Proceedings of the 90th European Study Group
Mathematics with Industry

SWI 2013

Leiden, January 28 – February 1, 2013

Editors:
Markus Heydenreich,
Sander Hille,
Vivi Rottschäfer,
Flora Spieksma,
Evgeny Verbitskiy

Contents

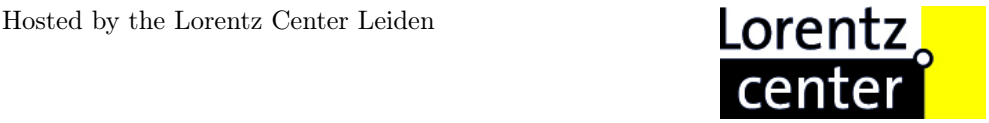
Preface

These are the proceedings of the 90th Study Group Mathematics with Industry (*Studiegroep Wiskunde met de Industrie*). It was held at the Lorentz Center in Leiden from January 28 to February 1, 2013.

The proceedings are provided in two different formats. In the current volume, the participants provide their own rendering of the week: they present the problems, the approach and the result, aimed at a scientific audience. A companion volume provides a different view on the week: science journalists Ionica Smeets and Bennie Mols describe the work for a general audience. The companion volume is written in Dutch.

The organizers of SWI 2013
M. Heydenreich, S. Hille, V. Rottschäfer, F. Spieksma, E. Verbitskiy

Hosted by the Lorentz Center Leiden



Oxygen transport and consumption in germinating seeds

Neil Budko (Delft University of Technology)
Alessandro Corbetta (Eindhoven University of Technology)
Bert van Duijn (Fytagoras) Sander Hille (Leiden University)*
Oleh Krehel (Eindhoven University of Technology)
Vivi Rottschäfer (Leiden University)
Linda Wiegman (Delft University of Technology)
Delyan Zhelyazov (Centrum voor Wiskunde en Informatica)

Abstract

Three mathematical models were formulated to describe the oxygen consumption of seeds during germination. These models were fitted to measurement data of oxygen consumption curves for individual germinating seeds of Savoy cabbage, barley and sugar beet provided by Fytagoras. The first model builds on a logistic growth model for the increasing population of mitochondria in the embryo during growth. The other two take the anatomy and physiological properties of the seed into account. One describes the oxygen uptake during the germination phase only. An extension of this model is capable of fitting the complete oxygen consumption curve, including the initial 'repair' phase in which the embryonic cells recover from their dormant state before extensive cell division and growth commences.

Keywords: Modelling, seed germination, cellular respiration, oxygen transport

1 Introduction

At the 90th Study Group Mathematics with Industry (SWI) held at Leiden University from 28 January to 1 February 2013 one of the questions was formulated by the company Fytagoras. Fytagoras is a company that is oriented on science with much expertise in the fields of sensor technology, seed technology and plant breeding. The question concerned the uptake and consumption of oxygen by germinating seeds. Our study considers the seeds of three particular species among the $\sim 250,000$ species of

*Corresponding author

flowering plants: barley (*Hordeum vulgare* L.), sugar beet (*Beta vulgaris*) and Savoy cabbage (*Brassica olerecea* var. *sabauda* L.).

Describing the dynamics of chemical compounds dissolved in liquid solvents (e.g. water) one typically uses the concept of concentration to describe the state of the system. Dealing with mixtures of gasses it is better to use the concept of *partial pressure*. It is the hypothetical pressure of a particular constituent of the gas mixture if the amount of it that is present would have occupied the total volume of the mixture alone, at the same temperature.[1] Gasses dissolve, diffuse and react according to their partial pressures, not their concentrations.

1.1 Some biology of germinating seeds

Seeds consists of at least the following three parts: (1) the *embryo*, which will grow out to become the new plant, (2) a supply of *nutrients* that the embryo can use in the early stage of germination, before it can use light and photosynthesis as main source of energy, and (3) the *seed coat* (or *testa*) that helps to protect the embryo from biotic and abiotic injury and drying out (see Figure 1). The embryo consists of one or two *seed leaves* ('*cotyledons*'), the *hypocotyl* that consists of an embryonic stem and the embryonic root, called the *radical*, and an embryonic shoot (epicotyl) above the point where the cotelydons attach.

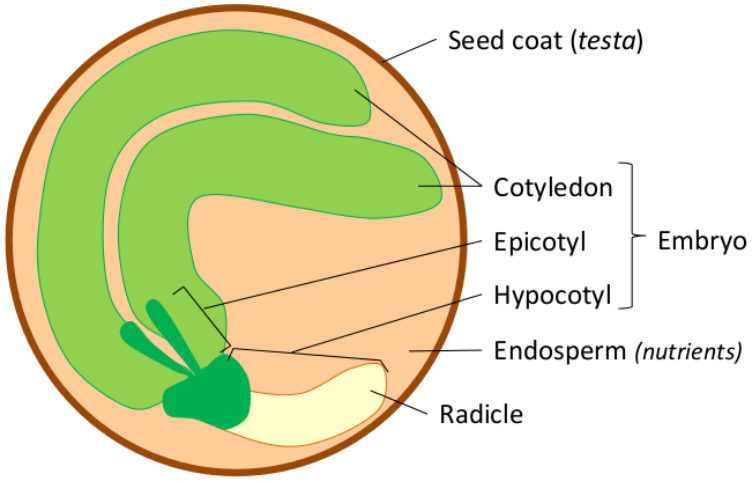


Figure 1: *Seed anatomy of a dicotyledon.*

The detailed internal structure of seeds varies highly among plant species. In some mature seeds the initial food storage tissue that results after fertilisation, the *endosperm*, is still present and contains the nutrients (mainly starch, but may also

[1] Recall the Gas Law: $pV/T = nR$, where $R = 8.314\text{J/mol K}$ is the universal gas constant.

contain oils and proteins) and forms with the seed coat an additional layer around the embryo. These type of seeds are called *endospermic*. Examples are barley and *Arabidopsis thaliana* (see Figure 3). In other, *non-endospermic* seeds, the cotelydons have absorbed all food in the endosperm during seed maturation. The endosperm is almost completely degraded in the mature seeds of this type and the cotelydons serve as sole food storage organ for the embryo. Examples of these are peas, sugar beet and Savoy cabbage.

The cells in a dormant seed are in a dehydrated state: most water content has been removed. Before germination can start the seed needs to get and take up water ('*imbibition*'). Its constituent cells will then take up water, restore and repair their internal biochemical structures and proceed to support the germination process. Typical vegetative cells consist of about 70% water, while dehydrated cells hardly contain any water. We shall call this the *repair phase* of germination. It is shown schematically in Figure 2.

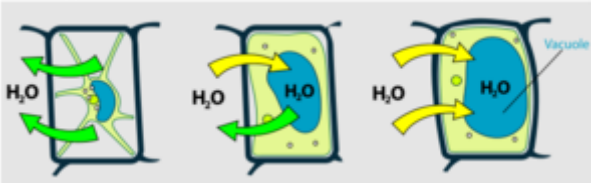


Figure 2: *Water uptake and cell repair.*

After imbibition the seed needs oxygen to germinate, both for repair and the subsequent growth phase in which the radical will start to grow first. The early stage of germination ends when the radical reaches the seed coat and breaks through: a germination event that is known as *testa rupture*. Whether and how much oxygen is available to the embryo for repair and growth depends on several factors, most importantly: *(i)* on the oxygen partial pressure outside the seed, *(ii)* on the consumption of oxygen of the cells inside the seed. i.e. the *respiration rate*, and *(iii)* on the oxygen transport through the seed coat and internal structures towards the embryo. Living seeds start respiration at the moment they start taking up water.

During the repair phase a bit of oxygen is being transported inside with the water. This oxygen is used to produce energy to facilitate the repair process. The repair is necessary because only cells with normal water content can divide and grow. After the repair process is finished and the seeds are fully saturated, the growth begins and the oxygen consumption increases.

The oxygen that the seed takes up has to be transported to the embryo in the middle of the seed because the embryo is the growing part of the seed, and growth can only take place when oxygen is available.

In the transport of oxygen from the seed coat to the embryo several aspects have to be considered that influence the permeability of the oxygen into the seeds. The specific structure of the coat affects the speed of oxygen transport into the seed. After

Figure 3: *Embryo of Arabidopsis thaliana.(M. Bayer; Max Planck Inst. for Developmental Biology)*

the oxygen has passed the seed coat it encounters a layer of starch or oil containing cells, depending on the type of seed. Cabbage, for example, is an oil seed. These cells block direct oxygen transport towards the embryo. It may be absorbed into these cells or pass through the channels in-between. See Figure 4 where a magnification of the layer of starch cells of two different seeds is given. The organisation of these cells varies per type of seed hence the permeability of oxygen differs per type. Moreover, if the space between the cells is very large and the cells lie in an orderly manner, it is much easier for molecules to pass than when the cells are very close together. In Figure 4 the difference in structure between poppy seed and gooseberry seed can be seen.

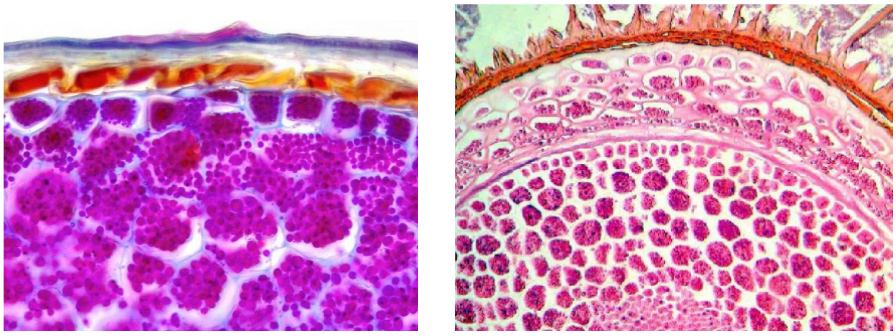


Figure 4: *Close-up images of the seed coat with the starch cells of poppy seed (left) and gooseberry seed (right). The different cell structures can be seen very clearly.*

After the oxygen has made its way through the layer of starch cells, it reaches the embryo. Inside the embryo cells the oxygen enters the mitochondria, which are membrane-enclosed organelles located in the cytoplasm of the cell. These mitochondria are sometimes described as 'cellular power plants' because their main task is to

produce energy that is needed for other processes inside the cell. They produce this energy in the form of ATP molecules by breaking down glucose to carbon dioxide using oxygen. This energy is used for growth, repair and transport in the cell. Note that the number of mitochondria per cell can vary from a few up to hundreds.

We are interested in the connection between oxygen availability and the growth of the embryo, so we will now take a closer look at this growth process. The embryo consists of two parts: an 'idle' part where hardly any growth occurs and the embryonic root (the *radicle*) that grows first to break the seed coat.

Growth of the root occurs by two processes: (1) *cell division* within the growing tip or *apical meristems* that consists of undifferentiated cells that are typically small, having a thin primary cell wall only and which are closely packed together. Their primary function is to divide. For each cell that divides in the meristem, one cell will leave the meristem. However, it need not be that one cell of a pair of daughter cells must become non-proliferative immediately (cf. [9], p.337). Cells that leave the proximal meristem (see Figure 5A) will differentiate into epidermal cells, cortex or stele and (2) *stretch*. Stretching effectively pushes the growth tip downwards. The newly formed cells from the distal meristem differentiate into root cap cells, that protect the apical meristems from rocks, dirt and pathogens. Between the proximal and distal meristems lies the so-called *quiescent center*. It has a much slower cell divison rate than the proximal and distal meristems. Its primary function seems to renew the cells in these meristems [9]. After division, the divided meristematic cell recover to their original size and start building up energy and resources until they have enough to divide again.

An illustration of this process is given in figure 5B. There, it can be seen that the growth tip is being pushed down as the embryo grows. During this process, the size of the growth tip doesn't change, only the idle part enlarges. This process keeps repeating itself until the root tip breaks through the seed coat and the seed has germinated.

1.2 The Q2 machine for non-intrusive oxygen measurements

Single seed oxygen consumption measurements can be made by the so-called '*Q2 machine*' for which Fytagoras developed the underlying measurement technology. It satisfies the criteria that the method is non-invasive, sensitive, fast and cost-effective. We now briefly describe this technique, see Figure 6.

Individual seeds are placed in cylindrical containers in a standard transparent plastic wells plate. A plate with 96 containers is shown in Figure 6. A single container is 10 mm high with a diameter of 5.0 mm, containing on one wall, on the inside, an oxygen sensitive fluorescent coating. When closed by the lid, these containers are almost air-tight. Not completely, because the small oxygen molecules are able to leak slowly through the plastic by diffusion. A 'completely' air-tight container would be too costly. When a light pulse in the blue range of the spectrum is shined on it, the coating will fluoresce a light pulse in the red range with a fluorescence life-time that is indicative of the oxygen level in the container. Outside the container a detector

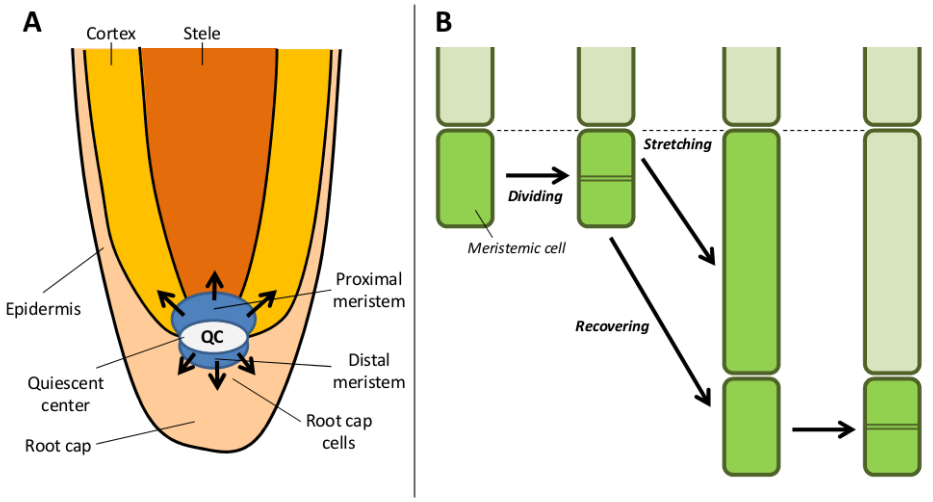


Figure 5: *Schematic presentation of the root apical meristems (growth tips) and differentiated tissues that result (Panel A - following [9]). Stretching of cells, the differentiated cells in particular, pushes down the growth tip (Panel B).*

is placed which measures the intensity and fluorescence life-time of the outgoing red light pulses. In this way one is able to measure regularly and automatically the level of oxygen within the container at high precision without opening it. The seeds are placed on top of filter paper soaked in water or agar to start imbibition.

Figure 7 shows three typical oxygen level curves for a single seed as a function of time. The oxygen level at each time is represented relative to the amount present in the container at the start of the experiment. Temperature is kept constant at room temperature (298 K) during an experiment. The volume in the container exterior to the seed may be assumed constant. That is, the volume of the seed will not expand substantially during early germination. Hence, the oxygen level curve can be interpreted as either change in partial pressure, concentration or total amount of oxygen in the exterior of the seed.

Computing from a container volume of $V_c = 1.9 \times 10^2$ mm^3 (see Table 1), an atmospheric pressure of 1 bar = 100 kPa at sea level and an oxygen content of 21% (by volume), one obtains an initial oxygen content of the container just after closing of

$$n^0_{O_2} = 0.21 \cdot \frac{pV_c}{TR} = 0.21 \times 7.7 \times 10^{-6} \text{ mol} = 1.6 \times 10^{-6} \text{ mol}.$$

Here $R = 8.315$ J/mol$\cdot$K is the gas constant. With a molar mass of 32.00 g/mol, this amounts to 52 μg of O_2 in an empty container. If a seed of e.g. cabbage is placed in the container this reduces slighly, to approximately 51 μg.

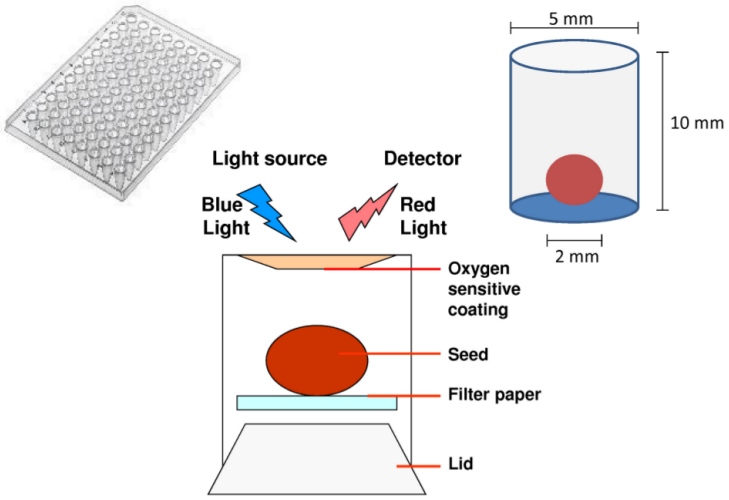


Figure 6: *A schematic presentation of the experimental set-up. A single container in the* 8×12 *well plate has a volume of approximately* 1.9×10^2 mm^3.

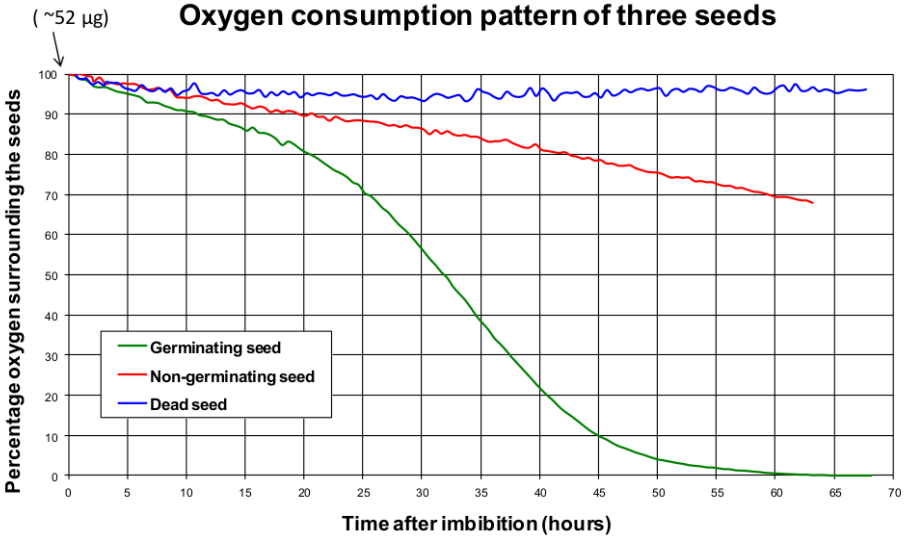


Figure 7: *Three typical oxygen consumption curves as measured by the Q2 machine. Oxygen level surrounding a seed in the container is expressed as fraction of the initial level at the start of the experiment.*

1.3 Problem description

Measured single seed oxygen consumption curves can be used to evaluate the quality of the seed since oxygen consumption is believed to be one of the main characteristics describing quality. However, the interpretation of the data in terms of physical, morphological and physiological properties and processes within the seed is still difficult, due to the lack of knowledge of the basic aspects of gas exchange in seeds and its role in the germination process.

> *The main goal of this study is to interpret characteristics of individual seed oxygen consumption curves as presented in Figure 7 in terms of underlying processes and seed properties.*

We focus on the first stage of germination until the point when the radical breaks through the seed coat, because this is the stage of seed germination in which the data curves are obtained.

1.4 Outline

During the study group we developed three mathematical models for the oxygen consumption of seeds that takes into account current biological knowledge of these processes. The measured oxygen consumption curves as provided by Fytagoras for barley (*Hordeum vulgare* L.), Savoy cabbage (*Brassica oleracea* var. *sabuada* L.) and sugar beet (*Beta vulgaris*) have been fitted to the corresponding curves predicted by these models.

In Section 2 we present a high-level phenomenological model that is able to describe the measured curves for Savoy cabbage and barley well. It results in a correlation among model parameters that can be interesting to look at in more detail (see Section 2.4). In Section 3 we develop a model starting from anatomical considerations in order to investigate the possibilities of relating characteristics of the oxygen consumption curve to seed structure and particular seed properties. In Section 3.4 the latter model is extended.

2 A logistic model for oxygen consumption

The experimental curves of oxygen consumption by germinating seeds of all three different plant species (barley, sugar beet and Savoy cabbage) provided by Fytagoras have a very characteristic 'sigma' shape, similar to the example shown in Figure 7. This indicates that this aspect of the curves originates from a universal underlying mechanism independent of the detailed seed morphology. In this section we propose a mathematical model for respiration of germinating seeds that does not require details on seed anatomy as input, but can reproduce the mentioned characteristic shape of the oxygen consumption curves.

Our central assumption is, that most of the oxygen that is taken up by the seed is consumed by the mitochondria within the cells that constitute the seed. According to the endosymbiotic theory, key organelles in eukaryotic cells, e.g. mitochondria, have evolved from bacteria and still largely behave as such. Bacterial replication under conditions of limited food/oxygen supply is well understood and is governed by the logistic equation.

We set up a general logistic equation for our problem, obtain its analytical solution, and fit it to the provided experimental data by tuning three independent parameters. It turns out that for the majority of the seeds that have been investigated in this study, two of these parameters, namely, the rate of growth of the mitochondria population and the final relative increase in this population, are proportional to each other. We propose a modification of the logistic equation that takes this effect into account.

2.1 Main assumptions

Let us summarize our major assumptions:

Assumption 1. Oxygen consumption happens mainly in mitochondria.

Assumption 2. Although groups of mitochondria are encapsulated inside cells with different biological function, morphology and growth rates, taken all together mitochondria behave as a colony of bacteria, i.e. 'multiplying' at a rate proportional to their number. As with usual bacterial colonies, the population is limited by the available resources, oxygen in particular.

Assumption 3. The rate of oxygen consumption by the seed is proportional to the rate of growth of biomass. In turn, the latter is proportional to the rate of change in the total number of mitochondria in all cells within the seed. There is a maximal number of mitochondria that the seed is able to produce within the closed container.

Assumption 4. The rate of diffusion of oxygen through the cell walls is much faster than the rate of oxygen consumption. Hence, we neglect the (constant) difference between the internal and external concentrations of oxygen.

2.2 Model derivation

We use the following notation:

- O_0 – the initial level of oxygen in a container
- O_c – the critical level of oxygen below which no growth can occur
- m_0 – the initial number of mitochondria
- m_c – the maximum sustainable number of mitochondria

If mitochondria behave as a colony of bacteria, then their number will be changing at the following rate:

$$\frac{dm}{dt} = \alpha m \left(1 - \frac{m}{m_\mathrm{c}}\right), \tag{1}$$

where α is the maximum rate of growth. This is a *logistic equation* with the well-known analytical solution

$$m(t) = \frac{m_\mathrm{c}}{1 + \left(\frac{m_\mathrm{c}}{m_0} - 1\right) e^{-\alpha t}}, \tag{2}$$

which satisfies the initial condition $m(0) = m_0$. *Assumption 3* can be written as

$$\frac{dO}{dt} = -\beta \frac{dm}{dt}. \tag{3}$$

It leads to the following equation for the amount of oxygen present in a container at time t:

$$O(t) = C - \beta m(t), \tag{4}$$

where the constants β and C must be such that

$$O(0) = O_0, \quad \lim_{t\to\infty} O(t) = O_\mathrm{c}. \tag{5}$$

Solving these equations for β and C gives:

$$C = O_0 + m_0 \frac{O_0 - O_\mathrm{c}}{m_\mathrm{c} - m_0}, \quad \beta = \frac{O_0 - O_\mathrm{c}}{m_\mathrm{c} - m_0}. \tag{6}$$

The final equation for the amount of oxygen is thus

$$O(t) = O_0 + \frac{O_0 - O_\mathrm{c}}{m_\mathrm{c} - m_0} \left[m_0 - \frac{m_\mathrm{c}}{1 + \left(\frac{m_\mathrm{c}}{m_0} - 1\right) e^{-\alpha t}} \right]. \tag{7}$$

Introducing the normalized quantities,

$$\tilde{O}(t) = \frac{O(t)}{O_0}, \quad \tilde{O}_\mathrm{c} = \frac{O_\mathrm{c}}{O_0}, \quad \tilde{m}_\mathrm{c} = \frac{m_\mathrm{c}}{m_0} \tag{8}$$

we arrive at the dimensionless result:

$$\tilde{O}(t) = 1 + \frac{1 - \tilde{O}_\mathrm{c}}{\tilde{m}_\mathrm{c} - 1} \left[1 - \frac{\tilde{m}_\mathrm{c}}{1 + (\tilde{m}_\mathrm{c} - 1)\, e^{-\alpha t}} \right]. \tag{9}$$

Note that the measured oxygen consumption curve as presented e.g. in Figure 7 shows $\tilde{O}(t)$. The functional expression (9) for the latter has three parameters that allow it to be fitted to the experimental data.

2.3 Fitting results

Figure 8 shows the normalized measured oxygen levels for 91 barley seeds and the corresponding curves obtained by fitting with expression (9). To determine these parameters we have used a standard nonlinear least squares optimization routine with the lower bounds set as: $\tilde{m}_c \geq 1$, $\tilde{O}_c \geq 0$, and $\alpha \geq 0$. Figure 9 shows the

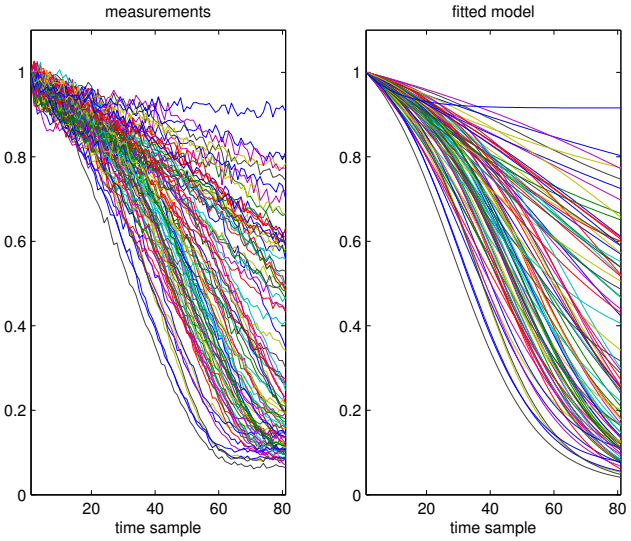


Figure 8: *Left: Measured oxygen consumption for 91 individual barley seeds at* 295 K *over a time period of 80 hours. Right: fitted logistic model.*

experimental curves and the corresponding fits for 24 Savoy cabbage seeds.

2.4 Conclusions and discussion

A closer look at the mutual relation between the three fitted parameters in different experiments reveals an interesting tendency. Namely, the growth rate parameter α appears to depend almost linearly on the relative total population increase $\tilde{m}_c = m_c/m_0$, so that when α is plotted against $1/\tilde{m}_c$ one gets a pronounced hyperbolic curve – see Figure 10. In other words the ratio $\alpha/\tilde{m}_c = k$ is close to a constant. This behavior indicates that instead of (1) the following form of the logistic equation should be used from the start:

$$\frac{dm}{dt} = km\left(\frac{m_c}{m_0} - \frac{m}{m_0}\right), \tag{10}$$

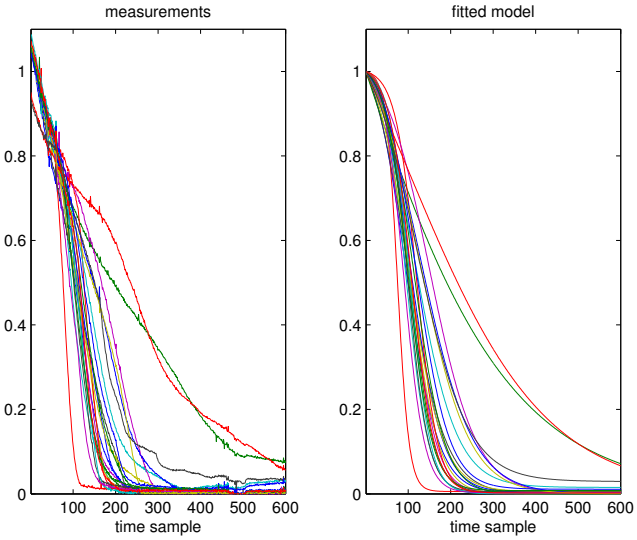


Figure 9: *Left: Measured oxygen consumption for* 24 *individual Savoy cabbage seeds at* 293 K. *Time samples were taken each 30 min. Right: fitted logistic model.*

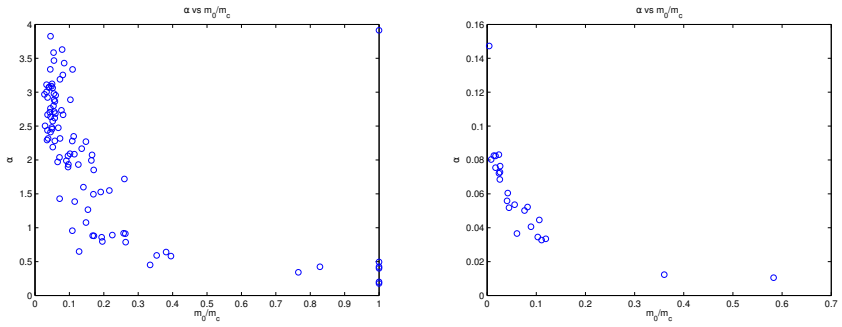


Figure 10: *Apparent hyperbolic dependence of the growth parameter* α *on the inverse of the relative total population growth* $1/\tilde{m}_c$, *i.e.,* $\alpha \approx km_c$ *for some constant* $k > 0$. *Left – barley , right – Savoy cabbage.*

so that the original standard logistic equation (1) is recovered under the assumption

$$\alpha = k\tilde{m}_c. \tag{11}$$

Of course, the actual values of all fitted parameters reconstructed in our simulations should not be taken too literally as we have used a normalized time axis. This fact, however, does not invalidate our conclusions about the mutual relation between the fitted parameters. In any case, further interpretation of the biological meaning of these parameters will require a much wider statistical analysis in an absolute time frame (say, in hours), as well as deeper understanding and an additional mathematical model of the apparent link between the growth rate and the final relative increase in the mitochondria population.

3 An anatomically structured model for oxygen uptake

As mentioned in Section 1.3, the main objective is to interpret the characteristics of the oxygen consumption curve in terms of physical, biological and morphological properties of the seed. The logistic model that we discussed in the previous section cannot be used for such explanatory purposes as it does not take any details of underlying shape or processes into account in its derivation.

In this section we take a first step in developing a mathematical model for oxygen uptake in a germinating seed that can be used to assess seed quality using oxygen consumption measurements. We have chosen to focus on a seed with the simplest morphology: the Savoy cabbage seed (*Brassica oleracea* var. *sabuada* L.). These have an almost perfect spherical appearance, with a radius of approximately 1.5 mm. Although its external shape is simple, internally the anatomy can still be quite complicated. Sugar beet seeds (*Beta vulgaris*) for example have a much more complicated morphology both externally and internally.

3.1 Driving processes and their time and spatial scales

We identify four processes that play a major role in the early germination stage of the seed.

P0. The process of water uptake (*imbibition*), that allows the embryonic cells to restore their water content, repair the internal molecular constitution and start functioning 'normally' in the early germination stage, i.e. targeted at supporting growth of the embryonic root (the *radical*).

P1. The process of *oxygen diffusion*: first through the seed coat and next through the seed's interior to reach the embryonic cells where it is finally used in metabolism.

P2. The process of *celullar respiration*, i.e. the oxygen uptake and consumption by the cells in the seed.

P3. The process of *asymmetric cell division* of the embryonic cells in the growth tip in the radical that results in growth. (We shall discuss this process in more detail below).

The initial repair process P0 is hard to describe in detail. In view of the oxygen consumption curve it corresponds to the first 15 hour time period in Figure 7, when oxygen consumption appears to happen at an approximately constant rate. We shall ignore this phase and focus on capturing the increased consumption rate and later steady decrease. Hence, we assume that P0 has been completed and only consider P1, P2 and P3.

As a first step toward the deduction of the model it is important to assess the characteristic time scales of processes P1–P3 (see e.g. [6]). We will use the available experimental data as much as possible. Otherwise, assumptions will be made. Awareness of these time scales motivates modelling decisions concerning its mathematical structure: i.e. whether the model will be formulated in terms of either ordinary differential equations or partial differential equations, depending on whether the processes involving the spatial variables can be neglected.

We now consider processes P1–P3 in further detail.

3.1.1 Oxygen uptake through seed coat and internal diffusion

The seed coat forms a barrier for oxygen transport. The physical and biochemical details of oxygen uptake through the seed coat, which is part of P1, are largely unknown. The application of artificial coatings to the seed coat are known to strongly influence oxygen uptake. Therefore, this process is taken as an important unknown in our model. Measurements have been performed on the permeability of the skin of fruits like apple, pear and nectarine though, cf. [8]. Such results for the seed coats of seeds of cabbage or sugar beet were not found in the literature.

The seeds of all species discussed in the report will have a complicated internal structure (cf. Figure 4). It is possible that there is free intercellular space in cell tissue through which oxygen and carbon dioxide can diffuse freely (see e.g. the free intercellular space volumes reported for flesh tissue in fruits like apple, pear and nectarine ranging from 17 to 2% respectively, cf. [8]). Similar structures are expected in seeds, but quantitive information was lacking. However, imbibition may have caused such channels to become water-filled. This makes a great difference for the effective diffusivity of oxygen through the cell tissue from the seed coat to the tissue where it is consumed most during germination. In fact, diffusion of oxygen in air is four orders of magnitude faster than that in water (cf. [2, 7]). The estimated diffusivity of oxygen in fruit flesh tissue is approximately 1.5×10^{-7} m^2/s on average [8].

Diffusivity:	Value:	Description:
$D_{\mathrm{O_2-water}}$	2×10^{-9} m^2/s	Diffusivity of oxygen in water at 298 K (Wilke & Chang [2])
$D_{\mathrm{O_2-air}}$	1.8×10^{-5} m^2/s	Diffusivity of oxygen in air at standard temperature (273 K) and pressure (1013 hPa; Massman [7], Table 8).
$D_{\mathrm{O_2-fruit}}$	$0.3 - 2.7 \times 10^{-7}$m^2/s	Estimated diffusivity of oxygen in fruit flesh (at 293 K, Rajapakse *et al.* [8]).

Collis-George & Melville [3] computed and discussed oxygen distribution profiles in four models for a spherical seed with specific assumptions on the properties of cellular respiration. In these models it is assumed that all seed tissue cells respire in the same fashion, i.e. the functional description for the local oxygen consumption rate is the same throughout the whole seed. In our study, we assume that most of oxygen uptake takes place in the dividing embryonic root cells. A steep oxygen gradient as predicted by the models in [3] in the tissue between seed coat and embryonic root is then expected to be more shallow. No measurements were available to support the theoretically reasonable oxygen gradient from seed coat to embryonic root by experimental data.

The time scale of convergence to a homogeneous steady state has been investigated numerically in a one-dimensional linear diffusion problem

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2} \qquad \text{on } [0, 1.7]$$

(in units mm and s for length and time). We took $D = 2 \times 10^{-3}$ mm^2/s, the minimal reported diffusivity for oxygen. In other cases equilibration will be faster. The initial condition is a step-wise normalized oxygen distribution u_0:

$$u_0(x) = \begin{cases} 0, & x \in [0, 1.5], \\ 1, & x \in (1.5, 1.7]. \end{cases}$$

Neumann boundary conditions (i.e. zero-flux) are imposed at both boundaries.

The computed evolution of the normalized oxygen concentration profile is shown in Figure 11. The system equilibrates on a time scale

$$T_d \approx 700 \text{ s} \approx 12 \text{ min} = 0.2 \text{ h}.$$

The partitioning of oxygen between water and air is given by Henry's Law,

$$[\mathrm{O_2}]_{\mathrm{water}} = k_{H,cp}\, p_{\mathrm{O_2}}, \tag{12}$$

where the Henry's Law constant $k_{H,cp}$, or oxygen solubility, equals 12 μM/kPa [5, 4] (at 293 K). Equivalently one may use the dimensionless Henry's Law constant k_H in terms of concentrations:

$$[\mathrm{O_2}]_{\mathrm{water}} = k_H\, [\mathrm{O_2}]_{\mathrm{air}},$$

where $k_H = 0.03$ at 293 K.

3.1.2 Cellular respiration: oxygen uptake and consumption

Detailed information on cellular respiration of (different types of) cells in seeds is hard to obtain. Rajapakse *et al.* [8] provide measured respiration rates for some fruits. Detailed measurement of changes in cellular respiration as a function of external oxygen pressure around mammalian cells (e.g. rat cardiomyocytes and human umbilical

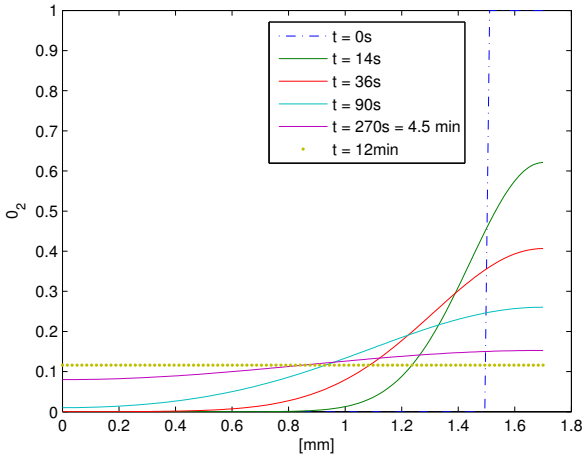


Figure 11: *Convergence to homogeneous steady state for oxygen diffusing in water at 298 K. Oxygen concentration has been normalized.* $D = 2 \times 10^{-3}$ mm^2/s.

vein endothelial cells) and isolated mitochondria from different types of rat cells have been reported in [5]. The drop in oxygen partial pressure within cells (myocytes) and a stagnant layer around cells is considered in [4].

Being aware that one cannot simply apply results for animal cells to the setting of plant cells, we nevertheless use the results on [5] as an indication of the order of magnitude of cellular respiration rates and a reasonable mathematical form for the functional dependence of this rate on partial oxygen pressure p_{O_2}. In fact, [5] finds that the oxygen flux J into uncoupled human umbilical vein endothelial cells in cell suspension can be fitted well by a hyperbolic curve

$$J = \frac{J_{\max}\, p_{O_2}}{p_{50} + p_{O_2}}, \tag{13}$$

with maximal flux per cell of $J_{\max} = 0.035$ fmol/s and an environmental oxygen pressure of $p_{50} = 0.023$ kPa at which the oxygen flux is at halve-maximal level (cf. [5], Figure 3, p.591; temperature is 37°C; single cell oxygen flux has been computed from the data provided in *loc.cit.* for a cell population). Coupled endothelial cells show a slightly increased maximal flux and 3-fold larger p_{50} (namely 0.068 kPa). An increase of the latter is expected: oxygen levels in the middle of a cluster will be lower than in cells at the outer end. Thus, a higher oxygen partial pressure is needed outside the cluster to reach the same (average) overall respiration rate.

3.1.3 Time scale of cell division and root growth

A central question in plant physiology is to determine the processes that regulate their growth. Growth rate is regulated by the combined activity of cell production by cell division and expansion of the cells that are already present [1]. Modelling root growth in detail is a complex project in itself, see [1, 9] and the references found there. The duration of the time between two consecutive cell divisions of the same meristematic cell (the *cell cycle*, recall Section 1.1, Figure 5 in particular) has been measured and is of the order of 10–15 hours [9].

We consider a very rudimentary model only. That is, the rate at which the root length r increases will be assumed proportional to the amount of oxygen consumed in the growth tip (i.e the root proximal meristem, recall Section 1.1), with proportionality constant ρ.

From the experimental oxygen consumption curves for Savoy cabbage (see e.g. the ensemble-averaged curve presented in Figure 14) one estimates that the duration T_g of the growth phase, from start to rupture of the seed coat, is approximately 60 hours. We assume that the radicle needs to extend over a length L equal to the diameter of the seed to do so. That is, in the setting of cabbage seeds, $L = 3$ mm.

A meristematic cell has length ℓ of approximately 15 μm and will stretch after differentiation with a factor σ_f to reach its final length. The length of the differentiated cell is a function of its distance to the quiescent center (cf. [1], Fig. 3). We ignore this effect and take an average value $\sigma_f = 4$ instead. Thus, in a time T_g, at least

$$N = \frac{L}{\sigma_f \ell} = 50$$

differentiated cells in a row must have been formed from the proximal meristem to realise an extension of the radicle over length L. Consequently, each

$$\Delta t = \frac{T_g}{N} = 1.2 \text{ h}$$

a new differentiated cell must appear in a linear array of differentiated cells in the growing root.

3.2 Model formulation

The closed container in which the seed is placed is represented by a perfect cylindrical domain of height h and radius R_c. The seed is modelled as a sphere of radius R_s, which is reasonable for seeds of Savoy cabbage (*Brassica oleracea* var. *sabuada* L.) of which we used the data provide by Fytagoras. The radical (embryonic root) inside the seed is cylindrically shaped too, with radius R_r. The cross sectional area of the radical, A, therefore equals πR_r^2. The radical is divided into a growth tip, of volume V_{tip} and the elongated, differentiated root cells, see Figure 3.2.

The oxygen partial pressure within the closed container, outside the seed, O_e, is considered homogeneously distributed. The oxygen partial pressure inside the seed is

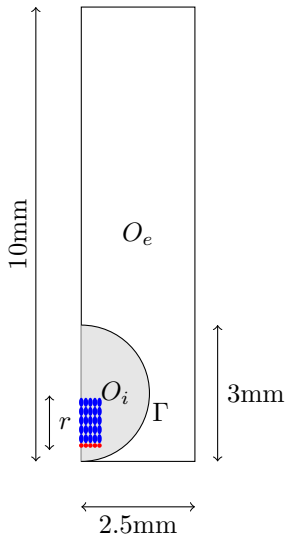


Figure 12: *The model geometry is axisymmetrical. Red cells: growth tip; blue cells: elongated and differentiated root cells that appeared by cell division after start of experiment (remainder of embryo is not shown).*

represented by O_i. As a first modelling approach and motived by the fast equilibration of the oxygen distribution in water on the time scale of cell division (see Section 3.1.1), we take O_i homogeneously distributed within the seed for simplicity, because of lack of more detailed information on oxygen gradients inside the seed.

Due to growth the radical will elongate. The elongation length of the radical at time t is $r(t)$. We ignore any bending in the radical when computing the additional root volume due to radical growth. No experimental observations were available for both elongation and shape of the radical during the early germination phase that could support more elaborate hypotheses. Thus, in the proposed model, the seed coat is the main oxygen barrier that separates the embryo from the free exterior oxygen in the container.

Variable:	Variable name:
External oxygen partial pressure	$O_e = O_e(t)$
Internal oxygen partial pressure	$O_i = O_i(t)$
Total elongation of radical due to growth	$r = r(t)$

Changes in volume of seed (or container) are ignored. Since experiments are performed under constant temperature, we may consider the oxygen partial pressures in the interior and exterior domains as proportional to concentrations, with constant proportionality everywhere. We shall do so from this point on.

The precise biophysical mechanisms of oxygen uptake through the seed coat at microscopic scale are largely unknown. In our model we take the simplest modelling

approach, by representing the seed coat as a barrier with unknown permeability a for oxygen (of dimension of velocity). The total flux of oxygen from the exterior into the seed through the seed coat is then given by

$$J_{sc} := aS_\Gamma(R^*O_e - O_i), \tag{14}$$

where R^* is the so-called *accumulation ratio* and allows for modelling asymmetric transport properties of the barrier. S_Γ is the total area of the spherical seed coat: $S_\Gamma = 4\pi R_s^2$. At steady state, i.e. when there is no flux through the barrier, $O_i = R^*O_e$. We have taken $R^* = k_H$, Henry's Law constant, thinking of oxygen dissolving into water inside the seed.

The uptake of oxygen by cells in the embryo and any remaining cells in the endosperm is considered to be due mainly to mitochondrial activity, which depends on the oxygen partial pressure (see [5]). We let $J(O_i)$ denote the average amount of oxygen consumed by cells per unit time per unit volume. It is of the form:

$$J(O_i) = \frac{J_{max}O_i}{J_0 + O_i} \tag{15}$$

(see Section 3.1.2). We allow cells in the growth tip to consume a factor α more than this average amount, because these cells are most active.

We thus arrive at the following model equations:

$$V_i\frac{dO_i}{dt} = -J(O_i)(\alpha V_{\text{tip}} + Ar + V_i) + aS_\Gamma(k_HO_e - O_i), \tag{16}$$

$$V_e\frac{dO_e}{dt} = aS_\Gamma(O_i - k_HO_e), \tag{17}$$

$$\frac{dr}{dt} = \rho J(O_i)V_{\text{tip}}, \tag{18}$$

with initial conditions

$$O_i(0) = O_{i,0}, \qquad O_e(0) = O_{e,0}, \qquad r(0) = 0. \tag{19}$$

The factor ρ in equation (18) is the oxygen-to-root biomass conversion factor. It measures the amount of oxygen needed in the growth tip to divide, differentiate and stretch to produce root biomass in terms of increased root length.

3.3 Fit to experimental data

The model equations (16)–(18) contain 12 parameters and two initial conditions. The dimension of this parameter space is too high to determine all of them from the experimental data that gives detailed time series of just one component of the solution only. The approach would be to solve an optimisation problem to minimize the distance between the solution of the model equations and the averaged experimental curve for a suitably chosen objective function. There may be difficulties in solving the

optimisation problem due to this limited information. Moreover, even if a sufficiently good solution is found, some of the parameters obtained may not be in a biologically reasonable range, causing interpretation issues.

However, some of the parameters and initial conditions can be estimated *a priori*, based on the given geometry of the experimental environment, anatomical information on the seeds and values found in the literature (often for other organisms however). In Section 3.3.1 we discuss these estimates that allow us to reduce the number of unknown parameters to two: the seed coat permeability a and the excess oxygen consumption factor α.

3.3.1 *A priori* estimating model parameters

Table 1 summarizes the parameter values that can be fixed by the geometric set-up of the experiments and seed morphology. We now discuss the constraints imposed on physiological parameter values using information found in the literature (see Sections 3.1.1–3.1.3) and the estimation of initial values.

The initial external oxygen concentration outside the seed just after the container has been closed, O_e^0 can be computed using an atmospheric pressure of 100 kPa, temperature of 293 K and oxygen content of 20%:

$$\begin{aligned} O_e^0 &= 0.2 \times \frac{p}{TR} = 0.2 \times \frac{100 \text{ kPa}}{293 \text{ K} \cdot 8.315 \text{ L/mol K kPa}} \\ &= 8.21 \times 10^{-3} \text{ mol/L} = 2.6 \times 10^{-4} \text{ mg/mm}^3. \end{aligned}$$

We assume that the initial internal oxygen concentration at that time is at equilibrium with the external initial oxygen concentration over the seed coat barrier:

$$O_i^0 = k_H O_e^0 = 0.03 \times 2.6 \times 10^{-4} \text{ mg/mm}^3 = 7.8 \times 10^{-6} \text{ mg/mm}^3.$$

The model (16)–(19) describes the oxygen consumption during the growth phase, following the repair phase. The start of the growth phase has been determined by visual inspection of average oxygen consumption curve (i.e. the consumption curve by averaging the individual curves of all seeds in the experimental run). It is taken at approximately time $t = 30$ h after the start of the experiment (see Figure 14). At that time, the oxygen level within the container has dropped to 79% of the initial level. At the end of the experiment, it has dropped to 19% of the initial level. The initial conditions $O_{e,0}$ and $O_{i,0}$ are computed from O_e^0 and O_i^0 by taking this reduction into account:

$$O_{e,0} = 0.79 \times O_e^0, \qquad O_{i,0} = 0.79 \times O_i^0.$$

Gnaiger *et al.* [5] report on various values for the oxygen partial pressure within the cells such that cellular oxygen consumption or mitochondrial oxygen consumption is at halve-maximal value. At 298 K, these values are in a range of roughly 0.005–0.08 kPa for mitochondria in animal cells, with a bias towards the lower values in the

Name:	Value:	Unit:	Description:
Geometric constants:			
R_c	2.5	mm	Radius of the cylindrical container
h	10	mm	Height of of the cylindrical container
R_s	1.5	mm	Radius of the spherical seed
R_r	0.5	mm	Radius of radical (root)
ℓ	1.5×10^{-2}	mm	Length of cells in growth tip
σ_f	4	–	Stretch factor
Computed geometric attributes:			
V_i	14.1	mm^3	Seed volume
V_c	196	mm^3	Volume of container
V_e	182	mm^3	Volume inside container exterior to seed
V_{tip}	0.0236	mm^3	volume of growth tip
S_Γ	28.3	mm^2	Seed coat area
A	0.79	mm^2	Cross sectional area of radical
Physical parameters:			
k_H	0.03	–	Henry's constant for oxygen in water
Estimated initial conditions (see Section 3.3.1):			
O_e^0	2.6×10^{-4}	$\mathrm{mg\,mm^{-3}}$	O_2 concentration outside seed at closure container
O_i^0	7.8×10^{-6}	$\mathrm{mg\,mm^{-3}}$	O_2 concentration inside seed at closure container
$O_{e,0}$	2.1×10^{-4}	$\mathrm{mg\,mm^{-3}}$	Initial oxygen concentration outside seed
$O_{i,0}$	6.2×10^{-6}	$\mathrm{mg\,mm^{-3}}$	Initial oxygen concentration inside seed
Estimated physiological parameters (see Section 3.3.1):			
J_{max}	3.6×10^{-3}	$\mathrm{mg\,mm^{-3}\,h^{-1}}$	Maximal oxygen consumption per tissue volume
J_0	3.2×10^{-7}	$\mathrm{mg\,mm^{-3}}$	O_2 level at $\frac{1}{2}J_{max}$ O_2-consumption
ρ	5.9×10^3	$\mathrm{mm\,mg^{-1}}$	Oxygen-to-root biomass conversion factor
L	3	mm	Elongation of the radical at time of testa rupture
Results of fitting: (see Section 3.3.2):			
a	7	$\mathrm{mm\,h^{-1}}$	Seed coat permeability for oxygen
α	2	–	Excess oxygen consumption factor

Table 1: *Parameter settings in the anatomically structured model for the growth phase and results of fit.*

range. We take 0.025 kPa. This yields an intracellular oxygen concentration of

$$\begin{aligned} J_0 &= \frac{p_{O_2}}{RT} = \frac{0.025 \text{ kPa}}{8.315 \text{ kPa L/mol K} \cdot 298 \text{ K}} = 1.0 \times 10^{-5} \text{ M} \\ &= 3.2 \times 10^{-7} \text{ mg/mm}^3. \end{aligned}$$

We ignore at this point the effects of an intracellular oxygen gradient from cell membrane towards the mitochondria.

Moreover, Gnaiger *et al.* [5] provide values for the maximal oxygen consumption of human umbilical vein endothelial cells of $80 - 100$ pmol/s per cm^3 of experimental medium that contained 2.6×10^6 cells per cm^3. For the value of 100 pmol/s this implies a maximal consumtion rate of 0.038 fmol/s per cell. For a cylindrical cell with a length of 15 μm and diameter of 10 μm (i.e. of volume 1.2×10^{-6} mm^3), this amounts to

$$J_{max} = 3.1 \times 10^4 \text{ fmol/mm}^3 \text{ s} = 3.6 \times 10^{-3} \text{ mg/mm}^3 \text{ h}.$$

The oxygen-to-biomass conversion factor ρ can be estimated in the following manner. The oxygen consumption function $J(O_i)$ is such that it has a quite sharp switch between maximal consumption and almost no consumption (cf. [5]). So we assume that it operates at maximal value for most of the growth phase. Therefore,

$$\rho \approx \frac{\Delta r}{\Delta t} \cdot \frac{1}{J_{max} \cdot V_{tip}} = \frac{3 \text{ mm}}{60 \text{ h}} \cdot \frac{1}{8.5 \times 10^{-5} \text{ mg/h}} = 5.7 \times 10^3 \text{ mm/mg}.$$

3.3.2 Fitting results

The remaining two parameters, a and α, in the model were determined by solving an optimisation problem to fit the experimental data. As targets in the optimisation process the following were considered:

- The ensemble-averaged time history $O_e^*(t)$ of the measured external oxygen level;
- The final root length $r_f^* := r(T_g) = L$.

The optimisation problem consisted of minimising the cost function

$$F = \|O_e - O_e^*\|_2^2 + w|r(T_g) - r_f^*|^2,$$

where $w > 0$ is a weight to give more or less importance to unevenness in the final root length. The results for parameters a and α are shown in Table 1. A simulation result of the model for these parameter values is shown in Figure 13.

The value for a, which is 7 mm/h $= 2 \times 10^{-6}$ m/s, seems to have a reasonable order of magnitude, when compared to values for the permeability of cell membranes for plant hormones, like auxin. However, no information on seed coat permeability for oxygen was avaible to compare this value with.

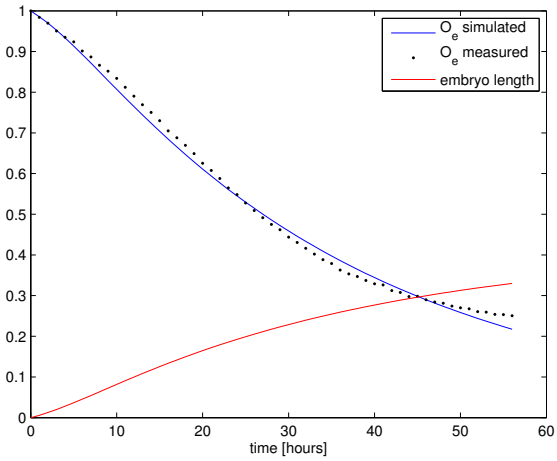


Figure 13: *Simulation result for the anatomically structured model* (16)–(19) *for the growth phase with parameter settings as summarised in Table 1. The data are obtained from Savoy cabbage.*

3.4 An extended model including the initial repair phase

The model (16)–(19) does not describe the repair phase that preceeds the growth phase. In this section we suggest an extension of the former model, which can be used to fit the entire oxygen consumption curve, including the initial repair process. Moreover, the experimental oxygen consumption curve is almost flat after approximately 80 hours. This fact seems difficult to account for directly with the oxygen consumption function $J(O_i)$ that was used in the previous model.

We suggest two modifications. Firstly, we realise that a higher oxygen level outside the growing cells is needed to create an internal oxygen concentration at the mitochondria to keep these 'maximally' functioning (see e.g. [4]). The simplest way to include this into the model, is to introduce a threshold value $\tilde{O}_i$ for the oxygen concentration outside the cells: if the oxygen concentration is below this value, the intra-cellular concentration near the mitochondria becomes so low that they stop functioning. This is realised mathematically by replacing O_i in $J(O_i)$ by $(O_i - \tilde{O}_i)_+$, where $(x)_+$ is the positive part of x:

$$(x)_+ := \begin{cases} x, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0. \end{cases}$$

Secondly, in model (16)–(19) we implicitly assumed that all cells in the embryonic root need the same, fixed, time to complete their repair phase and start the growth phase. A more relaxed assumption is that these times are normally distributed.

Thus, in our extended model, equations (16)–(18) remain the same except for

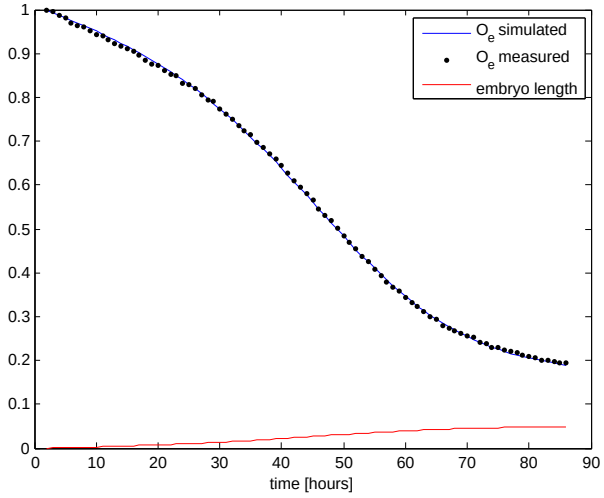


Figure 14: *The computed oxygen consumption curve using the extended model (blue line), compared to the experimental values (black dots). The growth phase starts at approximately* $t = 30$ h*, when the oxygen level has dropped to* 79%*. The length of the growing tip is shown as well. The data are obtained from Savoy cabbage seeds.*

replacing $J(O_i)$ in (16) by the time-dependent function

$$\tilde{J}(O_i, t) := f(t) \frac{J_{max}(O_i - \tilde{O}_i)_+}{J_0 + (O_i - \tilde{O}_i)_+}, \tag{20}$$

where

$$f(t) = 1 + \epsilon + \mathrm{erf}(c_1 t - c_2), \tag{21}$$

with $\mathrm{erf}(x)$ the *error function*:

$$\mathrm{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x e^{-\xi^2}\, d\xi.$$

The error function is a strictly increasing function with horizontal assymptotes at -1 and 1 when x tends to $-\infty$ and $+\infty$ respectively. It has a sigmoidal shape, is point symmetric with respect to 0 and takes values in $(-1, 1)$. As initial conditions in the extended model we take

$$O_e(0) = O_e^0, \qquad O_i(0) = O_i^0.$$

The two parameters c_1 and c_2 are related to the standard deviation and the mean of the normally distributed damage, respectively. Here ϵ is a small positive number,

Name:	Value:	Unit:	Description:
Physiological parameters:			
$\tilde{O}_i$	1.3×10^{-6}	$\mathrm{mg\,mm^{-3}}$	O_2 concentration threshold in the seed
J_{max}	2.7×10^{-2}	$\mathrm{mg\,mm^{-3}\,h^{-1}}$	Maximal oxygen consumption per tissue volume
J_0	3×10^{-8}	$\mathrm{mg\,mm^{-3}}$	O_2 level at $\frac{1}{2}J_{max}$ O_2-consumption
Results of fitting:			
a	25	$\mathrm{mm\,h^{-1}}$	Seed coat permeability for oxygen
α	1.8	–	Excess oxygen consumption factor
ϵ	1×10^{-3}	–	
c_1	1.8×10^{-2}	$\mathrm{h^{-1}}$	
c_2	2.37	–	

Table 2: *Parameter settings in the extended anatomically structured model and results of fit. Only new parameters and settings that differ from those mentioned in Table 1 are shown.*

reflecting the fact that there are some mitochondria that function from the very beginning.

The newly introduced parameter $\tilde{O}_i$ can be easily estimated from the right end of the measured oxygen consumption curve in Figure 14. The oxygen level has then dropped to 19% of the initial level, but still has not completely saturated. We therefore take $\tilde{O}_i$ at 17% of the initial level:

$$\tilde{O}_i = 0.17 \times O_i^0 = 1.3 \times 10^{-6}\ \mathrm{mg/mm^3}.$$

We estimated by hand the values of the remaining parameters. Similarly as described in Section 3.3.1. We had to increase J_{max} by a factor 7.6 and we used the refinement $J_0 \approx 3 \times 10^{-8}$ mg/mm^3. A summary of the modified and additional parameter settings together with the fitting results is given in Table 2. The resulting oxygen consumption curve, plotted together with the averaged experimental data is shown in Figure 14.

References

[1] G.T.S. Beemster and T.I. Baskin (1998). Analysis of cell division and elongation underlying the developmental acceleration of root growth in *Arabidopsis thaliana*. *Plant Physiol.* **116**: 1515–1526.

[2] C.R. Wilke and P. Chang (1955). Correlation of diffusion coefficients in dilute solutions. *AIChE Journal* **1**(2): 264–270. DOI: 10.1002/aic.690010222.

[3] N. Collis-George and M.D. Melville (1974). Models of oxygen diffusion in respiring seed, *J. Exp. Botany* **25**(89): 1053–1069.

[4] K.E. Dionne (1990). Oxygen transport to respiring myocytes, *J. Biol. Chem.* **265**: 15400–15402.

[5] E. Gnaiger, R. Steinlechner-Maran, G. Méndez, Th. Eberl and R. Margreiter (1995). Control of mitochondrial and cellular respiration by oxygen, *J. Bioenergetics and Biomembranes* **27**(6): 583–596.

[6] J. Mauseth (2011). *An Introduction to Plant Biology*, Jones & Bartlett Learning.

[7] W.J. Massman (1998). A review of molecular diffusivities of H_2O, CO_2, CH_4, CO, O_3, SO_3, NH_3, N_2O, NO, and NO_2 in air, O_2 and N_2 near STP, *Atmospheric Environment* **32**(6): 1111–1127.

[8] N.C. Rajapakse, N.H. Banks, E.W. Hewett and D.J. Cleland (1990). Development of oxygen concentration gradients in flesh tissues of bulky plant organs, *J. Amer. Soc. Hort. Sci.* **115**(5): 793–797.

[9] P.L. Webster and R.D. MacLeod (1980). Characteristics of root apical meristem cell population kinetics: a review of analyses and concepts, *Environmental Exp. Bot.* **20**: 335–358.

Estimates on Returnable Packaging Material

J. Bierkens (Radboud University Nijmegen)
H. Blok (Leiden University)
M. Heydenreich (Leiden University and CWI Amsterdam)
R. Núñez Queija (University of Amsterdam and CWI Amsterdam)
P. van Meurs (Eindhoven University of Technology)*
F. Spieksma (Leiden University) J. Tuitman (KU Leuven)

Abstract

When a beer company replaces its returnable packaging materials, for example when updating the design of a bottle, it needs to know in advance how much new material will be needed. Dutch beer brewer Heineken submitted the question of estimating the returnable packaging materials to the 2013 Studygroup Mathematics with Industry. In this report, we present both stochastic flow models and a queueing model to estimate the amount of returnable packaging material present in the market. Furthermore, we give recommendations on what data to collect, and how to sample this data in an unbiased way in order to increase accuracy of the estimation.

KEYWORDS: Modelling, Markov Chain, Stochastic Differential Equation, Queueing Theory

1 Introduction

Beer companies, like Heineken, use *returnable packaging materials* (i.e., bottles, cases and kegs) multiple times. To simplify our terminology, we will throughout refer to returnable packaging materials as *bottles*, keeping in mind that all results apply to other types of materials as well. In some markets, for example in the Netherlands, customers pay a *deposit* on bottles, which is returned to them when the bottles are returned. In other markets, for example in many African countries, a *full-for-empty* system is used instead. In this system, customers return empty bottles only when purchasing full ones. Unlike in the deposit system, in the full-for-empty system any purchase of new bottles is limited by the number of empty bottles available to the customers. Therefore, customers tend to keep a much larger stock of empty bottles in a full-for-empty system than in a deposit system.

*Corresponding author

Occasionally, the returnable packaging material is changed, for example because of a new bottle design, and then the beer company needs to know how many new bottles will be needed. This is an intrinsically difficult question, because bottles might be broken or stored away and reappear only many years after they are sold. At the moment, the companies miss an efficient model for estimating the number of bottles in the market, especially in the case of a full-for-empty system. In some packaging change operations, significantly more new bottles were needed than expected. Heineken requested the 2013 Study Group Mathematics with Industry to develop a model for estimating the number of bottles in the market more accurately, and asked for recommendations on what data to collect for use in such a model.

The structure of this report is as follows. First, in Section 2 we introduce terminology and notation, and explain what data is currently available. Moreover, we present an easy way of estimating the so-called break rate, which will be an important parameter in what follows. Then we develop two different models for the number of bottles in the market: in Section 3, Markov Chains and stochastic differential equations are used, while in Section 4 a queueing model is discussed. The difference of the two approaches is discussed in Section 4. Next, in Section 5, we elaborate on how to use the sample data to get reliable parameter estimates. Finally, we summarize our findings in Section 6.

2 Problem description

2.1 Modelling and notation

In a very simplified market model, bottles are sold at the distribution center, and arrive at the market. After remaining there for some time, the empty bottles are returned to the beer company. If the number of returned bottles is not sufficient to satisfy the beer demand, new bottles have to be produced (see Figure 1).

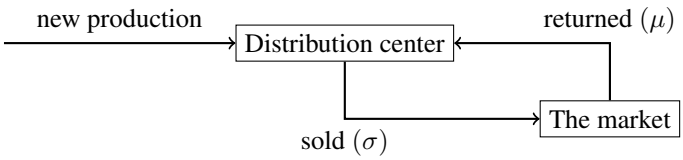


Figure 1: The flow of bottles

We aim at estimating the number of bottles currently in the market. In particular, we are interested in the number of bottles that are expected to be returned. To this end, we differentiate between different categories of bottles:

- **The returning bottles** $R(t)$
  The number of bottles at time t that will be returned even in the absence of a

packaging change. Typically, these bottles are in the market for a relatively short period of time before they return to the distribution center.

- **The sleeping bottles** $S(t)$
  The number of bottles at time t that will only be returned after a packaging change. These materials are temporarily stored away or used for other purposes and will turn up after a change of bottle.

- **The broken bottles** $B(t)$
  The number of bottles at time t that will never return to the distribution center. They could be broken, lost or stored away permanently.

The sum $M(t) = R(t) + S(t) + B(t)$ of these three numbers represents the total number of bottles present in the market at time t. The company is especially interested in estimating $S(t)$ more accurately.

2.2 Available Data

2.2.1 Volumes

The company has kept track of the volumes of the sold as well as the returned bottles for more than 20 years for different factories. These data are collected per product on a monthly basis. We introduce the following notation:

- **Sold items** $\sigma(t)$
  $\sigma(t) =$ the number of bottles sold in month t.

- **Returned items** $\mu(t)$
  $\mu(t) =$ the number of bottles returned in month t.

To obtain the number of bottles accumulated in the market since the beginning of measurements, we take the sum of the number of sold items and substract the number of returned items, that is,

$$M(t) - M(0) = \sum_{s=0}^{t} \big(\sigma(s) - \mu(s)\big).$$

2.2.2 Circulation Times

More recently, the company has started collecting samples of *circulation times* of bottles. The circulation time of a bottle is the time that elapses between leaving and returning to the distribution center. The data on the circulation times can be used to estimate $R(t)$, as explained in Section 3 and Section 4. As this data only represents returned bottles, it cannot be used to differentiate between sleeping and broken bottles. However, we can extract estimates for $M(t)$ and $R(t)$, and hence also for the sum $B(t) + S(t) = M(t) - R(t)$.

Currently the data is collected in a rather biased way, which results in unreliable estimates for the circulation time. In Section 5 we elaborate on how to improve the sampling.

2.3 Approximation of $R(t) + S(t)$

We are left with the question of how to estimate the total number of returning materials $R(t) + S(t)$. We proceed by assuming that $S(t)$ stabilizes for sufficiently large times t. This is a reasonable assumption since the storage capacity for bottles in the market is limited and hence the number of sleeping bottles cannot grow indefinitely. Moreover, we also assume that the fraction of sold bottles that ends up broken is constant. We call this constant β and refer to it as the *break rate*. Thus $\beta M(t) \approx B(t)$ for large times t. By our assumptions, the break rate can be estimated by the number of bottles that are not returned in a long time period $[t_0, t_1]$ divided by the number of bottles sold in that time period. That is,

$$\hat{\beta} = \frac{\sum_{s=t_0}^{t_1}(\sigma(s) - \mu(s))}{\sum_{s=t_0}^{t_1}\sigma(s)} = 1 - \frac{\sum_{s=t_0}^{t_1}\mu(s)}{\sum_{s=t_0}^{t_1}\sigma(s)},$$

for $t_0, t_1, t_1 - t_0$ large enough.

Now we can write

$$R(t) + S(t) = M(t) - B(t) = M(t) - \hat{\beta}M(t) = \frac{\sum_{s=t_0}^{t_1}\mu(s)}{\sum_{s=t_0}^{t_1}\sigma(s)}M(t).$$

Remark: The first assumption on the stabilization of the number of sleeping bottles in the market might not be very realistic in fast growing markets.

3 Analysis based on Markov models

In this section we propose a simple Markovian model to model the 'dynamics' of a single bottle while it is with a customer. It should be considered an example. More realistic models can be constructed by introducing more states and/or more complex dynamics (e.g., dynamics that are not homogeneous in time). The models in this section are especially suited for the deposit system, and less so for the 'full-for-empty' system, because we do not model the fact that a customer returns bottles at the same time that he buys new ones. See also the discussion in Section 3.5.

3.1 Markov model for single unit

First let us consider an individiual bottle at a customer. Suppose it behaves according to the following simple Markov dynamics. The bottle can be in either of four states, FULL, EMPTY, BROKEN and RETURNED. The states BROKEN and RETURNED are absorbing. Transitions (per unit of time) occur as in Figure 2.

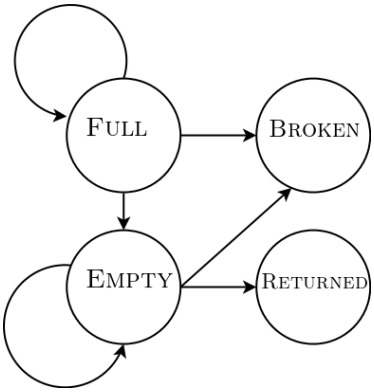


Figure 2: A Markov model. Transition rates between the states F(ull), E(mpty), B(roken), and R(eturned) are denoted by λ_{FE}, λ_{FB}, λ_{ER}, λ_{EB}.

3.2 Markov model for quantities of bottles

From the Markov model for a single bottle we can derive a model for the flow of bottles. Suppose, at time t there are a total number of F_t non-empty bottles and E_t empty bottles in the market. Let $h > 0$ be a small time step and suppose a number of $U_{t+h} - U_t$ bottles are sold during the time interval $[t, t + h)$. Denote the total number of returned bottles within $[t, t+h)$ by $Y_{t+h} - Y_t$. The variable names U for input and Y for output are chosen to correspond to the usual notation in systems theory. In systems theory, there also is the notion of state, denoted by X, corresponding in our case with the two-dimensional vector (E, F).

Remark: Note that the number of bottles sold per unit of time is $(U_{t+h} - U_t)/h$, which for small h is equivalent to the time derivative of U_t. The same holds for Y_t. This 'cumulative' notation for U and Y is helpful in the continuous time limit, where trajectories of U and Y will typically be non-differentiable.

If we denote by $(F \to B)_t$ and $(E \to B)_t$ the number of full and empty bottles broken within $[t, t + h)$, respectively, $(F \to E)_t$ the number of emptied bottles and $(E \to R)_t$ the number of empty bottles returned within this interval, we obtain the following balance equations:

$$\begin{aligned} F_{t+h} &= F_t + U_{t+h} - U_t - (F \to B)_t - (F \to E)_t, \\ E_{t+h} &= E_t + (F \to E)_t - (E \to B)_t - (E \to R)_t, \\ Y_{t+h} &= Y_t + (E \to R)_t. \end{aligned}$$

Now for all the transitions in the Markov model, assuming independent 'behaviour' of individual bottles, and using that h is small, we see that e.g. $(F \to B)_t \sim \mathrm{Bin}(F_t, \lambda_{FB}h)$. Using the normal approximation for the binomial distribution (assuming F_t and E_t are large), we find that $(F \to B)_t$ is approximately normally distributed

with mean $h\lambda_{FB}F_t$ and variance $F_t(h\lambda_{FB})(1-h\lambda_{FB}) \doteq h\lambda_{FB}F_t$. Therefore, we have approximately the following *discrete time Markov model*

$$\begin{aligned}
F_{t+h} &= (1 - h(\lambda_{FB} + \lambda_{FE}))F_t + U_{t+h} - U_t - \sqrt{h\lambda_{FB}F_t}\varepsilon_t^{FB} - \sqrt{h\lambda_{FE}F_t}\varepsilon_t^{FE},\\
E_{t+h} &= E_t + h(\lambda_{FE}F_t - (\lambda_{EB} + \lambda_{ER})E_t) + \sqrt{h\lambda_{FE}F_t}\varepsilon_t^{FE}\\
&\quad - \sqrt{h\lambda_{ER}E_t}\varepsilon_t^{ER} - \sqrt{h\lambda_{EB}E_t}\varepsilon_t^{EB},\\
Y_{t+h} &= Y_t + h\lambda_{ER}E_t + \sqrt{h\lambda_{ER}E_t}\varepsilon_t^{ER}.
\end{aligned}$$

where all the $\varepsilon_t^{\cdots}$ are normally distributed with mean 0 and variance 1. Assuming the fluctuations in F and E to be relatively small, the following model is more straightforward to analyse.

$$\begin{aligned}
F_{t+h} &= (1 - h(\lambda_{FB} + \lambda_{FE}))F_t + (U_{t+h} - U_t) - \sqrt{h}\sigma_{FB}\varepsilon_t^{FB} - \sqrt{h}\sigma_{FE}\varepsilon_t^{FE},\\
E_{t+h} &= E_t + h(\lambda_{FE}F_t - (\lambda_{EB} + \lambda_{ER})E_t) + \sqrt{h}\sigma_{FE}\varepsilon_t^{FE}\\
&\quad - \sqrt{h}\sigma_{ER}\varepsilon_t^{ER} - \sqrt{h}\sigma_{EB}\varepsilon_t^{EB},\\
Y_{t+h} &= Y_t + h\lambda_{ER}E_t + \sqrt{h\lambda_{ER}E_t}\varepsilon_t^{ER}.
\end{aligned} \tag{1}$$

3.2.1 Diffusion limit

By taking the $h \downarrow 0$ limit, we may write the Markov model (1) as the following system of stochastic differential equations (SDEs),

$$\begin{aligned}
dF_t &= -(\lambda_{FB} + \lambda_{FE})F_t\,dt + dU_t - \sqrt{\lambda_{FB}F_t}\,dW_t^{FB} - \sqrt{\lambda_{FE}F_t}\,dW_t^{FE},\\
dE_t &= [\lambda_{FE}F_t - (\lambda_{EB} + \lambda_{ER})E_t]\,dt + \sqrt{\lambda_{FB}F_t}\,dW_t^{FB}\\
&\quad - \sqrt{\lambda_{EB}E_t}\,dW_t^{EB} - \sqrt{\lambda_{ER}E_t}\,dW_t^{ER},\\
dY_t &= \lambda_{ER}E_t + \sqrt{\lambda_{ER}E_t}\,dW_t^{ER},
\end{aligned}$$

where $W_t^{\cdots}$ are independent Brownian motions.

This (non-linear) system is rather difficult to analyse, mainly because of the presence of square roots. As before, under the assumption that the fluctuations in F and E around their average values F_{average} and E_{average} are relatively small, we could work with the linearized system of stochastic differential equations in which the randomness is additive:

$$\begin{aligned}
dF_t &= -(\lambda_{FB} + \lambda_{FE})F_t\,dt + dU_t - \sigma_{FB}\,dW_t^{FB} - \sigma_{FE}\,dW_t^{FE},\\
dE_t &= [\lambda_{FE}F_t - (\lambda_{EB} + \lambda_{ER})E_t]\,dt + \sigma_{FB}\,dW_t^{FB} - \sigma_{EB}\,dW_t^{EB} - \sigma_{ER}\,dW_t^{ER},\\
dY_t &= \lambda_{ER}E_t\,dt + \sigma_{ER}\,dW_t^{ER}.
\end{aligned} \tag{2}$$

Here $\sigma_{FB} = \sqrt{\lambda_{FB}F_{\text{average}}}$, etc.

Assuming U is sufficiently smooth, we may write $dU = u(t)\,dt$. We may write (2) in abstract form as

$$\begin{aligned} dX(t) &= A_1 X(t)\,dt + Bu(t)\,dt + \Sigma_1\,dW_t, \\ dY(t) &= A_2 X(t)\,dt + \Sigma_2\,dW_t. \end{aligned} \tag{3}$$

where

$$\begin{aligned} X(t) &= \begin{bmatrix} F_t \\ E_t \end{bmatrix}, & Y(t) &= Y_t, \\ A_1 &= \begin{bmatrix} -(\lambda_{FB}+\lambda_{FE}) & 0 \\ \lambda_{FE} & -(\lambda_{EB}+\lambda_{ER}) \end{bmatrix}, & A_2 &= \begin{bmatrix} 0 & \lambda_{ER} \end{bmatrix}, \\ B &= \begin{bmatrix} 1 \\ 0 \end{bmatrix}, & \Sigma_2 &= \begin{bmatrix} 0 & 0 & 0 & \sigma_{ER} \end{bmatrix}, \quad \text{and} \\ \Sigma_1 &= \begin{bmatrix} -\sigma_{FB} & -\sigma_{FE} & 0 & 0 \\ \sigma_{FB} & 0 & -\sigma_{EB} & -\sigma_{ER} \end{bmatrix} \end{aligned}$$

and W is a four-dimensional standard Brownian motion.

3.2.2 ODE approximation / fluid limit

For the mean value behaviour of either of the models (1) or (2), we obtain the system of ordinary differential equations

$$\begin{cases} \dot{f}(t) &= -(\lambda_{FB}+\lambda_{FE})f(t) + u(t), \\ \dot{e}(t) &= \lambda_{FE} f(t) - (\lambda_{EB}+\lambda_{ER})e(t), \\ y(t) &= \lambda_{ER} e(t), \end{cases}$$

where $f(t) = \mathbb{E}F_t$, $e(t) = \mathbb{E}E_t$, but $y(t) = \frac{d}{dt}\mathbb{E}Y_t$. Let $\lambda_F := \lambda_{FB} + \lambda_{FE}$ and $\lambda_E := \lambda_{EB} + \lambda_{ER}$. The solution to this system of ordinary differential equations is given by

$$\begin{aligned} f(t) &= \exp(-\lambda_F t) f(0) + \int_0^t \exp(-\lambda_F (t-s)) u(s)\,ds, \\ e(t) &= \exp(-\lambda_E t) e(0) + \int_0^t \exp(-\lambda_E (t-s)) f(s)\,ds, \\ y(t) &= \lambda_{ER} e(t). \end{aligned}$$

This equation of f tells us the following. As $\lambda_F \in (0,1)$, the first term on the right hand side decays exponentially fast to zero. This means that the dependence on the initial value of the number of full bottles in the system will not matter. The second term says that the value of $f(t)$ depends on the history of the average number of the inflow of bottles (i.e. $u(s)$ for $s \in [0,t]$), where the dependence is the strongest at the most recent history.

To interpret e, we need some more calculation. Just as in the interpretation of f, we can conclude from the first term in the right hand side of the expression for $e(t)$ that the dependence on the initial data decays exponentially. For the second term, we separate 3 cases:

$\lambda_F = \lambda_E$, In this case, the probability that a full bottle leaves the FULL state in a certain time period, equals the probability that an empty bottle leaves the EMPTY state in the same period. We get

$$\begin{aligned} e(t) - \exp(-\lambda_E t)e(0) &= \int_0^t \exp(-\lambda_E(t-s))f(s)\,ds \\ &= t\exp(-\lambda_F t)f(0) + \int_0^t (t-s)\exp(-\lambda_E(t-s))u(s)\,ds, \end{aligned}$$

from which we learn that also $e(t)$ depends on the history of $u(t)$, but with a certain delay. This expression also shows that the dependence of $f(0)$ on $e(t)$ vanishes exponentially fast in the long run, but at the start there is an increasing dependence.

$\lambda_F \neq \lambda_E$. In this case, we obtain

$$\begin{aligned} &e(t) - \exp(-\lambda_E t)e(0) = \int_0^t \exp(-\lambda_E(t-s))f(s)\,ds \\ &= \frac{\exp(-\lambda_F t) - \exp(-\lambda_E t)}{\lambda_E - \lambda_F}\exp(-\lambda_F t)f(0) \\ &\quad + \int_0^t \frac{\exp(-\lambda_F(t-s)) - \exp(-\lambda_E(t-s))}{\lambda_E - \lambda_F}\exp((\lambda_E - \lambda_F)s)\,u(s)\,ds, \end{aligned}$$

from which we see again that the influence of $f(0)$ decays exponentially. The second term denotes the cumulative and delayed dependence on the input stream $u(t)$.

The interpretation of the number of bottles that is expected to be returned per time unit, i.e. $y(t)$, is just a fixed fraction of $e(t)$, of which we discussed its behaviour above.

3.3 State estimation: Kalman filtering

Let us consider the stochastic model again (see (1) for the time-discrete model and (2) for the continuous in time model). Before stating how we can get information out of the data by using these stochastic models, we would like to discuss the basic theory behind Kalman filtering.

Suppose a random variable Y has a conditional distribution depending on 'hidden state' X and 'input' U; loosely denoted as $p(Y|X,U)$. Furthermore suppose X and U are distributed according to some *'prior' distribution* $p(X,U)$. Bayes' formula gives

us that, given observations of U and Y, we may compute X as

$$p(X = x|Y = y, U = u) = \frac{p(X = x, Y = y, U = u)}{p(Y = y, U = u)} = \frac{p(Y = y|X = x, U = u)p(X = x, U = u)}{\sum_x p(Y = y|X = x, U = u)p(X = x, U = u)}.$$

In other words, based on the conditional distribution $p(Y|X, U)$ and the prior distribution $p(X, U)$, we may compute a *'posterior' distribution* $p(X|Y, U)$. This posterior distribution enables us to estimate the hidden state X based on observations of U and Y.

The same idea may be applied recursively to systems of the form (3), leading to the *Kalman-Bucy filter* [4, 7], or, in discrete time, the *Kalman filter* [1, 9]. Such a filter allows us in this example to obtain estimates $\hat{F}_t$, $\hat{E}_t$ of F_t and E_t, based on observations of U_t and Y_t. Kalman filters appear notationally invovled, but once the dynamic model (such as (1) or (2)) is identified, implementation of such a filter is relatively straightforward. It gives us estimates for $\mathbb{E}E_t$, $\mathbb{E}F_t$, $\operatorname{Var} E_t$ and $\operatorname{Var} F_t$, which become more accurate for larger t. It is therefore preferable to use a data set with a long time series.

3.4 Estimation of the model parameters

To complete our model, we need to estimate the $\lambda_{\dots}$ parameters. In this section we demonstrate an estimation method based on the data of the sampling.

3.4.1 Estimation through distribution of circulation times

Conditional on the eventual return of a bottle, we have to wait time $T_E \sim \exp(\lambda_{FE})$ before a bottle is being emptied, plus a time $T_R \sim \exp(\lambda_{ER})$ before the empty bottle is returned. The total waiting time $T = T_E + T_R$ is then the sum of two exponential random variables, and has a *hypoexponentially distribution* with parameters λ_{FE} and λ_{ER}.

For convenience, we write $\lambda_1 = \lambda_{FE}$ and $\lambda_2 = \lambda_{ER}$. Since the random variables T_E and T_R are independent, the mean of T is $\mathbb{E}T = \frac{1}{\lambda_1} + \frac{1}{\lambda_2}$ and the variance is $\operatorname{Var}(T) = \frac{1}{\lambda_1^2} + \frac{1}{\lambda_2^2}$. The distribution function of T may be computed (in case $\lambda_1 \neq \lambda_2$)

as

$$\begin{aligned} F_T(t) &= \mathbb{P}(T_E + T_R \leq t) = \int_0^t \mathbb{P}(T_E + T_R \leq t | T_e = s) f_{T_e}(s)\, ds \\ &= \int_0^t \mathbb{P}(T_r \leq t - s) \lambda_1 \exp(-\lambda_1 s)\, ds \\ &= \lambda_1 \int_0^t (1 - \exp(-\lambda_2 (t-s))(\exp(-\lambda_1 s)\, ds \\ &= 1 + \frac{1}{\lambda_1 - \lambda_2} \left(\lambda_2 \exp(-\lambda_1 t) - \lambda_1 \exp(-\lambda_2 t)\right), \quad t \geq 0. \end{aligned}$$

The density function is then

$$f_T(t) = \frac{d}{dt} F_T(t) = \frac{\lambda_1 \lambda_2}{\lambda_1 - \lambda_2} \left(\exp(-\lambda_2 t) - \exp(-\lambda_1 t)\right), \quad t \geq 0.$$

In case $\lambda_1 = \lambda_2$, a similar computation gives $F_T(t) = 1 - \exp(-\lambda_1 t) - \lambda_1 t \exp(-\lambda_1 t)$ and $f_T(t) = \lambda_1^2 t \exp(-\lambda_1 t)$ for $t \geq 0$.

Given n i.i.d. observations $t_1, \ldots, t_n$ of a hypoexponentially distributed random variable, we can estimate the parameters λ_1 and λ_2 in two ways:

(i) By maximizing the (log)-likelihood function

$$l(\lambda_1, \lambda_2) = \sum_{i=1}^{n} \ln \left(\frac{\lambda_1 \lambda_2}{\lambda_1 - \lambda_2} \left(\exp(-\lambda_2 t_i) - \exp(-\lambda_1 t_i)\right) \right), \quad \lambda_1, \lambda_2 > 0$$

with respect to λ_1 and λ_2. This will always provide estimates for λ_1 and λ_2, but needs to be performed numerically.

(ii) By the method of moments: choose λ_1 and λ_2 so that the sample mean and variance match the computed expectation and variance. Let $\hat{\sigma}^2$ denote sample variance and $\hat{\mu}$ denote sample mean. Write $a_1 = 1/\lambda_1$ and $a_2 = 1/\lambda_2$. By the above expressions for mean and variance of T, we find the conditions $a_1^2 + a_2^2 = \hat{\sigma}^2$, and $a_1 + a_2 = \hat{\mu}$. This results in the expression $a_{1,2} = \frac{1}{2}\hat{\mu} \pm \sqrt{\frac{1}{2}\hat{\sigma}^2 - \frac{1}{4}\hat{\mu}^2}$, so that

$$\lambda_{1,2} = \frac{1}{a_{1,2}} = \left(\tfrac{1}{2}\hat{\mu} \pm \sqrt{\tfrac{1}{2}\hat{\sigma}^2 - \tfrac{1}{4}\hat{\mu}^2} \right)^{-1}.$$

Note that these estimates become non-sensical in case $\frac{1}{2}\hat{\sigma}^2 - \frac{1}{4}\hat{\mu}^2 < 0$ or if $\sqrt{\frac{1}{2}\hat{\sigma}^2 - \frac{1}{4}\hat{\mu}^2} \geq \frac{1}{2}\hat{\mu}$. This means that we require

$$\tfrac{1}{2}\hat{\mu}^2 \leq \hat{\sigma}^2 < \hat{\mu}^2,$$

which will not hold in all cases. This is a limitation of the method of moments, whereas likelihood maximization will provides an estimate for all cases. It is

however also an indication that the proposed model does not need to be a perfect fit for the observed data. In Figure 3, the frequency data of bottle circulation times is compared with the best hypoexponential fit, using the method of moments.

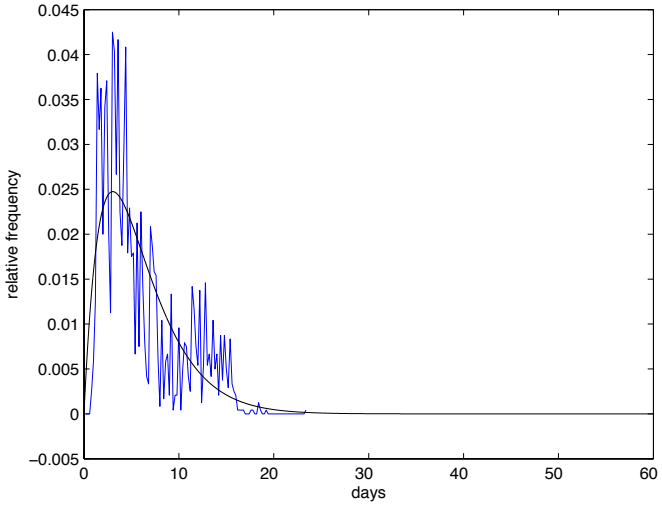


Figure 3: Frequency data of circulation times and the hypoexponential fit according to the method of moments.

3.4.2 Estimation based on stationarity assumption – 2 state model

Consider (1) in a stationary regime. For simplicity assume $h = 1$ and all transition probibilities $\lambda_{\dots} \ll 1$. We assume the number of bottles sold in a time interval $[t, t+1)$ equals $U(t+h) - U(t) \sim N(\mu_U, \sigma_U^2)$. Furthermore suppose $F \sim N(\mu_F, \sigma_F^2)$, $E \sim N(\mu_E, \sigma_E^2)$, and $Y(t+1) - Y(t) \sim N(\mu_Y, \sigma_Y^2)$. By the discrete time equations (1) (with $h = 1$), we immediately find

$$\mu_F = \frac{\mu_U}{\lambda_{FE} + \lambda_{FB}}, \quad \mu_E = \frac{\lambda_{FE}\mu_F}{\lambda_{EB} + \lambda_{ER}}, \quad \mu_Y = \lambda_{ER}\mu_E,$$

giving

$$\mu_Y = \frac{\lambda_{ER}\lambda_{FE}}{(\lambda_{EB} + \lambda_{ER})(\lambda_{FE} + \lambda_{FB})}\mu_U$$

and thus providing an equation for the unknown parameters $\lambda_{\dots}$ in terms of means μ_U and μ_Y, which can be estimated by the respective sample averages.

3.4.3 Estimation based on stationarity assumption – 1 state model

By a more involved analysis concerning covariances, extra equations may be obtained. We will explain this idea for a simplified model with only one recurrent state. It should in principle be possible, but more involved, to carry out the same analysis for the two-state model.

Consider the situation in which, within a single time step, a bottle can be broken (rate per unit time λ_B) or returned (rate λ_R). See Figure 4.

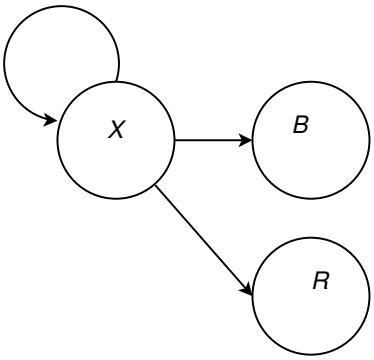


Figure 4: A simple Markov model with one absorbing state, used for determination of model parameters.

In a time interval $[t, t+1)$ a number of $U_{t+1} - U_t$ bottles is bought, independent of the number of unreturned bottles X_t. Assuming stationarity of the randomness as before, we have the following simple Markov system:

$$\begin{aligned} X_{t+1} &= (1 - \lambda_B - \lambda_R)X_t + U_{t+1} - U_t - \sigma_B \epsilon_t^B - \sigma_R \epsilon_t^R, \\ Y_{t+1} - Y_t &= \lambda_R X_t + \sigma_R \epsilon_t^R. \end{aligned}$$

We further simplify this model by assuming $\sigma_B = \lambda_B \mu_X$, $\sigma_R = \lambda_R \mu_X$, since the variance of a $\mathrm{Bin}(n, \lambda)$ random variable is proportional to $n\lambda(1-\lambda) \doteq n\lambda$ for small λ. This model has three unknowns $\lambda_B, \lambda_R, \mu_U$. Using the same argument as before, we can relate the empirical means $\hat{\mu}_Y$ and $\hat{\mu}_U$ of μ_Y and μ_U through the equality $\mu_Y = \frac{\lambda_R}{\lambda_B + \lambda_R} \mu_U$. Furthermore,

$$\begin{aligned} \sigma_X^2 + \mu_X^2 = \mathbb{E}X_{t+1}^2 &= \mathbb{E}\left[\left\{ (1 - \lambda_B - \lambda_R)X_t + U_{t+1} - U_t - \sigma_B \epsilon_t^B - \sigma_R \epsilon_t^R \right\}^2 \right] \\ &= (1 - \lambda_B - \lambda_R)^2 (\sigma_X^2 + \mu_X^2) + \sigma_U^2 + \mu_U^2 + \sigma_B^2 + \sigma_R^2, \end{aligned}$$

or equivalently

$$\left\{ 1 - (1 - \lambda_B - \lambda_R)^2 \right\} (\sigma_X^2 + \mu_X^2) = \sigma_U^2 + \sigma_B^2 + \sigma_R^2.$$

Therefore

$$\mathbb{E}(Y_{t+1} - Y_t)^2 = \sigma_R^2 + \lambda_R^2 \mathbb{E}X_t^2 = \sigma_R^2 + \frac{\lambda_R^2 \left(\sigma_U^2 + \sigma_B^2 + \sigma_R^2\right)}{1 - (1 - \lambda_B - \lambda_R)^2}$$
$$= \lambda_R^2 \left(\mu_X^2 + \frac{\sigma_U^2 + \mu_X^2(\lambda_B^2 + \lambda_R^2)}{1 - (1 - \lambda_B - \lambda_R)^2}\right).$$

Finally, by similar reasoning,

$$\rho = \mathbb{E}[(Y_{t+1} - Y_t)(U_t - U_{t-1})] = \lambda_R(\mu_U^2 + \sigma_U^2),$$

where the quantity on the lefthand side may be estimated from the data as

$$\hat{\rho} = \frac{1}{n} \sum_{i=1}^{n} y_i u_{i-1},$$

where u_i and y_i are the observed sales and returns in time period i, respectively. To summarize we have obtained three equations that relate the unknowns λ_B, λ_R and μ_X in terms of $\hat{\sigma}_U^2$, $\hat{\mu}_U$, $\hat{\sigma}_Y^2$, $\hat{\mu}_Y$, and $\hat{\rho}$.

3.5 Discussion

The above results are for illustration purposes. A more detailed analysis should be performed to determine which simple model might adequately describe the dynamics of the system. From such a model, equations should be derived as in the last section which estimate system parameters from observed statistics. Then state estimation may be performed on-line to compute actual estimates of number of full and empty bottles in the system, using Kalman filtering, based on observations of sold bottles and returned bottles, or perhaps using observations of circulation times.

In the full-for-empty system, in which full bottles are only sold once the same number of empty bottles is returned, extra modelling is necessary. A simple Markov model might than model the behaviour of a customer, who in a time period may drink a unit, do nothing, or buy a bottle and thus also return bottles.

4 A queueing model for the number of bottles in the market

In Section 3, we described a rather detailed modelling approach to the description of bottles in the market. We now discuss alternative models from queueing theory that have been thoroughly investigated in the literature and are by now well understood. These models stem from different modelling assumptions on the demand process (a compound Poisson process) and particularly focus on fluctuations of the demand rate in time. Another important difference is that these models only describe bottles that will be returned; the rate at which bottles break should be discounted in the demand

process. That is, if the demand rate is x and a fraction p of the bottles are broken, then $(1-p)x$ is the effective demand rate of bottles that will be returned.

In this section we describe the infinite server model from queueing theory that serves as a building block to model the number of bottles out in the market. The following is required as *input* for this model: The effective demand rate (function) and the distribution of d_t, the circulation time (cf. the definition in Section 2). The output is a distribution for the number of bottles that are simultaneously out in the market.

4.1 Constant demand rate, fixed circulation time distribution

We start by assuming that the demand has a constant rate and can be modeled by a Poisson process. Specifically, we start by assuming that the number of bottles purchased in an interval of t days has a Poisson distribution with an average of λt bottles, where λ is the daily average demand rate. Letting D denote a generic random variable having the distribution of the circulation time of a bottle we find that N –the number of bottles out in the market (in the stationarity regime)– has a Poisson distribution with mean $\lambda \mathbb{E}D$. The distribution of D enters the calculation only through its mean; this is usually referred to as "insensitivity" toward the circulation time distribution. The main requirement is that the time out in the market for each bottle is independent of that of all other bottles and shares the distribution of D. This result is standard in queueing theory and can be found in any textbook, see for example [5].

In reality, both the demand rate and the circulation time is season-dependent and/or may have a certain non-stationary trend. We discuss these in the following two paragraphs.

4.2 Time varying demand rate, fixed circulation time distribution

Assume now that the demand rate fluctuates over time. At time t it is $\lambda(t)$, i.e., the demand process is a time-varying Poisson process with rate function $\lambda(t)$. Still the number of bottles out in the market has a Poisson distribution, but this is no longer insensitive to the circulation time distribution. Now, the number of bottles out in the market at time t has a Poisson distribution with mean

$$\mathbb{E}N(t) = \int_0^\infty \mathbb{P}(D > v)\, dv = \int_0^\infty \left(\int_{t-v}^{t} \lambda(u) du \right) f_D(v)\, dv, \tag{4}$$

where $f_D(t)$ is the density of the distribution of the circulation time D; see for example Theorem 1 of [3]. For a constant demand rate, we recover from equation (4) that $\mathbb{E}N(t) = \lambda \mathbb{E}D$.

4.3 Time varying demand rate, time varying exponential circulation time distribution

If we restrict on the generality of the circulation time distribution and assume it has an exponential distribution, then a rather classical paper by [2] generalizes the previous to

the case where the mean of the circulation time distribution may fluctuate over time. Again the number of bottles in the market at time t is Poisson with mean $J(t)I(t)$ where

$$J(t) = e^{-\int_0^t \nu(u)du}, \tag{5}$$

with $1/\nu(t)$ representing the mean circulation time at time t, and

$$I(t) = \int_0^t \frac{\lambda(u)}{J(u)}du. \tag{6}$$

Of course, if we choose a fixed exponential circulation time by setting $\nu(t) = \frac{1}{\mathbb{E}D}$ in (5) and setting $f_D(t) = \frac{1}{\mathbb{E}D}e^{-\frac{1}{\mathbb{E}D}t}$ in (4), we obtain the same result.

For the data available from Heineken, this is probably the most useful model. In case more is known about the characteristics of the demand function and circulation time distribution, a useful extension to the above can be found in [6]. The model there allows for time varying non-exponential demand and circulation time distributions (specifically, they allow for phase type distributions).

5 Sampling the circulation time

A key quantity to understand is the expectation for the circulation time of bottles (see Section 2.2.2 for a definition of circulation time). We describe a method to obtain this value, together with a confidence interval depending on the sample size. We also discuss how other fluctuations in the beer market, like seasonality, can be incorporated in the method to improve the estimations.

The statistical theory of estimation, sampling, and confidence intervals is well-developed. In line with this theory, we consider the circulation times of individual bottles as *random variables* with *identical* distribution. Shape or parameters of this distribution are obtained by means of sampling, that is, computation of the circulation time for a small number of bottles (a sample), and extrapolation of the findings for this sample to the entire population of bottles.

The practical side of sampling is easy: When a bottle is returned, the expiry date on the label allows calculation of the production date, which in turn gives a sound estimate of the time of sale. Together with the time of return of the bottles, this allows a fairly exact computation of the circulation time. However, a high-volume or even continuous computation of circulation times in this way is expansive and impractical. Therefore, we first discuss in this section the required sample sizes in order to guarantee a certain confidence limit for the parameters. Subsequently, we discuss how seasonality and other artifacts of the beer market can be incorporated in order to improve the estimation.

A standard assumption to facilitate the statistical analysis is (mutual) independence of the circulation time of the bottles in a sample. Therefore, the choice of the sample should be made as random as possible.

5.1 Batch sizes

We address now the question how many bottles should be sampled in order to guarantee a certain accuracy of the circulation time. Suppose we have a sample of N different bottles with circulation times $X_1, X_2, \ldots, X_N$. Under the assumption that $X_1, \ldots, X_N$ are *independently and identically distributed* (i.i.d.), we use the sample to infer on the (unknown) distribution of circulation time. In parametric statistics, one assumes a certain *family of distributions* indexed by a finite-dimensional parameter space. It is then sufficient to estimate these parameters.

Usefulness of this approach is crucially relying on a decent choice of the family of distributions. The often used *normal family* identifies a 95%-confidence interval for the mean as all points at distance smaller than 1.96 times sample standard deviation from the sample mean.

In light of Section 3.4, it seems most reasonable to choose as model the hypoexponential distribution with two parameters. Derivation of confidence limits in closed form, such as for the normal family, seems impossible for this model. Nevertheless, *bootstrapping* provides a theoretically not very pleasing, yet very efficient, practical method to determine confidence interval by means of Monte Carlo simulation. This works as follows. Start by estimating the two parameters of the hypoexponential distribution by using either of the methods (i) or (ii) in Section 3.4. Use then a statistical software package to generate a high number, say 1000, of i.i.d. random variables with this distribution using the estimated parameters, and sort them from smallest to largest (call them $Y_1, \ldots, Y_{1000}$). The $(1-\alpha)\%$-confidence interval for the circulation time as given by the bootstrap is the interval $\left[Y_{\alpha/2\times 1000}, Y_{(1-\alpha/2)\times 1000}\right]$.

5.2 Seasonality

An artifact of the beer market is seasonality. Sales show a certain seasonal peak, typically located in summer, when people drink more beer than in other times of the year. Particularly in *full-for-empty* systems for returnable packaging materials, it is tempting to believe that customers operate with more bottles during peak time, and store some bottles elsewhere throughout the rest of the year. If this reasoning is true, then seasonality has an impact on the circulation time (at the start of the peak, customers bring the stored bottles, which yields a higher circulation time).

We propose to carry out a statistical test whether the null hypothesis H_0: "Seasonality has no significant effect on circulation time" can be rejected in favor of the alternative hypothesis H_1: "Seasonality does have a significant effect on circulation time". A possible test could go as follows. Gather data at several moments in the year, e.g. monthly, bi-monthly or quarterly. Call k the number of measurements per year, and K the total number of measurements. Record the following data:

Y_n	circulation time at time n
$\bar{Y} = 1/K \ \sum_{n=1}^{K} Y_i$	mean circulation time,
X_n	sales volume at time n,
$n = 1, 2, \ldots, K$	time.

We are now considering the general linear model in centralized form:

$$Y_n - \bar{Y} = \beta_1(X_n - X_{n-1}) + \beta_2\big(X_n - \sum_{i=1}^{k} X_{n-i}\big) + \varepsilon_n, \qquad n = 1, \ldots, K, \quad (7)$$

where the error terms ε_n $(n = 1, \ldots, K)$ are i.i.d. normally distributed. In this model, β_1 'explains' derivation in circulation time by an upward or downward sales trend (mimicking the beginning or end of a peak period). Further, $X_n - \sum_{i=1}^{k} X_{n-i}$ is large if we are in a peak, and small otherwise. Hence, β_2 simply relates circulation time to peaks. Of course, the above model could easily be adapted to account for other effects, for example, incorporating long-term trends in sales. Mind that the Y_n itself are sample means, which justifies our normality assumption for the ε_n.

With this model at hand, our earlier described null hypothesis can be sharpened as

$$H_0 \colon \beta_1 = \beta_2 = 0.$$

In order to test the hypothesis, we use multilinear regression obtaining regression sum of squares (RSS) and error sum of squares (ESS). Dividing both by their corresponding degree of freedom (2 for RSS, $K-2$ for ESS), and comparing the ratio of these two with the $F(2, n-2)$-distribution obtains the p-value associated with the data. This is known in the statistical literature as ANOVA ('analysis of variance'). Further tests could be imposed if the null hypothesis has been rejected at the desired level of confidence.

A somewhat simpler approach would concentrate only on the 'peak' phenomenon. This time, we take only two samples per year, one at the beginning of the peak, and the other one at the end of the peak. We estimate parameters for each of these two samples separately, and compare how 'distant' they are using the methods described in the next subsection. This method is simpler than the one described before, but sheds no light on other (possible) temporal dependencies.

5.3 Handling different distribution channels

Circulation time may very well depend on the distribution channel. The main difference we shall consider here are the channels bars/restaurants on the one hand, and private customers on the other hand. Similar to the discussion in the previous subsection, we suggest a statistical test to investigate the issue. The general setup is somewhat easier this time. We are in the situation where we have two samples, with two estimated sets of parameters, and now we want to test whether they are "significantly different" in order to reject the null hypothesis H_0: "There is no difference in circulation time parameters for different distribution channels." The F-test (ANOVA) is the right one under the assumption of normality. However, there are also generally applicable non-parametric tests, for details we refer to Section 11.2 in [8].

5.4 Unreadable labels – an indicator for a long circulation time

A practical problem that occurs at the sampling procedure, is that the expiration date is not readable on some bottles. This is due to damage to the label, or a completely

removed label. It is likely that these bottles have a longer circulation time than the bottles with readable expiration dates, so it would bias the statistics if one leaves these bottles out of the sample. Although there are methods available to reduce the bias compared to leaving these bottles out of the sample (for example, the EM algorithm), we did not look into this any further.

6 Conclusion

In Section 2.3, we discussed a way of estimating the break rate of bottles from data that is currently available. This is already an interesting result in itself, but can also be used to identify some parameters in the stochastic flow models (see (1) and (3)) and in the queueing model (see Section 4.3). In Section 3 and Section 4, we then studied two different kinds of models for the number of bottles in the market. All of our models are quite simple, and might therefore not be very accurate in practice. Further (statistical) research is required to test their accuracy. It may be necessary to increase model complexity by dropping some assumptions. However, we note that this may result in a) lack of explicit solutions of the models, b) more unknown parameters that need to be estimated and/or c) increased computational effort.

In Section 5 it is discussed how to obtain the expected circulation time of a bottle in the market from the data, together with a confidence interval. As this value may depend on the time in the year at which the sampling has been done, we also discussed a method to test whether it is reasonable to assume that this seasonality is of little importance.

References

[1] D. P. Bertsekas. *Dynamic programming and optimal control. Vol. I*. Athena Scientific, Belmont, MA, third edition, 2005. ISBN 1-886529-26-4.

[2] T. Collings and C. Stoneman. The $m/m/\infty$ queue with varying arrival and departure rates. *Operations Research*, 24(4):760–773, 1976.

[3] S. G. Eick, W. A. Massey, and W. Whitt. The physics of the $m_t/g/\infty$ queue. *Operations Research*, 41(4):731–742, 1993.

[4] W. H. Fleming and R. W. Rishel. *Deterministic and stochastic optimal control*. Springer-Verlag, Berlin, 1975. Applications of Mathematics, No. 1.

[5] D. Gross and C. M. Harris. *Fundamentals of Queueing Theory*. Wiley New York, 1998.

[6] B. L. Nelson and M. R. Taaffe. The $ph_t/ph_t/\infty$ queueing system: Part i – the single node. *INFORMS journal on computing*, 16(3):266–274, 2004.

[7] B. Øksendal. *Stochastic differential equations*. Universitext. Springer-Verlag, Berlin, sixth edition, 2003. ISBN 3-540-04758-1. An introduction with applications.

[8] J. A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury press, 2007.

[9] R. F. Stengel. *Optimal control and estimation*. Dover Publications Inc., New York, 1994. ISBN 0-486-68200-5. Corrected reprint of the 1986 original.

Value-at-Risk of coffee portfolios

S. Gugushvili J. Nowotarski C. W. Oosterlee
L. Ortiz-Garcia E. Verbitskiy*

Abstract

Coffee is the second most traded commodity in the world, and the coffee market can be very volatile even over very short periods of time. Nedcoffee was looking for better ways to access Value-at-Risk of their portfolio. Additional difficulty stems from the fact that portfolio comprises of contracts with delivery dates in January, March, May, July, September and November of each year. Our proposed solution was evaluated using historical market data.

KEYWORDS: coffee, market volatility, value-at-risk, modeling

1 Introduction

With sales volume equal to 106000 metric tons in the fiscal year 2012, Nedcoffee is a major coffee trader with headquarters located in Amsterdam, from which it trades and controls all its green coffee from its sourcing companies in Africa, Asia and South America. Coffee market is a volatile market with traders managing highly complex portfolios. A problem of paramount importance in risk management in general and for Nedcoffee in particular is estimation of a profit and loss distribution of a portfolio over a specified time horizon and the associated risk measures. Value-at-Risk (VaR) has become an important measure for estimating and managing portfolio market risk; see Jorion (2007) for a detailed exposition. VaR is defined as a certain quantile of the change in value of a portfolio during a specified holding period. While the basic concept of VaR is simple, many complications can arise in practical use. Of these part are statistical: VaR is not an absolute, but a model-dependent quantity. Choosing a right probability distribution for an adequate description of the profit and loss distribution is thus of great importance. However, models will typically depend on parameters, which have to be inferred from the data and uncertainty in which will propagate to estimates of VaR. A further complication is that when determining VaR, one is estimating a quantile far in the tail of the distribution, which is a notoriously difficult statistical task. There is also a problem of a different, conceptual nature, that is inherent in the definition of VaR: it is an incoherent risk measure, which in non-technical terms means that a diversified portfolio might have a higher VaR than

*Corresponding author.

prior to diversification, the fact which traders will be reluctant to accept; see Artner et al. (1999) for details. Despite this, the use of VaR is extremely widespread in practice.

Since trading decisions at Nedcoffee are to a considerable extent determined by VaR considerations, the company is greatly interested in 1) constructing better models to be used in VaR computations than currently used by the company, and 2) given a model, using statistically efficient tools for the actual computation of VaR.

The rest of the report is organized as follows: in Section 2 we recall the definition of VaR and introduce some notation. In Section 3 we briefly review the approach employed by Nedcoffee and indicate its shortcomings. Sections 4 and 5 outline some alternatives and present small scale simulation study results for one of them. Section 6 concludes with an outlook and some future work.

2 VaR

Assume a portfolio consists of positions in k different assets, and let $N_i(t)$ and $P_i(t)$ denote respectively the number of contracts and the price of one contract in the ith position, $i = 1, \ldots, k$, at time t. The price of portfolio at time t is then

$$S(t) = \sum_{i=1}^{k} N_i(t) P_i(t).$$

Let Δt be the the holding period of the portfolio, so that the portfolio composition remains constant over the time period $[t, t + \Delta t]$, i.e. $N_i(t) = N_i(t + 1)$. The value of the portfolio at time $t + \Delta t$ is $S(t + \Delta t)$. The change in the portfolio value during the holding period is

$$\Delta S = S(t + \Delta t) - S(t) = \sum_{i=1}^{k} N_i(t) \Delta P_i(t) = \sum_{i=1}^{k} N_i(t) R_i(t) = \langle \mathbf{N}(t), \mathbf{R}(t) \rangle, \qquad (1)$$

where $\langle \mathbf{N}(t), \mathbf{R}(t) \rangle$ is the scalar product of vectors $\mathbf{N}(t)$ and $\mathbf{R}(t)$ with components $N_i(t)$'s and $R_i(t)$'s, respectively. The VaR_α risk measure, associated with a given level $0 < \alpha < 1$, is defined by the relation

$$\mathbb{P}(\Delta S < -\mathrm{VaR}_\alpha | \mathbf{N}(t)) = \alpha. \qquad (2)$$

Thus

$$\mathrm{VaR}_\alpha = |F^{-1}(\alpha)|,$$

where F is the distribution function of $\Delta S(t)$ given $N(t)$. In practice Δt typically ranges from one day to two weeks and $\alpha \leq 0.05$, often $\alpha = 0.01$.

In the case of Nedcoffee the portfolio consists of various types of coffee futures and some options written on them. In this paper we will for simplicity assume that the Nedcoffee portfolio consists of futures only. Ignoring options, in principle k can be

as large as 10, which corresponds to two major coffee species, Arabica and Robusta, and five possible contract listings per species. Nedcoffee is primarily interested in estimating the 1-day VaR of their portfolio, so that $\Delta t = 1$ day. The level α they aim at is somewhat unrealistically set at $\alpha = 0.015$. In our argumentation we will use a general α.

3 Nedcoffee approach

From formulae (1) and (2) it is obvious that VaR depends on the choice of the model for the futures price process $P(t) = (P_1(t), \ldots, P_k(t))$. A number of possibilities are available here.

NEDCOFFEE employ currently an empirical formula for Value-at-Risk at confidence level $\alpha = 0.9985$ (0.15 percent) based on the assumption that underlying Arabica and Robusta coffee prices are separately normally distributed with constant covariance matrices over a period of 3 months (≈ 60 trading days). We are not going to discuss precise details of the method, but will use the NEDCOFFEE VaR estimate for benchmarking purposes.

Given the assumption of Gaussian distribution of prices holds, covariance matrices can be easily estimated using the available historical data on futures prices and then the VaR can be determined in a straightforward fashion. However, except for simplicity of computation, there is little empirical justification for assumptions made in this case. As an illustration of this, we produced a normal Q-Q plot based on returns of the 2-month futures from 15 November 1993 to 7 February 1994, which gives us in total 59 data points. Strong deviation from normality is visible in the plot. Also a formal test for normality, the Shapiro-Wilk test, performed on the same dataset yields the p-value equal to 0.002392, which is very strong evidence against the null hypothesis that the data originate from a certain normal distribution. We would reject the null hypothesis at level 0.05.

4 Possible alternatives

The results from the previous section indicate that one has to look for alternative models and VaR computation methods than those currently employed by Nedcoffee. Two natural options are: a continuous-time model, in which P is a solution to a (multidimensional) stochastic differential equation (SDE), or a time series model, such as a (multivariate) GARCH model. Models based on SDEs are attractive due to the fact that under suitable assumptions a streamlined theory for pricing financial derivatives (e.g. options) is available for them. Furthermore, they are capable of reproducing the mean reversion property one often sees in asset prices (this is achieved through appropriately choosing the drift coefficient of the equation), as well as fitting a wide range of return distributions (this is achieved by selecting a right diffusion coefficient, or by using a general Lévy process instead of the Brownian motion as a driving process of the equation). On the other hand a very fine level of detail provided by sample paths of

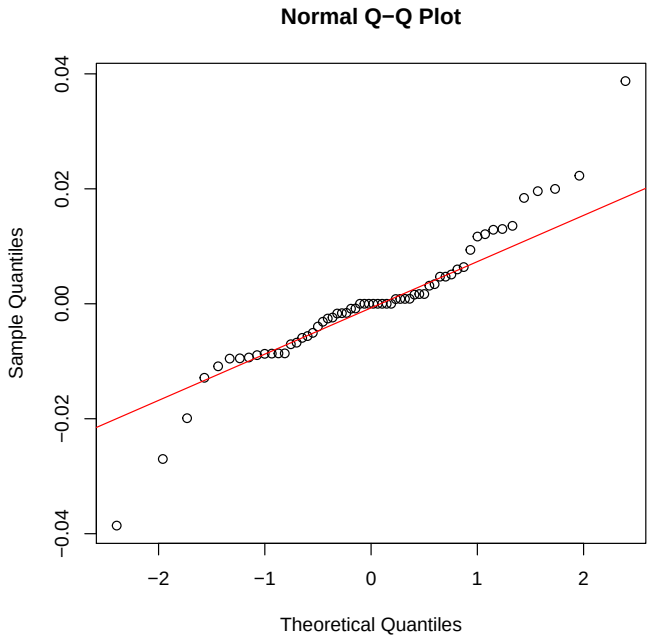


Figure 1: Normal Q-Q plot for the returns based on the 2-month futures data (15/11/1993 to 07/02/1994).

SDE models is not always warranted to be seen in actual financial time series; see e.g. Carr et al. (2002). When asset prices are observed at high frequency, microstructure noise becomes a problem. Moreover, parameter estimation in SDE models, especially in the high-dimensional case when both the dimension of the system of SDEs and of the parameter space are large, is computationally and statistically a very difficult task, unless one restricts attention to simple, but often not realistic models, such as e.g. the Black-Scholes model. In the case of the Black-Scholes asset price dynamics, provided the model parameters have been accurately estimated, VaR can be efficiently computed following the method described in Ortiz-Garcia and Oosterlee (2013) (an extra technical complication in our case would be the fact that we are dealing with futures prices). We refer to the same paper for additional references. As far as the time series models are concerned, multivariate generealisations of traditional univariate models are far from trivial due to the fact that the multivariate character of the model greatly increases the number of parameters required for its description, while a drastic cut of the number of parameters due to parsimony considerations might well render the model inadequate for data description purposes; see e.g. Silvennoinen and Teräsvirta (2009). GARCH process is not an only option here; one can e.g. also consider the AR processes (either the classical or the semiparametric ones), but the same remarks apply.

5 Present approach

Below we propose an approach to VaR computation that in our opinion strikes a good balance between being computationally easy and still beter than the one currently employed by Nedcoffee.

Assume that the underlying asset prices are jointly normally distributed with a constant covariance matrix over a period of 3 months. In this case, the returns $\mathbf{R}(t)$'s have normal distribution with unknown variance and the standard VaR estimate can be applied.

Time series analysis of the returns suggests that the Student's t-distribution is a better fit than the normal distribution. The standard Student distribution is given by the density

$$f(x) = \frac{1}{\sqrt{\nu\pi}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where ν is the number of degrees of freedom (shape parameter). For our purposes the non-standardized Student's t-distribution with the density

$$f(x|\mu,\sigma,\nu) = \frac{1}{\sigma\sqrt{\nu\pi}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(1 + \frac{\left(\frac{x-\mu}{\sigma}\right)^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where μ is the mean (location), σ is the scale parameter, ν is the the number of degrees of freedom (shape parameter).

Note that for $\nu \to \infty$, the the non-standardized Student's distribution $F_{\mu,\sigma,\nu}$ approaches the normal distribution $\mathcal{N}(\mu, \sigma^2)$. Therefore, the application of Student's t-distribution has an advantage over the normal law when the distribution of the returns has lighter tails, and can recover the normal law, when necessary.

Again, we assume that $\mu = 0$, the corresponding VaR value is the product of the scale parameter σ and the appropriate quantile of the distribution.

The VaR models described above have been evaluated (backtested) for a number of different fictitious portfolios in the following fashion: for each trading day the corresponding VaR value has been computed, and the number of trading days when the loss exceeded the predicted VaR has been computed. Ideally, the fraction of such days should be close to the chosen confidence level α.

Here are the results for several portfolios and confidence levels α. We have tested two types of portfolios. For the first type, positions in Arabica and Robusta are long, and for the second type, one is long in Robusta, and short in Arabica. This choice corresponds to test the performance of NEDCOFFEE's estimator, since it is constructed differently for long/long and long/short portfolios. Secondly, we tested the length of the past period used in estimation of the parameters $M = 60$ days and $M = 90$ days. Finally, we performed backtesting for confidence levels $\alpha = 0.15, 1.5,$ and 5 percent.

Tables below give the performance of the estimators, most accurate in bold.

N=[3000,2000,100, 2000, 600, 200], M=60

Confidence Level	NEDCOFFEE	GAUSS	STUDENT
0.15	0.33	0.66	**0.25**
1.50	0.66	2.40	**1.74**
5.00	2.23	**5.21**	5.87

N=[3000,2000,100, 2000, 600, 200], M=90

Confidence Level	NEDCOFFEE	GAUSS	STUDENT
0.15	0.34	0.51	**0.17**
1.50	0.76	2.46	**1.61**
5.00	2.12	**5.09**	5.85

N=[3000,2000,100, -2000, 600, 200], M=60

Confidence Level	NEDCOFFEE	GAUSS	STUDENT
0.15	0.83	0.58	**0.41**
1.50	2.15	1.65	**1.49**
5.00	5.77	4.58	**5.00**

N=[3000,2000,100, -2000, 600, 200], M=90

Confidence Level	NEDCOFFEE	GAUSS	STUDENT
0.15	0.76	0.59	**0.42**
1.50	2.29	**1.53**	1.44
5.00	5.77	4.58	**5.00**

6 Conclusions and outlook

The current VaR estimator for low α severely overestimates the true VaR for long/long portfolios, and underestimates the VaR for mixed portfolios. Suggested extensions demonstrate better performance. In particular, Student's t-distribution offers significant improvement. We also have found that Nedcoffee should consider various levels of confidence, e.g., 1 or 5 percent. The current level of 0.15 percent seems too small to provide accurate risk assessment.

In this report we primarily discussed construction of VaR estimators based on univariate time series $S(t)$ or $\Delta S(t)$. Multivariate modelling of the returns $\boldsymbol{R}(t)$ might provide a better insight into the dynamics of underlying assets. Moreover, multivariate modelling opens a possibility for portfolios optimisation. Another important direction for future work is incorporation of options in the analysis similar to the one performed above.

References

P. Artzner, F. Delbaen, J.M. Eber, D. Heath (1999). Coherent measures of risk. *Mathematical Finance*, 9:203–228, 1999.

P. Carr, H. Geman, D. B. Madan and M. Yor (2002). The fine structure of asset returns: an empirical investigation. *Journal of Business*, 75:305–332.

P. Jorion (2007). *Value at risk: the new benchmark for managing financial risk.* McGraw-Hill, 3rd edition.

L. Ortiz-Garcia and C. W. Oosterlee (2013). Efficient VaR and expected shortfall computations for non-linear portfolios within the delta-gamma approach. Preprint.

A. Silvennoinen and T. Teräsvirta (2009). Multivariate GARCH Models. In T. G. Andersen, R. A. Davis, J. P. Kreiß and Th. V. Mikosch (Eds.), *Handbook of Financial Time Series*, Springer-Verlag, Berlin-Heidelberg, 2009, pages 201–229.

The random disc thrower problem

Ted van der Aalst Dee Denteneer Hanna Döring
Manh Hong Duong Ross J. Kang* Mike Keane Janne Kool
Ivan Kryven Thomas Meyfroyt Tobias Müller Guus Regts
Jakub Tomczyk

September 3, 2014

Abstract

We describe a number of approaches to a question posed by Philips Research, described as the "random disc thrower" problem. Given a square grid of points in the plane, we cover the points by equal-sized planar discs according to the following random process. At each step, a random point of the grid is chosen from the set of uncovered points as the centre of a new disc. This is an abstract model of spatial reuse in wireless networks. A question of Philips Research asks what, as a function of the grid length, is the expected number of discs chosen before the process can no longer continue?

Our main results concern the one-dimensional variant of this problem, which can be solved reasonably well, though we also provide a number of approaches towards an approximate solution of the original two-dimensional problem. The two-dimensional problem is related to an old, unresolved conjecture ([6]) that has been the object of close study in both probability theory and statistical physics.

KEYWORDS: generating functions, Markov random fields, random sequential adsorption, Rényi's parking problem, wireless networks

1 Introduction

Various algorithms have been developed for effectively and efficiently maintaining and updating network information in wireless sensor networks. In order to analyse the performance of certain stochastic gossiping algorithms, Philips Research has posed the following two questions.

The first question is related to the maintenance of network information in wireless sensor networks. Consider an $n \times n$ grid of evenly spaced points (with spacing 1). A disc thrower sequentially distributes (closed) discs of fixed radius r, so that each

*Corresponding author.

disc is centred on a grid point randomly and uniformly chosen from grid points that were not covered by one of the previous discs. (Discs are allowed to overlap.) The disc thrower continues throwing discs until every grid point is covered by at least one disc. The question is, how many discs is the disc thrower expected to throw? More specifically, what is the probability distribution for the number of discs thrown?

A second question is related to the propagation of new network information in wireless sensor networks. Again consider an $n \times n$ grid. The disc thrower now throws discs in phases. Again he throws discs sequentially. During each phase though he is only allowed to throw discs so that each disc is centred on a grid point that he covered during the previous phase. Once he cannot throw any more discs, he starts his next phase. During the first phase he is only allowed to throw a disc on one of the corner points. The question of interest here is, what are the expected number of phases until the disc thrower covers the entire grid?

We will primarily focus on the first problem. As such, let us precisely define the parameters that we are interested here. Let us first note that there is a natural d-dimensional generalisation of the (first) random disc thrower problem. Consider a d-dimensional $n \times \cdots \times n$ grid with n^d equally spaced points with spacing 1. A sphere thrower sequentially distributes (closed) d-dimensional spheres of radius r such that each sphere is centred on a grid point chosen uniformly at random from grid points not contained within a previously thrown sphere. Let $N_d(n, r)$ be the number of spheres thrown by the end of this process. We are most interested in the *coverage ratio*

$$\theta_d(n, r) := \frac{\mathbb{E}[N_d(n, r)]}{n^d}.$$

Since our first interest is in the case $d = 2$ that corresponds to discs, we often drop the subscript and write $N(n, r)$ $(= N_2(n, r))$ and $\theta(n, r)$ $(= \theta_2(n, r))$.

The structure of the report is as follows. We start by discussing some of the related background literature, then present some simulation results that give us insights on how the expected number of thrown discs is related to the grid size n and the discs' radius r. This is followed by an analysis of the one-dimensional disc thrower problem, where exact results can be obtained. We also propose two approximation methods based on Markov chains and enumeration of all possible fillings of the grid and discuss them for the one-dimensional case. We then discuss the possibility of extending these methods to the two-dimensional grid, in the specific case of discs having radius $r = \sqrt{2}$. We leave the extension to larger discs for later work. At the end, we briefly look at the second problem and discuss how it can be solved in the one-dimensional case.

1.1 Random sequential adsorption and Rényi's parking problem

The (first) disc thrower problem is closely related to some models of physical chemistry. Perhaps the most relevant is that of random sequential adsorption (RSA) which we describe now. Suppose we have some surface and a sequence of particles land at random locations on the surface. Each particle adheres to the surface, or is "adsorbed",

if it lands at an exposed portion of the surface. In particular, each adsorbed particle covers a region of the surface, which prevents the adsorption of any particle that lands there afterwards. The process is irreversible, meaning that the system will eventually arrive at a *jammed state*, after which no new particles may be adsorbed. The standard models that incorporate these kinds of stochastic and geometric elements are known as *random sequential adsorption (RSA)*. For a comprehensive overview of RSA, the reader is referred to [4]. Most of the research done in this field is about determining the *coverage ratio* θ in the eventual jammed state: for adsorption onto lattices, θ is the ratio of adsorbed particles to the number of lattice points; in continuous space, θ is the density of points of adsorption in the surface. Clearly, the random disc thrower problem is captured by an appropriate lattice RSA model.

Although Philips is mainly interested in the lattice version of the problem, we mention here that the continuous RSA may be viewed as a useful limiting case. In lattice RSA, discs of small radius correspond to regions of the lattice that are jagged and far from round. On the other hand, taking r large and n even larger, the shapes more closely approach perfectly circular shapes in a continuous square. In particular, the random disc thrower problem can in this case be approximated by the continuous disc thrower problem, in which perfectly circular (closed) discs of radius r are randomly and sequentially thrown into the plane, so that their centres land inside a square of side length x, but do not land on a previously covered point, continuing until no area of the square is uncovered. Let us denote the analogously defined coverage ratio by $\tilde{\theta}(x, r)$ and write $\tilde{\theta}(r) = \lim_{x\to\infty} \tilde{\theta}(x, r)$. This problem is also known in the literature as continuous RSA with hard discs of radius $r/2$. Figure 1 depicts a jammed state in this model.

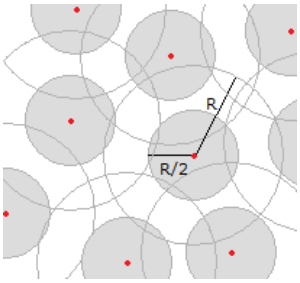


Figure 1: RSA with hard discs of radius $r/2$ in jammed state.

This leads us to another very relevant but older model, one that is related to the random car parking problem solved by [9]. Consider the d-dimensional cube $[0, x]^d$ for some $x \geq 1$. First we place a random unit-length d-dimensional cube so its centre point is uniformly distributed in $[1/2, x - 1/2]^d$. Subsequently, we place a new d-dimensional cube randomly and independently within $[0, x]^d$ in such a way that it does not intersect any previously placed d-dimensional cube. We repeat this until no more cubes can be placed. The question is, what is the expected eventual

density of cubes placed in $[0,x]^d$? This model may be referred to as the (continuous) *random sequential packing* of d-dimensional cubes, and such problems are discussed more generally in [3].

The random sequential packing problem for d-dimensional cubes is quite natural and is related to a wide variety of problems in statistical physics. By a suitable simple rescaling and ignoring boundary effects in the asymptotic limit, one may consider the $d = 2$ case as equivalent to a continuous and "square" version of our random disc thrower problem. Since the objects are simpler, one might expect that this problem is somewhat easier than the disc thrower problem; however, the process retains the same irreversibility property, which seems to be a fundamental difficulty. [9] solved the $d = 1$ case, known best as the random car parking problem, by proving that the expected density in the limit as $x \to \infty$ has the form

$$C_1 := \int_0^\infty \exp\left(-2\int_0^t \frac{1-e^{-u}}{u}du\right)dt \approx 0.748.$$

The evocative name of this model comes from imagining the intervals as cars of unit length that are parked randomly in a street of length x. [6] considered the d-dimensional problem in general and conjectured that the density should converge to ${C_1}^d$. From simulation results, this conjecture is widely believed to be false for every $d > 1$. Nonetheless, it remains open after more than half a century! Even the *existence* of the d-dimensional limit density was not proven until [7].

Following on that work, [8] proved a law of large numbers and central limit theorem for a general class of lattice RSA models, of which the random disc thrower problem is a member. Those results imply the following.

Theorem 1.1 ([8])**.** *For all $r \in \mathbb{R}^+$ there is a constant $\theta = \theta(r) \geq 0$ such that for all $p \in [1,\infty)$*

$$\frac{N(n,r)}{n^2} \xrightarrow{L^p} \theta \quad as \quad n \to \infty.$$

Theorem 1.2 ([8])**.** *For all $r \in \mathbb{R}^+$ there is a constant $\sigma = \sigma(r) \geq 0$ such that*

$$\frac{\mathrm{Var}[N(n,r)]}{n^2} \to \sigma \quad as \quad n \to \infty$$

and

$$\frac{N(n,r) - \mathbb{E}[N(n,r)]}{n} \xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma) \quad as \quad n \to \infty.$$

Hence for n large enough the distribution of $N(n,r)$ is approximately Gaussian. Moreover, the coefficient of variation tends to zero as n^{-1}. Thus a good approximation for grids of moderate size is $\mathbb{E}[N(n,r)] \approx n^2\theta(r)$.

1.2 Simulation estimates of $\theta(r)$

In this subsection, we discuss heuristic and simulation estimates for the coverage ratio $\theta(r)$ in terms of r. Let us consider first the case $r = 1$. A disc thrown on a grid point

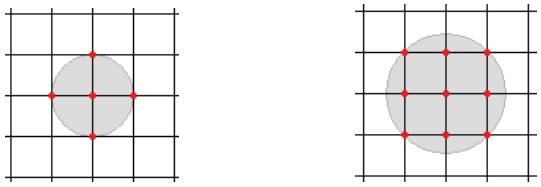


Figure 2: A disc with radius $r = 1$ or with radius $r = \sqrt{2}$.

will then cover four neighbouring nodes, see Figure 2, left. Hence, in this case, the problem is equivalent to randomly selecting grid points in sequence, subject to the restriction that each selected point does not neighbour a previously selected point. This problem is known as RSA of monomers on a square lattice with nearest-neighbour exclusion. An approximation for $\theta(1)$ derived from series analysis has been obtained (cf. [10]) with the estimate $\theta(1) \approx 0.36413$. Ignoring effects of the grid boundary, or taking n sufficiently large, we can then estimate $\mathbb{E}[N(n,1)] \approx 0.36413n^2$. Next let us consider the case $r = \sqrt{2}$. A disc thrown on a grid point will then cover eight nearby nodes, see Figure 2, right. This problem is known as RSA of monomers on a square lattice with next-nearest-neighbour exclusion. One can deduce that this problem corresponds to RSA of hard unit squares on a square lattice ([4], page 1310). The coverage for the latter problem is estimated as $\theta \approx 0.7476$. However, since a square covers four points and a square corresponds to one disc, we find for our problem $\theta(\sqrt{2}) \approx 0.7476/4 = 0.1869$, giving the estimate $\mathbb{E}[N(n,\sqrt{2})] \approx 0.1869n^2$.

We are not aware of published analysis to approximate the coverage ratio for other disc sizes. Simulations could help to determine the coverage. Moreover, methods using rate equations to estimate the coverage are discussed in [4] and [10].

Using continuous RSA as an approximation however, we can develop asymptotic estimates of $\theta(r)$ for large r. The continuous RSA problem has been studied extensively and the coverage has been estimated by $\tilde{\theta}(1/\sqrt{\pi}) \approx 0.5479$ (cf. [10]). Since each disc has area $\pi r^2/4$, we deduce that $\theta(r) \approx 0.5479/(\pi r^2/4) \approx 0.6976r^{-2}$ as r grows large. This approximation and some simulation results are plotted in Figure 3 for $n = 128$ against r, where for each r we average over 100 simulations. The approximation becomes more accurate for larger r as the discrete discs become more circular.

We end this section with a short remark about the effect of the grid boundary on the covering ratio, i.e. the difference between $\theta_2(n,r)$ and $\theta_2(r)$. (The boundary effect is also the essential difference between RSA and random sequential packing.) Points near the boundary are more likely to be covered than points in the interior of the grid. However, in [4] it is noted that in the one-dimensional case these boundary effects decay superexponentially. Furthermore, simulations suggest that for the two-dimensional case boundary effects tend to decay even faster than in the one-dimensional case. Thus, in moderate size grids, the boundary has only a small influence on the actual coverage ratio, when compared with the coverage ratio for

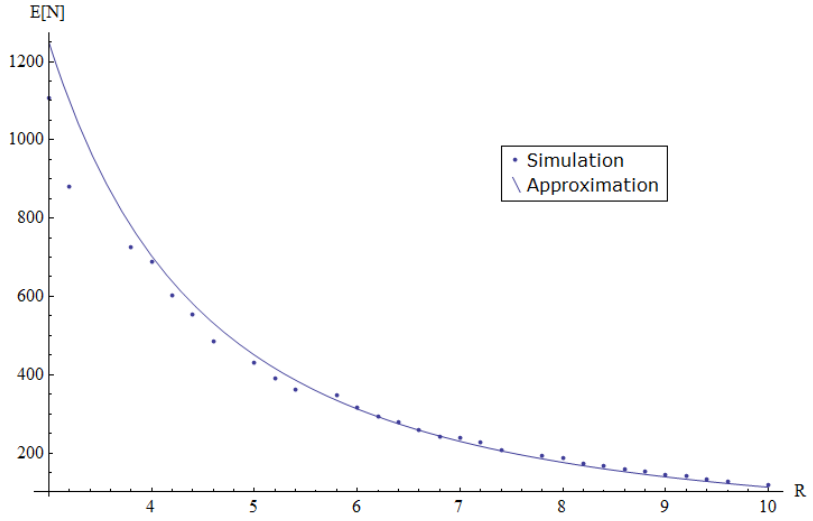


Figure 3: Asymptotic behaviour of $\mathbb{E}[N]$.

large n. In Figure 4, simulation results are shown for the coverage ratio with respect to the position of grid points. We average over 10^5 simulations on a 30×30 grid and discs with radius $r = 5$. We indeed see that the boundary effects decay rapidly, after approximately one radius length. It also appears as though the boundary effects propagate like waves to the interior of the grid.

2 One-dimensional disc throwing

In this section, we analyse the one-dimensional restriction of the random disc thrower problem. This amounts to throwing equal-sized line segments on a line with n grid points. For simplicity, we shall usually assume that the segments all have radius $r = 1$, that is they each cover exactly three points of the line.

2.1 Recurrence

We now describe a recursive approach to the one-dimensional random disc thrower problem. Each time a disc is thrown, the problem naturally splits into two smaller subproblems which are independent of one another. See Figure 5. This leads to the

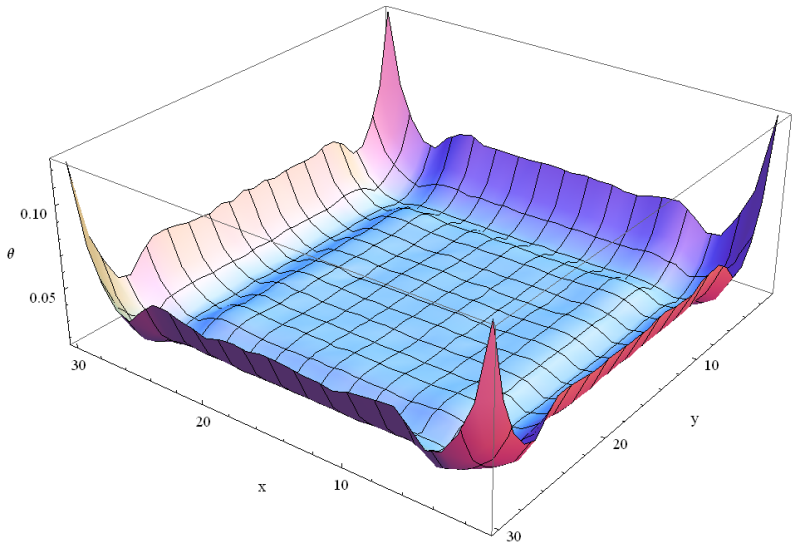


Figure 4: Simulation results of boundary effects.

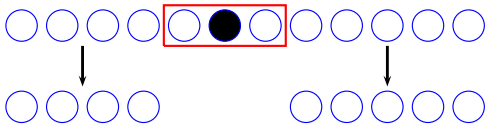


Figure 5: A depiction of the decomposition of the one-dimensional problem into two smaller independent subproblems.

following recursion:

$$\theta_1(n,1) = 1 + \frac{1}{n}\sum_{k=3}^{n-2}(\theta_1(k-2,1) + \theta_1(n-k-1,1)) + \\ + \frac{2}{n}(\theta_1(n-3,1) + \theta_1(n-2,1)). \tag{1}$$

We note that the last term on the right hand side corresponds to the case where the disc has been thrown close to one of the endpoints of the line. Though the above recurrence is written only for $r = 1$, this relation can be extended to arbitrary r. Hence for reasonably small n (and r) we can use the recurrence to compute $\theta_1(n,r)$ exactly.

To compute the limit of $\theta_1(n,1)$, we require analytical tools. This is done explicitly in Dutour Sikirić and Itoh [3, Chapter 2] for a closely related one-dimensional discrete random sequential packing problem, referred to as the Flory model. This process is

defined as follows: place at random a left-closed, right-open interval of length r in $\{0,\ldots,n\}$, with the left endpoint of the interval chosen uniformly from $\{0,\ldots,n-r\}$; then each subsequent interval (of length r) is chosen uniformly at random so that it intersects with no previously chosen interval. Here the problem is to study $v_r(n)$, the number of intervals that have been packed by the end of the process, and in particular to determine $\mathbb{E}v_r(n)/n$.

The relationship between the one-dimensional random disc thrower problem and the one-dimensional discrete random sequential packing is, for instance for $r = 1$, that

$$\lim_{n\to\infty} \frac{\mathbb{E}v_1(n)}{n} = \liminf_{n\to\infty} \theta_1(n,1). \tag{2}$$

It is possible to derive from a recurrence relation similar to ours in (1) a differential equation for the generating function

$$F(x) := \sum_{n=1}^{\infty} \frac{\mathbb{E}v_1(n)}{n} x^n.$$

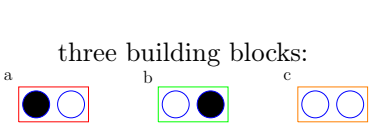
By solving the differential equation, we can then obtain an analytic expression for $F(x)$. Then reading off the coefficient of x^n in the Taylor series of $F(x)$, we obtain

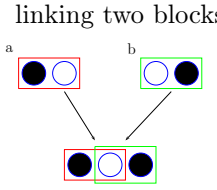
$$\lim_{n\to\infty} \theta_1(n,1) = \frac{1}{2}\left(1 - \frac{1}{e^2}\right) \approx 0.432$$

where we used (2).

2.2 A Markov chain approximation

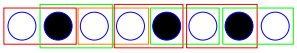
The following approach is suitable for the coverage for a very large grid compared to the size of the disc. Consider the one-dimensional case and a new random process with the following three building blocks, named a, b and c. Linking two building blocks we require the endpoints to overlap and to be either both occupied or both empty. The following pictures illustrate the building blocks, how to link them, and give an example of an admissible configuration.



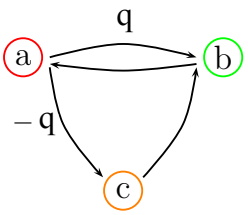


In order to approximate the one-dimensional random disc thrower problem, we would like to exclude on one hand the possibility of two centres being neighbouring points, and on the other hand the appearance of three (or more) unoccupied neighbouring grid points. Therefore, we consider the following transition matrix:

example of an admissible configuration:



$$\mathbb{M} := (\mathbb{M}_{ij})_{1\le i,j\le 3} = \begin{pmatrix} 0 & q & 1-q \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$



for $0 < q < 1$, q being the only parameter that determines the Markov chain. Which value of q should we choose? Our Ansatz is the following: we choose the value of q to maximise the entropy of the system.

The entropy of the Markov chain with transition matrix $\mathbb{M}$ is given by the entropy rate

$$\mathcal{H} = -\sum_{i,j} \pi_i \mathbb{M}_{ij} \log \mathbb{M}_{ij} = -\frac{1}{3-q}\big(q \log q + (1-q)\log(1-q)\big), \tag{3}$$

where $\Pi = (\pi_1, \pi_2, \pi_3)$ with $\pi_1 + \pi_2 + \pi_3 = 1$ is the stationary distribution, i.e. it satisfies the equation $\Pi\mathbb{M} = \Pi$. The latter two equations imply $\Pi = \big(\frac{1}{3-q}, \frac{1}{3-q}, \frac{1-q}{3-q}\big)$. Note that $\frac{1}{3-q}$ is an approximation for the coverage ratio $\theta_1(1)$.

Let us heuristically explain our Ansatz: the system is very likely not to be in an extreme event. A likely event is a steady state, i.e. a configuration that does not change much under small fluctuation. Entropy is the measure of the multiplicity of a configuration. So we can reformulate our task as to extremise the entropy. This Ansatz is consistent with the Second Law of Thermodynamics. The concept of entropy and the entropy maximisation principle also plays an essential role in information theory, see Cover and Thomas [1].

The mathematical justification is based on large deviations. Morally, we say that a sequence of random variables Y_n taking value on a (Polish) space Y satisfies a large deviation principle with a rate function $I\colon Y \to [0, +\infty]$ if for any event A

$$\mathbb{P}(Y_n \in A) \approx \exp[-n \inf_{x\in A} I(x)] \text{ as } n \to \infty.$$

The rate function I characterises the probability of observing an event: for an event A, the smaller values of $\inf_{x\in A} I(x)$ yield higher probabilities of observing A. The value of I is always non-negative and attains its minimum at the most probable event. Thus the large deviation principle explains that in order to have the most likely event, we need to extremise the rate function.

Now we consider the empirical pair measure of an irreducible Markov chain $X =$

$\{X_i\}$ where $X_i \in \Gamma, i = 1, \cdots, k$,

$$L_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, X_{i+1})}.$$

Sanov's Theorem (cf. den Hollander [2]) states that L_n satisfies a large deviation principle with the rate function

$$I_{\mathbb{M}}(\nu) = \sum_{ij} \nu_{ij} \log \left(\frac{\nu_{ij}}{\overline{\nu}_i \mathbb{M}_{ij}} \right)$$

for any probability measure ν on $\Gamma \times \Gamma$, where $\overline{\nu}_i = \sum_j \nu_{ij}$. Since the empirical measure counts the (average) number of realisations of an event, the theorem says that in order to have a maximum number of realisation of a Markov chain, we should extremise the relative entropy. Note that we can rewrite I as a relative entropy

$$I_{\mathbb{M}}(\nu) = \mathcal{R}(\nu || \overline{\nu} \otimes \mathbb{M}), \tag{4}$$

where $\overline{\nu} \otimes M$ is defined by $(\overline{\nu} \otimes M)_{ij} = \overline{\nu}_i \mathbb{M}_{ij}$. Now we can rewrite the function $\mathcal{H}$ in (3) in the form of (4). Let $u_{ij} = \frac{1}{|\Gamma|^2}$ be the uniform measure on $\Gamma \times \Gamma$.

$$\begin{aligned}
\mathcal{H} &= -\sum_{i,j} \pi_i \mathbb{M}_{ij} \log \mathbb{M}_{ij} \\
&= -\sum_{ij} \pi_i \mathbb{M}_{ij} \log \frac{\pi_i \mathbb{M}_{ij}}{u_{ij}} + \log |\Gamma|^2 + \sum_i \pi_i \log \pi_i \\
&= -\mathcal{R}(\pi \otimes \mathbb{M} || u) + \log |\Gamma|^2 + \sum_i \pi_i \log \pi_i.
\end{aligned}$$

Hence, up to sum of a constant and entropy of the stationary distribution, $\mathcal{H}$ can be written as the (negative) relative entropy. The argument above explains why it is reasonable to maximise the function $\mathcal{H}$.

Therefore, we are finding the value q that satisfies

$$\frac{d\mathcal{H}}{dq} = 0.$$

Using (3), we can calculate explicitly

$$\frac{d\mathcal{H}}{dq} = -\frac{3 \log q - 2 \log (1-q)}{(3-q)^2},$$

that yields the equation $q^3 - q^2 + 2q - 1 = 0$. Solving this equation, we find $q = 0.5698$ and $\theta_1(1) \approx \frac{1}{3-q} = 0.4115$.

This is close to the actual value (≈ 0.432, as we saw in the last subsection). We have hope to find better estimates of the coverage ratio by including longer building blocks and consequently studying larger transition matrices. This is a task for future research.

2.3 Configurations and weights

A final approach to understand the disc throwing problem is by an analysis of the direct calculation. Such a direct calculation of all possible disc throwings quickly becomes unfeasible when the number of grid points n is large compared to the disc size r. However, when we approach the problem from the end result of a random disc throwing, we see that the disc throwing process can be separated into two problems:

1. determine all possible configurations c_i of covering the grid.

2. determine the relative weight, i.e. the probability P_i of reaching a given configuration c_i by the disc throwing process.

Here $i \in \{1, \ldots, l\}$ runs over all possible configurations of covering the grid. If both problems are solved exactly, the expected value for the number $N_1(n, 1)$ of discs to cover the grid equals

$$\mathbb{E}N_1(n, 1) = \sum_{i=1}^{l} k_i P_i, \tag{5}$$

where k_i is the number of discs in configuration c_i.

The advantage of separating the original disc throwing process in this way, is that the two subproblems might be easier to calculate or estimate than the disc throwing process itself.

2.3.1 Configurations

Determining all possible configurations of a completely covered grid of length $n - 1$ (n grid points) by discs of size $r = 1$ is not hard in the one-dimensional situation. For this, we have to fill an interval of length $n - 1$ or $n - 2$ or $n - 3$ with intervals of length 2 or 3 and add 0, 1 or 2 intervals of length 1 at the boundary respectively to end up with a combined length equal to $n - 1$. These intervals represent the possible distances between neighbouring centers of discs. The following example with $n = 21$ grid points (Table 2.3.1) should clarify the methodology.

C	k_C	similarity class of configurations	a_C
1	11	$10 \cdot 2$	$\binom{10}{0}$
2	10	$1+9 \cdot 2+1$	$\binom{9}{0}$
3	10	$1+8 \cdot 2+1 \cdot 3$ or $8 \cdot 2+1 \cdot 3+1$	$2 \cdot \binom{9}{8}$
4	10	$7 \cdot 2+2 \cdot 3$	$\binom{9}{7}$
5	9	$1+6 \cdot 2+2 \cdot 3+1$	$2 \cdot \binom{8}{6}$
6	9	$1+5 \cdot 2+3 \cdot 3$ or $5 \cdot 2+3 \cdot 3+1$	$2 \cdot \binom{8}{5}$
7	9	$4 \cdot 2+4 \cdot 3$	$\binom{8}{4}$
8	8	$1+3 \cdot 2+4 \cdot 3+1$	$\binom{7}{3}$
9	8	$1+2 \cdot 2+5 \cdot 3$ or $2 \cdot 2+5 \cdot 3+1$	$2 \cdot \binom{7}{2}$
10	8	$1 \cdot 2+7 \cdot 3$	$\binom{7}{1}$
11	7	$1+0 \cdot 2+6 \cdot 3+1$	$\binom{6}{0}$

Table 1: List of all possible similarity classes. $n = 21$ and $r = 1$.

In the enumeration of possible configurations, we can restrict ourselves to enumerating different *similarity classes* C, due to the symmetry of the c_i. A similarity class C contains a_C configurations c_i, for which the analysis of the disc throwing process does not depend on the precise ordering of the different intervals of length 2 and 3. k_C stands for the number of discs used for each configuration in the similarity class C. As an example, the 5^{th} similarity class consists of configurations with 6 intervals of length 2 and 2 intervals of length 3. A representative configuration from this class is presented in Figure 6. Since the probability P_i of one representative c_i in a similar-

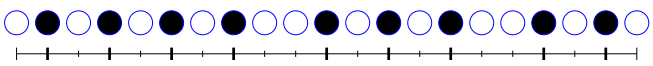


Figure 6: Representative configuration for $C = 5$.

ity class C equals the probability of any other representative in that class, the total probability of ending up in any configuration in a certain similarity class is $a_C P_i$.

2.3.2 Weights

First approximation The probabilities of occurrence for intervals of length 2 and 3 are not the same. From Monte Carlo simulations, we know that the probability of an interval of length 2 is higher than 0.5 and is length-of-grid-dependent. However, we can conclude that the error we make when we do not use the exact P_i, but rather assume that each configuration c_i is equally probable, is not large. In Figure 7, one can see the distribution of $N_1(100, 1)$ under assumption that each configuration c_i is equally probable.

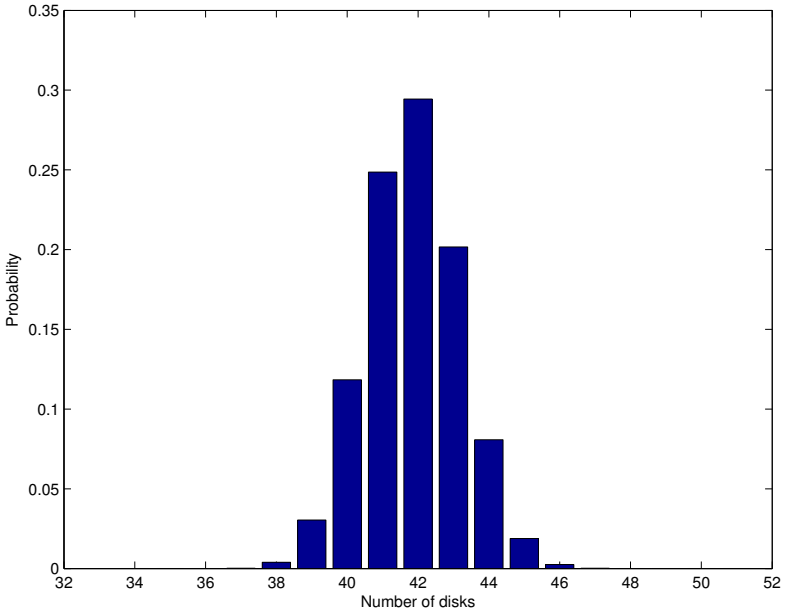


Figure 7: Distribution of discs needed to cover $n = 100$ grid points.

We obtained $\mathbb{E}N_1(100,1) \approx 41.8370$, $Var\ N_1(100,1) \approx 1.7773$ and $\sum_{i=1}^{N=51} a_i = 2066337330754$. This result is not far from 10000 Monte Carlo simulations where we obtained $\mathbb{E}N_1(100,1) \approx 43.5225$, $Var\ N_1(100,1) \approx 1.9155$. It would be useful to improve our results by approximating the P_i. The following section covers this subject.

Approximation of P_i Finding the probability P_i of ending up in a given configuration c_i brings us back to the original problem. If the number of discs in the configuration c_i is denoted by k_i, then there are $k_i!$ ways of ending up in this configuration. At this stage, we want to emphasise that the probability is *not* given by the ratio of multiplicities of the different configurations, $P_i \neq \frac{k_i!}{\sum_i^l k_i!}$. Rather, to find P_i one has to consider each $j \in \{1, \ldots, k_i!\}$ of the $k_i!$ possible ways to end up in c_i. Each of these disc throwings depends on the order of the discs being thrown and how many grid points were covered in previous steps,

$$P_{i_j} = \prod_{m=1}^{k_i} \frac{1}{n - \sum_{s=0}^{m-1} z_{i_{j_s}}}, \tag{6}$$

where $z_{i_{j_s}}$ is the number of additional grid points covered by the s'th disc thrown in the j'th process of covering configuration c_i and $z_{i_{j_0}} = 0$. Note that $\sum_{s=1}^{k_i} z_{i_{j_s}} = n$, since the grid is fully covered after all discs have been thrown.

Calculating P_{i_j} for all these different disc throwing processes is exactly what makes the disc throwing process hard. Therefore, in this method we suggest to approximate the additional coverings $z_{i_{j_s}}$ in a sensible manner, based on what typically happens in a disc throwing process. As evidenced by Figure 8, the values for $z_{i_{j_s}}$ are typically $2r + 1 = 3$ for the first discs thrown. Over the course of the disc throwing process, some discs only cover 2 additional points, while at the end, most newly thrown in discs cover only a single additional point. On average, 0.5 of the discs cover 3 points,

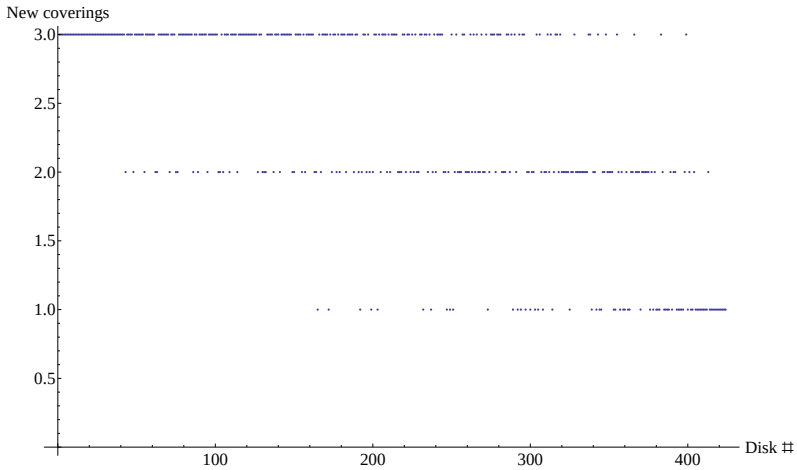


Figure 8: Additional points covered by newly thrown in discs, for an $n = 1000$ simulation.

0.31 of the discs cover 2 points and 0.19 of the discs cover 1 point.

As an approximation for P_{i_j} we assume that *all* initial discs cover 3 points, *all* next discs cover 2 additional points and *all* remaining discs cover a single point. Then, $P_{i_j} = P_{i_1}$ for all j and we do not have to worry about different throwing processes. Denote with ${}_pk_i$ the number of discs that cover p previously uncovered points, then from

$$ {}_3k_i + {}_2k_i + {}_1k_i = k_i, \qquad 3{}_3k_i + 2{}_2k_i + {}_1k_i = n, \tag{7} $$

we can solve ${}_2k_i$ and ${}_1k_i$ in terms of ${}_3k_i$. As a further approximation, we assume that $\frac{{}_2k_i}{{}_1k_i} = \frac{0.31}{0.19}$. This then fixes ${}_3k_i$ and allows us to write an approximate value for P_i as

$$ P_i = k_i! \prod_{m=1}^{{}_3k_i} \frac{1}{n - 3(m-1)} \times \prod_{m=1}^{{}_3k_i} \frac{1}{n - 3{}_3k_i - 2(m-1)} \times \prod_{m=1}^{{}_1k_i} \frac{1}{n - 3{}_3k_i - 2{}_2k_i - (m-1)}. \tag{8} $$

As a result of approximating the probabilities, we are not guaranteed to have a total probability equal to unity. These P_i should therefore be seen as *relative* probabili-

ties and, to find the expectation $\mathbb{E}N_1(n,1)$, one should use the rescaled probabilities $\hat{P}_i = P_i / \sum_{i=1}^{l} P_i$. By this method, we find approximations $\mathbb{E}N_1(21,1) \approx 9.0749$ and $Var\ N_1(21,1) \approx 0.24479$, which are in reasonable but not excellent agreement with 10000 Monte Carlo simulations that suggest $\mathbb{E}N_1(21,1) \approx 9.3769$ and $Var\ N_1(21,1) \approx 0.43549$.

3 Two-dimensional disc throwing

In this section, we analyse the original two-dimensional random disc thrower problem. We have chosen to restrict our attention to discs of radius $r = \sqrt{2}$. This in fact corresponds to discs which are squares each covering 9 points of the grid.

3.1 Recurrence approach

In Subsection 2.1, we outlined an explicit recurrence for the one-dimensional restriction of the problem. This recurrence relied on the fact that we could split into smaller independent subproblems. Here we would like to discuss the difficulties of extending this approach to the two-dimensional problem.

A natural idea is to split the problem at each stage into the four subproblems corresponding to the four regions of the grid determined by the randomly thrown disc, as depicted in Figure 9, left. Solving these as independent subproblems and then combining them provides a lower bound to a related disc throwing problem. We can solve this related problem by writing the corresponding recurrence relation, one that is similar to (1) but with four terms rather than two. As in Subsection 2.1, such a recurrence allows us to write down exact answers for small values of n, or to potentially solve it asymptotically using analytic methodology.

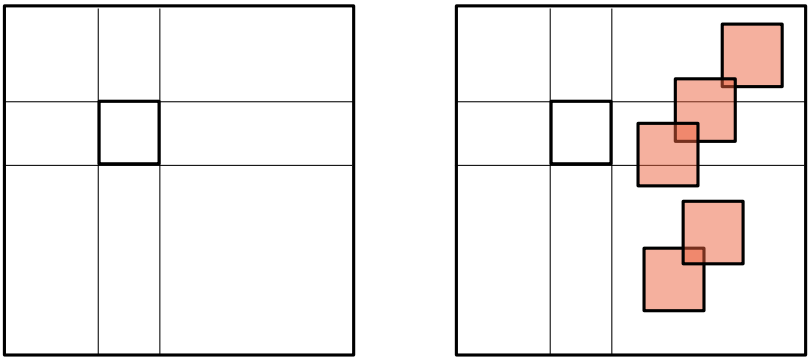


Figure 9: A depiction of the problem with decomposition of the two-dimensional problem into smaller subproblems.

Unfortunately, in attempting to decompose the problem in this way, there are possible long-range dependencies between the four regions, as depicted in Figure 9, right. Though it may be possible to generate high quality solutions for the related problem, it would require further investigation to determine if such solutions are at all related to the original problem. In particular, it would be worthwhile quantifying the error due to long-range dependencies. Perhaps they are negligible when n is large.

3.2 Markov chain approach

In two dimensions the Markov chain approach could be extended to the consideration of *Markov fields*, see e.g. [5]. Let us first recall the definition of a Markov random field on a regular two-dimensional grid.

Let $S = \{1, \cdots, n\} \times \{1, \cdots, n\}$ be the grid of n^2 points, which we call sites. For a fixed site s define its neighborhoods $N(s)$. For instance, for the site $s = (i, j)$, that is in the interior, the neighborhood could be $N(s) = \{(i-1, j), (i+1, j), (i, j-1), (i, j+1)\}$. For a site on the boundary, there are less neighbors.

A Markov random field $X(S)$ on S is defined via local conditionals

$$\mathbb{P}(X(s) = x_s \mid X(S \setminus s) = x_{S \setminus s}) = \mathbb{P}(X(s) = x_s \mid X(N(s)) = x_{N(s)}).$$

In other words, the full conditional distribution of $X(s)$ depends only on the neighbors $X(N(s))$.

To generalise our Markov approach to the two-dimensional case, we observe that the probability of a point in the lattice chosen to be a centre given all other outcomes is the same as the probability of being a centre conditioned only on the outcome of the neighbors. We see that the space of configurations as well as the above requirement in the two-dimensional disc thrower problem resemble with the definition of the Markov random field on the grid. For this reason, we believe that the Markov random field would be a good model for the two-dimensional disc thrower problem. The study of this model is out of the scope of this report and is a topic for future research.

3.3 Configurations and weights approach

Extending the approach of Subsection 2.3 to more than one dimension introduces additional complications. The feasibility of that method is determined by how well one can recover from the complications in each of the two subproblems.

Configurations Finding all configurations in a two-dimensional grid is hardly as straightforward as it is in the one-dimensional case. At least the separation of the disc throwing process into the two subproblems (finding configurations and weights) allows us to first only consider the problem of finding all possible configurations rather than all disc throwing processes. In principle, all configurations can be algorithmically enlisted by the following method: for a grid covered by $r = \sqrt{2}$ discs, we know that exactly one of the 4 upper left grid points should be covered. Depending on which grid point happens to be covered, there are again 4 possibilities for the disc to the

right of it. By continuing this process also in the vertical direction, one finds all configurations. However, for an $n \times n$ grid, there are a maximum of $\lceil n/2 \rceil^2$ discs, each with at most 4 possible center points, giving an upper bound for the number of configurations equal to $2^{2\lceil n/2 \rceil^2}$. To get a feeling for the number of configurations, taking square discs covering 25 grid points on a 30×30 lattice, there are about $9^{\lceil 30/3 \rceil^2} \approx 10^{95}$ configurations.

Weights Extending our approximation of the probability P_i of a given configuration to the two-dimensional case does not seem to be the weakest link in this approach. Once the method has been fine tuned in the one-dimensional case, one can generalise easily to the two-dimensional situation. The probability P_i only depends on the number of discs k_i in the configuration, leading to only a small number of required calculations. The difficulty is how to find reasonable values for the ${}_pk_i$. In the one-dimensional situation discussed above, only ${}_3k_i$ had to be determined by the assumption that $\frac{{}_2k_i}{{}_1k_i} = \frac{0.31}{0.19}$. In two dimensions, there are many more ${}_pk_i$ and still only two constraint equations $\sum_p {}_pk_i = k_i$ and $\sum_p p_pk_i = n^2$. However, this is not a problem inherent to the two-dimensional situation, as it also occurs in the one-dimensional case for $r > 2$. Therefore, an analysis of the applicability of our approximation for the probabilities in this more general one-dimensional situation, should also be conclusive for the two-dimensional system.

3.4 Hexagonal approach

One can consider other lattices than the square one. One of the most common lattices in nature is the hexagonal one. While in a square lattice every point can be covered up to four times, in the hexagonal lattice every disc (with $r = 1$) covers up to seven points and every point can be covered at most three times. This may simplify matters. Figure 10 depicts one of the possible configurations for a hexagonal lattice. Determining the expected number of randomly thrown discs needed for coverage of a given region is a problem that remains out of reach for the hexagonal lattice, yet we can still make some comparisons.

We have compared the square and hexagonal lattices directly, by performing 10000 Monte Carlo simulation on a 20×20 grid. Table 2 presents the outcomes.

Lattice	$\mathbb{E}N_2(20,1)$	$Var\ N_2(20,1)$	θ	area
square	149.1485	22.6434	0.3729	400
hexagonal	97.8310	7.4639	0.2824	≈ 346.41

Table 2: Comparison of characteristics of $N_2(20,1)$ for a square and hexagonal lattice for a 20×20 grid.

On the one hand, with 400 points in a hexagonal lattice we can cover an area which is approximately 12.5% smaller than an area covered in a square fashion. On

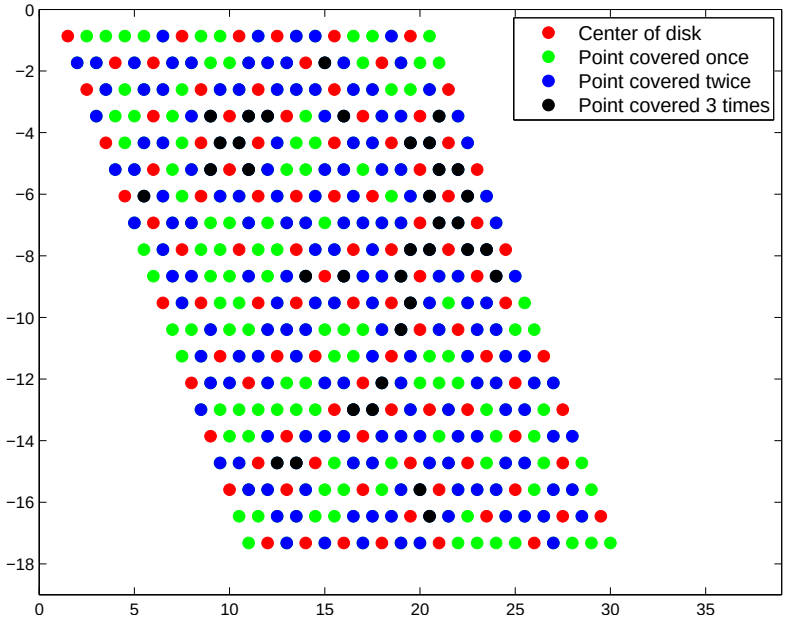


Figure 10: One realisation of a coverage process for a hexagonal lattice in a 20×20 grid.

the other hand, we need 34% fewer discs and the coverage is 24% smaller, which makes the hexagonal approach interesting for further investigations.

4 The second problem of Philips

As a second question, we shortly discuss a related but different disc throwing process that models information propagation on a lattice. We restrict ourselves to a one-dimensional grid with n grid points. In this second problem, at a given intermediate stage, the first m grid points serve as possible locations for the centre of a disc. Let us call these m grid points the *base for disc throwing*. During 1 *uts* (unit time step) the usual disc throwing process is applied to this base for disc throwing. After all discs have been thrown, m' discs will be covered, where $m' \geq m$ depending on the position of the (right) boundary disc. For the next *uts* the usual disc throwing process is applied to the new base for disc throwing, consisting of the m' grid points that were covered in the previous phase. The process continues until all n grid points are covered. The question is: how long, i.e. how many unit time steps, will it take on average to cover a grid of n grid points when one starts by throwing a disc centred on the first (left) grid point?

We consider discs of radius $r = 1$. At an intermediate stage — when the base

for disc throwing consists of m points — the signal is only propagated when the disc throwing process produces a configuration in which there is a disc centred on the last allowed grid point m, thereby expanding the base for disc throwing for the next phase to $m+1$. Instead, if there is a disc centred at $m-1$, there will be no additional points covered and the signal will not have propagated during this unit time step. With $P(k,m)$ we denote the probability that at the end of the usual disc throwing process on the base of m points, there is a disc centred at the k'th point. Note that $P(m-1,m)+P(m,m)=1$. The average speed of propagation is $v(m)=(0\cdot P(m-1,m)+1\cdot P(m,m))/(1\ uts)=P(m,m)/(1\ uts)$ when the base for disc throwing has size m. Hence, on average the time it takes to expand to a base of size $m+1$ is $t(m)=\frac{1}{1-P(m-1,m)}\ uts$ and the total time it takes to propagate information along a grid of size n is $T(n)=\sum_{m=1}^{n-1} t(m)\ uts$.

For small values of m, the probability $P(m-1,m)$ is quickly calculated: $P(0,1):=1-P(1,1)=0$, $P(1,2)=\frac{1}{2}$, $P(2,3)=\frac{1}{3}$. To find a closed expression for $P(m-1,m)$ for arbitrary m, we refer to ([4], page 1286) in which a problem isomorphic to the one presented here is studied. A recursive formula for $P(m-1,m)$,

$$P(m-1,m)=\frac{1}{m}\big(P(m-3,m-2)+P(m-4,m-3)+\ldots P(2,3)+P(1,2)+1\big), \quad (9)$$

follows from adding probabilities corresponding to the grid sizes to the right of the first disc thrown. This recursive relation is solved by $P(m-1,m)=\sum_{r=0}^{m}\frac{(-1)^r}{r!}$, which approaches e^{-1} for large m. Hence, the average time of propagating through a lattice of n grid points is

$$T(n)=\sum_{m=1}^{n-1}\frac{1}{1-\sum_{r=0}^{m}\frac{(-1)^r}{r!}}\ uts, \quad (10)$$

which is easy to compute for small n, e.g. $T(10)=14.007\ uts$. Furthermore, because $P(m-1,m)$ approaches e^{-1} rapidly for large m, we can find a linear propagation formula, $T(n)\approx 14.007+1.582(n-10)\ uts$, for $n\geq 10$.

In this analysis, we have only considered discs of radius $r=1$. Our methods can be generalised to larger discs, $r\geq 2$. The average propagation velocity is then given by $v(m)=(P(m-r+1,m)+2P(m-r+2,m)+\ldots+rP(m,m))/(1\ uts)$ and each of the $P(k,m)$ are given in terms of a recursive formula. Extending the one-dimensional situation to a two-dimensional setup requires more care, however. In two dimensions, more exotic topologies are possible and the expansion rate is a more subtle notion to define. Nevertheless, the one-dimensional situation described in this section is already interesting by itself for its corridor or street/highway applications.

References

[1] T. M. Cover and J. A. Thomas. *Elements of information theory.* Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006. ISBN 978-0-471-24195-9; 0-471-24195-4.

[2] F. den Hollander. *Large deviations*, volume 14 of *Fields Institute Monographs*. American Mathematical Society, Providence, RI, 2000. ISBN 0-8218-1989-5.

[3] M. Dutour Sikirić and Y. Itoh. *Random sequential packing of cubes*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2011. ISBN 978-981-4307-83-3; 981-4307-83-1. doi: 10.1142/9789814307840. URL http://dx.doi.org/10.1142/9789814307840.

[4] J. Evans. Random and cooperative sequential adsorption. *Reviews of Modern Physics*, 65(4):1281, 1993.

[5] R. Kindermann and J. L. Snell. *Markov random fields and their applications*, volume 1 of *Contemporary Mathematics*. American Mathematical Society, Providence, R.I., 1980. ISBN 0-8218-5001-6.

[6] I. Palásti. On some random space filling problems. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:353–360, 1960.

[7] M. D. Penrose. Random parking, sequential adsorption, and the jamming limit. *Comm. Math. Phys.*, 218(1):153–176, 2001. ISSN 0010-3616. doi: 10.1007/s002200100387. URL http://dx.doi.org/10.1007/s002200100387.

[8] M. D. Penrose. Limit theorems for monotonic particle systems and sequential deposition. *Stochastic Process. Appl.*, 98(2):175–197, 2002. ISSN 0304-4149. doi: 10.1016/S0304-4149(01)00152-1. URL http://dx.doi.org/10.1016/S0304-4149(01)00152-1.

[9] A. Rényi. On a one-dimensional problem concerning random space filling. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 3(no 1/2):109–127, 1958.

[10] J. Wang. Series expansion and computer simulation studies of random sequential adsorption. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 165(1):325–343, 2000.

Effective Water Storage as Flood Protection
The Rijnstrangen Study Case

Chris Budd, University of Bath
Joep Evers, Eindhoven University of Technology
Jason Frank, CWI
Sarah Gaaf, Eindhoven University of Technology
Ron Hoogwater, Leiden University
Domenico Lahaye, Delft University of Technology
Corine Meerman, Leiden University
Eric Siero, Leiden University
Tara van Zalen, Leiden University

January 2013

Abstract

Climate change is expected to cause higher discharge levels in the river Rhine at the Dutch-German border. In this study group project that was commissioned by *Rijkswaterstaat*, we investigate the possibility of flooding the Rijnstrangen area as a protective measure. We identify three subproblems. We first analyze the data recorded by *Rijkswaterstaat* and estimate the likelihood and the duration of extremely large discharges at the German border into the river. Next, we investigate how a change in discharge levels affects the water height in the first 35 kilometer section in the Netherlands. Finally we study the design of weirs and floodgates to allow diverting a sufficiently large amount of water flow from the river into the retention area. Our statistical analysis shows that an extreme discharge level is expected to occur once every 1250 years and to last for about three and a half days. Our numerical flow model shows the water height reaches equilibrium on a time scale that is much smaller than the one on which flooding occurs. The flow can thus be considered quasi-stationary. Passive weirs finally are shown to be too long to be feasible. Actively controlled floodgates are therefore recommended.

1 Introduction

The Netherlands is a low-lying country that hosts the estuaries of various rivers, among which the river Rhine. As a large number of people live and work in this delta, the protection of its population and its economic assets are administered at a national level. *Rijkswaterstaat*, part of the Dutch Ministry of Infrastructure and the Environment, is responsible for building and maintaining Dutch infrastructure.

Due to climate change, higher discharge volumes and higher water levels are expected. Certain safety standards specified by Dutch legislation might be violated. *Rijkswaterstaat* therefore has to take action. In this report we focus on the river Rhine. Just downstream from the border with Germany, this river (Upper Rhine or *Bovenrijn*) splits into the *Waal* and the Pannerdens Canal (*Pannerdens Kanaal*), the latter of which again splits into the Lower Rhine (*Nederrijn*) and the *IJssel* (see Figure 1 for a sketch). The ratios of the discharges in the final three branches are regulated by national policy..

The Netherlands has very little influence on the amount of water that enters the country. This is partly due to the uncontrollability of rainfall in upstream parts of the river. Moreover, Germany decides which measures are to be taken (or not) on German territory. The control of flooding of Dutch territory consequently must be achieved in the Netherlands. This SWI group was asked to study one of the possibilities to do so.

Near the cities of Nijmegen and Arnhem and adjacent to the river there is an area called *Rijnstrangen* that potentially serves as a retention area. A retention area is a region surrounded by dikes, intended to serve as a buffer to moderate extremely high discharge volumes. Under normal circumstances, this land is home to people who live and work there with the knowledge that in exceptional cases, they may have to evacuate due to flooding. The Rijnstrangen area could accommodate a maximum capacity of 150 million m^3.

Obviously, turning a region into a retention area has serious consequences. Before this decision can be made, the following issues need to be addressed:

- What is the minimum required capacity of the retention area given the uncertainty in the discharge of the Rhine entering the Netherlands? How does the minimum capacity depend on the peak discharge? *Rijkswaterstaat* would like the discharge peak to be reduced by 500 m^3/s. The retention area should not be completely filled before the extreme discharge event has passed.

- What should the outlets to the retention area look like? Important characteristics are: length, height, location and number. Three types of outlets have been distinguished. The first is a dike that is (partly) demolished in the event of a flood and that needs to be rebuilt afterwards. The second type uses a more sophisticated system of floodgates. For the first two types an authority deciding

on the opening of the outlets needs to be installed. The third type is just to have lower dikes locally, acting as a weir. If the water level exceeds the threshold, the retention area will start to fill automatically. In all cases *Rijkswaterstaat* is also interested in having estimates for the water velocities at the outlet;

- What are the consequences of the design of the retention area and the outlets on the discharge ratio in the final three branches of the river?
- How can the retention area be emptied following a flooding event?
- How can people be kept aware of the risks of living and working in the retention area over the course of several centuries? The retention area is only to be flooded under extreme conditions, i.e. very rarely. Most of the inhabitants will eventually perceive the situation to be less serious than it actually is.

These are the issues that the SWI group was asked to consider.

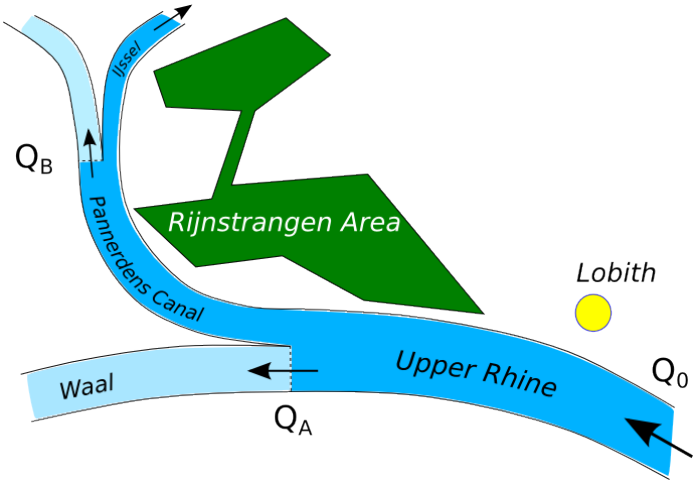


Figure 1: Rhine river system.

2 Solution approach

The following three subproblems were considered:

1. What is the likelihood and expected duration of an extreme flooding event and what is the total capacity needed to be retained in the Rijnstrangen area? Using statistical analysis of the data provided by *Rijkswaterstaat*, we attempt to estimate these numbers. Methods and results are discussed in Section 3.

2. What are the transient effects, within the Rhine region, of a change in discharge? This question was addressed using a PDE-based shallow water model and numerical simulations. The method and results are presented in Section 4 and Section 5 of this report. The main conclusion is that considering the time scales upon which flooding occurs, the transient effects are negligible.

3. What options exist for the design of the outlets such that the desired retention capacity is attained? These consist of passive flooding, floodgates and inflow from the bottom of the river. Demolition of a dike was not considered as an outlet option. Models are presented and analysed in Section 6.

3 Basic data analysis

Rijkswaterstaat provided us with a data set of daily discharge measurements at Lobith where the Rhine enters the Nethertlands. We want to derive from this data a model to describe and, more importantly, to predict extreme events. The extreme event in which we are interested is the occurrence of a very high discharge. In this section we will treat two approaches: a more theoretical investigation of extreme events (Section 3.1) and a hands-on interpretation of the data (Section 3.2).

3.1 Distribution of extreme events

In this section we will elaborate the theory of extreme events. We will first introduce a distribution that can be used to describe extreme events. Using this distribution and the available data we will be able to comment on the likelihood of an extreme event to occur.

3.1.1 Generalized extreme value distribution

Discharge Q in m^3/s at Lobith is recorded daily. The available data set of daily measurements of Q contains values from 1 January 1989 to 21 July 2012. To analyse the extreme discharge values we use the 23-year data from 1 July 1989–30 June 2012.

Now we define the vector $\hat{Q}_{\max}$ as the vector of length 23 containing the maximum discharge values at Lobith of each of the given 23 years.

We want to model the yearly maximum discharge at Lobith. This can be done, as described in [2], with the Generalized Extreme Value distribution (GEV distribution), which is defined as

$$G(x) = \exp\left[-\left(1+\gamma\frac{x-\mu}{\sigma}\right)_+^{-\frac{1}{\gamma}}\right], \tag{1}$$

where the notation $F(x)_+$ must be interpreted as

$$F(x)_+ = \begin{cases} F(x) & \text{if } F(x) > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

In this work we will assume that $\gamma \downarrow 0$. In the limit the GEV distribution converges to the Gumbel distribution. It is also called the Type I extreme value distribution and it is given by

$$F(x; \mu, \sigma) = e^{-e^{-z}}, \quad \text{where} \quad z = \frac{x - \mu}{\sigma}, \tag{3}$$

as is stated in [2] and the references therein. To this cumulative distribution function (cdf) belongs the probability density function

$$f(x; \mu, \sigma) = \frac{1}{\sigma} e^{-z - e^{-z}}, \quad \text{where again} \quad z = \frac{x - \mu}{\sigma}. \tag{4}$$

Other distributions are for example the Generalized Pareto Distribution (GPD) which is used together with the Peaks Over Threshold (POT). However, we will not work with these distributions for the following reason. In order to work with independent extreme values, [2] proposes that the distance between two peaks above some threshold should be at least 100 days. Assume that we extract from our relatively small amount of data a set of peaks that are at least a distance of 100 days apart. It then follows that these peaks nearly all coincide with the annual maxima. Therefore, due to our small data set, the GPD and the GEV distributions can be used almost interchangeably. In this paper we have chosen to work with the GEV distribution, considering only the yearly maxima, since the extreme values are those we are interested in.

To obtain a reliable estimate of the parameters it is important that the maximum values are independent and a sufficiently large number of extremes is used. We assume in this paper that the maximum values in $\hat{Q}_{\max}$ are independent, since these maximum values are taken over the period of a year. Seasonal trends, for example, will therefore have no influence on the maximum values. It is also important that we defined a year from July to June, since therefore the winter—the season in which extreme values are typically observed—is in its entirety contained in a single year. Notice that our analysis is based on a small set of extreme values as a consequence of a limited amount of available data. We observe here that our estimates are based on a small set of measurements and that they therefore might lack accuracy.

We use this data taking into account that our estimates are based only on a small set and therefore might not be as accurate as desired.

We fitted the Type I extreme value distribution to our measurements. We obtain the following estimates of the mean (the location parameter μ) and standard deviation (the scale parameter σ) from the fitted distribution

$$\sigma = 1\,625.4 \text{ m}^3/\text{s} \quad \text{and} \quad \mu = 5\,965.3 \text{ m}^3/\text{s}. \tag{5}$$

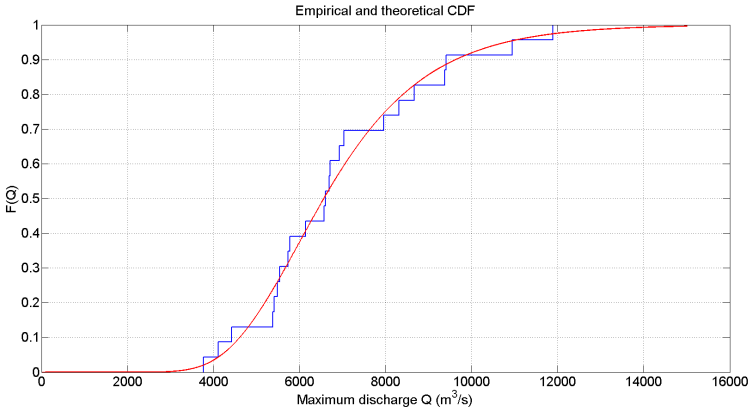


Figure 2: Empirical and theoretical (fitted) cumulative Type I distribution function for the yearly maximum discharge values of 23 years.

The estimates here are maximum likelihood estimators.

In Figure 2 we plotted the empirical cumulative distribution function for the vector $\hat{Q}_{\max}$ containing the yearly maximum discharge values. The empirical cdf is defined as the proportion of the yearly maximum discharge values Q_i in $\hat{Q}_{\max}$ less than or equal to a certain Q. We plotted also the theoretical cdf (3), using the approximate values μ and σ we obtained. Taking *Rijkswaterstaat*'s proposed threshold value $Q^* = 17500$ m^3/s, we also can calculate the probability that the yearly maximum discharge value exceeds Q^*:

$$\mathbb{P}(Q > 17500) = 8.2763 \cdot 10^{-4}. \tag{6}$$

Yet, this does not really tell us something practically. In the sequel, we will introduce the return period which will give a more explicit expression for the probability of the the occurrence of an extreme event.

As for now it can be seen that the empirical distribution nicely coincides with the theoretical distribution. Taking into account the limited amount of data, the agreement between the empirical cdf and the theoretical distribution is satisfactory.

However, plotting the corresponding probability density function (4) of the Type I extreme value distribution gives a less satisfactory outcome. In Figure 3 and Figure 4 we show the distribution of $\hat{Q}_{\max}$ in the form of a histogram together with the pdf scaled to our data vector $\hat{Q}_{\max}$ of length 23 and to the histogram intervals.

We already stated that the amount of data we use might not yield estimates of the satisfactory accuracy. In fact, the blocks of the histogram corresponding to the empirical data in Figure 3 follow the fitted and scaled pdf and can be considered acceptable, but the histogram in Figure 4 with interval lengths 1000, does not follow

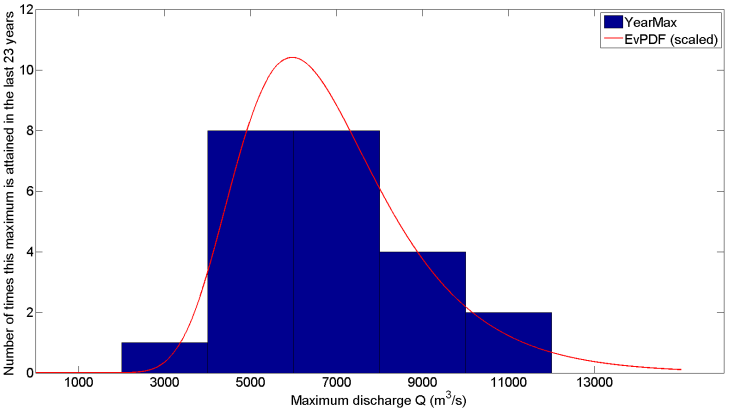


Figure 3: Histogram intervals of length 2000 and the probability density function scaled to our data vector $\hat{Q}_{\max}$ and the histogram intervals.

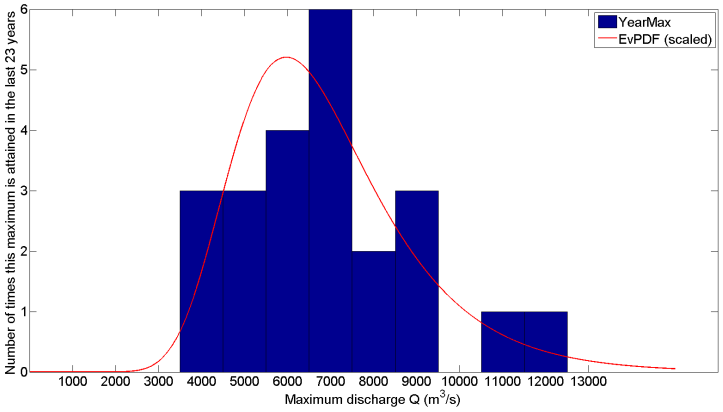


Figure 4: Histogram intervals of length 1000 and the probability density function scaled to our data vector $\hat{Q}_{\max}$ and the histogram intervals.

the fitted and scaled pdf in a way we would like. In order to obtain more satisfactory results we need more data.

3.1.2 Return level and return period

Other interesting information we can extract from our data relates to the return period. Let $0 < p < 1$. The return period $\frac{1}{p}$ is an estimate of the likelihood of a

flood event. The return level z_p associated with the return period is the value that the annual maximum will exceed with probability p. The return level z_p is defined as

$$z_p = \mu - \sigma \log(-\log(1-p)). \tag{7}$$

Remark: this z_p corresponds to the case in which $\gamma = 0$, i.e. to the case of the Type I extreme value distribution. For the exact derivation see [2] and the references therein.

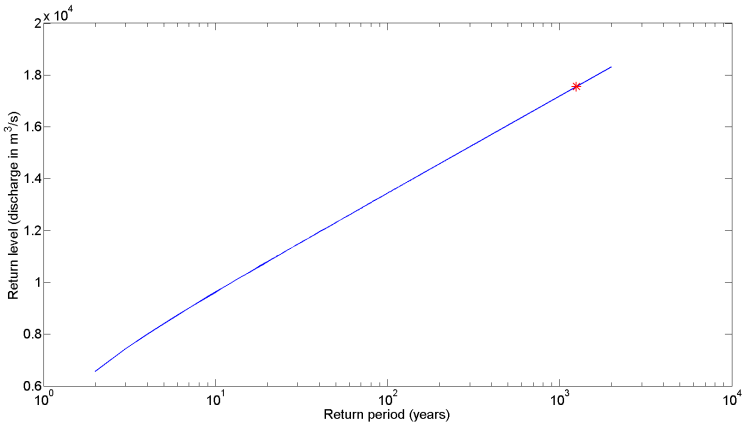


Figure 5: The return level $z_p = 5965.3 - 1625.4 \cdot \log(-\log(1-p))$ associated with the return period $\frac{1}{p}$. With probability $p = \frac{1}{1250}$, i.e. once every 1250 year, the discharge value $z_p = 17\,555$ m^3/s will be exceeded (red star).

In Figure 5 the return level is plotted against the return period with a logarithmic scale. According to statistical calculations of *Rijkswaterstaat*, the Dutch dikes have been constructed such that they can deal with exceptional discharges that occur once in 1250 years. The threshold value is recalculated every five years, based on currently available statistical data. In the last decades extreme discharges have become more frequent, and thus the the threshold was increased several times. In 2001 the level was adjusted to 16000 m^3/s (which is the current standard). Due to climate changes, it is expected that this exceptional water level will increase even more. Our model (based on data until 2012) predicts that the return level for the choice of $p = \frac{1}{1250}$ equals $z_p = 17\,555$ m^3/s, as indicated in the figure. This means that with probability $p = \frac{1}{1250}$, i.e. once every 1250 years, this discharge value $z_p = 17\,555$ m^3/s will be exceeded. Our prediction thus confirms that the threshold value will probably need to be adapted more in the coming years.

Up to now our starting point was the underlying theory for extreme events. In the following section, we explore what information can be deduced from the data in a more *ad hoc* manner.

3.2 *Ad hoc* data analysis

In the previous section we considered extreme events (i.e. high discharge) and their return times. In this section we ask the question: if such an extreme event occurs, what can we say about the amount of time the discharge remains above a certain threshold level Q^*? This information is very useful if at a certain moment we decide to influence the natural discharge. We can do this by allowing water to flow into the retention area Rijnstrangen. Now, consider the situation that we do not want the discharge to exceed some Q^*. This means we need to remove the excess of water effectively. Our aim in this section is to predict how much water the retention area must be able to contain.

3.2.1 Interval length and water volume above threshold

Here, we formulate more precisely what was said in the introduction of this section.

We fix a threshold Q^* and we open the outlets to the retention area if the current discharge Q exceeds Q^*. Let us assume that our actions work effectively enough to make sure that the discharge (in the river) directly downstream of the outlet to the retention area is $\min(Q, Q^*)$.

As said before, we use the daily discharge measurements at Lobith. After Q^* has been fixed, the data set contains a number of intervals in which the measured discharge exceeds Q^*. As in the previous section, we want to consider independent extreme events and to achieve this we only consider those intervals for which the corresponding maximum $Q_{\max}$ is also an annual maximum. The intervals we consider are of *maximal length*. By this we mean that the last measurement before the start of the interval and the first measurement after the interval are smaller than Q^*. In Figure 6 we show an example of such an interval. We determine its boundaries by calculating the intersection of the horizontal line $Q = Q^*$ with the linearly interpolated continuation of measurement data. The interval length L follows from these interval boundaries.

Linear interpolation of the measurements determines a graph (t, Q). The area under this graph but above the line $Q = Q^*$ equals the excess amount of water carried by the river during the period in which $Q > Q^*$. This is exactly the amount of water that needs to be stored in the retention area. We calculate this area by applying the trapezoidal rule to the measured data. In Figure 6 this is the shaded grey area.

Note that this integration procedure has to be executed with some care, since the units of Q (m^3/s) and t (days) do not match.

Now, consider those peaks for which the maximum corresponds to an annual maximum. Let us focus on one such peak (with maximum value $Q_{\max}$), and define $\tilde{Q} := Q_{\max} - Q^*$. From the measured data, we calculate $L(\tilde{Q})$ and $V(\tilde{Q})$ by the

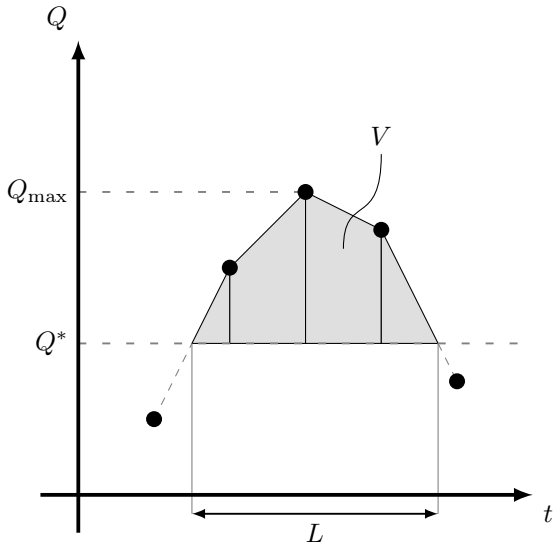


Figure 6: Schematic illustration of how we derive information about the interval length L and total volume V from the measurements. Measured data is indicated by the black circles.

procedure described above. By varying Q^* one can investigate the relation between $\tilde{Q}$ and L or V, respectively. In Figure 7 we collect this data for all peaks corresponding to annual maxima.

Let $\tau = 86400$ be the number of seconds in a day. One can show that theoretically $L(\tilde{Q})$ and $V(\tilde{Q})$ are related by

$$V(\tilde{Q}) = \int_0^{\tilde{Q}} \tau\, L(\bar{Q}) d\bar{Q}, \tag{8}$$

or, as a result,

$$\frac{dV(\tilde{Q})}{d\tilde{Q}} = \tau\, L(\tilde{Q}). \tag{9}$$

This implies that we can derive from the data relations $(\tilde{Q}, L(\tilde{Q}))$ and $(\tilde{Q}, V(\tilde{Q}))$, but these relations are not mutually independent. A first attempt might be to fit a parabola to the data in Figure 7, right. Then (9) implies that a linear relation should hold in Figure 7, left. The resulting (least squares) fits are indicated in Figure 8.

The linear fit in Figure 8, left, seems justifiable for large $Q_{\max} - Q^*$. However, for $Q_{\max} - Q^*$ near zero the fit is poor. Note that from a practical point of view, we are mostly interested in obtaining information around $Q_{\max} - Q^* \approx 500$. This is because

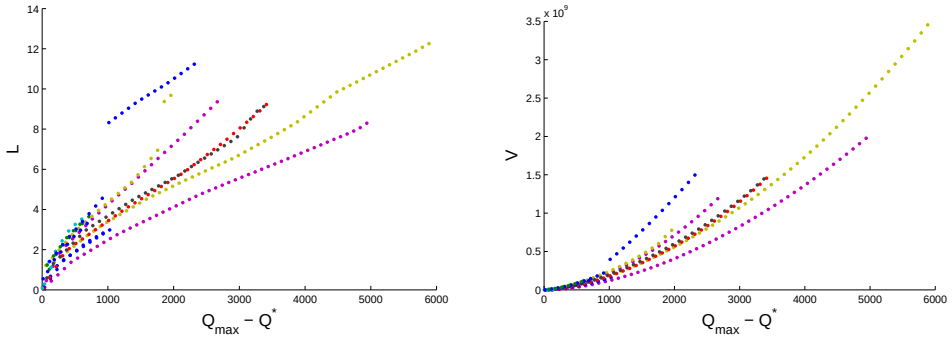


Figure 7: Left: Scatter plot of the length L of an interval in which the discharge Q is above a certain threshold Q^*. The values are given as a function of $\tilde{Q} = Q_{\max} - Q^*$. The aim is to draw conclusions that are independent of the peak height $Q_{\max}$. Data points corresponding to the same peak have the same colour. Right: Scatter plot of the volume V of water in a flood wave above a certain threshold Q^*. The values are given as a function of $\tilde{Q} = Q_{\max} - Q^*$. Data points corresponding to the same peak have the same colour.

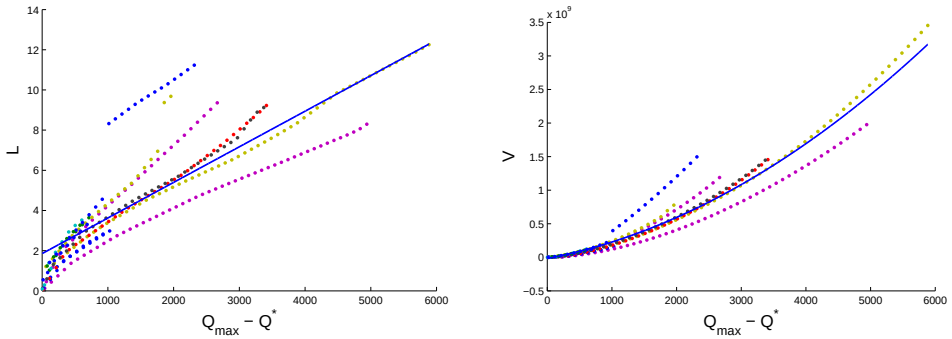


Figure 8: Left: A linear relation fitted to the data of Figure 7, left-hand side. Right: A parabolic relation fitted to the data of Figure 7, right-hand side. Note that slightly negative values occur in the graph due to the fitted parabola, not due to the data (cf. Figure 7, right-hand side).

*Rijkswaterstaa*t has in mind a situation where $Q_{\max} = 18000$ m^3/s and $Q^* = 17500$ m^3/s.

There is more to criticize about these fitted relations. From a physical point of view it is clear that $L(\tilde{Q}) = 0$ and $V(\tilde{Q}) = 0$ should hold if $\tilde{Q} = 0$. This is clearly not the case

in the fitted line in Figure 8, left. Moreover, (9) is violated. We have fitted:

$$L = a\,\tilde{Q} + b, \tag{10}$$

$$V = c\,\tilde{Q}^2 + d\,\tilde{Q} + f. \tag{11}$$

The fitted parameter values are

$$a = 0.0018, \tag{12}$$

$$b = 1.8533, \tag{13}$$

$$c = 60.7973, \tag{14}$$

$$d = 1.8422 \cdot 10^5, \tag{15}$$

$$f = -1.7661 \cdot 10^7. \tag{16}$$

It follows from (9) that $2c = \tau\,a$ and $d = \tau\,b$ should hold. The fitted coefficients satisfy

$$\frac{2c}{\tau\,a} = 0.7944, \tag{17}$$

$$\frac{d}{\tau\,b} = 1.1505, \tag{18}$$

which both do not really convince us that the above fits are appropriate.

In Figure 9 we show that using double logarithmic axes are of help here. The represented data is the same as in Figure 7.

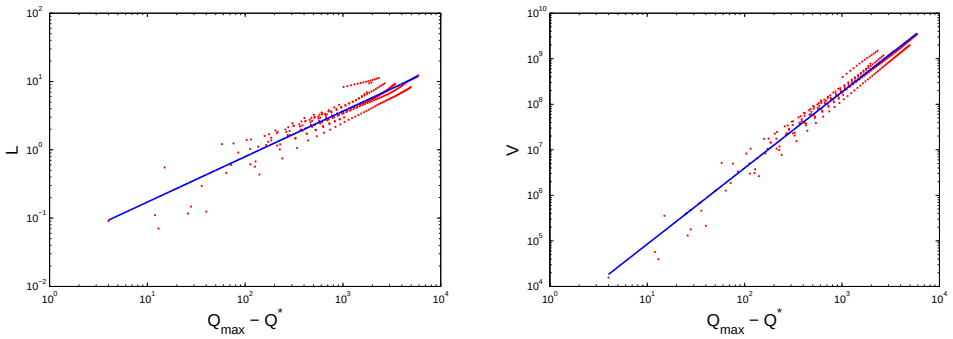


Figure 9: Left: The data of Figure 7, left, plotted in double logarithmic axes. A linear relation is added (i.e. linear in double logarithmic scaling). Right: The data of Figure 7, right, plotted in double logarithmic axes. A linear relation is added (i.e. linear in double logarithmic scaling).

In these axes the correlation between $\tilde{Q}$ on one hand and L or V on the other hand, is much clearer. To emphasize this, in both cases a linear fit is added. To show that

these fits are much better than the previous ones, we add them to Figure 8, the result of which is shown in red in Figure 10. Especially, we see in Figure 10, left, that the issues that arose for the linear fit are resolved. Both for small and for large $Q_{\max} - Q^*$ the fit resembles the data. Moreover, these fits pass the origin by definition. We have

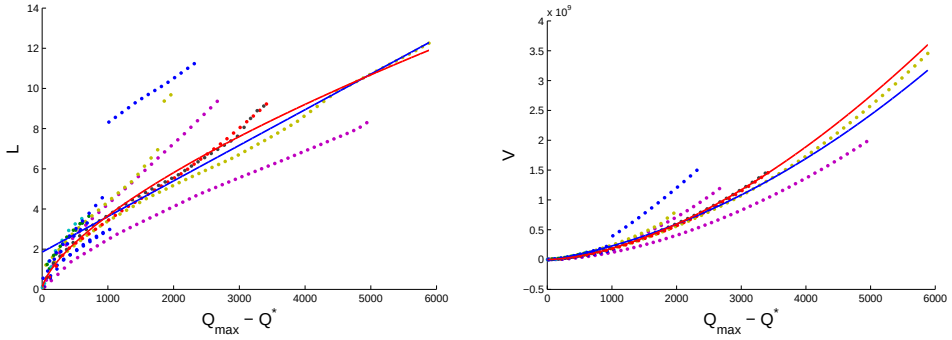


Figure 10: Left: The data of Figure 7, left, with fitted a linear relation (blue, cf. Figure 8, left) and an exponential relation (red, cf. Figure 9, left). Right: The data of Figure 7, right, with fitted a linear relation (blue, cf. Figure 8, right) and an exponential relation (red, cf. Figure 9, right).

fitted

$$\log L = A \log \tilde{Q} + B, \tag{19}$$

$$\log V = C \log \tilde{Q} + D, \tag{20}$$

or, equivalently,

$$L = e^B \tilde{Q}^A, \tag{21}$$

$$V = e^D \tilde{Q}^C, \tag{22}$$

which are positive for positive $\tilde{Q}$. The fitted parameter values are

$$A = 0.6646, \tag{23}$$

$$B = -3.2929, \tag{24}$$

$$C = 1.6708, \tag{25}$$

$$D = 7.5011, \tag{26}$$

for which (9) implies that $C\, e^D = \tau\, e^B$ and $C - 1 = A$ should hold. The fitted coefficients satisfy

$$\frac{C\, e^D}{\tau\, e^B} = 0.9423, \tag{27}$$

$$\frac{C-1}{A} = 1.0093. \tag{28}$$

This is more satisfying than our earlier attempt. Note that comparing these numerical values quantitatively to (17)–(18) does not make sense, since the coefficients in (17)–(18) and (27)–(28) have different meanings.

We use the exponential fit to make some predictions about what to expect for $Q_{\max} = 18000\ \mathrm{m}^3/\mathrm{s}$ and $Q^* = 17500\ \mathrm{m}^3/\mathrm{s}$. These numbers are the reference situation provided by *Rijkswaterstaat*. They serve as an illustration here, as similar computations can be made for other thresholds.

For $\tilde{Q} = 500\ \mathrm{m}^3/\mathrm{s}$ our fitted relations (21)–(22) predict the following values:

$$L = 2.3107 \text{ days}, \tag{29}$$

$$V = 5.8502 \cdot 10^7\ \mathrm{m}^3. \tag{30}$$

This means that the outlet to the retention area needs to be opened roughly two and a half days, and that in total nearly 59 million m^3 needs to be stored. According to these predictions, the Rijnstrangen area (capacity: 150 million m^3) would be more than sufficient as a retention area.

Two remarks about these predictions need to be made:

1. No data is available for discharges as large as $18000\ \mathrm{m}^3/\mathrm{s}$. This is because these values are currently not reached, but they are expected in the future due to climate change. Our predictions are only based on historic data. Extrapolation to changing situations in the future therefore requires caution. A positive point is however that the data used does contain information about 23 annual extreme events (represented by $Q_{\max}$ values). By introducing the translated coordinate $\tilde{Q}$, we account for the fact that the value of $Q_{\max}$ is different every year. As a result, the predicted value at $\tilde{Q} \approx 500$ in some sense contains information from the whole data set.

2. Our fitted relations reflect more or less the average trend of the data. It might be of more interest to know about the extremes (e.g. extremely high $Q_{\max}$ and simultaneously rarely high $L(\tilde{Q})$). We wish to state explicitly that these cases are not treated here.

3.2.2 An alternative approach

There is also an alternative approach to which there are some disadvantages. These will be commented on later in this section. In fact these negative aspects drove us to the approach described in Section 3.2.1.

The alternative approach is based on the assumption that a peak can be approximated locally by a parabola.[1] If its maximum $Q_{\max}$ is attained at time t_0, then the discharge

[1] This is a strong assumption, but still the results turn out to be useful.

Q as a function of time t is assumed to be given by

$$Q = Q_{\max} - \frac{1}{2}\,\alpha\,(t - t_0)^2. \tag{31}$$

Note that

$$\frac{d^2Q}{dt^2} = -\alpha, \tag{32}$$

where α is a parameter still to be identified. The second derivative is (a measure for) the curvature of the graph at the peak. Again we have to be careful about the units here. The second derivative d^2Q/dt^2 should have the unit of volume/(time^3). Since Q is measured in m^3/s and t in days, the unit of d^2Q/dt^2 is m^3/(s $\cdot$ days^2). At first sight this might seem a bit artificial. In the sequel, we write Q'' instead of d^2Q/dt^2, keeping well in mind this issue about the units.

An encouraging observation in this approach is the following. There appears to be a correlation between the value of Q_{max} and the second derivative Q'' in that point. In short: higher peaks are more narrow. This is shown in Figure 11 where we plot a numerical (second-order) approximation of Q'' against $Q_{\max}$ for all local maxima from July 1989 to June 2012. A linear fit is added to emphasize the correlation. We use all local maxima, instead of just annual maxima, since our fit would otherwise depend on only 23 data points. From this fit we are able to extrapolate (e.g. to $Q_{\max} = 18000$)

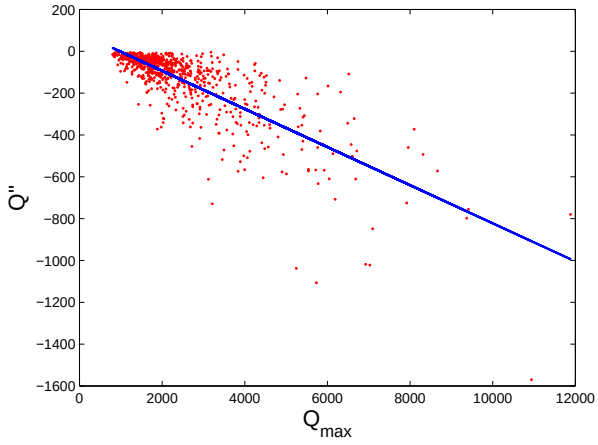


Figure 11: Scatter plot of the (numerical) second derivative in the top, against the discharge value in the top $Q_{\max}$. The linear correlation is indicated by the blue line. The units are: $[Q] = \text{m}^3/\text{s}$, $[t] = \text{days}$. To base the fit on a sufficiently large data set, all local maxima are used; not just the annual maxima.

and find the corresponding value for Q''. Moreover, exact calculations lead to the

following expressions

$$L = \sqrt{2\frac{Q_{\max} - Q^*}{\alpha}}, \tag{33}$$

$$V = \frac{4}{3}\sqrt{\frac{2}{\alpha}}\,\tau\,(Q_{\max} - Q^*)^{3/2}, \tag{34}$$

for the interval length L and volume V, respectively. From these expressions predictions for L and V are easily made, once we provide e.g. $Q_{\max} = 18000$ m^3/s, $Q^* = 17500$ m^3/s and the extrapolated value of $Q'' = -\alpha$. We find:

$$L = 0.8028 \text{ days}, \tag{35}$$

$$V = 4.6241 \cdot 10^7 \text{ m}^3. \tag{36}$$

These predictions imply that Rijnstrangen area (capacity: 150 million m^3) is sufficiently large to contain the amount of water that needs to be stored. Compared to the prediction of Section 3.2.1, the value of L is very low.

The latter statement about L indicates that we should be cautious here. In Figure 12 we have a closer look at the quality of the fits provided by (34). We can see that sometimes (e.g. for 27 March 2001) the fit very poorly represents the actual data. For other peaks (e.g. 17 February 2005) the fit is appropriate. We emphasize

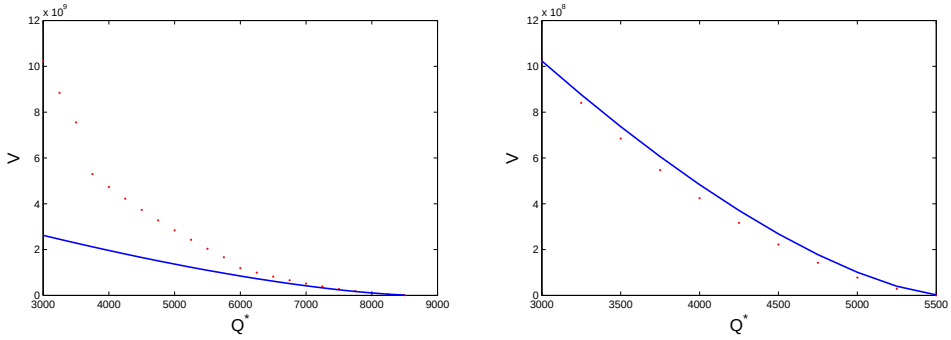


Figure 12: Left: Data, in red, and fit (34), in blue, for V corresponding to the annual maximum discharge which was attained on 27 March 2001. The fit deviates significantly from the data. Right: Data, in red, and fit (34), in blue, for V corresponding to the annual maximum discharge which was attained on 17 February 2005. The fit agrees quite well with the data.

that we should concentrate on the error for relatively small $\tilde{Q}$, since $\tilde{Q} = 500$ is the reference situation. Figure 12 strongly suggests to have a closer look at how well the approximations (33)–(34) perform, and especially for $\tilde{Q} \approx 500$.

Let L_{m} and V_{m} denote the actual *measured* quantities. The quantities fitted according to (33)–(34) are L_{f} and V_{f}. We consider the relative errors $|L_{\mathrm{m}} - L_{\mathrm{f}}|/L_{\mathrm{f}}$ and $|V_{\mathrm{m}} - V_{\mathrm{f}}|/V_{\mathrm{f}}$

in Figure 13. Intentionally, only relatively small $\tilde{Q} = Q_{\max} - Q^*$ are shown, since our predictions are for $\tilde{Q} = 500$. In the graphs we see no particular trend, but it is important to note that the order of magnitude of the error in L is 3.5 and of the error in V is 0.9. We wish to use these estimates for the error to improve our predictions

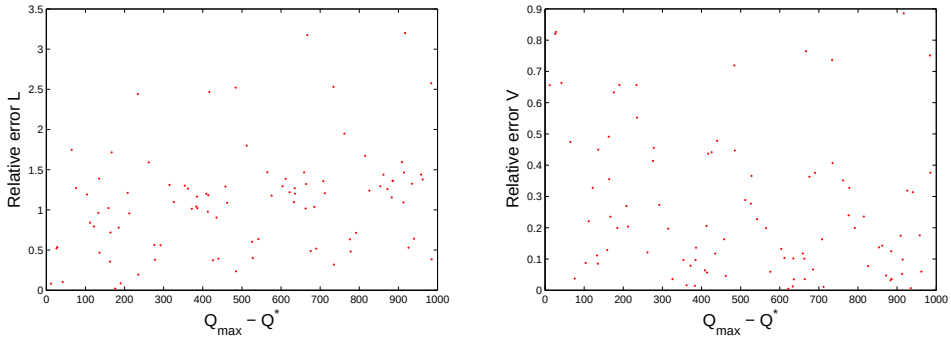


Figure 13: Left: Relative error in the prediction (33) of L, compared to the actual data, as a function of $\tilde{Q}$. Right: Relative error in the prediction (34) of V, compared to the actual data, as a function of $\tilde{Q}$.

(35)–(36). Note that if e.g. $|L_{\mathrm{m}} - L_{\mathrm{f}}|/L_{\mathrm{f}} < \beta$, then $L_{\mathrm{m}} < (1+\beta)L_{\mathrm{f}}$. By multiplying our predictions by "$1 + \mathcal{O}(\text{error})$" we thus get an upper bound for the quantity we want to predict.
We multiply the predictions in this case by $1+3.5 = 4.5$ and $1+0.9 = 1.9$, respectively. What we obtain are the improved predictions:

$$L = 3.6126 \text{ days}, \tag{37}$$

$$V = 8.7859 \cdot 10^7 \text{ m}^3. \tag{38}$$

Note that these values still imply that the retention area is large enough.

For comparison, we also show the error plots for the approach of Section 3.2.1. The relative errors of L and V are given in Figure 14. An important point is that there is a trend. The relative error decreases for increasing $\tilde{Q}$ as indicated by the blue dashed line (which estimates the maximum error). It levels off to approximately 0.5 for L and 0.6 for V.

Again, we use these estimates of the relative error to improve our predictions. We thus multiply (29)–(30) by 1.5 and 1.6, respectively. Our new predictions are

$$L = 3.4661 \text{ days}, \tag{39}$$

$$V = 9.3603 \cdot 10^7 \text{ m}^3. \tag{40}$$

We remark that the predictions for L are, after correction, more in agreement: compare (37) and (39). Moreover, the improved prediction (40) remains lower than the capacity of the Rijnstrangen area.

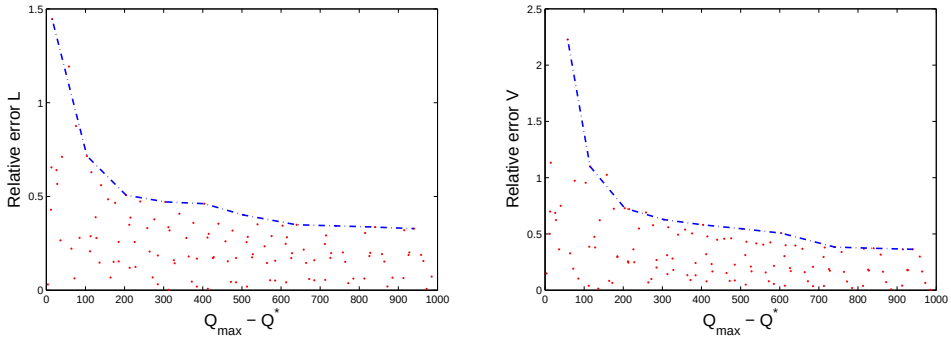


Figure 14: Left: Relative error in the prediction (21) of L, compared to the actual data, as a function of $\tilde{Q}$. The blue dashed line is an indication of the maximum error. Right: Relative error in the prediction (22) of V, compared to the actual data, as a function of $\tilde{Q}$. The blue dashed line is an indication of the maximum error.

3.3 Summary

We summarize the results of Section 3:

- The annual maxima of our data set obey a Type I extreme value distribution (or: Gumbel distribution), whose parameters are $\mu = 5965.3$ m^3/s and $\sigma = 1625.4$ m^3/s; cf. (5). Knowing this distribution we can e.g. estimate that $\mathbb{P}(Q > 17500) = 8.2763 \cdot 10^{-4}$; cf. (6).
- According to the aforementioned distribution, statistically, once every 1250 years a discharge level of 17 555 m^3/s will be exceeded; see Section 3.1.2.
- We expect a peak with maximum discharge 18000 m^3/s to remain above the level of 17500 m^3/s for about three and a half days. In this prediction we have taken into account a correction for the error in our approximation; see (39) and (37). The capacity of the Rijnstrangen area is large enough to contain the excess of water if we open the outlet at $Q^* = 17500$ m^3/s and a maximum discharge of 18000 m^3/s is eventually attained; cf. (40) and (38).

The following remarks need to be made:

- Our predictions and claims are based on data, and as such describe the current situation. We are not able to account for future changes in climate and the consequences thereof.
- The fitted probability distributions and relations between quantities more or less describe the *typical behaviour* of extreme events. One could imagine that, given the fact that a certain peak discharge has an extreme value $Q_{\max}$, its other characteristics (like duration L) are also variable. Our claims are then about the average duration of the extreme event, and we do not have information about

the distribution of L (i.e. the conditional distribution, given the maximum value $Q_{\max}$) or its dispersion. Thus, we cannot predict anything about how probable it is (for example) that such extreme event also lasts extremely long.

- We have shown that the retention area will have to be used only very rarely. Having two such events one shortly after the other is therefore (even more) unlikely to occur. From this point of view, there is no need for emptying Rijnstrangen very quickly after inundation, in order to be able to use it again. We do not address the issue of emptying in this report. However, due to the above arguments we believe that it is a minor factor in the overall decision process. Of course, we understand there might be other reasons (social, economic, geological,...), but these are beyond the scope of the present study.

4 Rhine system discharge model

One option to be considered for peak discharge retention relies on static weirs installed in the winter dikes that permit runoff of extreme discharge peaks into the Rijnstrangen region. Because the height of these weirs is only a few centimeters lower than the channel free surface height at extreme flood level, and to attain the desired retention rate of $\Delta Q = 500\text{m}^3/\text{s}$, the static weirs must extend over several kilometers atop the winter dike. For the design of the weirs we must estimate the extreme flood level free surface height in the Rhine within the model region. To this end we construct a 1D hydrological model in this section. The spatial domain of the model consists of a 35 km section of channel beginning 5 km upstream from Lobith and extending along the Upper Rhine, Pannerdens Canal and IJssel. The bifurcations at the Waal and Lower Rhine are modeled as geometric discontinuities at which outflow occurs.

Denote by x the distance downstream from Lobith, $x \in [x_0, x_0+L]$, where $x_0 = -5\text{km}$ (location of the German border) and $L = 35\text{km}$. The channel is assumed to have vertical walls with width given by $w(x)$, which is a piecewise continuous function with jump discontinuities at bifurcation points x_A and x_B. The free surface height $h(x,t)$ is defined with respect to a mean bottom orography. The fluid model is given by the St. Venant or shallow water equations for flow in a channel

$$wh_t = -(whv)_x + q, \tag{41}$$

$$v_t = -vv_x - gh_x + g(S - S_f), \tag{42}$$

where $v(x,t)$ is the mean cross-sectional velocity, $q(x,t)$ represents the lateral inflow per unit length (i.e. negative for outflow over a weir), g is the gravitational acceleration, $S(x)$ is the slope of the bottom, and $S_f(x)$ is the friction slope, which encodes the combined forces of friction. For future reference we also introduce the cross-sectional flow area $A(x,t) = h(x,t)w(x)$

The St. Venant equations can be simplified using scaling assumptions [1]. The most appropriate of these are the kinematic wave approximation, for which the momentum

equation (42) is replaced by the balance relation

$$S_f = S, \tag{43}$$

and the diffusion wave approximation in which (42) is replaced by

$$S_f = S - h_x. \tag{44}$$

The friction slope is defined as

$$S_f = \frac{n^2}{k^2}\frac{Q^2}{A^2 R_h^{4/3}}, \tag{45}$$

where $k = 1$ for SI units, n is the Gauckler-Manning coefficient, $Q = Av = hwv$ is the flow rate, and R_h is the hydraulic radius:

$$R_h = \frac{hw}{w + 2h} \approx h$$

These equation express a depth-averaged flow in a thin, incompressible fluid layer [1]. Inserting (45) in (43) and solving for v yields the flow speed

$$v = \frac{1}{n} h^{2/3} S^{1/2}, \tag{46}$$

from whence the kinematic wave model is reduced to

$$wh_t = -(\kappa(x) w(x) h^{5/3})_x + q, \tag{47}$$

where $\kappa(x) = S(x)^{1/2}/n(x)$. The Gauckler-Manning coefficient has been estimated for flow over surfaces of different roughnesses. We take $n(x) \equiv 0.03$ uniformly here.

Similarly, the diffusion wave approximation (44) can be solved for v to yield

$$v = \frac{1}{n} h^{2/3} (S - h_x)^{1/2} \approx \frac{1}{n} h^{2/3} (S^{1/2} - \frac{1}{2} S^{-1/2} h_x),$$

which in turn gives a nonlinear advection-diffusion equation for h.

The kinematic wave model (47) assumes the form of a hyperbolic conservation law for the height $h(x, t)$. To specify a unique solution, it must be equipped with appropriate initial and boundary conditions. We will assume that all characteristics are monotone in the downstream direction, meaning that it is only necessary to specify the inflow at the upstream boundary. Given an inflow discharge $Q_0 = Q(x_0, t)$, we can determine the inflow layer depth $h_0(t)$ from

$$h_0 v_0 = \frac{Q_0}{w(x_0)}$$

by substituting (46) in the above expression and solving for h_0. An initial condition can be obtained by solving

$$-(\kappa(x)w(x)h^{5/3})_x = 0$$

with this initial condition:

$$h(x,0) = h_0(0) - \int_{x_0}^{x} \kappa(s)w(s)h(s)^{5/3}\,ds.$$

Numerically, one can simply choose a constant layer depth $h(x,0) = h_0(0)$ and solve with constant inflow until steady state is reached.

Note that the velocity in the Rhine system is on the order of 1 m/s. This means that any given fluid parcel passes out of the system within about 8 hours. If variations in the inflow $h_0(t)$ are slow on this time scale, then the system is quasi-steady state, with height given by the previous expression. In this case too, the differences between the kinematic wave and diffusive wave models will be small.

5 Numerical unsteady flow computation

In this section we present a simple numerical code to solve the 1D model developed in the previous section, with the goals of (1) providing a tool that can be used for studying more complex retention scenarios, such as the placement of outlets at strategically chosen points around Rijnstrangen (not considered in this report), and (2) to show that for the flooding scenario predicted in §3, with excess peak discharge lasting more than 72 hours, the flow can be assumed to be in quasi-equilibrium state. At the end of the section, two computational scenarios reflecting these goals, are included, the first illustrating that the code can be used for transient calculations if so desired, and the second illustrating that even for a flooding scenario lasting < 48 hours, quasi-equilibrium is a good approximation.

Since our model includes only that part of the total discharge that flows through the trajectory Upper Rhine, Pannerdens Canal, IJssel, it is crucial for the numerical model to properly treat the outflows into the Waal and Lower Rhine, to avoid reflected waves or other numerical artifacts. We consider a class of finite volume methods with the necessary properties. We define grid points $x_i = x_0 + i\Delta x$, $i = 0, \ldots, N$, $N\Delta x = L$. The channel width at grid point i is denoted w_i. We allow the width to change discontinuously at a grid point, and denote the upstream and downstream values by w_i^- and w_i^+, respectively. The bottom slope S_i and the Gauckler-Manning coefficient n_i are also specified at grid points. The outflow per unit length q is approximated at the midpoint of the interval $q_{i+1/2} \approx q(x_{i+1/2})$.

To determine the layer depth $h_i(t)$ at grid point i, we discretize (47) using a compact one-parameter class of finite volume schemes. This class of discretizations has the

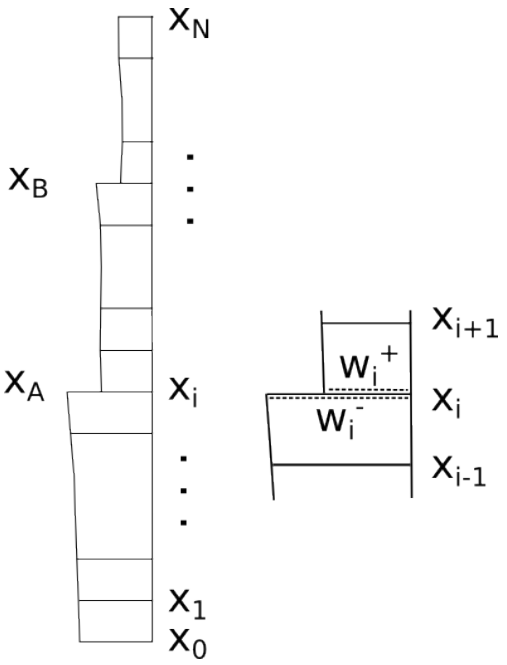


Figure 15: Discrete geometry. Left: one-dimensional grid in x, with channel widths w_i specified at grid points x_i and indicating discontinuities at bifurcation points x_A and x_B. These are the points A and B indicated in Figure 1. Right: a discontinuity in channel width w_i at grid point x_i, modelling outflow at a bifurcation.

property that it preserves the dispersion relation of the underlying PDE, in the sense that sign of group velocity is always correct. This means that information always flows downstream when it should, allowing us to model outflow at bifurcation points and at the end of the domain without incurring numerical side effects like reflected waves. The discretization is defined for a generic grid interval (x_i, x_{i+1}):

$$\theta w_i^+ \frac{\partial h_i}{\partial t} + (1-\theta) w_{i+1}^- \frac{\partial h_{i+1}}{\partial t} = -\frac{1}{\Delta x}(Q_{i+1}^- - Q_i^+) + q_{i+1/2}, \tag{48}$$

where the discrete fluxes are defined by

$$Q_i^\pm = \frac{S_i^{1/2}}{n_i} w_i^\pm h_i^{5/3}.$$

The parameter θ may take values in $(0, 1/2)$. For $\theta > 1/2$ the scheme is unstable for flow in the positive x direction. The scheme is implicit for $\theta > 0$ due to the weighted average on the left side. The choice $\theta = 0$ corresponds to upwind differencing, which is first order accurate, monotone, and highly diffusive. The choice $\theta = 1/2$, which is

symmetric and the only second order choice, yields the Preissman box scheme, well known in hydrology. However it is implicit, so for efficiency we have only implemented the case $\theta = 0$ here.

It can be checked by summing both sides of (48) that total mass $\sum_i h_i w_i$ is conserved locally when there is no inflow or outflow ($q_{i+1/2} = 0$, $w_i^+ = w_i^-$, for all i).

To model a bifurcation (points A and B in Figures 1 and 15), we choose the grid point closest to the bifurcation point, say x_i. Define w_i^- to be the width of the channel directly upstream of the bifurcation, and w_i^+ to be the width of the branch whose flow is to be included in the model. Let $w_i^{\text{out}} = w_i^+ - w_i^-$, then balance of flux at x_i requires $Q_i^- = Q_i^+ + Q_i^{\text{out}}$. Hence a specific outflow-to-inflow ratio requires

$$\gamma = \frac{Q_i^{\text{out}}}{Q_i^-} = \frac{Q_i^- - Q_i^+}{Q_i^-} = 1 - \frac{w_i^+}{w_i^-} = \frac{w_i^{\text{out}}}{w_i^-}.$$

Consequently a discontinuous change in channel width of ratio γ yields a corresponding change in discharge of ratio γ.

For $\theta = 1/2$ it is recommended to integrate (48) in time with the implicit midpoint rule. In this paper, we choose $\theta = 0$ and integrate in time with the well-known fourth order explicit Runge-Kutta method.

It is a straightforward matter to adapt the 1D Matlab code, given accurate geometrical channel data from the Upper Rhine region, as well as discharge data for Lobith. During the Study Group Week, a geometrically simple scenario was computed, using uniform channel widths between each bifurcation point, with the widths chosen to achieve the correct relative discharge rates in each branch of channel model.

We take domain parameters $x_0 = -5$km (German border) and $L = 35$km, such that Lobith is at $x = 0$. Moreover, we take $\Delta x = 100$m and uniform canal segments of width $w_0 = 1500$m in the Upper Rhine, dropping discontinuously to $0.36w_0$ at the Pannerdens Canal bifurcation point and to $0.15w_0$ at the IJssel bifurcation point. We assume uniform values of $n = 0.03$ and $S = 1 \times 10^{-4}$. In all simulations, the time step was taken to be $\Delta t = 50$s, and the initial condition was taken to be a stationary flow with discharge at Lobith corresponding to 5000 m^3/s.

We considered two flooding scenarios, based on the current peak discharge design capacity of 16000 m^3/s and the proposed new design capacity of 17500 m^3/s. In both cases a Gaussian (in time) discharge profile was defined with peak flow rate $Q_{\max} = 18000$ m^3/s. It is assumed that all discharge exceeding the design capacity is removed instantly at Lobith ($x = 0$), meaning the inflow profiles are capped at 16000 and 17500 m^3/s, respectively. In the first scenario, the peak was attained in 2 days to illustrate transient effects. In the second scenario the peak was reached in 11.5 days.

Figure 16 illustrates the results for the transient case with design discharge capacity 16000 m^3/s. We observe a rapid increase in flow rate, especially in the Upper Rhine

segment, at 30 and 40 hours, but a steady state has been reached at the peak flow after 50 hours. The total retention in the Rijnstrangen area is between 60 and 70 million cubic meters.

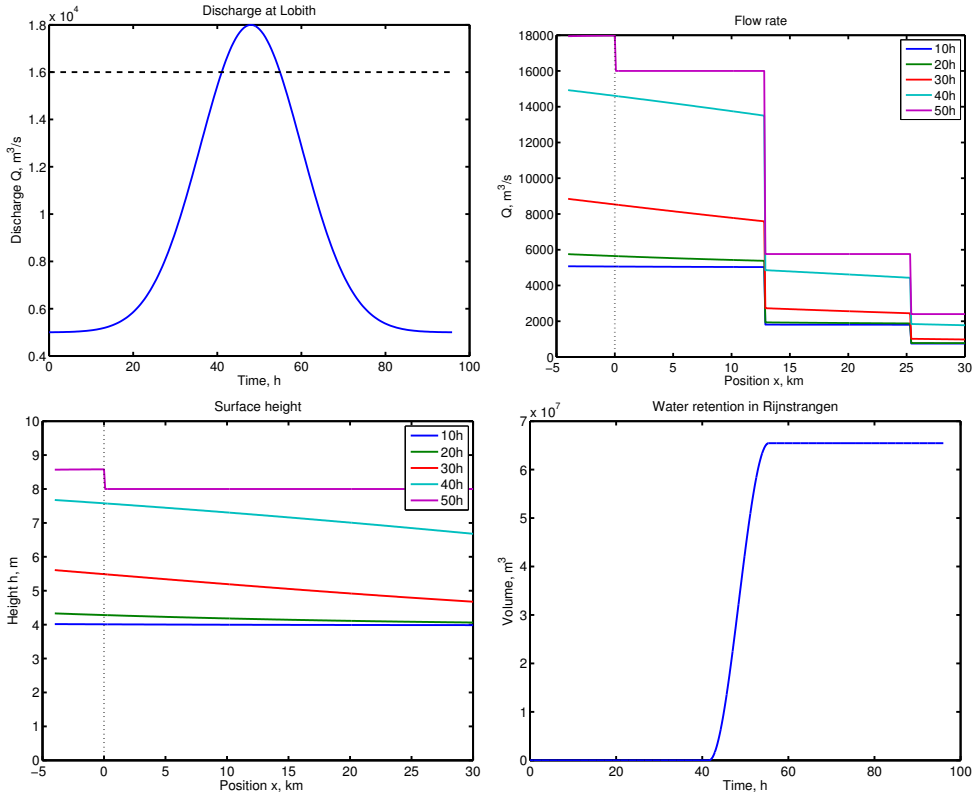


Figure 16: Flooding scenario reaching 18000 m^3/s in 2 days at current design peak discharge of 16000 m^3/s. Top left: discharge scenario at Lobith. Top right: flowrates throughout the domain at sequential times; discontinuities occur at channel bifurcation points. Bottom left: free surface height in the domain at sequential times; discontinuity at $x = 0$ due to removal to Rijnstrangen retention area. Bottom right: volume of water retained in Rijnstrangen area as a function of time.

Figure 17 illustrates the results for the longer event with peak after 11.5 days. In this case the flow may be assumed to be quasi-steady state, with the levels being nearly stationary at any given time. For the proposed design discharge capacity of 17500 m^3/s, the required retention in the Rijnstrangen area is 45 million cubic meters.

We stress that these computations are illustrative only! An accurate computation of

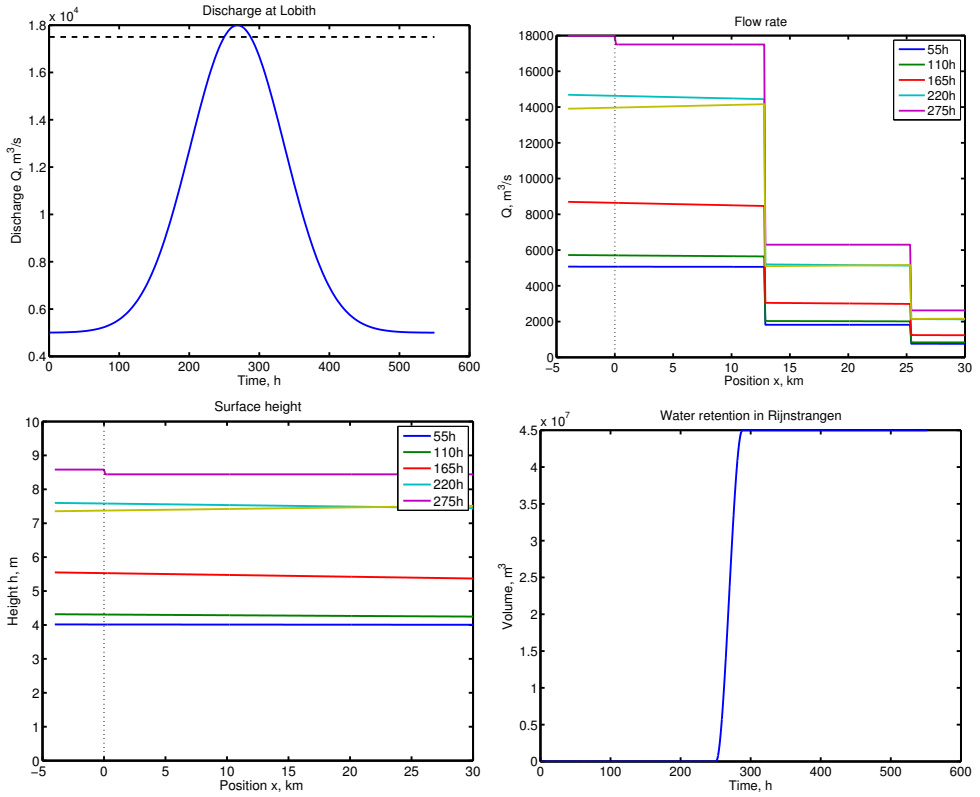


Figure 17: Flooding scenario reaching 18000 m^3/s in 11.5 days at proposed design peak discharge of 17500 m^3/s. Top left: discharge scenario at Lobith. Top right: flowrates throughout the domain at sequential times; discontinuities occur at channel bifurcation points. Bottom left: free surface height in the domain at sequential times; discontinuity at $x = 0$ due to removal to Rijnstrangen retention area. Bottom right: volume of water retained in Rijnstrangen area as a function of time.

transient flow would require correct measurements of the outer dike geometry. Furthermore the effects of sloping dike walls and the inner dike geometry have been neglected in the model. In general, the bottom slope $S(x)$ and Gauckler-Manning coefficient $n(x)$ should be properly estimated, and generally vary in x. Finally, the precise transient dynamics depend on the inflow discharge profile $Q_0(t)$ and Rijnstrangen outlet configurations.

6 Outlet configurations

We restrict ourselves to a single outlet before the first bifurcation of the river Rhine. The advantage is that a single outlet could suffice, while this would not be possible for an outlet after the first bifurcation without affecting the discharge ratio at the first bifurcation. For this case the following values for the discharge are relevant.

Let Q_{before} and Q_{after} denote the discharge of the river before and after the outlet. Also let Q_{outlet} be the discharge through the outlet so that the following conservation law holds: $Q_{\text{after}} + Q_{\text{outlet}} = Q_{\text{before}}$. To prevent flooding downstream it is required that $Q_{\text{after}} \leq 17500$. Here Q is measured in units of m^3/s, as in the rest of this section. Unnecessary use of the retention area comes at a high cost and may also limit uptake in the near future, so the preferred outlet has the property that:

$$Q_{\text{after}} = \min(Q_{\text{before}}, 17500). \tag{49}$$

We analyse several ideas for constructions of the outlet to the reservoir. After introducing our simple model we first study the construction of a weir, where water flows over the weir into the reservoir. A weir is a passive construction: water only flows if it has reached a critical level. Second we study the construction of a floodgate, by which we mean that a gate has to be lowered to reach the required outflow of water. In both cases, we investigate which size the outlet should have. Third we study an outlet where the water flows underneath a floodgate or through a pipe near the bottom of the river. Our analysis leads to estimates on the size of the constructions involved and gives insight into various other characteristics that need to be taken into account. These are all collected in the discussion at the end of this section.

6.1 A model for the flow of water over a weir

To calculate the flow over a weir, we use a so-called 'dam-break' model from [3, Section 10.5]. In this model the one-dimensional flow of water due to pressure is calculated on each location $x \in \mathbb{R}$ at time $t \geq 0$, where we assume the initial condition in Figure 18. That is, we assume that a dam is instantaneously removed at $x = 0$ and time $t = 0$ and water starts flowing from right to left.

In [3, Section 10.5] an analytical solution is found using the method of characteristics. The model only depends on the initial height of the water above the weir, which we denote by η. The height of the water as a function of x at some time t is sketched in Figure 19. It turns out that the height η_0 at $x = 0$ is constant for all time $t > 0$ and is given by $\eta_0 = \frac{4}{9}\eta$. The velocity of the water u at $x = 0$ is also constant (over time and height) and is given by

$$u = \frac{2}{3}\sqrt{g\eta}, \tag{50}$$

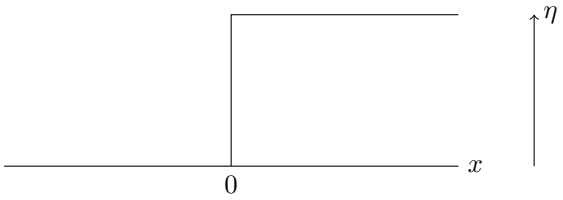


Figure 18: The initial condition at $t = 0$ in the dam-break model. The initial height of the water along the x-axis is denoted by η.

where g is the gravitational constant. So the flux through $x = 0$ is

$$q = u\eta_0 = \left(\frac{2}{3}\right)^3 \sqrt{g}\eta^{\frac{3}{2}}.$$

We obtain the total flow over the weir Q_{outlet} by multiplying with the length of the weir, which we denote by l:

$$Q_{\text{outlet}} = \left(\frac{2}{3}\right)^3 \sqrt{g}\eta^{\frac{3}{2}} l. \tag{51}$$

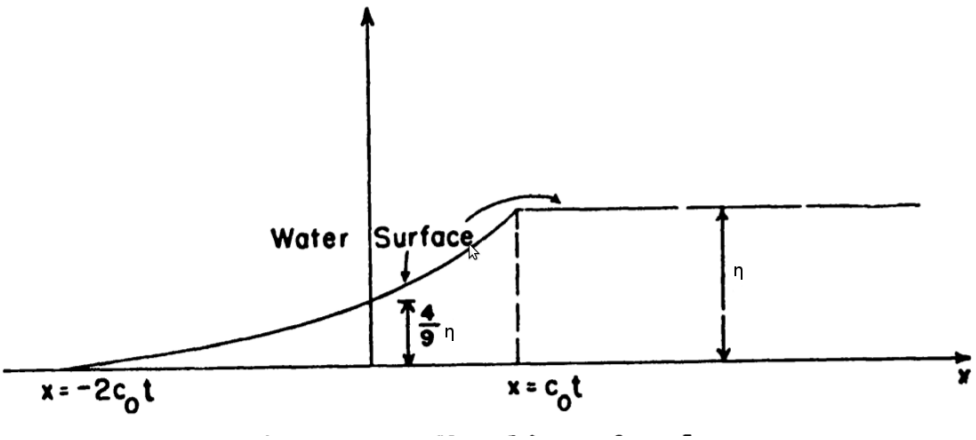


Figure 19: Dam-break model: the height of the water is sketched against the one-dimensional spatial variable x at time t. This figure is from [3, page 313], with adjusted variables.

We now apply this model to the situation where a weir is built alongside a river. We take a cross-section of the river and assume that this section moves with the velocity of the river, see Figure 20. If this cross-section arrives at the weir, then—from the

perspective of the cross-section—the dike is instantaneously lower, and we apply the dam-break model.

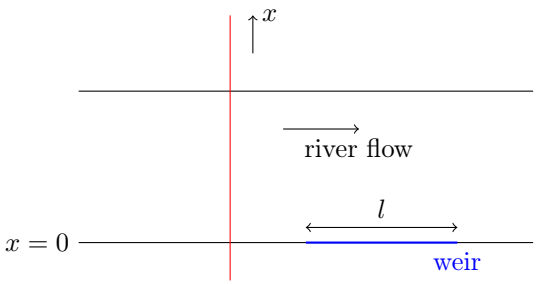


Figure 20: Depiction of the river from above with the cross-section (red) mentioned and the weir (blue). With this perspective we argue that we can apply the dam-break model.

If we denote the height of the weir by d, then the initial height of the water in the river is $h = d + \eta$. This situation is sketched in Figure 21. The x-axis in Figure 21 corresponds to the x-axis in Figures 18, 19 and 20. We take the weir to be flat in both the x-direction and the direction of the river flow. We neglect the effects of water in the river deeper than the top of the weir. We also neglect any effects resulting from the flow of water perpendicular to the cross-section (e.g. in the direction of the flow of the river).

Since d is fixed, the model depends on the height of the water in the river h which we obtain from the model derived in Section 4:

$$h = \left(\frac{nQ}{w\sqrt{S}} \right)^{\frac{3}{5}}, \tag{52}$$

where n is a constant depending on the river bed, Q is the discharge of the river (the flux in m^3/s), w is the width of the river and s is the slope of the river.

It follows from (51) that the power law $Q_{\text{outlet}} \propto \eta^{\frac{3}{2}}$ holds with proportionality constant $\left(\frac{2}{3}\right)^3 \sqrt{g} l$. In hydrology this power law is well-established. For instance in [4] section 5.1 the same power law is acquired from rough estimates for water flowing over a weir like in Figure 21. Assuming subcritical (laminar) flow before and supercritical (turbulent) flow after the weir, it is reasoned that at the weir the flow must be critical. Applying Bernoulli's principle (see equation (56) below) yields the same power law (but with a different proportionality constant).

6.2 Constructions with water flowing over a weir

We first consider the situation of water flowing over a weir where d is constant over the length l of the weir. A weir with length l is constructed with a fixed height d,

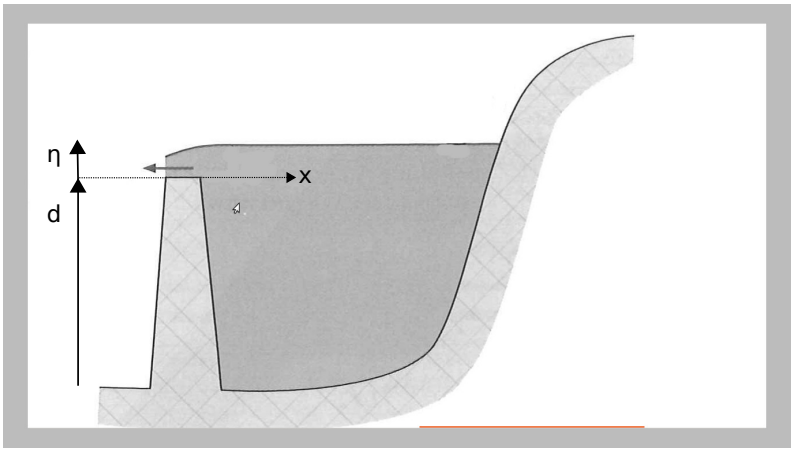


Figure 21: Cross-section of the river, with indication of the variables used. We assume the slope of the weir to be horizontal, so in contrast to what the picture suggests, the top of the weir continues to the left following the dotted line. This figure is from [4].

such that:

$$Q_{\text{outlet}} = 0, \text{ if } Q_{\text{before}} = 17500, \qquad \text{and} \qquad Q_{\text{outlet}} = 500, \text{ if } Q_{\text{before}} = 18000. \quad (53)$$

This weir is in essence a place where the dike is lowered, see also Figure 20 and Figure 21. We also require that water starts flowing over the weir only when $Q_{\text{before}} > 17500$ and investigate the implications of this requirement.[2].

The height d can be explicitly calculated from (52) and the first part of (53). It follows that the weir should be 7.10m. When $Q_{\text{before}} = 18000$, the height of the water level is $h = 7.22$m by (52). Consequently, we know the height of the water above the weir: $\eta = h - d$. Using (51) we can determine the length of the weir from the second part of (53). It follows that the weir has to be approximately 13 kilometers long. This is a disadvantage since there is no place in the Netherlands where it is possible to construct such a long outlet.

Suppose that there would be an area long enough for this weir, then there is another disadvantage. One requirement of *Rijkswaterstaat* is that Q_{after} has to remain be below 17500 to avoid flooding downstream. From (52) and (51) an expression can be given of Q_{after} as a function of Q_{before} by:

$$\begin{aligned} Q_{\text{after}} =& Q_{\text{before}} - Q_{\text{outlet}} \\ =& Q_{\text{before}} - \left(\frac{2}{3}\right)^3 \sqrt{g} \left(\left(\frac{n Q_{\text{before}}}{\sqrt{S} w} \right)^{\frac{3}{5}} - d \right)^{\frac{3}{2}} l. \end{aligned} \quad (54)$$

[2]This assumption ensures that the weir is no longer (and thus higher) than necessary.

In Figure 22, this function is plotted in blue for $Q_{\text{before}} \geq 17500$ (for $Q_{\text{before}} < 17500$, Q_{after} coincides with the black line.) The figure shows that this passive construction does not satisfy the condition of *Rijkswaterstaat* for intermediate values.

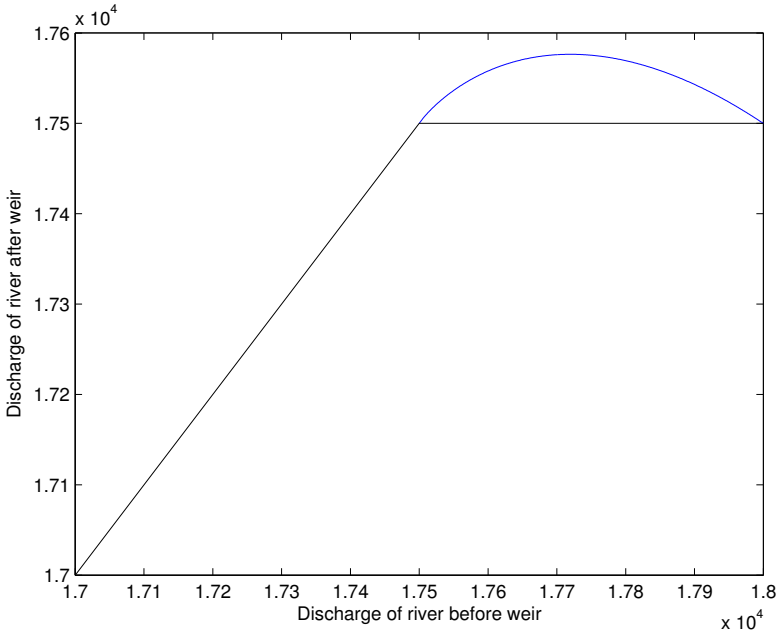


Figure 22: The discharge after the weir is plotted in blue against the discharge before the weir when using a weir of 13 kilometers long. The black line refers to the preferred situation (49). The requirement from *Rijkswaterstaat* that the discharge be reduced to $Q_{\text{after}} \leq 17500$, is not met.

To avoid the unwanted behavior at intermediate values of the previous weir we now replace one of the conditions of (53) by an equation on the derivative:

$$\frac{dQ_{\text{outlet}}}{dQ_{\text{before}}} = 1, \text{ if } Q_{\text{before}} = 18000, \qquad \textit{and} \qquad Q_{\text{outlet}} = 500, \text{ if } Q_{\text{before}} = 18000. \tag{55}$$

Because of the apparent concavity of (54) the condition on the derivative will preclude $Q_{\text{after}} \geq 17500$ and it will also minimize the amount of unnecessary water entering the retention area for these kinds of weirs. Again employing (52) and (51) with (55) leads to a relation which can be numerically solved for l and d, $l \approx 7$ kilometer and $d \approx 7.04$m. In Figure 23, Q_{after} is plotted against Q_{before}, with the formula in (54). The figure shows that Q_{after} never exceeds 17500, but it also shows that from $Q_{\text{before}} = 17250$ onward, there is already water flowing over the weir. But the

superfluous amount of water entering the retention area stays below 100 m^3/s all the time.

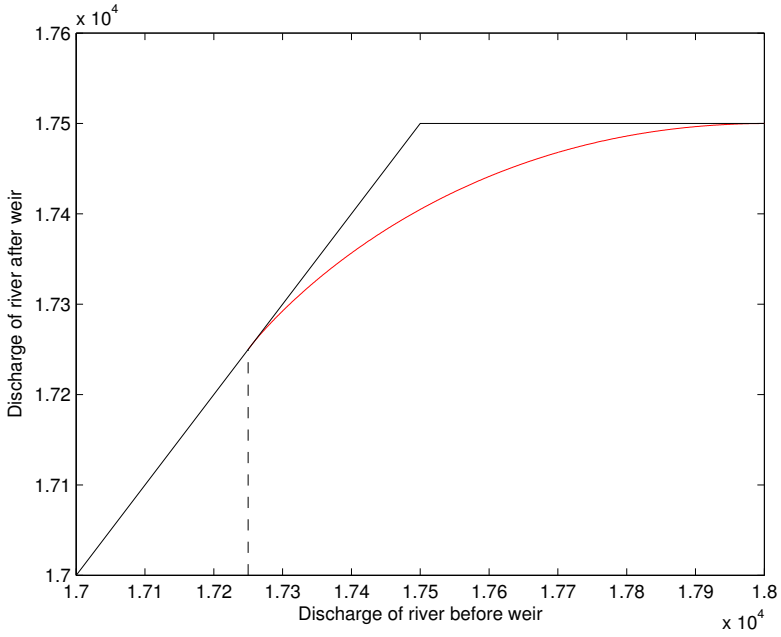


Figure 23: The red curve indicates the discharge after the weir when a weir construction is used, with a length of 7 km. The black line again refers to the preferred situation in (49). The condition $Q_{\text{after}} \leq 17500$ holds for all values of Q_{before}, but water starts flowing into the retention area already at $Q_{\text{before}} = 17250$, indicated with the dashed line.

6.3 Constructions with water flowing over a floodgate

As we saw in the previous section passive constructions need to be very long. In this section we look at controllable floodgates; essentially these are weirs of variable height. So water is assumed to flow over the floodgate, this allows us to use the dam-break formulas derived previously.

First we consider a floodgate whose height can be continuously adjusted to allow for precisely the correct amount of water entering the retention area. We give a description of how this height needs to be adjusted. Second we look at a floodgate which can be either open or closed.

6.3.1 Continuous control

The preferred outlet (49) for $Q_{\text{before}} \geq 17500$ is equivalent to $Q_{\text{outlet}} = Q_{\text{before}} - 17500$. Using this, (51), (52) and $\eta = h - d$ we obtain an equation for the height of the floodgate as a function of Q_{before} and l:

$$d = \left(\frac{nQ_{\text{before}}}{\sqrt{S}w}\right)^{\frac{3}{5}} - \frac{9}{4}\left(\frac{Q_{\text{before}} - 17500}{\sqrt{g}l}\right)^{\frac{2}{5}}.$$

For several choices of l we have plotted the graph in Figure 24.

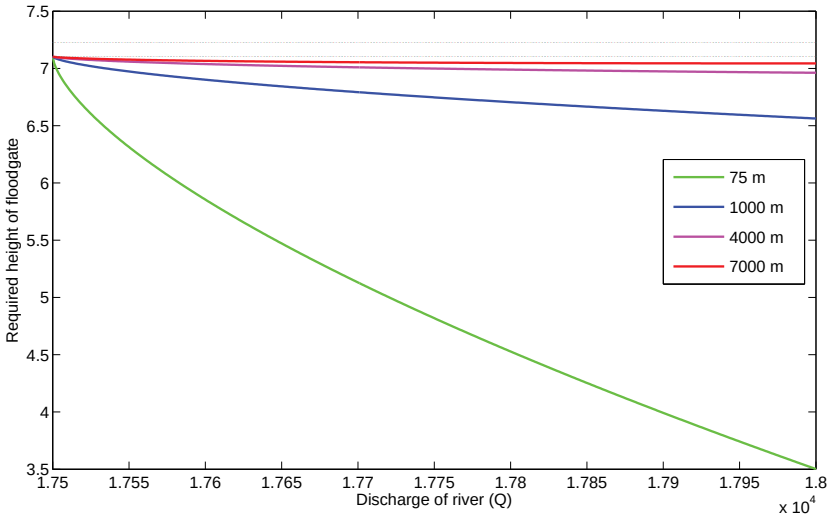


Figure 24: Height of floodgate in meters above the bottom of the river to reduce Q_{after} to 17500. Smaller floodgates (in length) need to be lowered more to allow enough flow through. When floodgates are controlled accordingly, each floodgate has the preferred property (49). The dotted horizontal lines at height 7.10 and 7.22 are the heights of the river at $Q_{\text{before}} = 17500$ resp. $Q_{\text{before}} = 18000$.

In theory it is now easy to build the preferred outlet. But implementation may be complicated as the height needs to be continuously adjusted and calculated from a known discharge. This relatively complicated procedure may be prone to errors and since high discharge is assumed to be a rare event it may be hard to tune the floodgate correctly.

To partially circumvent these problems we next look at a floodgate that can either be open or closed.

6.3.2 Open/closed floodgate

Here the idea is to lower the floodgate when $Q_{\text{before}} \geq 17500$ and to close it again when $Q_{\text{before}} < 17500$. The floodgate is lowered to a fixed level such that $Q_{\text{outlet}} = 500$ when $Q_{\text{before}} = 18000$. This last property yields a relation between d and l. After choosing l (and thus d) we can now plot the function (54) again, for values $17500 \leq Q_{\text{before}} \leq 18000$. Below $Q_{\text{before}} = 17500$ the floodgate is closed and no water leaves the river.

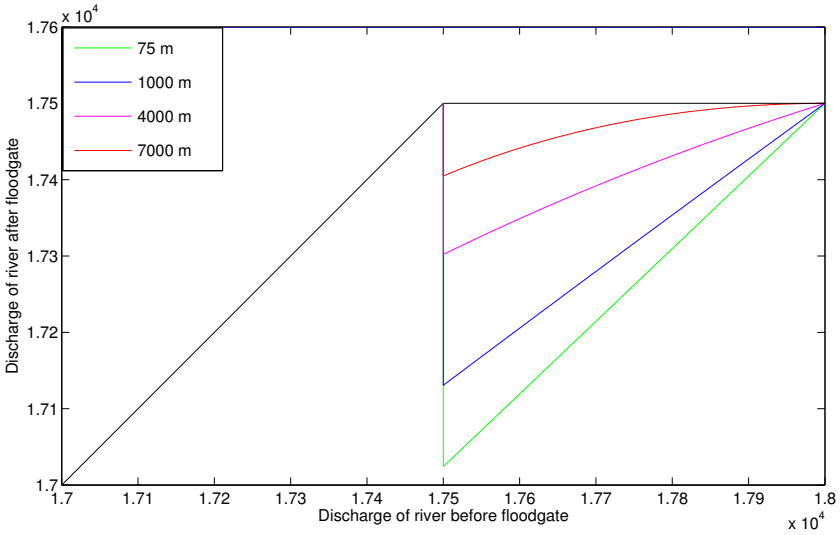


Figure 25: Effect of opening floodgate on discharge in river. For intermediate discharge too much water enters the retention area. This effect is larger for shorter floodgates, which need to be lowered more (see Figure 24). Below Q_{before} all functions coincide with the black line.

It has to be mentioned that for $l < 7000$, Q_{after} will grow beyond 17500 when $Q_{\text{before}} \geq 18000$, so the corresponding floodgates do not decrease discharge sufficiently in case of an extremely high Q_{before}. In fact $l \approx 7000$ corresponds to the only weir with

$$Q_{\text{after}} = 17500 \text{ when } Q_{\text{before}} = 18000,$$

and

$$Q_{\text{after}} \leq 17500 \text{ for all } Q_{\text{before}}.$$

So it is the only weir that meets the requirement in (49) and also behaves properly for $Q_{\text{before}} \geq 18000$.

6.4 Constructions with flow underneath

Each of the previous configurations had water flowing over the construction, which is necessary for options without control. We now shortly look into the option of flow underneath a closable floodgate or through a pipe, as sketched in Figure 26.

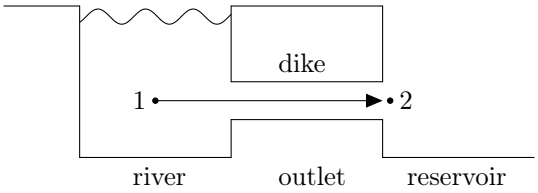


Figure 26: A cross section of the construction with a pipe in the dike. When the pipe is opened water from the river flows through the outlet, e.g. from 1 to 2.

All calculations will be based on Bernoulli's law, which is essentially an energy conservation law along streamlines:

$$\frac{\rho v_1^2}{2g} + \rho z_1 + \frac{p_1}{g} = \frac{\rho v_2^2}{2g} + \rho z_2 + \frac{p_2}{g}, \tag{56}$$

with v the speed of the water flow perpendicular to the river flow, g the gravity acceleration, z the elevation, p the pressure and ρ the density of water. The subscripts correspond with different locations, see Figure 26. The velocity v_1 (perpendicular to the flow of the river) can be assumed to be almost zero. Since the water is horizontally flowing through the pipe, $z_1 = z_2$. Furthermore, p_2 equals the air pressure and the pressure at position 1 may be written as $p_1 = p_2 + \rho g h$, with h the distance between the opening of the pipe and the water surface. Therefore, equation (56) reduces to

$$v_2 = \sqrt{2gh}, \tag{57}$$

which was already derived experimentally by Torricelli in the 17^{th} century.

If the outlet is located at the bottom, the height of the river is given by equation (52). For $Q = 18000$, this means $h = 7.22$. Using equation (57), the velocity is determined,

$$v_2 \approx 12\text{m/s} \approx 43\text{km/h}.$$

For comparison, we estimate the velocity of water flowing through the outlet in both the case of a weir using condition (55) ($h = 0.18$) and a small floodgate as discussed in section 6.3.1 with a length of 75 metre ($h = 3.7$). Using equation (50), we obtain:

$$v_{\text{weir}} = 0.89\text{m/s} \qquad v_{\text{floodgate}} = 4.0\text{m/s}.$$

The speed of the water flowing underneath a floodgate or through a pipe is much larger compared to the situation where the water is flowing over a weir or a floodgate. This may cause more damage for constructions inside the reservoir, but the size of the outlet can be made as small as $\frac{500}{12} \approx 40\text{m}^2$.

6.5 Discussion

Because of the nonlinearity of Q_{after} as a function of Q_{before}, it is not possible to construct a weir with preferred characteristic (49). One is naturally led to weaker requirements ((53) and (55)), of which the latter is preferred because it at least has sufficient inflow into the reservoir. In fact, for a weir, requiring that water starts flowing into the reservoir only at a certain discharge Q_{before}, is not the correct way to view the problem.

In theory, the continuously controlled floodgate fulfills the preferred characteristics of equation (49). Though this construction may be hard to implement in practice, Figure 24 gives an indication of how much a floodgate should be lowered at all possible discharges.

For long weirs or floodgates ($l = 13000$, $l = 7000$), the water levels at $Q_{\text{before}} = 18000$ are only about 12 and 18 centimeter above the weir or floodgate respectively, hence implementation would need a high level of precision. The authors doubt whether any model would be able to predict the required height of the weir with sufficient accuracy. This will certainly be the case for the model we used, as it implicitly assumes the width of the river to be large compared to the length of the outlet.

Compared to the weir, the open/closed floodgate prevents superfluous inflow for $Q_{\text{before}} \leq 17500$. This allows for smaller constructions, although the disadvantages of redundant inflow for $17500 < Q_{\text{before}} < 18000$ and insufficient inflow for $Q_{\text{before}} > 18000$ are more significant.

Even smaller constructions are possible with an outlet at the bottom of the river. However, one needs to take into account the speed of the water exiting the outlet.

7 Conclusions

Discharge levels into the river Rhine at the Dutch-German border are expected to rise due to climate change. Assets and livelihood in the Netherlands are thus at risk unless appropriate measures are taken. The study group working on this problem looked into three aspects of flooding the Rijnstrangen area as a protective measure. First we performed a statistical analysis of the recorded water levels. We found that an extreme discharge is expected to occur every 1250 years and to last for about

three and a half days. The Rijnstrangen area was found to be large enough to serve as a buffer for the excess of water. Next, we developed a partial differential equation model for the water flow in the river. Numerical results indicate that the time scale of transient phenomena is small compared to that on which discharge rises and falls. Transients can thus be neglected in the model. Finally the design of passive weirs and active floodgates was investigated. We advise the implementation of the latter option if indeed the Rijnstrangen area is to be used as a retention area.

References

[1] United States Army Corps of Engineers, "Engineering and Design - Flood-Runoff Analysis", Publication number EM 1110-2-1417, Chapter 9, 31 Aug. 1994.

[2] R. Vitolo, P. M. Rutti, A. Dell'Acqua, M. Felici, V. Lucarini and A. Speranza, "Accessing extremes of mid-latitudinal wave activity: methodology and application", Volume 61, Issue 1, pages 35–49, Tellus, 2009.

[3] Stoker: *Water waves: The Mathematical Theory with Applications*, Wiley, 1957.

[4] Hendriks: *Introduction to physical hydrology*, Oxford University Press, 2010.

Stress distribution during neck formation: An approximate theory

Bas van 't Hof Lotte Sewalt Keith Myerscough
Nicodemus Banagaaya Björn de Rijk Johan Dubbeldam*

15 March 2013

Abstract

In this paper we investigate the effects of deformation of a metal specimen, which is either a plate or a cylindrical rod in our case. In particular we study neck formation in tensile loading of a plastic metal. We try to generalize the work of Bridgman, who considered a purely two-dimensional geometry, to an effective theory that takes into account some essential three dimensional characteristics. That extending the description of neck formation to three dimensions is necessary was illustrated by recent experimental findings of [1].

We have studied existing models from the literature that describe necking for plates and cylinders to identify the consequences of the crucial assumption of uniform in-plane stress. We also developed a new model that we have not yet been able to analyze. Finally, using work of [4] in which a power law relation between the von Mises stress and the effective strain is used, a perturbation analysis for a simple flat geometry was performed. The perturbation analysis offers a good starting point for generalizing the work of Bridgman to three dimensions.

Keywords: Neck formation, von Mises stress, tensile pulling, plane stress assumption

1 Introduction

In many daily life situations materials are deformed. If deformations are very small the material will respond elastically. However, for metals deforming in collisions the plasticity regime is entered for relatively small deformations and the material will therefore not return to its initial state. This phenomenon can also be observed in uniaxial tension experiments of a specimen, which is in our case a metal bar or cylinder. If in an experiment the length of the bar is continuously increased by exerting a pulling force sufficiently large to accomplish elongation of the metal, then for certain

[0]e-mail: j.l.a.dubbeldam@tudelft.nl
*Corresponding author

loading the metal will yield and enter the plasticity regime. Still further enlarging the load gives rise to so-called "necking". A picture of typical neck occurring in a cylindrical specimen is depicted in Fig. 1.

The stress distributed in the metal has been shown in [1] to become fully three-dimensional, that is, the two-dimensional models originally proposed by Bridgman in the nineteenfifties will not be appropriate to model necking. The goal of this study during the SWI 2013 is to extend the 2D description of Bridgman, so that the findings in [1] are effectively incorporated. Such a description would be very useful in finite element codes for collisions of ships, as full 3D models are computationally very expensive and so an approximate incorporation of the stresses in 2D could drastically improve the computation time needed to analyze such situations. To find an extension of the existing models a good review of the literature and the most important concepts in continuum mechanics were required.

The paper is organized as follows. We first introduce some concepts from continuum mechanics needed for the description of the problem. Next we describe the model that we employed for both a cylindrical and a plane geometry. In section 3, we discuss our preliminary findings. Finally, in section 4, we summarize our results and make recommendations for future research.

Figure 1: A neck appears after applying a critical load to the cylinder.

2 Model

In this section we discuss three different models that were investigated. The first model was constructed from some special assumptions using the general theory and a modelling assumption about the strain rate. We discuss two possible choices for the strain rate. The first possibility was to assume a constant strain rate, the other was found in a paper [4]. The other two models we studies were both taken from

the literature: the original model of Bridgman in [2] and a more recent version by Kaplan in [5]. Before discussing the models we start with reviewing some concepts of continuum mechanics.

2.1 Concepts from continuum mechanics

To understand the problem of necking, we need some concepts of continuum mechanics that we here present. If a material is deformed a *displacement field* results, which is denoted as $\mathbf{u}(\mathbf{x})$. The *strain* ϵ is defined as

$$\epsilon = \frac{1}{2}\left(\nabla\mathbf{u} + (\nabla\mathbf{u})^T\right). \tag{1}$$

It is a symmetric tensor that is related to the stress tensor $\sigma(\mathbf{x})$, which assigns a value of the force per unit area to each point $\mathbf{x}$ in the material, by a constitutive relation. In the elastic case the constitutive relation between σ and ϵ is linear. In the regime where the material yields and the deformation is plastic, the situation is much more difficult. However, for the one-dimensional case an empirical relation between stress and strain still exists as we will see.

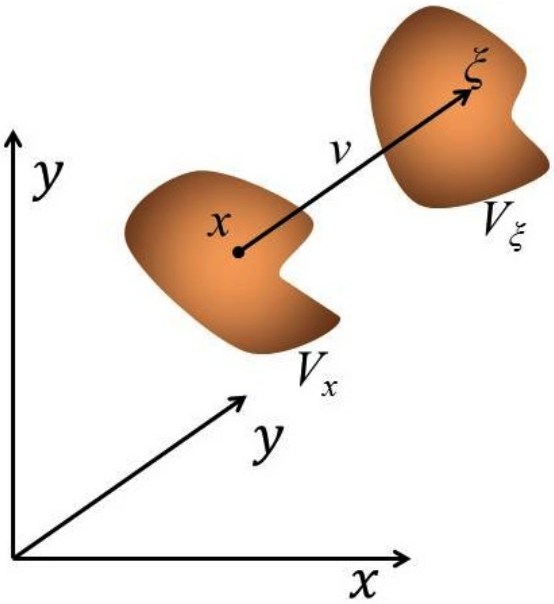


Figure 2: A deformation of a material may lead to a change in volume and stresses throughout the solid.

If we assume that the yielding is unaffected by moderate hydrostatic pressure or

tension, which is correct to a first approximation it follows that the yielding condition only depends on the principal components, or eigenvalues, of the deviatoric stress tensor, σ', defined by

$$\sigma'_{ij} = \sigma_{ij} - \frac{1}{3}\mathrm{Tr}(\sigma)\delta_{ij}. \tag{2}$$

The eigenvalues of σ', $\{\sigma'_1, \sigma'_2, \sigma'_3\}$ are not independent since they satisfy

$$\sigma'_1 + \sigma'_2 + \sigma'_3 = 0,$$

as follows immediately from the definition of deviatoric stress. If we further assume that the material isotropic, the condition for which yielding will occur only depends on the eigenvalues, of which only two are independent. So we can write the equation for yielding

$$F(\sigma'_1, \sigma'_2) = 0, \tag{3}$$

with F an arbitrary function.

Finally, we use the von Mises proposal (1937) which has been verified in a number of experiments that the yielding condition depends quadratically on σ'_1, σ'_2, σ'_3. Using symmetry this gives

$$\sigma'^2_1 + \sigma'^2_2 + \sigma'^2_3 = \frac{2}{3}\bar{\sigma}^2, \tag{4}$$

where $\bar{\sigma}$ is called the *von Mises* stress. This nonlinear relation can be expressed in term of the eigenvalues of the original stress tensor σ as

$$\bar{\sigma} = \sqrt{\frac{(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2}{2}}. \tag{5}$$

It turns out that in the plastic regime the von Mises stress is related to the strain by a power law. For the one-dimensional case this relation is

$$\bar{\sigma} = C\epsilon^N, \tag{6}$$

where C is a material constant and N is a power law exponent whose value is in the range $[0.1, 0.2]$. To generalize the constitutive relation (6) to three dimensions different approaches are possible, that we will discuss in Model I. The original Bridgman model and related to it the model by Kaplan, will be explained in the subsections Model II and Model III.

2.2 Model I

1. *Steady motion*
   In this approach we try to generalize (6) to 3D by defining an effective scalar strain, $\bar{\epsilon}$ such that (6) remains valid when ϵ is replaced by $\bar{\epsilon}$. The problem is that

we need to know the effective strain, which in general depends on the loading history. Even though we do not know how the strain evolves, the strain rates $\dot{\epsilon}$ obey [3]

$$\dot{\epsilon} = \sqrt{\frac{2}{9}\left((\dot{\epsilon}_1 - \dot{\epsilon}_2)^2 + (\dot{\epsilon}_1 - \dot{\epsilon}_3)^2 + (\dot{\epsilon}_3 - \dot{\epsilon}_2)^2\right)}, \tag{7}$$

where the dot denotes differentiation with respect to time. Assuming constant time derivatives, Eq. (7) is also valid for the strains, which implies that the dots in (7) can simply be left out.

In order to close the equations we need two more equations. To this end we invoke the Levy-Mises flow rules which state

$$\frac{\epsilon_1 - \epsilon_2}{\sigma_1 - \sigma_2} = \frac{\epsilon_1 - \epsilon_3}{\sigma_1 - \sigma_3} = \frac{\epsilon_3 - \epsilon_2}{\sigma_3 - \sigma_2}. \tag{8}$$

and give the two necessary conditions to close the system.

2. *Hutchinson theory*
In a paper by Hutchinson et al. [4] it was proposed to generalize relation (6) in the following natural fashion

$$\dot{\epsilon}_{ij} = \frac{3}{2}\alpha\bar{\sigma}^{n-1}\sigma'_{ij}, \tag{9}$$

with α a material constant and n the strain hardening exponent that is typically larger than 1 and directly related to N. We remark that relation (9) is similar to (6) if we require in addition that the the strain rate is time independent and that the strain and the stress have a common set of eigenvectors.

We next discuss two existing models in the literature. One is the orginal model of Bridgman for necking. The other model is a model introduced by [5].

2.3 Model II: The Bridgman model for necking

Bridgman discusses neck formation in a cylindrical tensile specimen. The distribution of stress across a transverse section is, however, not necessarily uniform. Measurements generally only provide data about the mean stress through the neck. Since the shape of the neck is not known beforehand calculating the stress distribution is extremely difficult and determining its shape from first principles requires tracing the time evolution of the dynamical process of neck formation as is done by [4]. Bridgman made the assumption, based on his own experimental data, that at the neck minimum the stress is uniformly distributed. From the area reduction at the position of the neck, which we call $x = 0$, the strain is known. Furthermore the strain rate can be shown to be proportional to the radial distance r. For equation for equilibrium is again given by

$$\frac{\partial \sigma_{rr}}{\partial r} + \frac{\partial \sigma_{rz}}{\partial z} = 0, \quad \text{at } z = 0. \tag{10}$$

The yield condition is also very much simplified in this case, because $\sigma_{rr} = \sigma_{\theta\theta}$ as $\dot{\epsilon}_{rr} = \dot{\epsilon}_{\theta\theta}$ and the strain rate is proportional to r. This gives the yield condition

$$\sigma_{zz} - \sigma_{rr} = Y. \tag{11}$$

Bridgman then lets his principal stress direction in a meridian plane to the axis as in Fig. 2.1. We have

$$\sigma_{zz} \simeq \sigma_3, \qquad \sigma_{rr} \simeq \sigma_1, \qquad \sigma_{rz} \simeq (\sigma_3 - \sigma_1)\psi. \tag{12}$$

This implies that the yield condition is

$$\sigma_3 - \sigma_1 = Y + \mathcal{O}(\psi), \tag{13}$$

from which it immediately follows that

$$\left(\frac{\partial \sigma_{rz}}{\partial z}\right)_{z=0} = Y \left(\frac{\partial \psi}{\partial z}\right)_{z=0} = \frac{Y}{\rho}. \tag{14}$$

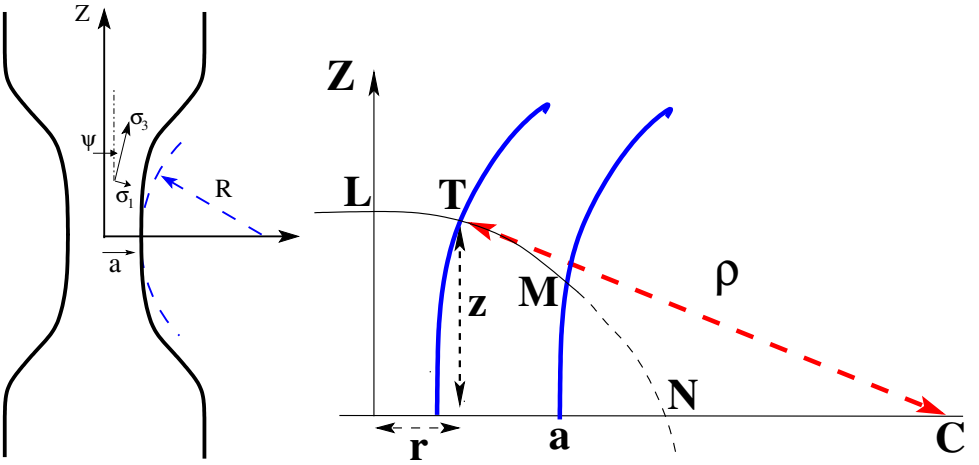


Figure 3: The geometry for the Bridgman model. Neck formation in a cylindrical geometry. Plane and uniform stress assumption at $z = 0$.

This leads to the following partial differential equation for σ_{zz}

$$\frac{\partial \sigma_{zz}}{\partial r} + \frac{Y}{\rho} = 0, \quad \text{for } z = 0. \tag{15}$$

Using the symmetry in Fig. 3, we can deduce that

$$\rho^2 \simeq CT^2 = OC^2 - ON^2 \simeq (r + \rho)^2 - ON^2, \tag{16}$$

hence for any point on OA we have

$$\rho = \frac{a^2 + 2aR - r^2}{2r}. \tag{17}$$

From (17) [2] obtained the formula

$$\frac{\sigma_{zz}}{Y} = 1 + \ln\left(\frac{a^2 + 2aR - R^2}{2aR}\right). \tag{18}$$

2.4 Model III: The Kaplan model

In 1973, M.A. Kaplan [5], extended the analysis of Bridgman. His analysis deals with necking in bars of mild steel. One assumption that is made here is that the displacement will be axially symmetric in a symmetric bar. This reduces the number of unknowns in this analysis drastically. Another important assumption is that the deformation will be produced entirely by plastic flow, so the elastic deformation is neglected. This can be supported by the fact that the elastic contribution to the total axial strain in ductile metals is of the order of 1 percent at the onset of necking, and decreases as necking proceeds.

Kaplan uses the work of Bridgman to calibrate his own model. Experimentally, Bridgman concluded that during necking, the ratio between the external radius and core radius remains constant during necking from the point on where necking has started. This means that $r/a = r_0/a_0$ where r and r_0 are the deformed and initial radial positions of a particle respectively. The radius of the profile, a is measured on a plane on which the particle lies after deformation. The initial radius of the bar is a_0. This fixed quantity implies an equation for the radial displacement in terms of the profile of the bar a, which is assumed to be a function of the z-direction (see Fig. 1) and time t. Kaplan then performs a formal analysis to derive the strain fields and strain rates in terms of this profile.

From the strain field, the Levy-Mises plasticity equations and use of the radial displacement uniformity, the stress field could be determined. This is a closed system of eight equations and unknowns. After some formal derivations the predicted profile of the neck, $a(z,t)$ is determined and has the shape of a parabola.

The analysis of [5] is valid throughout a significantly larger portion of the plastic flow region than the analysis of [2]. The analysis holds only for a necked cylindrical mild steel tensile specimen. However, experiments show a high similarity for ductile metals and Kaplan argues that his results apply to those materials as well.

A nice feature of Kaplan's work is that all parameters used in his model can be measured very well. Also, his results have good agreement with experiments and tensile tests.

3 Results

To understand the necking process in time we studied a thin sheet as sketched in Fig. 4. In this study we use the formalism as presented in [4]. The sheet is initially in

rest and then a force P per unit length is applied at the ends, which are initially at $x = -L$ and $x = L$. For simplicity we assume $L \gg 1$. The width of the sheet is given by $h(x,t)$. The stress is linearly related to P by at each section $x_1 = constant$

$$P = \sigma_{11} h \tag{19}$$

We next make the plane stress assumption, that is, all quantities are independent of x_2. Moreover only the σ_{22} and σ_{11} do not vanish and these are taken to be *uniform* over a section with constant x_1. Finally, we assumed symmetry with respect to the x_3 coordinate and imposed volume conservation, that is,

$$\dot{\epsilon}_{11} + \dot{\epsilon}_{22} + \dot{\epsilon}_{33} = 0, \tag{20}$$

with the additional condition that $\dot{\epsilon}_{22} = 0$, as there is no x_2 dependence. If we use Eq. (9), we find after calculating the von Mises stress

$$\dot{\epsilon}_{11} = -\dot{\epsilon}_{33} = \frac{\alpha\sqrt{3}}{2} \left(\frac{\sqrt{3}P}{2h} \right)^n . \tag{21}$$

We can now find the evolution equation of $h(x_1,t)$, by calculating the time derivative of h with respect to time, keeping in mind that there will also be a convective contribution, that is

$$\dot{h} = \frac{\partial h}{\partial t} + v_1 \frac{\partial h}{\partial x_1} = \dot{\epsilon}_{33} h. \tag{22}$$

In Eq. (22) we introduced $v_1(x_1)$, which is the velocity in the x_1-direction. If we now use Eq. (21), we have derived an evolution equation for $h(x_1,t)$, which however includes the velocity v_1.

3.1 Perturbation analysis

To find an approximate solution to Eq. (22) we perform a perturbation analysis. We will repeat here the reasoning of [4]. First we assume that the sheet is perfect and therefore $h(x_1,t)$ only depends on t, next we add a small sinusoidal perturbation, so we can write

$$h(x_1,t) = h_0(t) \left(1 - \xi \cos\left(\frac{2\pi x_1}{l} \right) \right), \tag{23}$$

where ξ is a small parameter. The solution to zero order in ξ satisfies

$$\dot{h}_0 = -\frac{1}{2}\sqrt{3}\alpha \left(\frac{\sqrt{3}P}{2} \right)^n h_0^{1-n} = -h_0 f(h_0), \tag{24}$$

where we introduced $f(h_0) = \dot{\epsilon}_{11}$ for notational convenience. Eq. (24) is easily solved as

$$h_0(t) = \left[h_0^n(0) - \frac{n}{2}\sqrt{3}\alpha \left(\frac{\sqrt{3}P}{2} \right)^n t \right]^{\frac{1}{n}}. \tag{25}$$

The first order contribution in ξ can be obtained by substituting Eq. (21) in Eq. (22) and next differentiating with respect to x_1. This yields

$$\frac{\partial^2 h}{\partial x_1 \partial t}\frac{\partial h}{\partial x_1} - \frac{\partial^2 h}{\partial x_1^2}\frac{\partial h}{\partial t} = \left(\frac{\partial h}{\partial x_1} \right)^2 [-f'(h) - f(h)] + f(h)\frac{\partial^2 h}{\partial x_1^2}, \tag{26}$$

with f as defined in (24). We could try to solve the nonlinear equation (26) numerically, but we have to keep in mind that that this equation is only valid for in-plane stress and the strain rates only depend on x and not on z . If we do make such an assumption then solving Eq. (26) would determine how a perturbation $h(x,t)$ would evolve in time. For reasons of time we have not numerically solved (26), but rather delved deeper in the theory behind neck formation closely following [4].

Of course, like in [4] it is possible to substitute the sinusoidal expression for h (23) and see keeping only terms linear in ξ to calculate the linear variation of the $h(x_1,t)$ in time as a consequence of the convective term. We will not repeat this calculation here, but rather try to determine the functional form of h when a perturbation is introduced.

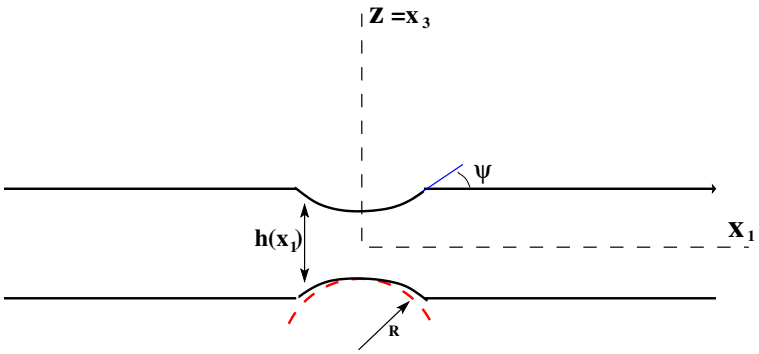


Figure 4: A perturbation analysis can help to calculate the initial neck shape, without assumptions on the curvature.

Assume now that the wavelength of the perturbation is very large and define $X = \beta x_1$, where β is of the order of the inverse wavelength as in [4]. We use X and $z = x_3$ as coordinates. In the creeping flow approximation we can use $\operatorname{div}\sigma = 0$, as

an equilibrium condition. In components this reads

$$\beta\frac{\partial\sigma_{11}}{\partial X}+\frac{\partial\sigma_{13}}{\partial z}=0 \tag{27}$$

$$\beta\frac{\partial\sigma_{13}}{\partial X}+\frac{\partial\sigma_{33}}{\partial z}=0 \tag{28}$$

The boundary condition at $z=h(X)/2$ is given by

$$\begin{aligned}-\sigma_{11}\sin\psi+\sigma_{13}\cos\psi&=0\\ -\sigma_{13}\sin\psi+\sigma_{33}\cos\psi&=0\end{aligned} \tag{29}$$

where $\tan\psi=\beta h'(X)/2$ and the prime denotes differentiation with respect to X. The boundary conditions can be expanded up to order β^2 as well as the stresses

$$\begin{aligned}\sigma_{11}&=\sigma^{(0)}(X)+\beta\sigma_{11}^{(1)}+\beta^2\sigma_{11}^{(2)}+\cdots\\ \sigma_{33}&=\beta\sigma_{33}^{(1)}+\beta^2\sigma_{33}^{(2)}+\cdots\\ \sigma_{13}&=\beta\sigma_{13}^{(1)}+\beta^2\sigma_{13}^{(2)}+\cdots\end{aligned} \tag{30}$$

and the strain rates

$$\begin{aligned}\dot\epsilon_{11}&=-\dot\epsilon_{33}=\dot\epsilon(X)+\beta\dot\epsilon_{11}^{(1)}+\beta^2\dot\epsilon_{11}^{(2)}+\cdots\\ \dot\epsilon_{13}&=\dot\epsilon_{13}^{(0)}+\beta\dot\epsilon_{11}^{(1)}+\beta^2\dot\epsilon_{11}^{(2)}+\cdots,\end{aligned} \tag{31}$$

where $\epsilon_{13}^{(0)}=0$, but is kept for clarity as in [4].

The strain rates are related to the flow velocity in the following way

$$\dot\epsilon_{11}=\beta\frac{\partial v_1}{\partial X}\qquad\qquad\dot\epsilon_{33}=\frac{\partial v_3}{\partial z}\qquad\qquad 2\dot\epsilon_{13}=\frac{\partial v_1}{\partial z}+\beta\frac{\partial v_3}{\partial X}. \tag{32}$$

To go beyond the in-plane plane stress assumption we would need to take into account $\dot\epsilon_{22}$, which could be achieved in a perturbative approach. Of course, this would make the equations much more difficult to solve, but in this way a good estimate of non in-plane effects can be given.

We next continue with the Hutchinson analysis. A major simplification of $\dot\epsilon_{22}=0$ is that we can express σ_{22} in terms of σ_{11} and σ_{33} as

$$\sigma_{22}=\frac{\sigma_{11}+\sigma_{33}}{2}.$$

If we write all expressions up to order β^2, we find the following values of the stress and strain rates

$$\begin{aligned}\sigma_{11}&=\sigma^{(0)}\left[1+\frac{\beta^2(n-2)hh''}{12n}\left(1-12\frac{z^2}{h^2}\right)\right]\\ \sigma_{33}&=\sigma^{(0)}\frac{\beta^2hh''}{8}\left(1-4\frac{z^2}{h^2}\right)\\ \dot\epsilon_{11}&=\dot\epsilon^{(0)}\left(1-\frac{\beta^2hh''}{24}\right)\left(n+4+12(n-4)\frac{z^2}{h^2}\right)\end{aligned} \tag{33}$$

From Eqs. (33) it can be seen that only when the strain hardening exponent n equals 4, $\dot{\epsilon}_{11}$ will be uniform across the neck.

Furthermore, we can now compare the result in [4] with that in [2] by introducing the radius of curvature as

$$\frac{1}{R} = \frac{\beta^2}{2} h''. \tag{34}$$

Buy eliminating h from the expression for σ_{11} and σ_{33} we obtain

$$\begin{aligned} \frac{\sqrt{3}\sigma_{11}}{2\bar{\sigma}} &= 1 + \frac{h}{4R}\left[1 - \frac{z^2}{h^2}\right] \\ \frac{\sqrt{3}\sigma_{33}}{2\bar{\sigma}} &= \frac{h}{4R}\left[1 - \frac{z^2}{h^2}\right], \end{aligned} \tag{35}$$

which agree exactly with the Bridgman expressions to order $\frac{z^2}{h^2}$.

4 Conclusions and recommendations

We conclude that the problem of neck formation is far from trivial. In order to find a generalization of the Bridgman result we studied 3 different models. The model of Kaplan is interesting and may prove very useful, however, we have not been able to generalize this to more dimensions. We constructed a model using the assumption of constant strain rates, which makes it much easier to take the convective terms into account. Finally, we found a study of [4], which appears to be a good alternative to the Bridgman theory. This model takes into account time dependent strain rates and can indeed be generalized to cases in which there is no assumption made about the stresses all being in-plane. Unfortunately, time has not permitted to do the complete analysis, but a perturbation analysis along similar lines as that in [4] would open new avenues for the resolution of the problem of taking the necking problem to three dimensions.

Another direction which may be fruitful, is to start with one of the constitutive models proposed in this study and investigate them numerically. Comparison between experimental data and modeling results would indicate which constitutive relation would be best. Next, complementary to the perturbation analysis, a numerical study of the nonlinear model could be performed at reasonable computational costs, so that a good estimate of the errors resulting from the assumptions such as a uniform stress distribution across the neck can be obtained.

5 Acknowledgement

We thank Carey Walters for his patience and help and very fruitful discussions during the days we have been working on the problem.

References

[1] Y. Bai and T. Wierzbicki. Applications of extended mohr-coulomb criterion to ductile fracture. *Int. J. Fract.*, 161:1,20, 2010.

[2] P. Bridgman. *Studies in Large plastic flows and fracture.* Harvard University Press, 1964.

[3] J. Chakrabaty. *Theory of Plasticity.* Elsevier Butherworth-Heinemann, 2006.

[4] J. Hutchinson, K. Neale, and A. Needleman. Sheet necking i: Validity of plane stress assumptioms of the long wavelength approximation. *Mechanics of sheet metal forming*, 8:111–126, 1978.

[5] M. Kaplan. The stress and deformation in mild steel during axisymmetric necking. *J. Appl. Mech.*, 8:271–276, 1973.

Acknowledgements

The generous financial support from NWO and STW together with the contributions from the Lorentz Center, KWG, and ECMI and by the participating industries (Fytagoras, Heineken, Nedcoffee, Philips Research, Rijkswaterstaat, TNO) made the SWI 2013 possible. The success of the meeting is due to both the active involvement of the mathematicians and the transparence as well as close collaboration of the industrial partners. We thank you all. We are indebted to the Lorentz Center for hospitality and kind support.

The organizers of SWI 2013