+1)} \to \mathbb{R}^{2(N+1)}$ defined by

$$F\left(t, \bar{c}\right) := \begin{bmatrix} Q \cdot \bar{c}_1 + \Psi\left(t, \bar{c}_1, \bar{c}_2\right) \\ Q \cdot \bar{c}_2 - \Psi\left(t, \bar{c}_1, \bar{c}_2\right) \end{bmatrix},$$

where $\bar{c} := \begin{bmatrix} \bar{c}_1 \\ \bar{c}_2 \end{bmatrix}$. Then (52) can be rewritten as

$$\begin{cases} \dfrac{d\bar{c}}{dt} = F\left(t, \bar{c}\right), \quad t \in [0,T], \\ \bar{c}(0) = c^0, \end{cases} \tag{53}$$

where

$$c^0 = \begin{bmatrix} c_1^0\left(s_0\right) & \dots & c_1^0\left(s_N\right) & c_2^0\left(s_0\right) & \dots & c_2^0\left(s_N\right) \end{bmatrix}^T.$$

Finally, the solution of (53) is approximated at the times $\left(t_k\right)_{k=0}^{m}$ by using the θ-method:

$$\bar{c}\left(t_{k+1}\right) = \bar{c}\left(t_k\right) + \delta_k \left[\theta F\left(t_k, \bar{c}\left(t_k\right)\right) + \left(1-\theta\right) F\left(t_{k+1}, \bar{c}\left(t_{k+1}\right)\right)\right],$$

where $0 \le k \le m-1$, $\delta_k = t_{k+1} - t_k$, and $\theta \in [0,1]$ is a fixed parameter.

7 Numerical results

In this section we investigate the behavior of the 1-d model developed in Section 5 with the aid of numerical simulations. There are many interesting aspects of the proposed model to investigate; both from a numerical point of view and from a modeling point of view. Here we restrict our attention to studying the influence of the coupling term ψ. More precisely, we investigate the dependence of the model on the height $h \in (0, D)$ of the layer interface for two different scenarios. To accomplish this, we fix all other parameters throughout this section. We note that there are many more interesting parameter dependencies to investigate and leave this as a topic of future research.

Physically relevant parameters We set the length of the pipe and the radius of its cross-sections equal to $L = 2$ and $R = 1$, respectively. The constant volume fractions of methanol pumped into the upper and lower regions of the pipe are set to $\alpha_1 = 1$ and $\alpha_2 = \frac{1}{10}$, respectively. For the sake of simplicity, we choose the initial distribution of methanol in both the upper and lower part of the pipe to be constant throughout the pipe, i.e., $c_1^0, c_2^0 : [0, L] \to \mathbb{R}$ are constant. Therefore, due to the Dirichlet boundary conditions at the left-end of the pipe, we must necessarily set $c_i^0 \equiv \alpha_i$. We impose a Neumann boundary condition at the right end of the pipe by setting $\beta_1 = \beta_2 = 0$. Finally, we choose the horizontal and vertical diffusion coefficients to be the same in each coordinate direction: $D_w = D_1 = D_2 = 10^{-2}$.

Computational parameters We use the same computational parameters in all numerical simulations (see Section 6 for the implementation details). The parameters associated to the discretization sizes in time and space are set to $\delta_k \equiv \delta = 10^{-3}$ and $N = 200$, respectively. The latter corresponds to a uniform spatial discretization of size $\frac{L}{N}$. Furthermore, we use $\theta = 0$ to perform the time integration, which corresponds to a backward Euler scheme. Finally, in each numerical simulation, we set the integration time to $T = 10$. This particular choice was motivated by the observation that in all numerical experiments the solutions approached a steady state within this time frame.

We consider the following two scenarios:

(i) The velocity in the upper part of the pipe is smaller than in the lower part: $u_1 = \frac{1}{10}$, $u_2 = 1$.

(ii) The velocity in the upper part of the pipe is larger than in the lower part: $u_1 = 1$, $u_2 = \frac{1}{10}$.

For each scenario, we have performed numerical simulations for different choices of the height of the layer interface; we have considered

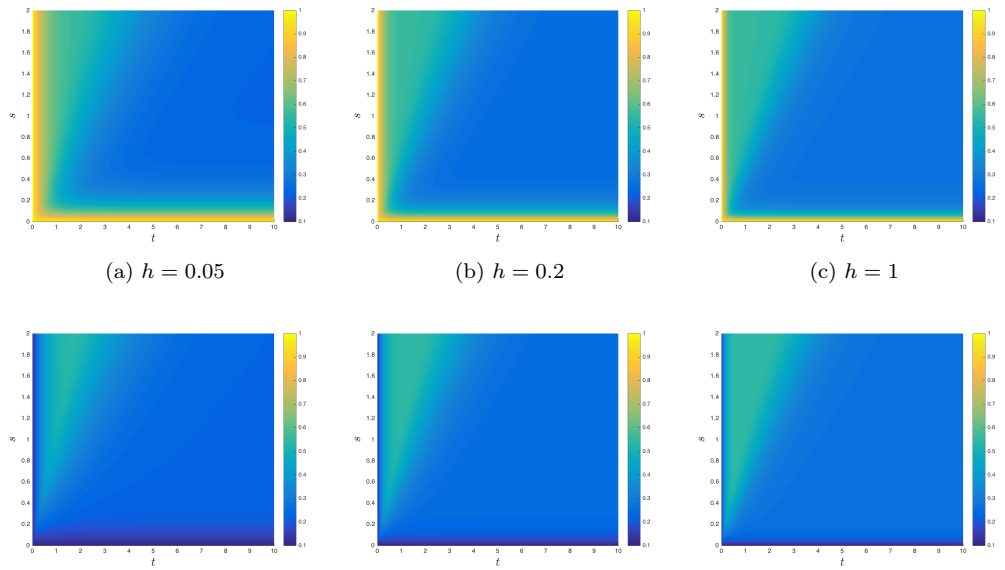
$$h \in \Delta := \left\{ h_j := \frac{5 + j}{100} : 0 \leq j \leq 95 \right\}.$$

7.1 Case (i)

We start with the case in which $u_1 < u_2$. In all scenarios, i.e., for all $h \in \Delta$, the volume concentrations of methanol in the upper and lower part of the pipe converged to a steady state. We have shown the typical behavior of c_1 and c_2 in Figure 3 for three different choices of h. The results show that the volume fractions of methanol in the upper part of the pipe evolved into a decreasing function of s as time progressed. In particular, the concentration profiles transitioned more quickly into these decreasing "states" as the height of the layer interface increased. Furthermore, the time in which c_1 approached a steady state decreased as h increased.

The volume fraction of methanol in the lower part of the pipe evolved into an increasing function of s as time progressed. For relatively small t, the fractions were relatively high "near" the right-end of the pipe (the part of the pipe which corresponds to the green regions in Figures 3d, 3e and 3f). Furthermore, the fractions in these regions decreased as time progressed. In particular, the rate at which this decrease occurred (with respect to time) slightly decreased as h increased.

We have depicted the steady states to which c_1 and c_2 converged in Figure 4 for $h \in \Delta$. In all scenarios the steady states were constant in a relatively large region of the pipe. Furthermore, the size of these regions increased as h increased. Moreover, the "final" volume fractions of methanol in these parts of the pipe were (approximately) the same in both the upper and lower region and increased as h increased from $h = 0.05$ to $h = 1$.



(a) $h = 0.05$ (b) $h = 0.2$ (c) $h = 1$

(d) $h = 0.05$ (e) $h = 0.2$ (f) $h = 1$

Figure 3: Case (i): (a), (b), (c) The values of c_1 on $[0, T] \times [0, L]$ for various choices of h. (d), (e), (f) The values of c_2 on $[0, T] \times [0, L]$ for various choices of h.

To quantify the assertion that c_1 and c_2 approached a steady state more quickly as h increased, recall that we approximated the volume fractions at the following discrete moments in time: $t \in \mathcal{T} := \left\{k\delta : 0 \leq k \leq 10^4\right\}$, where $\delta = 10^{-3}$. Let $\varepsilon > 0$ be a given tolerance and set

$$T_i(h) := \min\left\{t \in \mathcal{T} : \left\|c_i\left(t, \cdot\right) - \hat{c}_i\right\|_{L^2([0,L])} < \varepsilon\right\}, \quad i \in \{1, 2\},\ h \in \Delta,$$

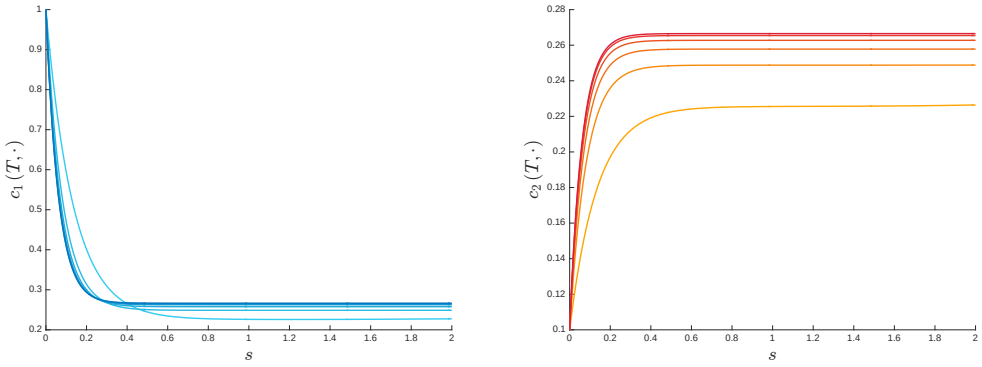


Figure 4: Case (i): numerical approximations of the steady states to which c_1 and c_2 converged for various values of $h \in \Delta$. (a) The steady states associated to the upper part of the pipe. For h close to $h_0 = 0.05$, we have colored the corresponding steady states in light blue. As h increased to 1, we have used increasingly darker shades of blue. (b) The steady states associated to the lower part of the pipe. For h close to $h_0 = 0.05$, we have colored the corresponding steady states in orange. As h increased to 1, the color of the steady states transitioned from orange to red.

where $\hat{c}_i : [0, L] \to \mathbb{R}$ is a numerical approximation of the steady state to which c_i converged. In practice, we set $\hat{c}_i = c_i(T, \cdot)$. We remark that it would be more accurate to determine an approximation $\hat{c}_i$ by directly solving the steady state equation (an ODE). In any case, if $\hat{c}_i$ is a sufficiently accurate approximation of the steady state in question (which we are assuming) and $\varepsilon > 0$ is sufficiently small (but not too small), then T_i can be used to substantiate the above assertion. More specifically, if $T_i(h_1) < T_i(h_2)$, then we have numerical evidence for the claim that the solution associated to h_1 approached a steady state more quickly than the solution associated to h_2.

We have depicted the points $\{(h, T_i(h)) : h \in \Delta\}$ on the graph of T_i for $\varepsilon = 10^{-5}$ and $i \in \{1, 2\}$ in Figure 5. The results support our claim and show that c_1 and c_2 approached a steady state more quickly as h increased.

7.2 Case (ii)

Finally, we consider the case in which $u_1 > u_2$. The typical behavior of the fractions is shown in Figure 6. In each scenario, the observed behavior was similar (but not entirely the same) as in the previous case. In particular, c_1 and c_2 both approached a steady state as time progressed. The steady state associated to the upper part of the pipe was decreasing in s and the one associated to the lower part was increasing. Furthermore, both steady states were constant in a relatively large part of the pipe.

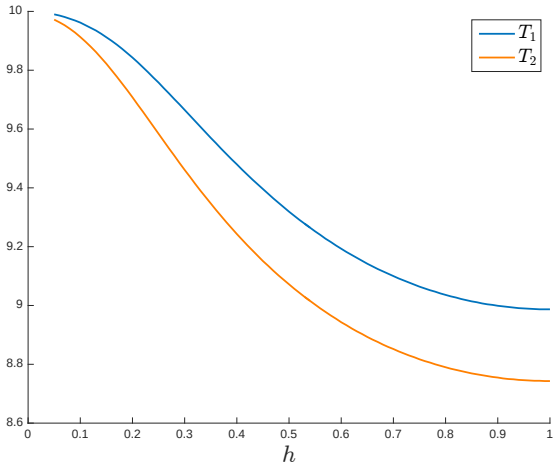


Figure 5: The dependence of T_i on the height of the interface for $i \in \{1, 2\}$. The depicted curves were obtained by sampling T_i on Δ.

A key difference in this case is that the volume fraction of methanol "near" the right-end of the pipe (the "upper" green regions in Figure 6) increased as time progressed, whereas in the previous case it decreased. Furthermore, on average, the methanol fraction throughout the pipe was higher than in the previous case. Another noticeable difference is that the values of the steady state solutions decreased as h increased in those regions of the pipe where the "final" fractions were constant, see Figure 7.

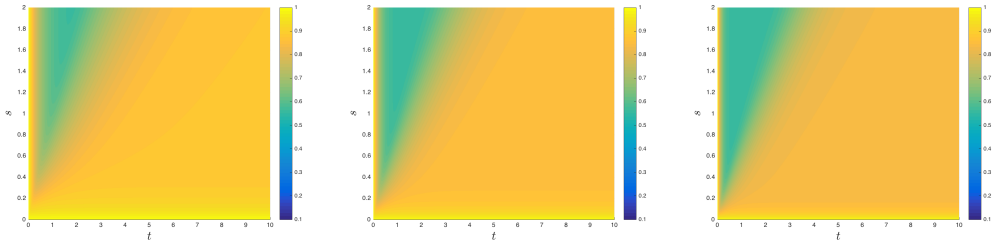
Acknowledgments
We thank Patricio Rosen Esquivel (Shell) and Benjamin Sanderse (CWI and Shell) for useful discussions and suggestions, as well as for formulating this interesting SWI problem.

References

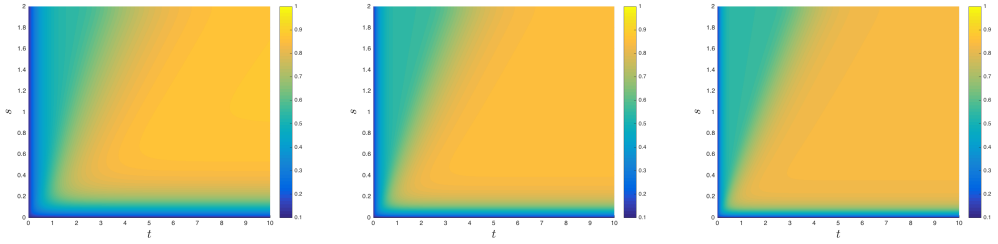
A. Chorin and J. Marsden. *A Mathematical Introduction to Fluid Mechanics*. Springer, 1979.

K. Oldham, J. Myland, and J. Spanier. *An atlas of functions, 2nd edition*. Springer, 2009.

N. Temme. *Special Functions: An Introduction to the Classical Functions of Mathematical Physics*. John Wiley & Sons, 1996.



(a) $h = 0.05$ (b) $h = 0.2$ (c) $h = 1$



(d) $h = 0.05$ (e) $h = 0.2$ (f) $h = 1$

Figure 6: Case (ii) (a), (b), (c) The values of c_1 on $[0, T] \times [0, L]$ for various choices of h. (d), (e), (f) The values of c_2 on $[0, T] \times [0, L]$ for various choices of h.

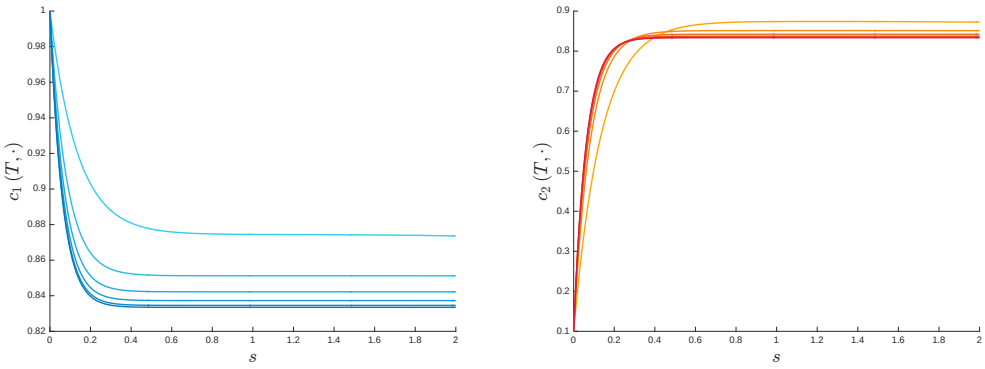


Figure 7: Case (ii): numerical approximations of the steady states to which c_1 and c_2 converged for various values of $h \in \Delta$. (a) The steady states associated to the upper part of the pipe. For h close to $h_0 = 0.05$, we have colored the corresponding steady states in light blue. As h increased to 1, we have used increasingly darker shades of blue. (b) The steady states associated to the lower part of the pipe. For h close to $h_0 = 0.05$, we have colored the corresponding steady states in orange. As h increased to 1, the color of the steady states transitioned from orange to red.

Spillback Effects from Traffic Accidents

Jacobien Carstens, Bart Litjens, Verena Schamboeck, Tineke School, Veerle Timmermans, Jacob Turner

Contents

1 Introduction

Understanding how traffic incidents affect the flow of traffic is a problem of great importance. It is of interest to both the general public as well as to many private companies that rely heavily on transportation. Disturbances to traffic come in an enormous number of varieties and are very sensitive to a large number of different situational and environmental factors. A complete understanding is therefore far from feasible at this present time as many individual aspects are still not well described.

In this note, we focus on a very specific problem inside of the general theory. Given a traffic incident that inhibits vehicles to travel freely, as in normal conditions, how does the gradual build up of slow moving vehicles congesting the road behave? There are several basic questions to investigate in relation to this situation:

(a) How quickly does traffic congestion build upstream from the incident given the nature of the disruption?

(b) If this back up progresses all the way back to an intersection, will it cause back up on other roads? (When traffic backs up onto another road, it is called *spillback*)

(c) If this is the case, can the roads that will also suffer significant congestion be predicted?

(d) How many vehicles can be expected to avoid the road on which the disturbance has occurred and which alternative roads will thus see an inflow of traffic as rerouting begins to occur?

(e) How will an individual driver make the decision to wait in traffic versus finding an alternative route?

As evidenced by the large number of complex questions above, even this simple situation is unlikely to have a single consistent pattern. We now explicitly draw attention to two distinct effects that occur when a traffic disturbance is present, as can already be seen in the questions above. Namely:

Effect 1. Those drivers who will remain in the congested area until they can proceed along their originally intended route, contributing to a back up on the affected motorway.

Effect 2. Those drivers who will seek to avoid the affected road altogether and deviate from the initial route onto different routes.

The latter phenomenon is very dynamic and difficult to predict. The former effect can be studied easily with some simplifying assumptions. This effect will be present if there is essentially no choice for the drivers in their route given their origin and destination. For example, one might expect a long stretch of road connecting different cities to be more prone to backup. If such a road becomes affected by a traffic incident, it is not uncommon that any other route linking the two cities will be a significant deviation in time and distance, likely involving travel to a completely different city. This is because of the relative sparsity of roads between cities in contrast to roads within a city area that makes long traffic jams more likely.

However, such intuition may not be reflected in reality and the aforementioned roads need not be the only ones on which drivers will feel that any alternative route would be such a deviation that the only realistic choice is to wait in traffic. Roads with this property will be called *vulnerable*. Leaving out circumstantial causes that may affect the drivers choice, it should be clear that *vulnerability* of a road is a property of the road network itself.

The problem then becomes how to identify such roads and to formulate some measure of road vulnerability. If an incident occurs on a very vulnerable road, then we should expect Effect 2 to be negligible. In this case, understanding how traffic behaves becomes less complex. The follow-up problem is then to describe how traffic behaves in this simpler type of scenario.

This paper is organized as follows. In Section 2 we address the notion of link vulnerability. In order to do so, we first describe the model of the road network that we use in Section 2.1. Then, in Section 2.2, we define several vulnerability measures for roads in the network. Some definitions of road vulnerability have been considered

before in Freeman et al. (1991); Jenelius (2009, 2010); Knoop et al. (2008). Their primary focus was on roads on which an incident causes the maximum disruption of traffic in the whole network. Our notion of vulnerability however is orthogonal to the amount of traffic flow on the road. It captures how much choice a driver taking that road has in choosing an alternative route.

Subsequently, in Section 3, we use vulnerability to make some actual predictions. In Section 3.1 vulnerability as well as some additional time-dependent parameters are used to estimate the rate of people rerouting in case of an event on a fixed road. Lastly, in Section 3.2 local order-destination information is used to predict spillback on highly vulnerable roads.

2 Link vulnerability

2.1 The model

We consider the Dutch road network to be a weighted undirected graph $G = (V, E)$, where each edge (or link) represents a part of the motorway and each vertex (or node) represents a junction of motorways. Only the motorways, which in the Netherlands are indicated by the letter A followed by a number, and a few provincial roads, which are important for the global structure of the road network, are taken into account. In this paper, we will refer to the chosen network as the 'motorway network'. A more comprehensive model would also include all provincial and city roads. We assume that at each node one has the possibility to move to any motorway incident with that node. For $e \in E$, let $\ell(e)$ denote the time it takes to travel from one endpoint of e to the other. In this paper these times are computed using Google Maps at a specific time (2pm on a weekday without traffic incidents). A more accurate weight is obtained by averaging over different times on several days. The weighted graph is shown in Figure 1.

The reason for restricting the network to motorways and a few important roads is that we have access to detailed data on the traffic on these roads. There are thousands of sensors throughout this part of the Dutch road network, recording the number of cars passing and their velocity every minute of the day. In Section 3.2 we use this data to analyse how quickly traffic backs up after an incident occurs.

For any path $P \subseteq E$, let $\ell(P)$ be *length* of P, i.e., $\ell(P) = \sum_{e \in P} \ell(e)$. Whenever we speak of a path, it is assumed to be simple, i.e, without repeated edges or vertices. For $i, j \in V$, we define $P(i, j)$ to be the set of paths connecting the vertices i and j. Then we define the length $c(i, j)$ of a *shortest path* between i and j as

$$c(i, j) := \min\{\ell(P) \mid P \in P(i, j)\}.$$

As we are also interested in alternative routes, for any $e \in E$ we furthermore define $c(i, j, e)$ to be the length of the shortest path from i to j in the graph G when the edge e is missing,

$$c(i, j, e) := \min\{\ell(P) \mid P \text{ a path from } i \text{ to } j \text{ in the graph obtained by removing } e\}.$$

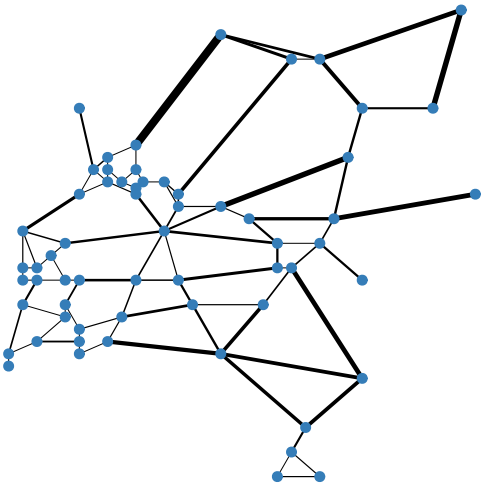


Figure 1: The Dutch motorway network.

Let $i, j \in V$ and $k \in \mathbb{R}_{\geqslant 1}$. We define $P(i, j, k)$ as the set of all paths from i to j whose length is at most k times the length of the shortest path between i and j,

$$P(i, j, k) := \{P \in P(i, j) \mid l(P) \leqslant k \cdot c(i, j)\}. \tag{1}$$

For $e \in E$, we are also interested in the subset of $P(i, j, k)$ consisting of paths that contain e,

$$P(i, j, k, e) := \{P \in P(i, j, k) \mid e \in P\}. \tag{2}$$

We define the set of order-destination pairs that suffer from the fact that the link e becomes inaccessible,

$$S(e) := \{(i, j) \in V^2 \mid \text{ there exists a shortest path from } i \text{ to } j \text{ that contains } e\}. \tag{3}$$

Lastly, whenever we will speak of *free flow* on an edge e, we mean that all lanes at e are open and that the average speed of the cars on e is at least 10 km/hr.

2.2 Vulnerability measures

In this section a vulnerability measure is assigned to each link that indicates whether or not there are good alternative routes available if a link becomes inaccessible. We define these measures to satisfy the following properties:

1. The vulnerability should be a number between 0 and 1. A rate of 0 implies that many alternative routes are available. A rate of 1 implies that no alternative roads are available.

2. The vulnerability can be computed using the network topology, taking into account the travel time on each link assuming free flow. Hence, this rate does not depend on the current traffic situation, and only needs to be computed once using the graph G defined in Section 2.1.

We will define two different vulnerability measures that all satisfy the above properties. For definitions and notation, see Section 2.1. The first measure generalizes the notion of *(edge-)betweenness centrality*, a network theoretic concept that has been formally defined first by Freeman (1977).

Definition 2.1 (Vulnerability measure 1, based on edge-betweenness centrality)**.** Let $e \in E$ and $k \in \mathbb{R}_{\geqslant 1}$. Then we define

$$V_1(k, e) := \frac{1}{|V|(|V|-1)} \sum_{i \neq j \in V} \frac{|P(i, j, k, e)|}{|P(i, j, k)|}, \tag{4}$$

where the sum runs over all distinct vertices i and j.

The factor in front of the sum normalizes the sum of the ratios to ensure the measure V_1 is a number between zero and one. In practice, only the cases $1 \leqslant k \leqslant 2$ are interesting, as we do not expect drivers to reroute if the alternative route would take more than twice as long as usual.

The second vulnerability measure that we define considers drivers that suffer from the closing of link e. We compute the average fraction of time that is lost by closing link e. Note that when e is not on any shortest path, we define $V_2(e) = 0$, as no one suffers from deleting this link. On the other hand, if deleting e would disconnect the network (i.e, if e is a *bridge*), we set $V_2(e) = 1$.

Definition 2.2 (Vulnerability measure 2, based on edge deletion I)**.** Let $e \in E$. Then we define

$$V_2(e) = \begin{cases} 0 & \text{if } e \text{ is not in any shortest path,} \\ 1 & \text{if } e \text{ is a bridge,} \\ \frac{1}{|S(e)|} \sum_{(i,j) \in S(e)} \frac{c(i,j,e) - c(i,j)}{c(i,j,e)} & \text{otherwise.} \end{cases} \tag{5}$$

Note that $V_2(e)$ is well-defined as both $|S(e)|$ and $c(i, j, e)$ are nonzero if e is on a shortest path and not a bridge. The two different vulnerability rates are depicted in Figure 2.

Notice that the vulnerability measures V_1 and V_2 take on completely different values on bridges that are on the fringe of the network. Our current implementation of V_1 is too slow to compute the vulnerability rates of the complete motorway network of the Netherlands. Figure 3 depicts the vulnerability rate V_2 for this network.

One could improve these measures by considering weighted sums, where the weights are defined as the number of times an order-destination pair (i, j) is traveled. One difficulty there is that these order-destination pairs are hard to determine from data. Locally, however, this can be done and this method is exploited in Section 3.2. An easier approach is to define the weights proportionally to the travel time, assuming more people drive shorter routes.

Edge	Unweighted V_1	Unweighted V_2	Weighted V_1	Weighted V_2
$\{1,2\}$	0.104	0.271	0.087	0.043
$\{1,3\}$	0.104	0.500	0.101	0.645
$\{2,3\}$	0.104	0.271	0.106	0.353
$\{2,4\}$	0.250	1.000	0.250	1.000
$\{4,5\}$	0.180	0.327	0.168	0.258
$\{4,7\}$	0.180	0.252	0.164	0.347
$\{5,6\}$	0.180	0.194	0.178	0.218
$\{6,7\}$	0.180	0.172	0.178	0.323
$\{6,8\}$	0.111	1.000	0.111	1.000
$\{7,9\}$	0.111	1.000	0.111	1.000

Table 1: The values of the vulnerability measures V_1 with $k = 2$ and V_2 for the network in Figure 2. The unweighted values are computed for the network with all edge weights equal to 1, the weighted values are computed using the edge weights as shown in Figure 2.

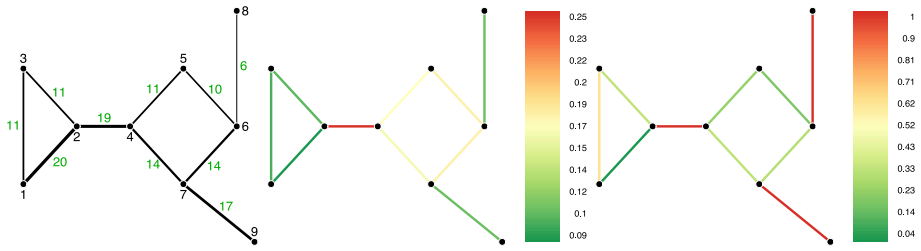


Figure 2: From left to right: weighted example network where edge thickness corresponds to edge weights, the edges of the network are colour coded by the vulnerability rate V_1 for $k = 2$, the edges of the network are colour coded by the vulnerability rate V_2.

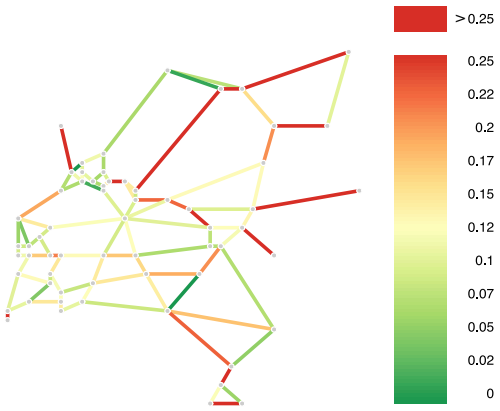


Figure 3: The vulnerability rate V_2 of the motorways network of the Netherlands.

3 Vulnerability in practice

3.1 Predicting rerouting

In this section we estimate the amount of people that will take an alternative route, and the number of commuters that will stick with their initial route in case of a traffic incident. In order to do so, we need more than only the vulnerability measure. The percentage of drivers deviating from their original routes is also affected by the current flow in comparison with the capacity of the link, and the size of the accident (in terms of the number of lanes that are closed).

Let $e \in E$. Then by $f_t(e)$ we denote the number of cars on e that at time t are in free flow. We call $f_t(e)$ the *flow of* e at time t. The maximum number of cars in free flow on e is called the *capacity of* e and is denoted by $c(e)$. By $\text{lanes}(e)$ we denote the number of lanes on e. We write $\text{open}_t(e)$ for the number of open lanes on e at time t.

Using the notions defined above we will now describe a function F that is an estimate of the percentage of people on a road e that will reroute in case at time t an incident happens and causes $\text{lanes}(e) - \text{open}_t(e)$ lanes to close, given a flow that equals $f_t(e)$ at that time. The function F furthermore depends on the capacity and the vulnerability measure. Fix a $k \in \mathbb{R}_{\geqslant 1}$ and set $V_1(e) := V_1(k, e)$ for $e \in E$. Given $e \in E$ and a time t, we first define the function

$$h_t(e) := \begin{cases} 1 \text{ if } f_t(e) \leqslant c(e), \\ 0 \text{ else.} \end{cases}$$

Then we define F as follows

$$F_t(e, i) := \frac{\alpha_1 \cdot V_i(e) + \alpha_2 \cdot (\text{open}_t(e)/\text{lanes}(e)) + \alpha_3 \cdot h_t(e)}{\alpha_1 + \alpha_2 + \alpha_3}, \tag{6}$$

where $i \in \{1, 2, 3\}$ and where $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}_{>0}$. The parameters α_1, α_2 and α_3, which at present do not depend on time, correlate the variables involved in the function F and will need to be determined from the actual data. The function F then returns the rate of people that will stay on their original route. In Section 4 we discuss ways of refining equation (6).

3.2 Predicting spillback

In this section, we assume that we have correctly identified a link as highly vulnerable. How do we expect back-up and spillback to happen? This is the question we now seek to address.

The first problem is to identify how quickly traffic will back up once an incident happens. The detection of a disturbance can occur within a matter of minutes from the induction loop detectors placed on the motorway. A sudden and significant drop in average speed is sufficient for this purpose.

If one plots a graph with axes corresponding to position and time, and each point color-coded based on the average speed of traffic as given by the induction loop

detectors, it is a well known empirical phenomenon that traffic incidents cause a parallelogram shape to appear in the colors of low speeds. This tells us that backup on a road accumulates linearly. As such, the rate of backup can be quickly determined using the first few minutes of incoming data after the accident has been detected.

An example of such a parallelogram is shown in Figure 4. The slope of side of the parallelogram running roughly in the vertical direction indicates the rate of backup of the traffic.

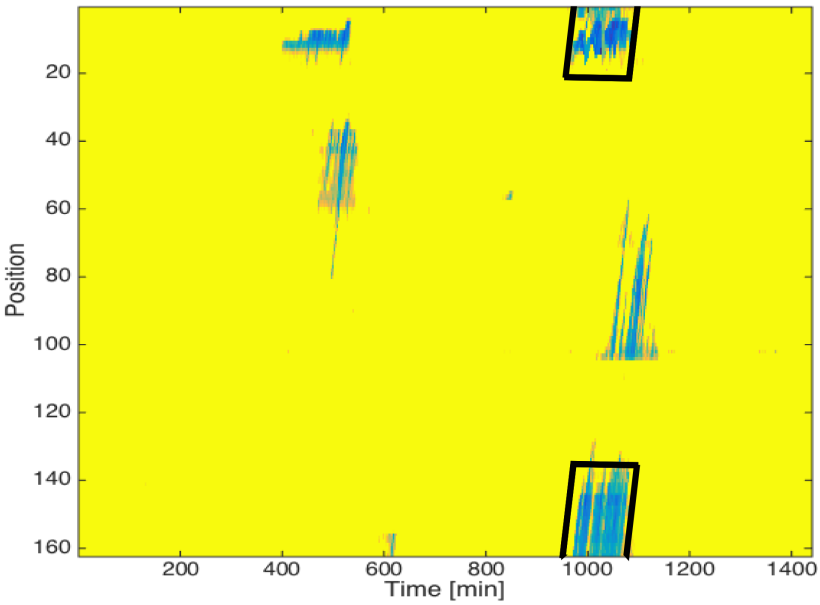


Figure 4: In this picture, the color blue represents a reduced speed. The parallelogram corresponding to the traffic congestion resulting from an incident has been framed. In this particular example, the road in question was a ring and so the graph should be viewed as on a cylinder, hence why the parallelogram is split in the picture.

Of course, it is an altogether different question to try and predict how long such an incident will take place and whether or not backup will reach an intersection. On this point, we make no comment. Instead, let us focus on what we expect to happen in the event that the backup does reach the nearest intersection. To make this prediction, we will define the *local traffic matrix* for an intersection which will depend on empirical data concerning the typical traffic behavior. In fact, one should have many matrices associated to an intersection, one for each time of day/week/year as these conditions can greatly affect what would be considered "normal traffic".

Now let vertex v denote the intersection in question. Let I be the set of those

edges incoming to v and O those that are outgoing. Then the local traffic matrix, $M(v)$, will be a matrix with rows indexed by I and columns indexed by O. For $i \in I$ and $o \in O$, the entry in $M(v)_{i,o}$ will be the average percentage of traffic that turns from road i to road o at v. Such a matrix should be computable given sufficient data and if there are sensors placed on entry and exit ramps between motorways.

Given that traffic on a vulnerable road o backs up to the intersection v, we would expect those roads i such that $M(v)_{i,o}$ is large to also experience congestion. Indeed, one would expext that if $x\%$ of traffic on i turns onto o at v, it will experience $x\%$ of the rate of backup that road o is experiencing.

If the traffic spills back onto another vulnerable road, then after enough time it may reach another intersection and the same method of prediction is possible. However, it seems naive to expect this spillback to continue indefinitely given enough time. One would expect that eventually people would begin canceling trips altogether as news of such a major accident spread. Additionally, if a truly large amount of spillback is occuring, officials may close the road altogether, again forcing people to cancel their trip. These effects would mitigate spillback onto more motorways even if the motorway on which the incident occured was very vulnerable.

4 Discussion

In this section we address the assumptions that were made throughout the paper and discuss ways to verify them from the data. Furthermore, we investigate how to improve the vulnerability measures defined in Section 2.2 and the function (6) defined in Section 3.1.

4.1 Improvements on the vulnerability measures

In Section 2 we defined the graph that represents the Dutch road network. As mentioned there, all motorways are included but the provincial and city roads have not been included. This results in the fact that for instance the motorway $A2$ between Weert and Maastricht (in the southern part of the Netherlands) is a bridge, and therefore maximally vulnerable in our model. However, in practice there is a very good alternative for that piece of road in the event of a traffic accident, namely the $N276$ (this is a provincial road).

In order to account for these kind of alternatives, we strongly recommend to include the provincial and city roads in further research that uses our model. From a graph theoretic point of view, this would make the graph far more complex. The vulnerability measure 2, as defined in equation (5), can still be computed efficiently, as there exists a fast algorithm to compute shortest paths in graphs. The time needed to compute vulnerability measure 1 (see equation (4)), would increase exponentially. However, the vulnerability measure only need to be computed once (for every edge). Therefore, we consider it still worthwile to explore this extended graph.

With respect to the function $F_t(e, i)$, defined in equation (6), computationally

nothing changes. The only real-time data it depends on is the number of open lanes and the current flow, both of which can be computed quickly from the data. There are however some refinements that we want to address. For instance, note that in computing $F_t(e, i)$, preferably one would also take the time after the accident in consideration. Drivers will only reroute if they are aware of the accident and if there is still time to take the alternative route. A more sophisticated approach would be to consider a dynamical system in which the value of the function F at a specific time depends on the value F at an earlier time and, in turn, serves as input for computations of F at later times.

Another way to improve upon the function $F_t(e, i)$ is to investigate quadratic or higher-order dependencies. In the current formula, the function depends on $V_i(e)$, $\text{open}_t(e)$ and $h_t(e)$ linearly. This may be a good first-order approximation but higher-order terms certainly will make the function more accurate.

Another potential issue is the disparity between objective understanding of the Dutch road network and the perception of drivers. While the measure of vulnerability should be solely a network measure, its definition fundamentally hinges on the notion of driver choice. As such, there are subjective factors at play and the network that should be measured should be, in some sense, the network as people imagine it, as opposed to how it actually is. If this difference is great, then it seems unlikely to craft a measure simply from geospatial information and a closer investigation of driver behavior will have to be taken into account.

4.2 Implicit assumptions and testable hypthoses

The analysis in Section 3.2 rested on some silent assumptions that should not be simply taken as axioms. We outline these assumptions here as testable hypotheses, to be confirmed or denied using available empiral data.

(a) We have tried to divorce our notion of vulnerability from the amount of traffic flow typical on a given motorway. While it seems clear that the ability to reroute is indeed independent of such considerations, the perceived ability to reroute may not be. It may be that roads most susceptible to back up are very short stretches of road that are very heavily traveled. Even though there may be many alternative routes, the shortness of the stretch of road could make people believe that they can push through in a short amount of time.

(b) We expect that incidents are 1) more common at intersections and 2) the accidents occuring near intersections will cause the greatest amount of spillback because of their proximity to other roads in the network. If this is true, then instead of focusing on the vulnerability of links, it may be more prudent to consider the vulnerability of intersections.

(c) If spillback occurs, how frequently does it occur across two or more upstream intersections? Our hypothesis is that this is an incredibly rare occurence and that after traffic has spilled back across one intersection, the knowledge and increased

visibility of the accident will cause significant rerouting, mitigating the upstream backup. If this is the case, it makes predicting spillback much simpler, although the question of rerouting related congestion remains complicated.

(d) Can the severity of spillback be dichotomized according to intercity versus intracity incidents? Or do highly vulnerable roads exist in both situations?

References

L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

L. C. Freeman, S. P. Borgatti, and D. R. White. Centrality in valued graphs: A measure of betweenness based on network flow. *Social networks*, 13(2):141–154, 1991.

E. Jenelius. Network structure and travel patterns: explaining the geographical disparities of road network vulnerability. *Journal of Transport Geography*, 17(3):234–244, 2009.

E. Jenelius. Redundancy importance: Links as rerouting alternatives during road network disruptions. *Procedia Engineering*, 3:129–137, 2010.

V. Knoop, H. van Zuylen, and S. Hoogendoorn. The influence of spillback modelling when assessing consequences of blockings in a road network. *EJTIR*, 4(8):287–300, 2008.

Acknowledgments

The main sponsor for the SWI 2017 was The Netherlands Organisation for Scientific Research (NWO). We gratefully acknowledge their generous support for this event, as well as their continued support for these events in the Netherlands. Further financial support was provided by the companies who submitted the problems (CPB, De Bank, Equalis, Marin, Shell and TNO). Our institutions (the University of Amsterdam and Centrum Wiskunde & Informatica (CWI), the Dutch Centre for Mathematics and Computer Science) provided organizational support. In particular we would like to thank Marieke Kranenburg and Monique Onderwater of the University of Amsterdam and Peter Hildering and Danielle Kollerie of CWI. Last but not least we would like to thank all participants for creating an inspiring week of industrial mathematics.

The organisers of SWI 2017

Netherlands Organisation
for Scientific Research

UNIVERSITY
OF AMSTERDAM

Centrum Wiskunde & Informatica