Grups d'Estudi de Matemàtica i Tecnologia

Barcelona, July 2010

Edited by

Aureli Alabert (UAB)
Tim Myers (CRM)
Jordi Saludes (UPC)

ESGI 78

European Study Groups in Industry

Centre de Recerca Matemàtica
Campus de Bellaterra, Edifici C
08193 Bellaterra (Barcelona)

First edition: March 2012

ISSN 2014-2323 (printed edition)
ISSN 2014-2331 (electronic edition)

Legal deposit:

Preface

The 7th *Grups d'Estudi de Matemàtica i Tecnologia* (GEMT) was held at the Centre de Recerca Matemàtica (CRM) and the Universitat Autònoma de Barcelona (UAB) from 6th to 8th July 2010. This was the first GEMT to be held under the auspices of the European Study Groups in Industry (ESGI) and it was the 78th in the ESGI series.

The problems studied at the meeting covered a wide range of fields. Cisco Systems presented a problem on bandwidth consumption; the group Sistemes Avançats de Control proposed one on flood prevention, and Sabirmedical presented one on monitoring blood pressure. This final problem led to a follow-up meeting entitled *Mathematical Modeling of Blood Flow and the Baroreflex System*, held at the CRM in December 2010. The final report of this volume describes the results of that meeting.

Participants came primarily from the Barcelona region, although this year there was stronger than usual international contingent through the participation of researchers from the Oxford Centre for Collaborative Applied Mathematics (OCCAM). Attendance at the meeting was free for both companies and academics. Funding came primarily through Ingenio Mathematica (i-MATH), and OCCAM participants were supported by the KAUST Global Research Partnership. The organisers wish to acknowledge both sources of financial support.

Tim Myers, Barcelona 2010

Contents

Analysis of the Baroreflex Model for Automatic Interpretation of the Plethysmograph

Problem presented by
Vicent Ribas Ripoll (Sabirmedical)

Report prepared by
Tim Myers (CRM), Maria Bruna (OCCAM), Joan Solà-Morales (UPC)

Study group contributors
Maria Bruna (OCCAM), Antoni Guillamon (UPC), Adam Mahdi (North Carolina State University), Tim Myers (CRM), Joan Solà-Morales (UPC), Jeff Springer (OCCAM), Amy Smith (OCCAM)

1.1 Introduction

A pulse oximeter is a device that measures oxygen saturation in blood. Typically it functions by shining two lights of different wavelength (but both close to infra-red) through a translucent part of the body. The different wavelength lights are absorbed to differing degrees by the oxygenated and deoxygenated haemoglobin and so the ratio of oxygenation to deoxygenation may be calculated. Since arterial blood vessels respond to pressure changes (due to the heart pumping), the signal is time-dependent and so the output of the pulse oximeter may also be used to monitor the heart rate. In fact this variation in the signal is essential to the functioning of the device, since the pulse oximeter only uses the varying part of the signal to distinguish light absorption from blood and the surrounding tissue.

Standard uses for the pulse oximeter include medical monitoring of oxygenation and heart rate and also the diagnosis of sleep disorders. However, it is recognised that the output signal (the photoplethysmograph or *pleth*

for short) contains a wealth of information that may be exploited for monitoring or diagnosis of other conditions. The pleth also exhibits very similar behaviour to the standard blood pressure signal.

Until recently there was no way to correlate the pleth to blood pressure. However, this problem has been solved by Sabirmedical. One of their goals now is to develop a model of the cardiovascular system and relate this model to the pressure predicted by the pleth, with the aim of automatically diagnosing certain conditions. Consequently, the research goal at GEMT was to develop an appropriate mathematical model.

1.2 Mathematical model

To understand the output of the pleth, mathematical models of the cardiovascular system are required. The modelling of this sytem can be tackled in various ways. Currently at Sabirmedical the output is interpreted through the baroreflex model of Ottesen [6]. This is a simple ODE model describing the interplay between arterial and venous pressure and heart rate. A more detailed approach involves modelling the flow in blood vessels through a PDE description. Although not immediately obvious, there is a relation between these approaches. Indeed, the Ottesen model may be considered an extension of the classical windkessel ODE model which describes the relation between excess pressure and flow rate in the circulatory system [4, p. 471]. The windkessel model may be obtained from standard flow equations in the limit of plug flow and zero fluid density. The relation between ODE and PDE models is described in more detail in [5]. However, in the following report we will focus solely on the ODE approach.

The heart pumps blood through the body. Heart contraction usually begins in the sino-atrial node. The action of this node is controlled by electrical pulses travelling through two major systems of nerves, the sympathetic and parasympathetic systems. The parasympathetic system is relatively fast acting but the sympathetic system is slow. The sympathetic system works in three ways. First it starts the contraction of blood vessels by the release of vasoconstrictors. At the same time, hormones are released into the blood and carried throughout the body to also cause contraction of blood vessels. Finally, it acts to increase the heart rate. This triple action leads to a much slower response than the parasympathetic system and so simple mathematical models require incorporating a delay term. More complex systems may not need the delay term, which really reflects the fact that the sympathetic system is not correctly modelled. Consequently, lumped models of the cardiovascular system typically involve delay differential equations. A summary of simple models and methods for deriving them may be found in [3, 5].

1.3 Baroreflex model

The basis of the current study is the baroreflex model of Ottesen,

$$\dot{P}_a(t) = -\frac{1}{c_a R} P_a(t) + \frac{1}{c_a R} P_v(t) + \frac{V_{str}}{c_a} H(t), \tag{1.1a}$$

$$\dot{P}_v(t) = \frac{1}{c_v R} P_a(t) - \left(\frac{1}{c_v R} + \frac{1}{c_v r} \right) P_v(t), \tag{1.1b}$$

$$\dot{H}(t) = f(P_a(t), P_a(t - \tau_H)), \tag{1.1c}$$

where P_a, P_v are the arterial and venous pressures, H is the heart rate, c represents the vessel compliance, R is the resistance, and V_{str} is the stroke volume. The function f is given by

$$f = \frac{\alpha_H}{1 + [P_a(t - \tau_H)/\alpha_s]^{\beta_s}} - \frac{\beta_H}{1 + [\alpha_p / P_a(t)]^{\beta_p}}. \tag{1.2}$$

Typical parameter values are provided in Table 1.1; see [6].

Constant	Value	Units
c_a	1.55	$\mathrm{ml\,mmHg^{-1}}$
c_v	519	$\mathrm{ml\,mmHg^{-1}}$
R	1.05	$\mathrm{mmHg\,s\,ml^{-1}}$
r	0.068	$\mathrm{mmHg\,s\,ml^{-1}}$
V_{str}	67.9	ml
α_0	93	mmHg
α_s	93	mmHg
α_p	93	mmHg
α_H	0.84	mmHg
β_0	7	1
β_s	7	1
β_p	7	1
β_H	1.17	1
P_0	93	mmHg
τ	4	s

Table 1.1: Typical parameter values for a baroreflex model

To better understand the system, we first non-dimensionalise by setting

$$P_a = \bar{P}_a x_a, \quad P_v = \bar{P}_v x_v, \quad H = \bar{H} x_h, \quad t = \bar{t} T, \tag{1.3}$$

where the terms with overbars represent constant, typical values. The equation for the arterial pressure (1.1a) then becomes

$$\frac{\bar{P}_a}{\bar{t}} \dot{x}_a = -\frac{\bar{P}_a}{c_a R} x_a + \frac{\bar{P}_v}{c_a R} x_v + \frac{V \bar{H}}{c_a} x_h. \tag{1.4}$$

This rearranges to give

$$\dot{x}_a = -\frac{\bar{t}}{c_a R} x_a + \frac{\bar{P}_v \bar{t}}{c_a R \bar{P}_a} x_v + \frac{V \bar{H} \bar{t}}{c_a \bar{P}_a} x_h. \tag{1.5}$$

The driving mechanisms for changing the arterial pressure are the pumping of the heart and the arterial pressure itself. We therefore choose the corresponding coefficients in equation (1.5) to be the unity, which leads to the following time and pressure scales:

$$\bar{t} = c_a R, \qquad \bar{P}_a = \frac{V \bar{H} \bar{t}}{c_a} = V \bar{H} R. \tag{1.6}$$

Finally, the arterial pressure equation (1.5) may be written

$$\dot{x}_a = -x_a + \frac{\bar{P}_v}{\bar{P}_a} x_v + x_h. \tag{1.7}$$

Ottesen [6] chooses the venous pressure scale to equal the arterial pressure scale, $\bar{P}_v = \bar{P}_a$. A more rational approach is to choose the scale through the appropriate governing equation. The non-dimensionalised version of the venous pressure equation (1.1b) is

$$\dot{x}_v = \frac{\bar{P}_a c_a}{\bar{P}_v c_v} x_a - \left(\frac{c_a}{c_v} + \frac{c_a R}{c_v r} \right) x_v. \tag{1.8}$$

The venous pressure is driven by the arterial pressure and so we see that the correct scale is $\bar{P}_v = c_a \bar{P}_a / c_v$. Since $c_a/c_v = 1.55/519 \approx 0.003$, it is clear that $\bar{P}_v \ll \bar{P}_a$ and Ottesen's scaling is inappropriate. One obvious consequence of this observation is that the venous pressure term in equation (1.1a) is negligible (note that the same conclusion was reached in [3]). Taking values from Ottesen's paper (see Table 1.1) we also note that

$$\epsilon_2 = \left(\frac{c_a}{c_v} + \frac{c_a R}{c_v r} \right) \approx 0.01, \tag{1.9}$$

which shows that the contribution of the venous pressure to (1.8) is negligible.

To determine the scale $\bar{H}$, we consider the non-dimensionalised version of the heart rate equation (1.1c),

$$\dot{x}_h = \frac{\alpha_H \bar{t}/\bar{H}}{1 + [\bar{P}_a x_a(T - \tau^*)/\alpha_s]^{\beta_s}} - \frac{\beta_H \bar{t}/\bar{H}}{1 + [\alpha_p/\bar{P}_a x_a(T)]^{\beta_p}}, \tag{1.10}$$

where $\tau^* = \tau/\bar{t}$. Since $\alpha_H \approx \beta_H$, the choice of $\bar{H}$ can come from either term on the right hand side of (1.10). Given that β_H is slightly larger than α_H, we choose $\bar{H} = \beta_H \bar{t} = \beta_H c_a R \approx 1.8$ and then

$$\begin{aligned} \dot{x}_h &= \frac{\alpha}{1 + [\lambda_1 x_a(T - \tau^*)]^{\beta_s}} - \frac{1}{1 + [\lambda_2/x_a(T)]^{\beta_p}} \\ &= \tilde{f}(x_a(T), x_a(T - \tau^*)), \end{aligned} \tag{1.11}$$

where $\alpha = \alpha_H/\beta_H, \lambda_1 = \bar{P}_a/\alpha_s, \lambda_2 = \alpha_P/\bar{P}_a$. Defining $\epsilon_1 = \bar{P}_v/\bar{P}_a \approx 0.003$, we may write the governing equations as

$$\dot{x}_a = -x_a + \epsilon_1 x_v + x_h, \tag{1.12a}$$
$$\dot{x}_v = x_a - \epsilon_2 x_v, \tag{1.12b}$$
$$\dot{x}_h = \tilde{f}(x_a(T), x_a(T - \tau^*)). \tag{1.12c}$$

Since $\epsilon_i \ll 1$, we may neglect these terms without losing accuracy and so the venous pressure equation uncouples from the system. The problem then simply reduces to solving, for x_a and x_h,

$$\dot{x}_a = -x_a + x_h, \tag{1.13a}$$
$$\dot{x}_h = \tilde{f}(x_a(T), x_a(T - \tau^*)). \tag{1.13b}$$

In fact we could simply differentiate equation (1.13a) and then replace $\dot{x}_h$ to solve a single second-order equation involving only x_a.

Note that, with this scaling, all the model parameters are contained in the function $\tilde{f}(x_a(T), x_a(T - \tau^*))$ and the form of the solution is determined through the values of

$$\alpha = \frac{\alpha_H}{\beta_H}, \quad \beta_s, \quad \beta_p, \quad \lambda_1 = \frac{VR^2\beta_H c_a}{\alpha_s}, \quad \lambda_2 = \frac{\alpha_p}{VR^2\beta_H c_a}. \tag{1.14}$$

That is, the choice of f and the values of the parameters within it are key to the success of the model. Given that f is defined in a rather ad hoc manner, it would be sensible to investigate the accuracy of this form and consequently determine whether it could be improved.

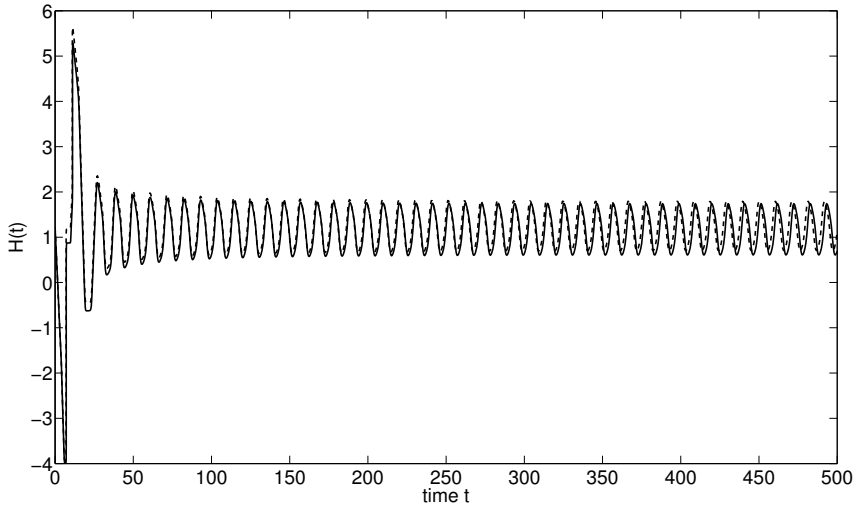


Figure 1.1: Comparison of $H(t)$ for reduced (dashed) and full (solid) model

In Figures 1.1 and 1.2 we show comparisons of the dimensional heart rate and arterial pressure obtained through the reduced system, consisting of system (1.13) (dashed line) and the full system (1.12) (solid line). The numerical solutions were obtained using modified versions of the Matlab routine `ddex2`.

Note that the high values obtained initially are due to inaccurate guesses in the initial conditions. However, these quickly settle to more sensible values which endure over a long time scale. The correspondence between the curves makes it clear that the small terms identified through non-dimensionalisation can be neglected for a certain length of time —in the figures this is true up to around 250 s. However, if the simulation is allowed to run for sufficiently large times then the curves start to move out of synch.

This is termed the *secular effect*. It will be noticeable when $\epsilon_1 t = \mathcal{O}(1)$ or $t = \mathcal{O}(100)$, a figure consistent with the observation that the curves start to diverge around 250 s. We could improve accuracy through a multiple scale (or, in this case, two-time scale) analysis; see Bender & Orszag [1] for example. However, given the short time of GEMT, this route was not explored.

1.4 The dicrotic notch

In real measurements of the arterial pressure P_a, at least in healthy patients, a secondary bump known as the *dicrotic notch* is observed in the descending phase of the signal. These bumps may be seen in Figure 1.3, which shows

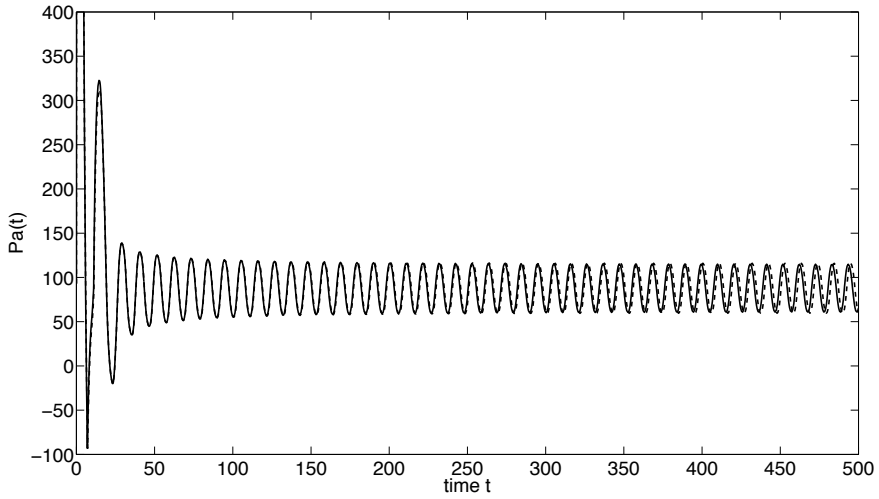


Figure 1.2: Comparison of arterial pressure for reduced (dashed) and full (solid) model

the arterial pressure P_a obtained by using an intra-arterial catheter in a real patient. In Figure 1.4 we show a close-up of a dimensional result produced by the mathematical model of equations (1.1). Whilst we are clearly able to predict the general variation of blood pressure, our model currently misses the dicrotic notch.

The dicrotic notch is caused by the closure of the aortic valve, which is due to the pressure difference between ventricle and aorta. When the ventricle pressure is greater than that in the aorta, the valve stays open, whereas when the ventricle pressure falls below that of the aorta then the valve slams shut. At this point, some blood which was situated on the ventricle side is rapidly pushed through to the aorta. This small, rapid injection of mass produces the pressure pulse observed as the dicrotic notch. *Therefore, we could say that the valve closure occurs at a certain (decreasing) value of P_a and ends soon afterwards.*

In this section we present a first attempt to modify the full non-dimensional model (1.12) to reproduce, at least phenomenologically, these peaks. As mentioned above, the origin of this bump is the closure of the aortic valve. The mathematical model, so far, does not contain a term to describe this effect. Consequently, we introduce a new term G into the non-dimensional equation for the arterial pressure (1.12a):

$$\dot{x}_a = -x_a + \epsilon_1 x_v + x_h + G. \tag{1.15}$$

To place the notch in the correct position, this forcing function has to

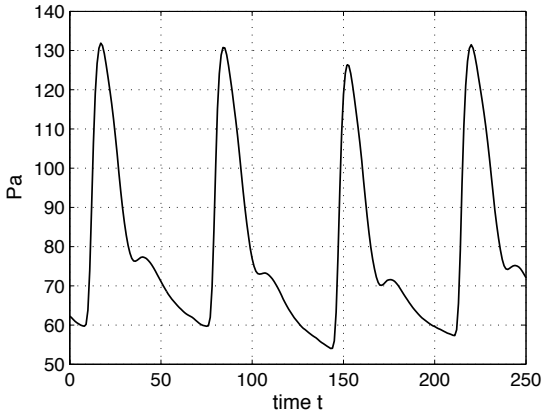


Figure 1.3: Arterial pressure measurement using an intra-arterial catheter

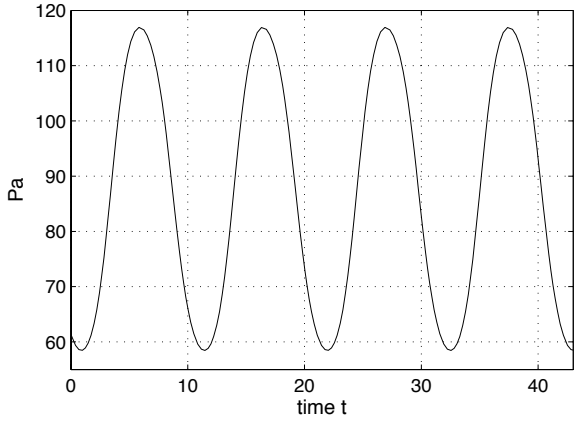


Figure 1.4: Arterial pressure P_a obtained from solving (1.1)

detect when $P_a(t)$ crosses the value $x_a = x^*$, where x^* is a specified constant indicating the non-dimensional ventricle pressure (non-dimensionalised with $\bar{P}_a$), but also when the crossing occurs in the decreasing direction.

Given that our model already uses a time delay, we found it sensible to discern if the function is in the increasing or decreasing phase at the time of the crossing using another time delay Δ, rather than a new time derivative. Furthermore, we note that the aortic valve does not close instantaneously but it takes a certain time, although small, which we call the *closing time*. Hence, we choose the forcing term G to be a function of $x_a(T)$ and $x_a(T-\Delta)$; the existence of this closing time will show up in the fact that G is nonzero over a small pressure range near x^*.

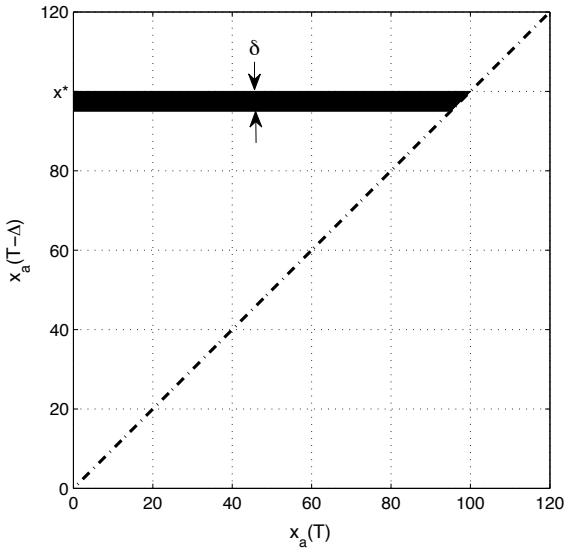


Figure 1.5: The black solid area represents the support of the forcing term $G(x_a(T), x_a(T-\Delta))$

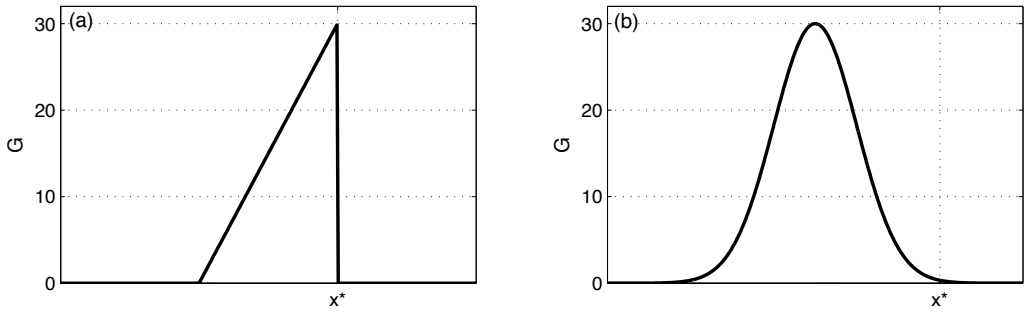


Figure 1.6: Two possible shapes for the forcing term $G(x_a(T), x_a(T-\Delta))$

Figure 1.5 shows the action of this forcing function G. We take G to be zero everywhere except at the strip shown in the figure. Note that this strip is placed when $x_a(T) \leq x_a(T-\Delta)$ and that it has a width $\delta > 0$ to account for the finite closing time of the valve. The health of the aortic valve is then quantified by the parameters Δ, δ and also by the form of the function G.

In the following calculations we use two different forcing terms G in the strip, as shown in Figure 1.6: a triangular ramp —Figure 1.6(a)— and a Gaussian —Figure 1.6(b). The numerical results obtained using these forcing

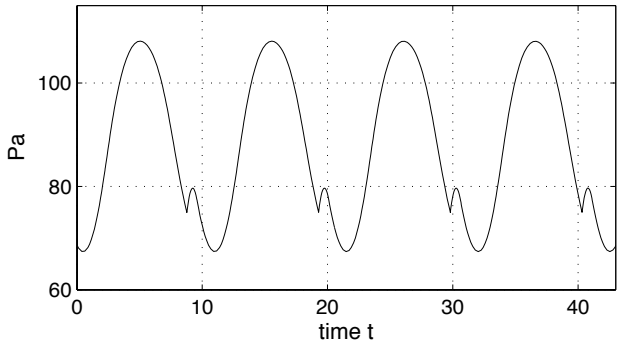


Figure 1.7: Dimensional arterial pressure P_a obtained from solving (1.1) with the ramp profile forcing

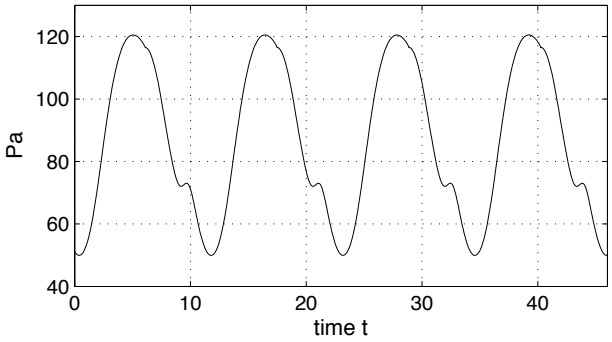


Figure 1.8: Dimensional arterial pressure P_a obtained from solving (1.1) with the Gaussian profile forcing

functions are shown in Figures 1.7 and 1.8. Figure 1.7 shows the arterial pressure when a triangular ramp profile is used, and Figure 1.8 shows the arterial pressure when a Gaussian profile is used. In both cases we see a reasonable approximation to the dicrotic notch. Following from this, the next goal is clearly to adjust the parameters carefully to ensure good matching between the artificial notch and that observed in practice. Again we did not follow this route due to the limited time of the GEMT.

1.5 Conclusions

During the GEMT we were able to understand and analyse the baroreflex model of Ottesen. Through non-dimensionalisation we noted that for time scales of the order of 100 s it is possible to uncouple the venous pressure from

the system and so to solve a simpler two ODE model. (It was later found that the same conclusion was reached in [3].) However, as time goes on, the effect of neglecting venous pressure accumulates until the arterial pressure moves out of phase with the solution obtained through solving the full system. For continuous monitoring one should then either include all terms in the equations or carry out a more detailed analytic solution that includes multiple time scales. The dicrotic notch was accounted for by including a source term in the arterial pressure equation.

Finally we note that Ottesen's model is reliant on a number of assumptions. The form of the pressure curve is determined to a large extent by the rather arbitrarily chosen function $f(t, \tau)$. The inclusion of a delay term τ may be interpreted as a neglect of the correct model for the sympathetic system. Consequently, we may conclude that an alternative approach may reproduce the blood pressure curve more accurately. This is now the focus of a current investigation and was the subject of the Workshop on Mathematical Modeling of Blood Flow and the Baroreflex System, held at the CRM in December 2010 (see Chapter 4 of this volume).

Bibliography

[1] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, Springer, New York, 1999.

[2] M. Canneson et al., *Relation between respiratory variations in pulse oximetry plethysmographic waveform amplitude and arterial pulse pressure in ventilated patients*, open access, `http://ccforum.com/content/9/5/R562`.

[3] A. C. Fowler, M. J. McGuinness, *A delay recruitment model of the cardiovascular control system*, J. Math. Biol. **51** (2005), 508–526.

[4] J. Keener, J. Sneyd, *Mathematical Physiology*, Springer, 1998.

[5] M. S. Olufsen, A. Nadim, *On deriving lumped models for blood flow and pressure in the systemic arteries*, Math. Biosci. Engng. **1** (2004), 61–80.

[6] J. T. Ottesen, *Modelling of the baroreflex-feedback mechanism with time delay*, J. Math. Biol. **36** (1997), 41–63.

Bandwidth Consumption and Invoicing Models

Problem presented by
Enric Folch (Cisco Systems)

Report prepared by
Gustavo Chavez (KAUST), Sara Costa (UdG), Michelle De Decker (CRM), Jonathan Low (CRM), Elena Rodríguez (UAB), Jesús Rosado (UAB)

Study group contributors
Aureli Alabert (UAB), Daniel Balagué (UAB), Gustavo Chavez (KAUST), Sara Costa (UdG), Michelle De Decker (CRM), Ruslan Krenzler (MathMods), Jonathan Low (CRM), Fernando Martínez (UPC), Xavier Muñoz (UPC), Elena Rodríguez (UAB), Jesús Rosado (UAB)

Problem statement

Internet service providers charge their customers according to a volume utilisation scheme or by a percentile billing scheme. Customers usually opt for volume billing, but as their business grows and start to solicit more traffic to their websites, they change to the percentile billing scheme. Currently, there is no logical link or explanation between the two schemes, especially when the differences in cost could possibly be large. Cisco Systems wish to find/define a relationship between the volume and percentile invoicing models. This may include linking both billing schemes to the cost of the bandwidth, taking into account the constraints of bandwidth supply or capacity and measuring the fairness/unfairness of the billing schemes, either from the point of view of a provider or a customer.

2.1 Introduction

Transit Internet Service Providers (ISPs) and other web hosting companies acquire the capability to host websites onto the Internet by buying bandwidth capacity and in turn selling this bandwidth service to customers that need their websites to be online. A diagram of this is shown in Fig. 2.1. This industrial project examines the billing aspect between the ISP (provider) and the customers.

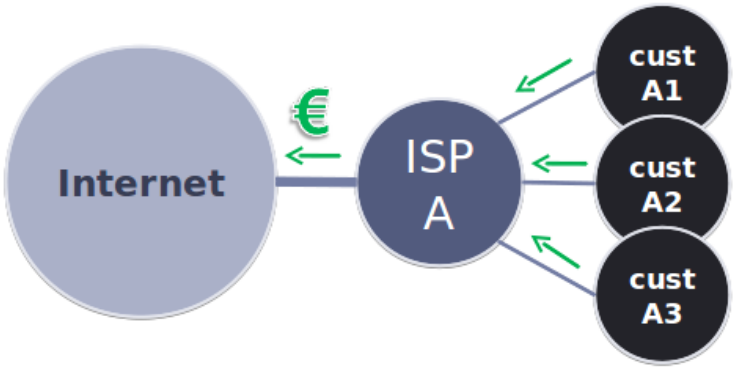


Figure 2.1: Diagram of an ISP provider buying Internet capacity and selling the bandwidth to customers

We will examine two current price models for commercial Internet billing and attempt to define sound mathematical relationships between them. This aims to help customers compare easily the schemes and decide on the suitability of each one for them. Other billing schemes exist such as those based on average billing utilisation or purchasing a *Committed Information Rate* value from an ISP but are not in the scope of this study. In addition, this study will focus on this simple model between a single tier ISP and its customers; we do not consider the more complicated scenario of a multi-tier network where the ISP can buy or resell its bandwidth capacity to other transit ISPs [1].

The two billing schemes under consideration between Internet Service Providers and customers are:

- Volume-based.
- Percentile-based.

Each has advantages and drawbacks, seen differently between ISPs and website customers. Both these methods are briefly described and the problems/issues are presented in the next three subsections.

2.1.1 Volume billing method

The volume billing method allows the customer to pay for Internet access based on the amount of data transmitted. Typically, the customer pays for a fixed amount of data transfer per month and this is paid before the start of service. This method is perhaps more deterministic than the other, as the customer:

- Knows how much data can be transferred in that period.
- Allows pre-payment for service.

However, when the customer's quota is exhausted, no more data transfer is possible and thus customers suffer a denial of service due to exceeding their bandwidth limit. This billing method is suitable for amateur or personally managed website, where they do not expect to attract a lot of traffic or are not providing a critical service where a denial of service would be catastrophic. In addition, because of the pre-payment method, the customer will typically use up less than the allowed amount of volume within the billing period if we assume that no denial of service takes place by not exceeding the data limit.

2.1.2 Percentile billing method

Also known as burstable billing [2], this is perhaps the most common billing method for professional and corporate ISPs and webhosts. The scheme is seen as a compromise between a customer paying volume and paying peak bandwidth utilisation. It works by sampling the traffic in a time window, typically five minutes and each of these interval samples determines a bandwidth rate for that particular period. Over a period of thirty days, these five minute samples are collected and sorted from highest to lowest. A percentage of the highest samples are then discarded and the customer is billed on the bandwidth rate sample at that percentile mark. A popular percentile figure in the market is 95; so, in an example of a billing period of thirty days, a total of 36 hours is discarded, which will contain the highest bandwidth rates (usually measured in bits per second). Fig. 2.2 shows an example of a bandwidth graph and marking out the 95 percentile mark. This scheme offers advantages over the volume billing method:

- Accommodates for occasional 'bursty' traffic without the extra cost for peak utilisation or having to pay more for a higher Committed Information Rate from the ISP provider.
- No quota limit, hence no limit of service.

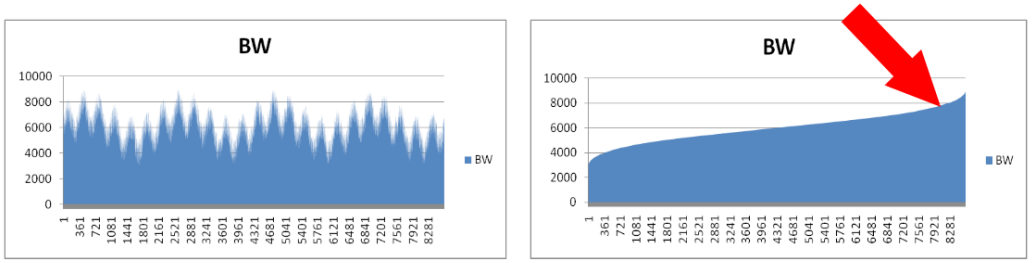


Figure 2.2: Example of a bandwidth utilisation graph (left) sorted from lowest to highest (right). For a 95 percentile billing scheme, the top 5% samples are discarded and the next highest value, marked by the arrow, becomes the billable utilisation for the entire billing period.

Unlike the volume method, the customer pays for the Internet service after the billing period, since the 95 percentile value can only be determined after the billing period.

We describe a possible scenario such that the customer pays more under this pricing scheme compared to the volume or an average throughput billing method: if a website proves popular over a 48 hour period, such as a weekend, it would experience high bandwidth data throughput for more than the permitted 36 hours. Consequently the 95 percentile value would be far higher than usual and so the customer pays more for that billing period, even though the website may experience lower rates of traffic than on average within that same period. Other events that could push the unusual peak rate over the 36 hour window include distributed denial-of-service (DDoS) attacks or excessive traffic due to backups.

2.1.3 Issues and fairness for providers and customers

Having described two billing methods for Internet service provision, we will now look at the fairness of these methods. This issue stems from the fact that providers and customers view these billing schemes differently from one another with regard to fairness.

The maximum throughput of a data channel conduit is called the *capacity* and this is what providers pay for, which in turn is served to customers to host their websites. In the simple model depicted in Fig. 2.1, the ISP (provider) buys a certain amount of capacity and the three customers host their websites through this ISP and are billed either by volume or percentile described earlier.

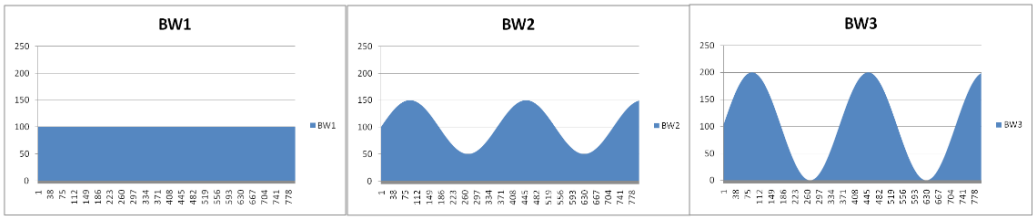


Figure 2.3: Three examples of bandwidth utilisation by three different customers. Under the volume billing scheme, all three would pay the same. But for the provider, customers BW2 and BW3 require higher capacity pipes in order to deliver full, uncapped bandwidth service to their visitors. This would incur higher costs than simply paying for 3 times the capacity of customer BW1.

Capacity of an ISP's data pipeline

Throughout this study, we work on the presumption that transit ISPs acquire their Internet hosting capability using a capacity based method. Referring to Fig. 2.1, we define this to be the ISP acquiring a fixed amount of bandwidth capacity from a network provider in exchange for money. For example, the ISP pays a network provider for a 30 Mbps capacity link, meaning that the ISP can transfer a maximum of 30 Mbps at any one time. This property is important because the ISP's quality of service can depend on the website traffic patterns of its three customers. If each customer's peak usage does not exceed 10 Mbps at any time, the ISP will have no problem providing service since the total usage of the three customers would not exceed 30 Mbps at any one time. However, if two customers send data at 20 Mbps each and the other at 30 Mbps at the same time, the ISP will have a problem as it cannot provide the data service at full speed for all three customers since the bandwidth capacity of the pipe is reached. To provide full service, the ISP needs to buy the data link with a 70 Mbps capacity. This leads to several issues for the provider to consider, especially the subject of fairness between the ISP and its customers.

Several issues for the provider are the following:

- Suppose that the provider wants to provide uninterrupted/uncapped bandwidth to all three customers. This means buying enough capacity that is wide enough to accommodate the sum of the customers' peak utilisation. But this is very expensive and is not cost-effective, since the data pipe becomes under-utilised most of the time.

- This leads to the question of predicting peak occurrences. Different business outlets have different spiky behaviour such as news sites tending to peak on weekday mornings or travel agencies getting most of their traffic on evenings and weekends. Should bandwidth be sold at different prices to customers that are the odd one out, e.g., sell differently to a travel agent when all your other customers are newspapers?

- Suppose that the provider bills all three customers with the volume method. Because Internet provision is by capacity of that data pipe, all three customers can possibly use the same amount in a billing period but have different peak utilisation values, as shown in Fig. 2.3 as an example. Customers B and C have higher utilisation peaks, so the provider has to buy a wider data pipe for them than that for customer A, which will cost more, but charging the same price for all three. Hence this is seen as an unfair scenario for the provider.

An issue common to both provider and customer is the non-deterministic nature of both pricing schemes and the fairness/unfairness for either party depending on the actual bandwidth utilisation outcome. For example, customers could virtually have all their bandwidth for little cost if most of their Internet traffic happens within a 36 hour period for the case of 95 percentile billing, though unlikely. Extreme scenarios of this nature are addressed in Section 2.2. On the other hand, the provider could get most of the capacity cost paid for by one customer whose utilisation graph shows a very high 95 percentile mark but statistically show under-utilisation of the pipe's capacity most of the time, an example shown in Fig. 2.4. For volume billing, customers have to predict how much data they are likely to transfer before suffering a denial of service by their ISP.

2.1.4 Aims

It has been said that the 95 percentile billing method has not been designed on any sound mathematical theory or optimisation techniques. This was used by a small number of large corporate providers and over time the billing method was adopted by many other companies and became a standard billing scheme [3]. In addition, scenarios are possible where customers discover that the difference in cost/payment between the schemes can be large with no sound explanation as to the reason why.

The objective of this report is to provide an insight into the relative cost between the two pricing schemes and to give further information such that either the provider or customer can make decisions on pricing and cost. The report is divided into the following sections. Section 2.2 expands on

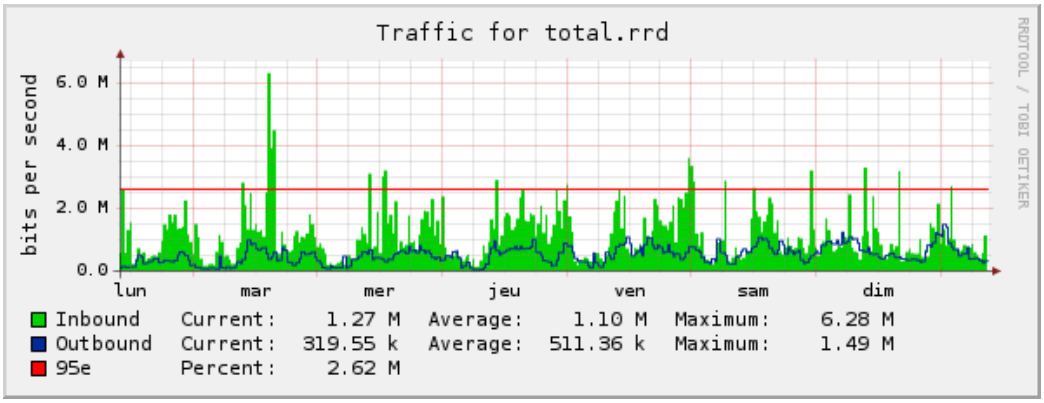


Figure 2.4: Example of bandwidth utilisation during a week. If this was the typical scenario for the whole billing period, this customer would be charged for 2.62 MBits/sec whereas an alternative billing scheme based on average use of 1.1 MBits/sec or by volume could work out cheaper. (Graphic taken from [2].)

the fairness issue that this project tries to address. Section 2.3 outlines three mathematical models that we pursue, and their results are presented in Section 2.4. Finally, the last section gives a summary of the study.

2.2 Improving the fairness of the billing

We concentrate on the case of a customer with high needs of bandwidth, such as big companies. Nowadays, the billing method used by the provider is the percentile method (see 2.1.2). As already commented in the introduction, a major problem with this method is the high variability of the volume billed due to the isolated peaks of traffic that may appear. As a consequence, we find two main problems: on the one hand, we have the impossibility of predicting a priori the expenses that a customer who moves from pre-payment to post-payment will have to pay for and on the other hand the possible unfairness, both for the customer and the provider, that may appear depending on the situation. The first problem reveals itself to be an intractable problem without adding extra assumptions, so here we concentrate on giving a solution to the second one and propose three models which try to reduce the unfairness in extreme cases. These models reduce the variability of the 95 percentile method, thus smoothing out the first problem as a by-product.

Before presenting the models, let us show examples of the two possible extreme situations we mentioned above, which will help us clarify this point.

- **Case 1:** The customer pays less volume than its actual consumption. In this case, the traffic generated by the customer for a whole month consists of very high peaks within a period of aggregated time corresponding to less than 5% of the total.

  An extreme (unreal) example of this situation would be the following. Suppose that the customer generates a traffic rate of 5 Mbps during 29 days but during the other day it is 100 Mbps. In this case, the 95 percentile mark would be 5 Mbps, so the customer would pay a total volume of 12960 Gb ($5 \times 30 \times 86400$), whilst the actual volume consumed is 21168 Gb.

- **Case 2:** The customer pays much more volume than its actual consumption. In this case, the traffic generated by the customer for a whole month consists of very high peaks of traffic for a period of time that exceeds just over 5% of the total.

  Again, let us show an extreme (unreal) example of this situation. Suppose that the customer generates a traffic rate of 5 Mbps during 28 days but during the other two days it is 100 Mbps. In this case, the 95 percentile mark would be 100 Mbps, so the customer would pay a total volume of 259320 Gb, whilst the actual volume consumed is only 29376 Gb.

Our goal is to propose new methods that decrease the unfairness of such cases. To do this, first of all we need to make it clear what do we mean by "fairness" of the billing method. One would think that the ideal billing method would be when the customer pays for the exact volume consumption (like the billing method used by phone companies). The problem with this method is that the hosting provider needs to buy enough bandwidth to supply the demand of all the customers, so the isolated peaks of traffic that may appear should be penalized. In general, looking at the real examples, we can see that the typical customer pays for about twice the real volume consumed, so we will consider this value as *normal*. For the three models that we describe here, first we need to introduce some notation related to the data sampling we consider here. Let us assume that we have a set of data as described in the introduction, that is, a set of T samples equally distributed in a time window of size l seconds, typically every 300 seconds (i.e., 5 minutes), so that the total billing time is $T \cdot l$ seconds. In the following we shall refer to this set as $S = (x_j, d_j)_{j=0}^{t}$, where $x_j = j \cdot l$ is the time associated to sample j and d_j is the mean bandwidth used during this time. For the time $j \cdot l$ we have the data d_j. Note that without loss of generality we can also assume that the sequence d_j is non-decreasing, i.e., that the samples are ordered like the way in Fig. 2.2.

2.3 Proposed models

The following solutions were proposed taking into account some constraints:

- Both the provider and the consumer should get maximum benefit in terms of service and money.
- The model should be easy to communicate.
- The model should be comparable with existing models in the market.

2.3.1 Convex combination model

As we have just explained, it is not always the best choice to use the percentile model, nor paying for the real consumption, so we propose a simple convex combination between both methods. Let us consider a set of samples in a billing period. Then we can compute the actual volume consumption, V, given in gigabits (Gb) and the percentile 95, P, given in megabits per second (Mbps), associated to this sample. We have to take into account that, in order to compare these two values, we need to scale them so that they are expressed in the same units. Therefore, let us denote by s the scaling factor associated to the billing period, which is the number of seconds in that period divided by 1000. Then the convex combination, CCM, is given by

$$CCM = (1 - \lambda)\frac{V}{s} + \lambda P. \tag{2.1}$$

It is important to remark the differences between this model and the weighted percentile model (see 2.3.3): In this case we make the convex combination between the *whole* volume data and the 95 percentile, while in the other only the volume of data less than the 95 percentile is taken into account. Moreover, as we show in the sequel, we give an explicit formula to compute the parameter λ implied in Eq. (2.1). Another difference is that here we do not consider the unit prices of the models, because we are only interested in the comparison of the amount of data that is paid for (as already discussed before, this could be the real amount, more, or even, in some cases, less).

Now, the problem consists in finding the best λ in order to minimize the unfairness for both the provider and the customer. Looking at Fig. 2.5 below, one can see that the unfairness of the 95 percentile method comes from the relation between the areas A and B, where A corresponds to the unused volume that the customer pays for and B is the volume he obtains for free. Mathematically, let us denote by $N_{95} = t \cdot l$ the time corresponding to the 95 percentile sample consisting of t time samples of length l; by $V_{N_{95}}$ the volume

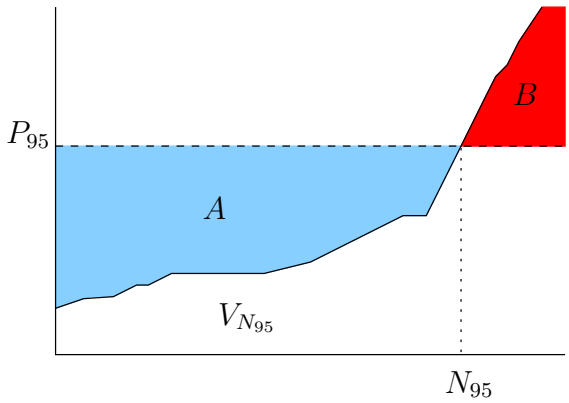


Figure 2.5: Convex combination model

up to that time; by T the total time of the billing period; by V_T the total volume for the whole billing period, and by P_{95} the bandwidth rate value at the 95 percentile mark. Then we can write

$$A = t \cdot l \cdot P_{95} - V_{N_{95}} \quad \text{and} \quad B = V_T - V_{N_{95}} - P_{95} \cdot (T - t) \cdot l.$$

These two quantities allow us to see if the customer is paying more or less than the real volume of data consumed by looking at A/B. Taking $\lambda = A/B$ could be a good choice here. We have three possibilities: either $A/B > 1$ when the customer is paying more than he consumes, or $A/B < 1$, which represents the opposite case, or $A/B = 1$, that would correspond to rare situations in which the customer is billed exactly for the real volume consumed. It is important to remark that in the first case if we take $\lambda = A/B$ we could obtain negative CCM values; to avoid this, we require that $\lambda \in [0, 1]$ by taking $\lambda = (A/B)^{-1} = B/A$ when $A > B$.

Note that, in general, the CCM billing model tends to benefit the customer against the provider. However, it will always charge for more than the actual volume consumption, except in those extreme cases where the percentile billing model itself does not. This is a coherent thing to do taking into account everything we have already discussed about how a fair billing method should be.

2.3.2 Weighted mean model

Another approach to reduce the unfairness of the percentile billing method would be taking into account not the number of peaks but the distribution of the total consumption. In the percentile model it is not considered whether

the volume corresponding to the discarded data is big or not, or if the traffic prior to this point (when we consider sorted data) is moderate or even non-existent. One may argue that a high peak which is not representative of the behaviour can be neglected whilst the same peak, when it corresponds to the main consumption, cannot. In the next paragraph we shall explain in detail one method to achieve this, but the idea is to distribute the area under an appropriate graph in several blocks, for which we shall choose a representative height, and then compute the mean of these values. Doing this, in the first case, the weight corresponding to high peaks will be low, whence their influence in the average is small. On the contrary, in the second case, the peak will deserve a high weight, since there will be many blocks with a high representative.

In order to do this, let $n \geq 1$ be an integer, and consider the total volume V given in Mb for convenience of notation. Let $\{x_i\}_{i=0}^{n}$ be such that the area of the step function between x_{i-1} and x_i for $1 \leq i \leq n$ is V/n and $x_0 = 0$. Note that $x_n = (t_T + 1) \cdot l = T$ where t_T is the total number of samples in the billing period. Then the weighted mean WMM is given by

$$WMM = \frac{V}{n} \sum_{i=1}^{n} \frac{1}{x_i - x_{i-1}}.$$

In Fig. 2.6 we apply it to a simple function as an example of how to use it. In that case, we see that

$$WMM = \frac{V}{2T} \left(1 + \frac{1}{\sqrt{2} - 1} + \frac{1}{\sqrt{3} - \sqrt{2}} + \frac{1}{2 - \sqrt{3}} \right).$$

It is clear that the value of WMM depends on the choice of the parameter n. Observe for instance that, when $n = 1$, the volume associated to WMM coincides with the real volume that the customer has consumed. However, this dependence is neither linear nor monotonic and we should expect that, for some choices of n, WMM gives results that are too high in relation to V. Because of that, an appropriate value for this parameter has to be chosen in order to get a result that we can consider acceptable in as many different situations as possible. After the results in our set of examples, we propose taking $n = 4$, but this value can be modified by the provider depending on sales policy.

2.3.3 Weighted percentile model

The third approach that we suggest incorporates some aspects of both the volume and percentile model; we refer to this new model as the weighted

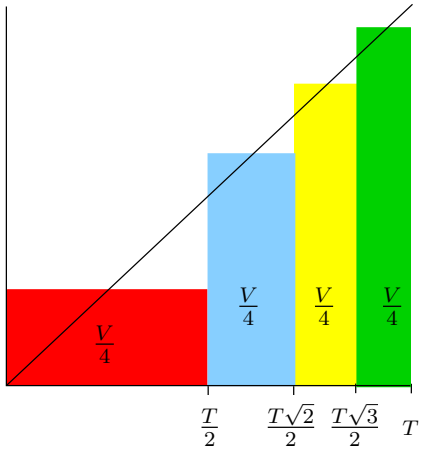


Figure 2.6: Weighted mean model for $n = 4$

percentile model. The provider is interested in providing the service for larger companies that have localized bandwidth peaks, but also has to consider an extra charge for the peaks they use. This model aims to find subcategories of profiles that describe bandwidth usage amongst the clients. This way it is easy to identify which companies have concentrated bandwidth usage within peaks and the provider can bill them in such a way that is fair for both customer and provider.

In summary, the weighted percentile model will use the 95 percentile with an additional penalty term. The penalty term takes into account historic data of the customer. This way the bill is based on the bulk bandwidth usage, whilst still having the attraction of not being penalized for the highest peaks over the 95 percentile. The following equations give the costs by the weighted percentile billing (WB) and the cost by volume billing (VB):

$$WB = \left(\lambda \sum_{i=1}^{N_{95}} V_i + (1 - \lambda) P_{95} N \right) \times p_w, \tag{2.2}$$

$$VB = \sum_{i=1}^{N} V_i \times p_v, \tag{2.3}$$

where

WB = Weighted percentile cost,
VB = Volume cost,
p_w = Unit price of weighted percentile model,
p_v = Unit price of volume model,

$$
\begin{aligned}
V_i &= \text{Volume in the interval } i,\\
N_{95} &= \text{Interval in which the 95 percentile is observed},\\
P_{95} &= \text{Value at the 95 percentile mark},\\
N &= \text{Number of billing intervals.}
\end{aligned}
$$

With $\lambda = 0$ the weighted percentile model is equivalent to that of the current percentile model employed by providers today. By increasing the value of λ, one decreases the penalty of the 95 percentile. As $\lambda \to 1$ the weighted percentile model takes on the form of the volume model. The idea of the weighted percentile model is to find a value for λ that is fair for all. The value of λ is determined by equating volume cost and weighted percentile cost. Different datasets will result in different λ values, making up the different profile categories sought after.

2.4 Results

2.4.1 Comparisons between the convex combination and weighted mean models against the percentile billing method

In this section we present some examples in order to compare the proposed models with the currently most used model, the 95 percentile (P). Note that P, CCM and WMM can be compared easily since all of them are given in Mbps. Since we also want to measure the unfairness of the method, we will also take into account the real volume V given in Gb in relation to the volume billed for by each of these three methods, V_P, V_{CCM} and V_{WMM}.

In the following examples we will consider sets of 720 equidistant samples taken along one month of 30 days, that is, each sample corresponds to one hour, and fix the parameter $n = 4$ for the WMM. Example 2 is based on real data. Instead of showing all the data, we will introduce the examples with a graph where we can see its distribution once ordered. Finally, the line marks the 95 percentile point.

Example 1. In this example, we can see that the volume consumed is way lower than the volume charged for with the percentile method. With both the CCM and the WMM methods, we would have charged the customer for a volume that is much closer to the real volume consumed. It can be seen that the CCM charges for a volume that is really close to the real one, whilst the WMM charges for a volume that is closer to the one given by the percentile billing method.

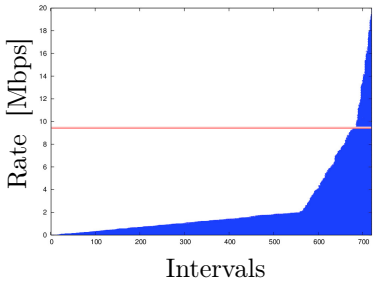


Figure 2.7: Data for Example 1

Example 1 Parameters	
$V = 6485.29$ Gb	$CCM = 2.75$ Mbps
$P = 9.45$ Mbps	$V_{CCM} = 7123.74$ Gb
$V_P = 24499.58$ Gb	$WMM = 7.09$ Mbps
$\lambda = 0.035$	$V_{WMM} = 18382.75$ Gb

Table 2.1: Parameter values for Example 1

Example 2. Here we present three sets of real data which, once sorted, display different typical profiles: a gentle slope in (2.a), a smoothered jump in (2.b) and a step slope in (2.c). We can see in all of them that the volume consumed is around half of the volume charged for with the percentile method. This is what we could call a normal situation.

In all three cases, the CCM and the WMM methods charge the customer for a volume that is between the real one and the volume charged with the percentile method, albeit closer to the real volume, especially the CCM.

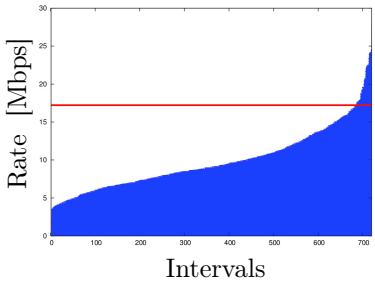


Figure 2.8: Data for Example 2.a

Example 2.a Parameters	
$V = 25714.04$ Gb	$CCM = 10.09$ Mbps
$P = 17.2$ Mbps	$V_{CCM} = 26158.71$ Gb
$V_P = 44582.4$ Gb	$WMM = 11.32$ Mbps
$\lambda = 0.02$	$V_{WMM} = 29351.35$ Gb

Table 2.2: Parameter values for Example 2.a

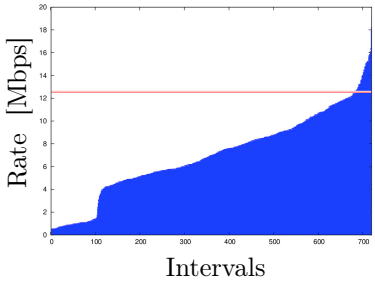


Figure 2.9: Data for Example 2.b

Example 2.b Parameters	
$V = 18245.52$ Gb	$CCM = 7.14$ Mbps
$P = 12.6$ Mbps	$V_{CCM} = 18499.80$ Gb
$V_P = 32659.2$ Gb	$WMM = 8.55$ Mbps
$\lambda = 0.02$	$V_{WMM} = 22153.31$ Gb

Table 2.3: Parameter values for Example 2.b

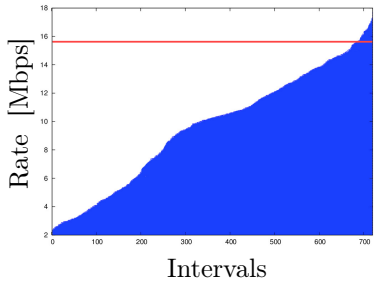


Figure 2.10: Data for Example 2.c

Example 2.c Parameters	
$V = 24936.80$ Gb	$CCM = 9.66$ Mbps
$P = 15.6$ Mbps	$V_{CCM} = 25039.80$ Gb
$V_P = 40435.2$ Gb	$WMM = 11.07$ Mbps
$\lambda = 0.01$	$V_{WMM} = 28693.92$ Gb

Table 2.4: Parameter values for Example 2.c

Example 3. In this example, we can see that the volume consumed is even more than the volume charged for with the percentile method. With both the CCM and the WMM methods, we would charge more than with the percentile method. However, with the CCM method we would still charge the customer for less than its actual consumption, while with the WMM method the customer would pay more than the consumed volume, which would be desirable.

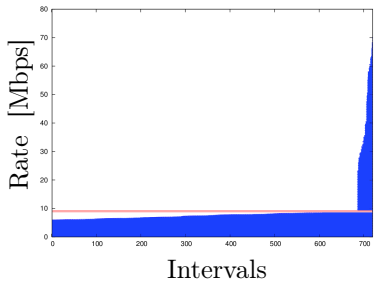


Figure 2.11: Data for Example 3

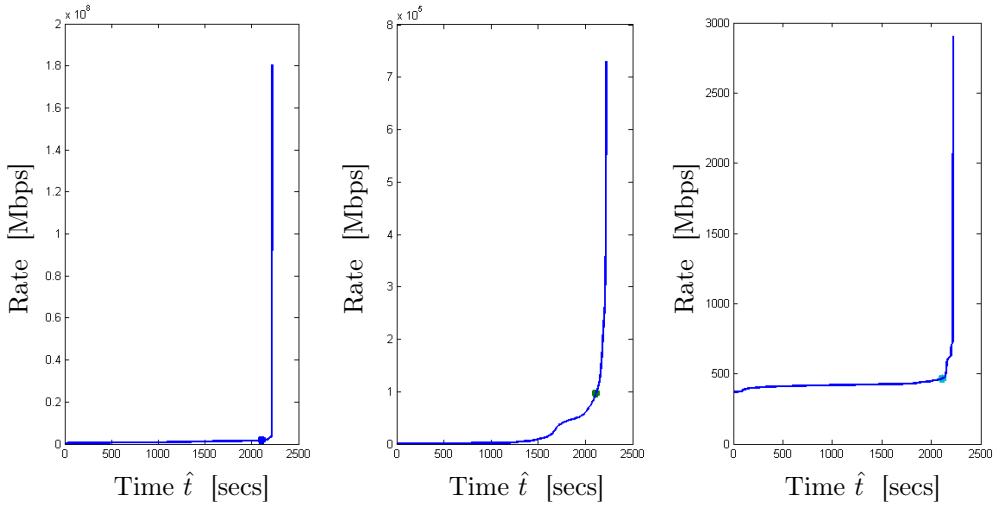
Example 3 Parameters	
$V = 24044.58$ Gb	$CCM = 9.04$ Mbps
$P = 8.99$ Mbps	$V_{CCM} = 23434.98$ Gb
$V_P = 23312.45$ Gb	$WMM = 13.94$ Mbps
$\lambda = 0.83$	$V_{WMM} = 36126.41$ Gb

Table 2.5: Parameter values for Example 3

2.4.2 Results for the weighted percentile model

We have applied this analysis to three sets of data; see Figs. 2.12(a)–2.12(c). In the finding of λ in this paper, we assumed $p_w = p_v = 1$ for ease of computation. In reality, p_w will be higher than p_v as the customer will be paying for the advantage of not being penalized for the high peaks of bandwidth usage. The value of λ specified in the title is that found by equating volume cost and weighted percentile cost. WB is the weighted percentile cost as calculated with λ, while WB_1 is the 95 percentile billing cost when $\lambda = 0$.

In all three cases the 95 percentile billing cost is considerably higher than that of the fair weighted percentile price, suggesting that the current billing scheme is overcharging customers.



(a) Data set 1: $\lambda = 0.77$, $WB = 2.4 \times 10^9$, $WB_1 = 3.93 \times 10^9$.

(b) Data set 2: $\lambda = 0.89$, $WB = 4.92 \times 10^7$, $WB_1 = 2.13 \times 10^8$.

(c) Data set 3: $\lambda = 0.57$, $WB = 9.5 \times 10^5$, $WB_1 = 1.04 \times 10^6$.

Figure 2.12: Three sets of data for the weighted percentile model

Each data set used here is distinctly different and each returns a different value of λ. This suggests that we should select few profiles of bandwidth usage, each having a different λ value. Upon looking at a customer's historic bandwidth usage, the customer can be assigned to a λ value that best suits their profile. The customer will be billed using the model specified in Eq. (2.2) calculated with their assigned λ value.

2.4.3 Fair percentile

Perhaps the 95 percentile is not the optimal percentile for charging the largest companies, so the objective of this analysis is to determine which percentile should be the optimal.

To determine a new percentile, we equate the amount of bandwidth that the client is paying for but not using (A_1) with the amount of bandwidth that the provider is not charging for but has already given up (A_2). This is shown in Fig. 2.13.

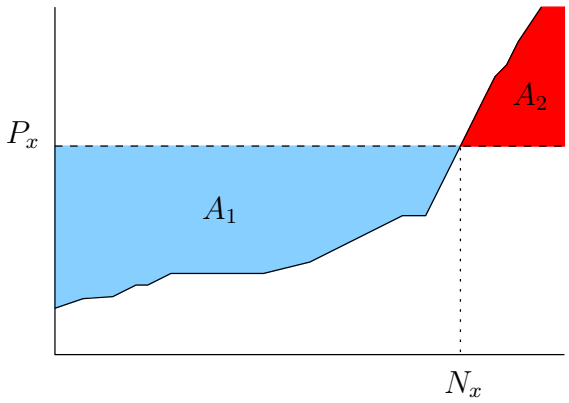


Figure 2.13: Fair percentile —the amount of bandwidth that the client and the provider lose is the same

The equations used are as follows:

$$A_1 = xP_x - \sum_{i=1}^{P_x} V_i, \tag{2.4}$$

$$A_2 = \sum_{i=P_x}^{N_x} V_i - (1-x)P_x, \tag{2.5}$$

$$A_1 = A_2,$$

where

A_1 = Amount of bandwidth that the client is not using,

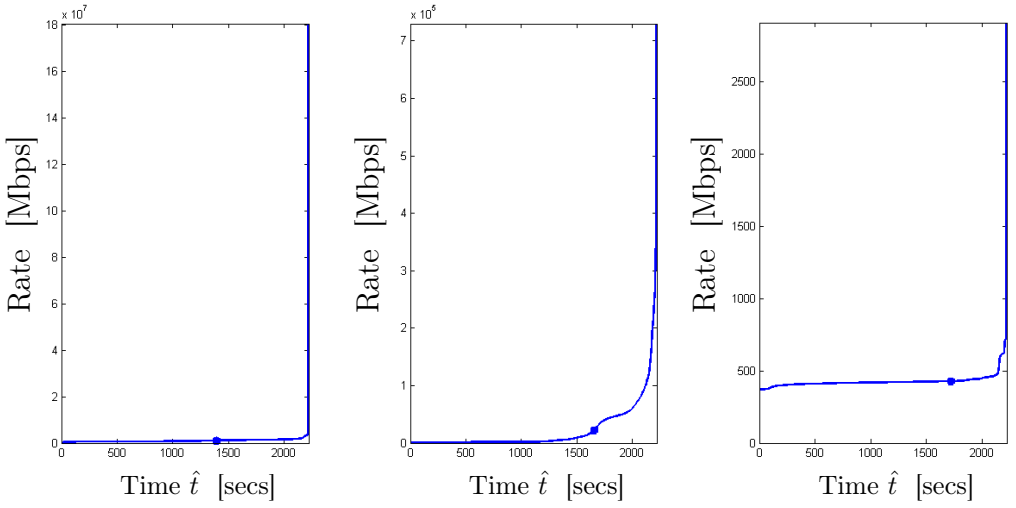
A_2 = Amount of bandwidth that the provider has to guarantee,

N_x = Fair percentile,

P_x = Bandwidth rate at percentile N_x.

The idea of this analysis is to find the percentile mark that guarantees areas A_1 and A_2 on the graph to be of the same size. This is equivalent to saying that both the client and the provider are exposed to the same risk. The value of P_x can be found by equating A_1 and A_2.

The results from three different sets of data are shown in Figs. 2.14(a)–2.14(c). The value of P_x is reported in the figure titles. As it happens, the most favoured value for P_x is roughly 70%. For different profiles, the fair percentile values differ. These can then be adjusted by the ISP according to commercial policy, perhaps depending also on the particular client or client profile.



(a) Data set 1: $P_x = 0.629$ (b) Data set 2: $P_x = 0.747$ (c) Data set 3: $P_x = 0.775$

Figure 2.14: Three sets of data for the fair percentile analysis

2.5 Summary

- This project studies attempts to address the issue of linking two billing methods commonly used by ISPs to charge their customers for Internet access. These are:

  1. Volume method.
  2. Percentile method.

- The two billing schemes raise the issue of fair billing between ISPs and their customers. This mainly results from how these ISPs acquire their ability to provide Internet bandwidth for their web hosting customers. Namely, ISPs acquire data pipes of a certain capacity from network providers which have a maximum rate transfer at any one time.

- Three mathematical models have been proposed in this study. All models assume the scenario of a simple network where an ISP buys capacity from any network provider and sells that bandwidth capacity to its customers in whatever way.

- The convex combination model links the entire amount of volume that a customer uses and the 95 percentile model through a parameter λ. The value of this parameter is calculated whilst taking into account the

amount of data the customer does not pay for against the amount of data obtained for free in a whole billing cycle.

- The weighted mean model attempts to take into account both the number and distribution of a customer's peak traffic spikes within a whole billing cycle. This is done by ordering the data into several blocks and computing the mean value of each of them.

- Comparisons were made between the convex combination and weighted mean models for five sets of data. Both models have shown to bill customers in between real volume billing and the 95 percentile billing and this has often been the case, making them consistent and predictable billing schemes for both providers and customers, with less chance of extreme variations.

- The weighted percentile model is similar to the convex combination model except that the data consumed before the 95 percentile mark in a billing cycle is taken into account, rather than the entire amount of volume consumed. This model includes a parameter, λ, regarded as a penalty term. Tested against three sets of data, this model shows that the classical 95 percentile billing scheme overcharges customers. With λ chosen based on the customer's usage history, this weighted percentile model can tailor the fairness of billing.

- Whilst analysing a (fair) percentile value that exactly equates the data volumes used in the volume and percentile billing schemes, it was often shown that a percentile value lower than 95 was required to equate the two schemes. This suggests that a percentile value lower than 95 would often make things fairer.

Acknowledgements

The authors are grateful to Enric Folch for providing the figures in the introduction section of this report, and to the study group contributors for their participation in this industrial project.

Bibliography

[1] M. Winther, *Tier 1 ISPs: What They Are and Why They Are Important*, IDC White Paper, 2006. Available from `http://www.ntt.net/english/library/pdf/IDCTier1-Whitepaper.pdf`.
Last accessed 25-10-2010.

[2] `http://en.wikipedia.org/wiki/Burstable`.
Last accessed 03-09-2010.

[3] X. Dimitropoulos, P. Hurley, A. Kind, M. Stoecklin, *On the* 95*-Percentile Billing Method*, Lecture Notes in Computer Science, vol. 5448, 2009, 207–216.

Flood Prevention in the Ebro Basin

Problem presented by

Joseba Quevedo (Sistemes Avançats de Control, UPC)

Report prepared by

Abel Gargallo (UPC), Manuel Quezada de Luna (OCCAM), Jordi Saludes (UPC), Jeff Springer (OCCAM)

Study group contributors

Abel Gargallo (UPC), Manuel Quezada de Luna (OCCAM), Jordi Saludes (UPC), Adrià Simon (UPC), Jeff Springer (OCCAM), Yi Ming (OCCAM)

Problem statement

Under heavy rain, some rivers used to have a discharge large enough to flood into nearby urban areas. A way to cope with this problem is to designate some fields adjacent to the river as *floodable*: at a cost, it is allowed to divert part of the flow into these fields by way of large gates which exist alongside the river. The area of the gate opening can be modified remotely. One wants to know the best control strategy for opening the gates, given the hydrologic profile of the flood which is gathered upstream some hours before it reaches the control point.

3.1 Motivation

Flooding resulting from excessive precipitation and surface runoff is a principal cause of significant damage, loss of property, and human suffering throughout the world. During the GEMT 2010 study group at the Centre de Recerca Matemàtica in Bellaterra (Barcelona), our team was given the task of determining a mathematical model to optimize gate operation along the Ebro River in Spain. Finding an optimal way to manage these

gates is a prominent goal of the CTP projects PREGO (2008ITT00007) and GECOZI (2010CTP00043).

We designed a simple model in which we assumed that the flood wave would maintain its original shape as it propagated downstream. This simplified model of the situation allowed us to look more closely at the mathematics involved in river flooding and gain some insight on a basic strategy for regulating flood gates. In addition to deriving this model for gate regulation, we also investigated the costs associated to the situation when flooding cannot be avoided. In this situation it is important to minimize the costs in terms of human suffering and property damage. We investigate in this paper two methods for approaching the situation where flooding cannot be avoided: strategy A —opening the gates when the wave is about to overflow the gates; and strategy B —opening the gates when the typical maximal level is achieved. This paper discusses the analysis of both strategies in detail. As a conclusion of our analysis, we state some future problems that might be taken up by our group or by other researchers with an interest in this problem.

3.1.1 States

The initial data is the forecasted avenue hydrogram (the graph of the flow as a function of time) computed from observations at a point upstream of the control point. Two parameters are relevant: the *maximal flow* and the *total volume* of the avenue.

We assume that the water height at the control point is an increasing function of the flow q at that point.

Let us consider three states related to the flow q:

- The *steady state* corresponds to $q < q_{min}$, where q_{min} is the flow that brings the water at the control point high enough for the gates to open.
- A *typical avenue* state corresponds to $q_{min} < q < q_{mao}$. The flow q_{mao} is the maximum flow for a recurrent avenue, one that happens every two or three years.
- The *high avenue* state happens when $q_{mao} < q < q_{max}$, where q_{max} is the flow that makes the water level spill over the floodable areas.
- Finally, when $q > q_{max}$.

For this, the following control strategies are defined:

- If the maximum forecast flow q_i is less than q_{mao}, the gates are not opened.

- When $q_i > q_{mao}$ but it is possible to level out the wave to a maximum q_f, and this value is less than q_{max}, then the goal is to minimize the peak discharge after leveling. To compute this, one cuts from the hydrogram an area equivalent to the capacity of the floodable areas, to get a maxim plateau of q_f flow.

3.2 First simple model

First we will discuss a simple control strategy:

> Given a set point q_f, open the gates completely when the discharge reaches this threshold.

As a simple first approximation, we ignored the geometry of the river bed and assumed that the flood wave would maintain its shape as it propagated down the river. We also assumed that the height of the river at a given point depends only on the discharge at this point by an increasing function. In this way we could use height and discharge interchangeably.

Given that opening a gate to a floodplain can reduce the flood by a given volume W, we look for the set point q_f to open and close the gate in order to reduce the flood by this area optimally.

Mathematically speaking, we define the discharge of a flood wave at the gate point as $q(t)$ and choose a reference interval $[t_0, t_1]$ containing the flood episode. For simplicity, we assume that q is unimodal and has a forecast maximum discharge q_i expected at time $t_m \in [t_0, t_1]$. Define, for $0 < y \leq q_i$,

$$V(y) = \int_{t_0}^{t_1} [q(t) - y]_+ \, dt \leq W, \tag{3.1}$$

where $[x]_+ = \max\{x, 0\}$.

This expression could be written on the time interval defined by the condition $q(a) = q(b) = y$ with $t_0 \leq a \leq t_m \leq t_1 \leq b$ using an iterated integral as

$$V(y) = \int_a^b \left(\int_y^{q(t)} dq \right) dt.$$

By inverting the order of the integration, we get

$$V(y) = \int_y^{q_i} \left(\int_{t_-(q)}^{t_+(q)} dt \right) dq = \int_y^{q_i} (t_+(q) - t_-(q)) \, dq, \tag{3.2}$$

where $t_-(q)$ (respectively $t_+(q)$) is the unique $t < t_m$ (respectively $t > t_m$) with $q(t) = q$. Since $V'(y) = t_-(y) - t_+(y) \leq 0$, $V(y)$ is decreasing with $V(q_i) = 0$ and it is easy to find the unique point q_f such that $V(q_f) = W$ by numerical integration followed by bisection or inverse interpolation.

3.2.1 A numerical experiment

We tried this idea with a mock flood wave made by adding a Rayleigh function to a constant flow (we chose such a function because real flood waves are typically asymmetrical —see [2]):

$$q(t) = 1 + 40\,\frac{[t]_+}{\sigma^2}\,e^{-t^2/(2\sigma^2)}, \tag{3.3}$$

with $\sigma = 10$ and a goal capacity of $W = 22$. The discharge is sampled between $t = -10$ and $t = 100$ for 1000 points (assuming one hour as time unit, which means roughly a sample every 6 minutes). The procedure was as follows:

1. The maximum discharge q_i is located at point t_m by inspecting the first differences of the sampled data (M data points). This partitions the time samples into two sets: the domains of t_- and t_+.

2. The difference $t_+(q) - t_-(q)$ is approximated by inverse interpolation of $q(t)$ data, for a uniform sampling of N values between the steady flow ($q = 1$) and the maximum flow ($q = q_i$).

3. By cumulative addition we approximate $y \mapsto \int_y^{q_i} (t_+(q) - t_-(q))\,dq$.

4. Again by inverse interpolation of the above table, we compute $q = q_f$ for $y = W$.

For $M = 1000$ and $N = 100$ and $W = 22$, we got $q_f \approx 1.70$, $t_-(q_f) = 1.782$ and $t_+(q_f) = 22.613$. To check the accuracy of this simple procedure, we compare the integral (which could be computed exactly in this case as $F(t) = t - k(1 - e^{-t^2/(2\sigma^2)})$) for $k = 40$ and $\sigma = 10$, giving

$$\int_{t_-(q_f)}^{t_+(q_f)} q(t)\,dt = F(t_+(q_f)) - F(t_-(q_f)) = 57.09, \tag{3.4}$$

while $W + q_f(t_+(q_f) - t_-(q_f)) = 57.44$.

3.3 Heavy flooding

When the flooding is particularly heavy, the capacity of the floodplains may be insufficient. In this case, even when the peak of the flood is reduced by W, it is still higher than q_{max} the levee height, causing floods. In this situation we have two options for controlling the gates. The first strategy is that the

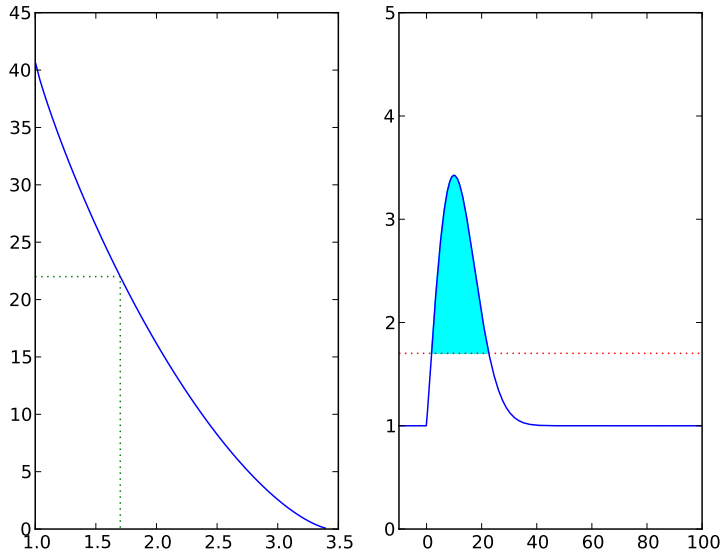


Figure 3.1: *Left:* $V(q)$ for the test function with the cut volume $W = 22.0$ corresponding to $q = q_f$. *Right:* The cut volume is the part of the graph of the test function above $q = q_f$.

gates are opened as soon as the water level reaches q_{max} and the reservoir is filled to capacity, hence reducing the front of the flood wave. This then gives us more time to deal with the flood, e.g., preparing to open more gates downstream (see Figure 3.2).

In the case of flows with multiple flood waves it is always ideal to open the gates at q_{max} instead of q_f, as it allows some capacity of the floodplain to be reserved for later flood waves. Thus when we have multiple flood waves we will never open the gates before $q = q_{max}$. See Figure 3.3 for a situation involving multiple waves and Figure 3.4 for a case where we can do even better.

3.4 Cost modelling of the main strategies in a flooding

In this section we assume that a flood that cannot be totally prevented is about to happen. Recall that a flooding corresponds to the scenario $q_{mao} < q_f$ and then any possible gate opening strategy can avoid that the water level of the Ebro surpasses at the analysed point the constructed walls and that

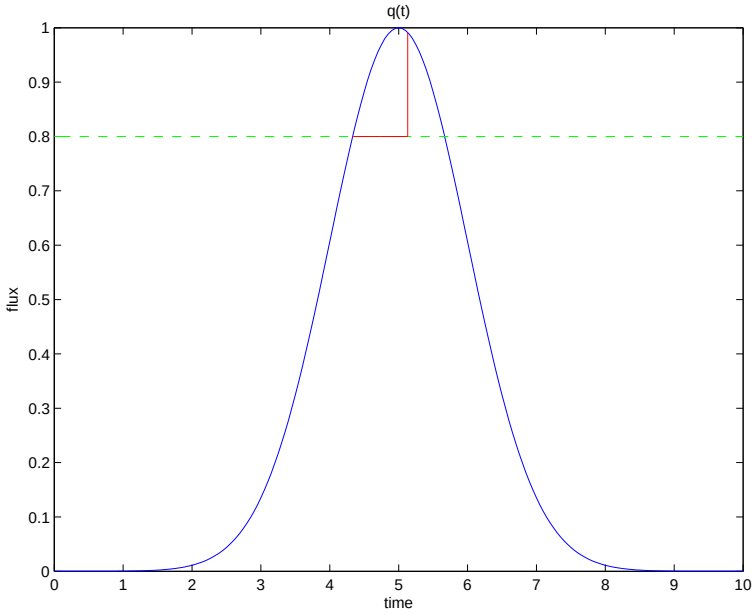


Figure 3.2: Flux versus time for a peak where flooding is inevitable

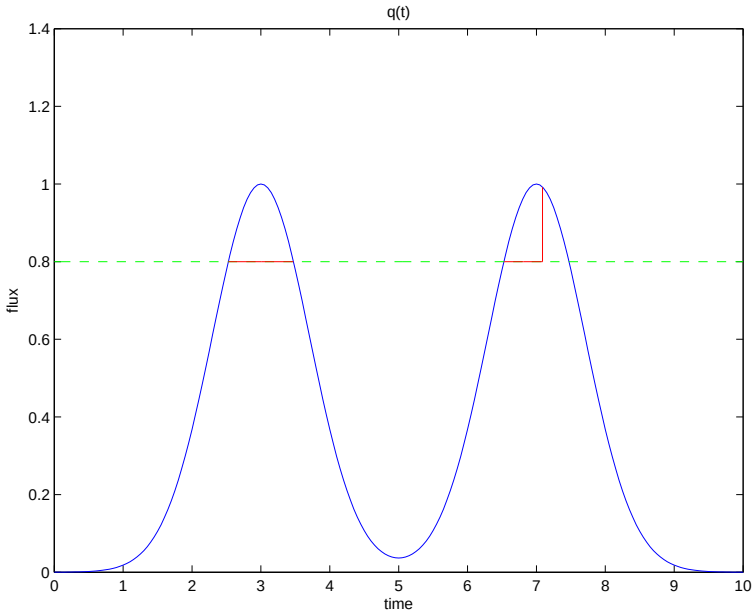


Figure 3.3: Opening the gate at q_{max} allows maximum remaining capacity for the second flood wave

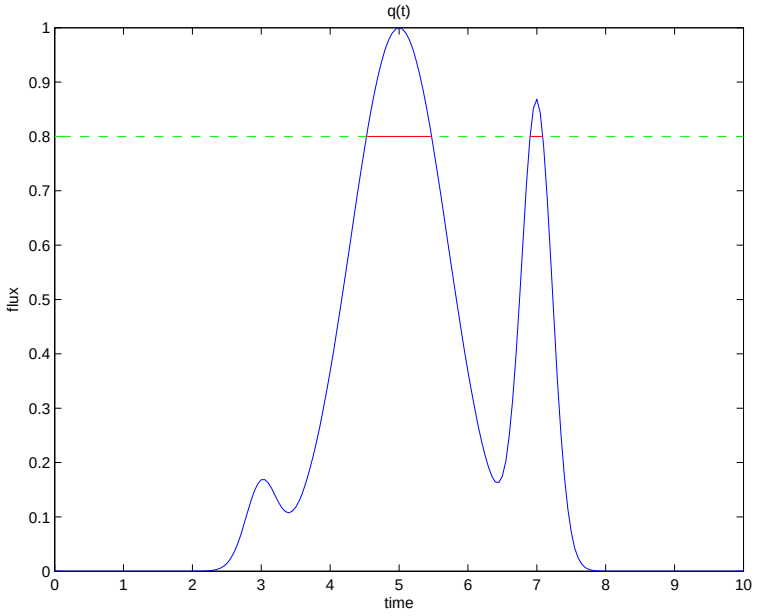


Figure 3.4: Opening the gate at q_{max} allows us to prevent flooding entirely in this case

floods the surrounding lands. When this happens, the controller of the gate opening must choose when to open the gate in order to reduce as much as possible the damage caused by the flooding. In this case, note that two main strategies can be raised: the gate can be either opened when the flow q achieves q_{mao} or it may be opened when q_f is achieved.

The objective of this section is to extract a numerical criterion, based on the observations that have been carried out in the previous sections, that allows us to choose the best strategy in each possible situation.

As we are going to see, each strategy has advantages and drawbacks:

- On the one hand, if the gate is opened at q_{mao}, the flood is going to be delayed for as long as possible. However, once the flooding capacity is surpassed, we are not going to have any control of the amount of water that the river carries from this point on. Hence, when the flooding arrives to the most sensitive areas the greatest intensity of the flood will not be mitigated. Figure 3.5 shows schematically the inherent idea that defines strategy A. The water coloured in light blue is the amount of water that is able to be drained through the gates. The water coloured in red corresponds to the amount of water that is going to flood the area of interest where we analyse the scenario. Hence, in the presented strategy, once the water overpasses the acceptable level

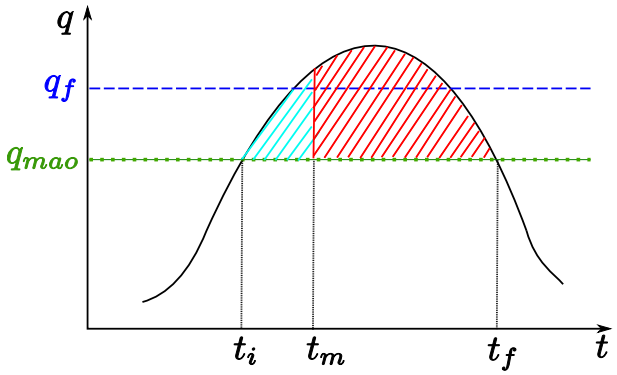


Figure 3.5: Flow diagram for strategy A

of the river, the door is opened. Then, while the flooding area is able to drain water, the flood is prevented. However, once we are not able to drain more water through the gates, all the remaining flooding water will continue on its way, flooding freely the area of interest.

- On the other hand, the controller of the gate can wait to open it until q achieves q_f. Then, the flooding of the sensitive areas starts earlier, whereas the most dangerous peak in the water amount that the flooding carries on is going to be cut out. This way, the main effects of the flooding will be reduced, because the intensity of the flow is going to be lower. However, we are going to allow less time for the affected population to leave the affected areas. Figure 3.6 shows a water profile of the river channel where strategy B is applied. The colouring scheme is the same as the one presented on Figure 3.5. Hence, Figure 3.6 clearly presents the main drawback and the main advantage of this strategy in front of strategy A. First, note that the red profile starts earlier

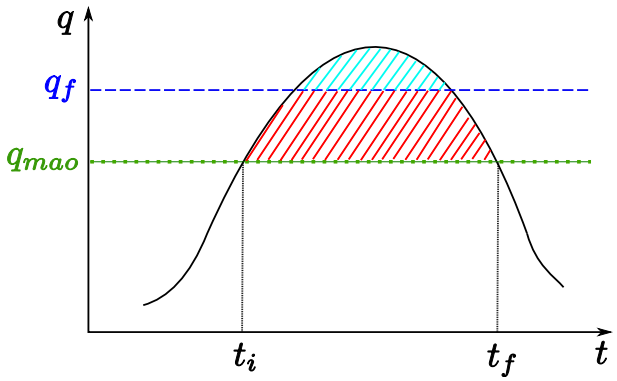


Figure 3.6: Flow diagram for strategy B

than in the previous strategy, and hence the flooding is not delayed as it happened before. However, the red profile does not reach the high values that strategy A achieved. Hence, strategy B avoids the flood in the peak where it has more intensity.

3.4.1 Damage cost function

Accounting for a correct selection of the strategy to follow in each possible scenario, we aim to define a cost function in order to measure the human, economical, etc. cost of each strategy. The defined cost function takes into account two main factors:

- the number of affected people depending on the height of the water in the flood, and
- the time that the strategy is able to delay the flood.

Assuming that the flooding starts at a given time t_0, the proposed cost function of the flood at a certain time t is

$$C(t) = \int_{t_0}^{t} \frac{\rho(h(q(t)))}{t}\, dt, \tag{3.5}$$

where ρ is the density of population that lives below a certain height h. Note that the height of the flooding is a function of q (the height that the flooding reaches depends on q), and that both q and h are known in any possible scenario. We have divided the density by t in order to "decrease the cost" when increasing the time, since we want to minimize the effects of the flood but we also want to penalize the time that a strategy gives to the population to leave the affected zone.

Note that, in this cost function definition, it can be decided if more importance is given to the amount of affected people or to the delaying of the flooding. For instance, if we want to give more importance to the delay of the flood, we can change the modelling cost and use some other t^r, $r > 1$, penalizing then a strategy that does not let enough time to the population to evacuate. Analogously, if we choose not to involve the delaying time in the cost penalization, we can also take $r = 0$.

3.4.2 Finding the strategy that minimizes the cost function

Given a scenario, we know all the variables that are involved in the definition of the cost function (3.5). Hence, in order to execute the better strategy, we

are only required to compute the integral (3.5) and take the strategy with a lower cost. The most valid strategy is the one that results in the minimum cost at the end time of the flood, t_1.

However, it may be interesting to be able to know a priori bounds about the possible cost of each situation. Hence, we could have information before the flooding occurs about how each strategy behaves and, for example, we could be able to fix a default strategy that is known to be safer through the a priori bounds. Furthermore, if we are able to find lower and upper bounds for both strategies we may find interesting conclusions about the procedure to follow when a flood occurs.

Strategy A: Opening the gates when q_{mao} is achieved.

With strategy A we find that, despite reaching higher flow values, the time in which q_{mao} is surpassed is delayed, and thus there are no costs for a longer time than in strategy B. However, we are just able to delay the flow until a known time t_m where the auxiliary flooding area is full and we cannot take more water from the river. Thus, the cost of strategy A is

$$C(t_1) = \int_{t_0}^{t_1} \frac{\rho(h(q(t)))}{t}\, dt = \int_{t_m}^{t_1} \frac{\rho(h(q(t)))}{t}\, dt,$$

where we must recall that all the data required in the integral are known, and thus in a particular case it is straightforward to compute the cost. Hence,

$$C_A = \int_{t_m}^{t_1} \frac{\rho(h(q(t)))}{t}\, dt \tag{3.6}$$

is the cost of strategy A.

Note that an upper bound of this expression can be found, namely

$$\begin{aligned} C_A &= \int_{t_m}^{t_1} \frac{\rho(h(q(t)))}{t}\, dt \leq \int_{t_m}^{t_1} \frac{\rho_*}{t}\, dt \\ &= \rho_* \ln\left(\frac{t_1}{t_m}\right) =: \hat{C}_A, \end{aligned}$$

where $\rho_* = \rho(h(q_*))$, being $q_* = \max_{t\in(t_m,\, t_1)} q(t)$ the peak value of q. Recall that this bound may be slightly coarse, but it is useful in the sense that if the cost that the bound assigns is acceptable compared to B, then the strategy is valid for our purposes.

Strategy B: Opening the gates when q_f is achieved.

For this strategy we can easily find an accurate upper bound for its cost. Note that, except at the beginning and at the end of the flow diagram, the flow that is flooding the area of interest is known, since we cut it using the gates at q_f value. Hence, we control that $q(t) = q_f$ in all the diagram except in a small region, where $q(t) \leq q_f$ for all t. Thus,

$$\begin{aligned} C(t_1) &= \int_{t_0}^{t_1} \frac{\rho(h(q(t)))}{t}\, dt \leq \int_{t_0}^{t_1} \frac{\rho(h(q_f))}{t}\, dt \\ &= \int_{t_0}^{t_1} \frac{\rho_{q_f}}{t}\, dt = \rho_{q_f} \ln\left(\frac{t_1}{t_0}\right), \end{aligned}$$

where we have denoted by ρ_{q_f} the density of population that lives below $h(q_f)$. Thus,

$$C_B = \int_{t_0}^{t_1} \frac{\rho(h(q(t)))}{t}\, dt \tag{3.7}$$

is the cost of strategy B, and it can be bounded as

$$C_B \leq \rho_{q_f} \ln\left(\frac{t_1}{t_0}\right) =: \hat{C}_B, \tag{3.8}$$

where ρ_{q_f}, t_0 and t_1 are all known.

Depending on the scenario that is placed, we can compute both C_A and C_B and have a reference of which one of the strategies will bring up more advantages. The bounds are simplified expressions that require less data and give useful information, since they tell the limit of the cost that the strategy may result in.

However, note that, judging just by the a priori found bounds of both strategies, no conclusions can be drawn about if there is one strategy that is always better than the other one just judging by either the affected population or the maximum flood capacity. No a priori conclusions can be extracted from the comparison $\hat{C}_A \leq \hat{C}_B$.

Summarizing, we first have been able to develop two main strategies to deal with a flood. Moreover, we have defined a cost function that can be used as a key to choose the ideal strategy to select in each scenario. Two upper bounds of the cost of both strategies have been extracted. However, no conclusions about the selection of the best strategy can be drawn just by judging the a priori bounds.

3.5 Conclusion and further work

In this paper we have stated an algorithm for determining the opening and closing times of a single flood gate. In addition we have studied how cost functions can be used to minimize damage and hardship caused by flooding in highly populated areas along the Ebro River.

We propose that in future research studies the case of gate control is considered more closely, perhaps in conjunction with the idea of minimizing costs in the case that flooding cannot be avoided. The most important extension of our ideas is to work with multiple gates and multiple flood waves. This case will be more complex but also will more closely impact actual implementation of gate control along the Ebro River. By assigning a cost to the flooding of particular floodplains, it will be possible to obtain the minimum amount of damage in case that we cannot prevent a flood with the gate control scheme.

Bibliography

[1] *Directive 2007/60/EC of the European Parliament and of the Council of 23 October 2007 on the assessment and management of flood risks*, `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:288:0027:0034:EN:PDF`.

[2] A. K. Lohani, N. K. Goel, K. K. Bahtia, *Real time flooding forecasting using fuzzy logic*, International Conference on Hydrological Perspectives for Sustainable Development (HYPESED 2005), Roorkee, India.

Modelling the Cardiovascular System for Automatic Interpretation of the Blood Pressure Curve

Problem presented by

Vicent Ribas Ripoll and Anna Sáez de Tejada (Sabirmedical)

Report prepared by

Tim Myers (CRM), Mark McGuinness (Victoria University of Wellington), Sarah Mitchell (MACSI, University of Limerick)

Study group contributors

Michelle De Decker (CRM), Francesc Font (CRM), Andrew Fowler (MACSI, University of Limerick), Jonathan Low (CRM)

This report stems from a follow-up meeting entitled *Mathematical Modeling of Blood Flow and the Baroreflex System*, held at the CRM in December 2010. The work continues from the Sabirmedical problem described in the first report of this book.

4.1 Introduction

The cardiovascular system is in charge of conveying nutrients and oxygen to the tissues and maintaining the gas exchange between tissues (CO_2 and O_2) necessary for homeostasis. The main components of the cardiovascular system are the heart, arteries and veins. It includes the pulmonary circulation —a closed loop through the lungs where blood is oxygenated— and the systemic circulation. Oxygenated blood enters the systemic system at the left heart and is then pumped into the aorta. The aorta branches into smaller arteries, arterioles and capillaries, where oxygen exchange takes place, and blood enters the systemic veins through which it flows in vessels of progres-

sively increasing size toward the right heart. The right heart pumps CO_2 rich blood into the lungs.

The term blood pressure refers to the force that blood exerts on the walls of blood vessels. This parameter changes both in time and with distance from the aortic arch. Systolic pressure is the highest surge of pressure during ventricular contraction whilst diastolic pressure is the lowest/baseline pressure reached during ventricular relaxation (diastole) [10]. Currently there exist two principal methods for measuring blood pressure:

- The sphygmomanometer —this is the standard cuff which is usually inflated on the upper arm. In general it is manually operated and requires a quiet environment. It is an old technology (from the 1800s) that is prone to operator error. It also provides data only for a short period. Whilst digital cuffs exist, these may be highly inaccurate.

- The catheter —this is inserted into an artery and can provide continuous data. However, it is an invasive technique which has a number of associated risks and so is primarily used on bed-ridden patients.

Obviously there is a clear need for a non-invasive, continuous monitoring technique and various research groups have attacked this problem by different methods. One such approach involves the use of the pulse oximeter [16]. This has particular appeal since the pulse ox is already standard equipment in most medical practices.

The pulse oximeter is a device that measures the oxygen saturation of the blood. Typically it functions by shining two lights of different wavelength (but both close to infra-red) through a translucent part of the body. The different wavelength lights are absorbed to differing degrees by the oxygenated and deoxygenated haemoglobin and so the ratio of oxygenation to deoxygenation may be calculated. Since arterial blood vessels respond to pressure changes, the obtained signal is time-dependent and so the output of the pulse oximeter may also be used to monitor the heart rate. In fact this variation in the signal is essential to the functioning of the device, since, to distinguish the light absorption from blood and the surrounding tissue, the pulse oximeter only uses the varying part of the signal.

The output from the pulse oximeter is termed photoplethysmograph (or *pleth* for short). The pleth closely resembles the blood pressure curve. Currently the main uses of the pulse oximeter are monitoring of oxygenation and heart rate and diagnosis of sleep disorders. However, the detailed features of the blood pressure curve and hence the pleth contain a wealth of information useful for diagnostic purposes. For example, the blood pressure curve may be used for:

1. Detection of cardiac arrhythmia;
2. Measuring the heart output;
3. Measuring blood loss;
4. Monitoring respiratory variation (which in turn may be related to fluid responsiveness in ventilated patients with circulatory failure);
5. Depending on different base pathologies such as sepsis or severe respiratory distress, the peripheral pulse waveform may be attenuated/ dampened and may be considered as an important measure for the assessment of microcirculation and tissue perfusion.

More comprehensive lists may be found in [3, 15, 17, 18]. Until recently there was no way to accurately relate the pleth to the blood pressure (the scaling can depend on patient age, obesity, gender and many other factors). The problem was solved by researchers at Sabirmedical using a random forest algorithm [16]. Unfortunately the random forest approach does not provide an understanding of the mechanisms behind the pleth (although, of course, physicians are able to interpret it). At two meetings, held in the Centre de Recerca Matemàtica in 2010, mathematicians were challenged to develop an accurate model of the cardiovascular system in order to better understand the pleth and so extract further information. This would allow the pulse oximeter to be used as an automatic diagnostic tool (so eliminating operator error and allowing less highly trained operators to perform preliminary diagnoses). The following work results from those meetings.

The specific goal of this paper is thus to produce a mathematical model capable of accurately reproducing the dominant features of the blood pressure curve and in particular the dicrotic notch and the variation due to respiratory sinus arrhythmia (RSA). To allow for easy interpretation and to minimise the number of parameters, effort was made to keep the model as simple as possible. In the next section we describe the compartment model approach that was used. Subsequent sections deal with model refinements and parameter estimation. Finally we compare the results of our model to data.

4.2 Compartment models

Differential equation models of the cardiovascular system vary from a pair of first-order ordinary differential equations [12] to systems of more than 40 coupled differential-delay equations [8, 20, 22]. Ottesen's approach [12] is to simply split the system into arterial and venous compartments, with non-pulsatile flow and the left ventricle adding a source term, whilst Grodins and

Ursino et al. [8, 20, 22] describe the pulsatile flow of blood through multiple compartments including the lungs, compliant arteries and veins and driven by a heart that is regulated by a sophisticated nervous feedback control system that responds to blood pressure and chemistry in a variety of ways.

Despite this level of complexity, in general compartment models are conceptually quite simple. They involve splitting the cardiovascular system into a number of compliant zones or compartments and as the blood passes through each zone, blood mass should be conserved. The change in volume in each zone is simply the difference between the flux entering from upstream and that leaving downstream. The heart drives this flux and the flow is resisted by the vessels through the shear stress at vessel walls.

In the present study we have tried to take the simplest approach possible, along the lines of Ottesen's model. Initially we employed a three-compartment model, involving the arteries, veins and left ventricle. However, one of our main goals was to capture the dicrotic notch. The dicrotic notch occurs due to a pressure pulse when the aortic valve closes and closure is caused by the pressure drop across the valve becoming negative. The size and location (in time) of the notch provides information about the health of the valve. In the three-compartment model the arterial pressure is an average across the whole arterial system, so that using this as a measure of when the valve closes will lead to late closure in the model. To improve this behaviour, we introduced a new compartment to describe the exit region close to the valve. Mathematically speaking, there is no problem in dividing the cardiovascular system into any number of compartments (as is done with finite element computations) but to give a physical meaning to this new compartment one could think of it as representing the aortic arch. Our basic model is therefore described by the following four-compartment system:

$$\dot{V}_e = Q_{LV} - Q_e,\ \dot{V}_a = Q_e - Q_a,\ \dot{V}_v = Q_a - Q_v,\ \dot{V}_{LV} = Q_v - Q_{LV}, \quad (4.1)$$

where V_i represents the volume and Q_i the flux of blood. The subscripts e, a, v, LV represent exit, arterial, venous and left ventricle, and dots indicate derivative with respect to time. The first equation indicates that the change in volume in the exit region depends on the difference between the fluid flowing in from the left ventricle and the fluid flowing out of the aortic arch. In the arteries the volume increases due to fluid flowing in from the arch and decreases as it flows out of the arteries (into the veins), etc. Hence these equations express conservation of blood mass. Summing the four we find $\dot{V} = \dot{V}_e + \dot{V}_a + \dot{V}_v + \dot{V}_{LV} = 0$, so the total volume is constant. Note that we assume the blood to be incompressible (see [14]) —changes in pressure are associated with the compliance of blood vessels rather than the relatively small compressibility of blood itself.

In a compliant elastic vessel we may relate the pressure to the volume via $V = V_0 + Cp$, where C is a constant, termed the compliance, and V_0 is the (constant) volume at ambient pressure (the value of the compliance is discussed in detail later). This equation serves to define compliance.

Special attention should be paid to the left ventricle, where the pumping of the heart is driven by changes in the elastance (the inverse of compliance). Consequently, in the left ventricle we write $V = V_0 + p/E_{LV}$. In fact here we see a loose definition of the elastance, E, which is the change in pressure divided by the change in volume (it is analogous to the spring constant in Hooke's law, which is defined by the change in force divided by the change in length). Differentiating the compliance and elastance definitions, we can relate volume and pressure as

$$\dot{V}_e = C_e \dot{p}_e, \quad \dot{V}_a = C_a \dot{p}_a, \quad \dot{V}_v = C_v \dot{p}_v, \quad \dot{V}_{LV} = \frac{d}{dt}\left(\frac{p_{LV}}{E_{LV}}\right). \tag{4.2}$$

Note that we assume that the compliance of the blood vessels is a constant whereas the elastance, representing the contraction of the heart muscle, varies with time. We may relate the fluxes to the pressure by considering standard, uni-directional pressure driven laminar flow (Poiseuille flow) in a pipe which leads to a relation of the form $Q \propto \Delta p$. This may be expressed as $Q = \Delta p/R$ where R is termed resistance [10]. Obviously the cardiovascular system does not consist of a single straight pipe and blood flow is often turbulent, so this definition of Q is rather approximate and the resistance R must represent the many intricacies of the system, rather than simply the viscous resistance from the classical Poiseuille flow model. With the fluxes written in terms of pressure drop, our initial system of differential equations can be expressed as

$$C_e \dot{p}_e = \frac{p_{LV} - p_e}{R_e} - \frac{p_e - p_a}{R_a}, \tag{4.3}$$

$$C_a \dot{p}_a = \frac{p_e - p_a}{R_a} - \frac{p_a - p_v}{R_c}, \tag{4.4}$$

$$C_v \dot{p}_v = \frac{p_a - p_v}{R_c} - \frac{p_v - p_{LV}}{R_v}, \tag{4.5}$$

$$\frac{d}{dt}\left(\frac{p_{LV}}{E_{LV}}\right) = \frac{p_v - p_{LV}}{R_v} - \frac{p_{LV} - p_e}{R_e}. \tag{4.6}$$

4.3 Model refinements

The above system, equations (4.3)–(4.6), constitutes our basic set of equations but still requires certain refinement: the driving mechanism for the flow is not defined, neither is there a mechanism to describe the dicrotic notch or the aortic valve.

The driving mechanism for the flow comes through the definition of the elastance. Modelling of the elastance is discussed in a number of papers. Whilst there is some difference in the fine detail, the general form is of a sequence of roughly Gaussian curves when contraction occurs separated by flat regions denoting the relaxation [5, 11, 13]. In [4] the elastance has an approximately square waveform. However, Suga [19] points out that an instant rise cannot account for the Fenn effect of the skeletal muscle, whereas a gradual rise model can. The elastance also exhibits a longer term variation and this leads to Respiratory Sinus Arrhythmia (RSA), which is a term for the observed changes in heart rate associated with respiration. Heart rate is usually observed to increase during inspiration and decrease during expiration. This is primarily due to a coupling through the vagal (nervous) system, and to a lesser extent it is also due to the effect of breathing on the pressure in the chest cavity near the heart. Denervated (transplanted) human hearts do still exhibit RSA but at around 7.9% of normal levels [2]. Dat [5] defines the elastance as

$$E_{LV} = E_d + a(t)(E_s - E_d), \tag{4.7}$$

where the diastolic elastance E_d is constant and $a(t) = \sin^2(\omega t)$.

We employ the same form as Dat but with two important refinements. Firstly, $a(t)$ should switch off for $2/3$ of the heart cycle. If T represents the period of one heart beat, then we want the $\sin^2$ term to rise from zero and fall back to zero once as t varies from a starting value t_0 to $t_0 + T/3$. So we set $\omega = 2\pi/T$ and define

$$a(t) = \begin{cases} \sin^2(3\omega(t - t_0)/2) & \text{if } t - t_0 \in (0, T/3), \\ 0 & \text{if } t - t_0 \in (T/3, T). \end{cases}$$

The period T is set to be time dependent to model RSA with heart rate varying with respiration, by choosing

$$\omega = \omega_0 + c_3 \sin\left(\frac{\omega_0 t}{c_2}\right). \tag{4.8}$$

Secondly, the peak in the elastance height varies over a longer time scale (related to heart beats per respiration). To account for this, we take the systolic elastance

$$E_s = E_{s0} + c_1 \sin\left(\frac{\omega_0 t}{c_2}\right). \tag{4.9}$$

The constants c_1, c_2, c_3 in the above definitions represent half the variation of elastance height, the number of heart beats per respiration (typically around 5) and half the variation of ω respectively. The constant ω_0 is an

angular frequency $\omega_0 = (2\pi/60) \times HR$ where HR is an average heart rate in beats/minute.

A typical form for the elastance is shown in Figure 4.1. The minimum value $E_d = 0.06$ mmHg/ml, while the maximum value varies according to equation (4.9) with an average value $E_{s0} = 3$ mmHg/ml. We took a heart rate of 72 beats/minute to give $\omega_0 \approx 7.54$. Other parameter values used to generate this graph are given in Table 4.1.

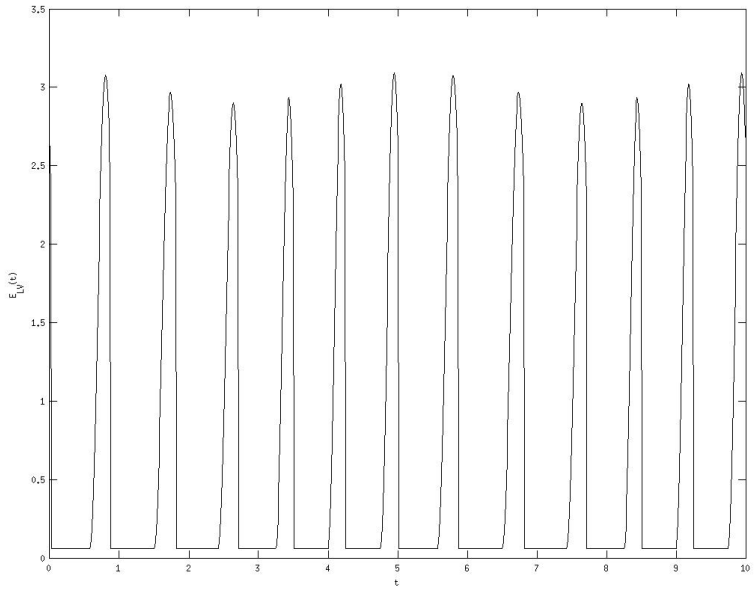


Figure 4.1: Variation of elastance with time

The resistances R_a, R_c, R_v are constant while R_e accounts for the aortic valve that closes when the pressure drop becomes negative. Consequently, R_e must be time-dependent. Since closure depends on the pressure difference, we write

$$R_e = R_{e0}\left[1 + \epsilon_1(\exp(-A_1(p_{LV} - p_e)))\right], \tag{4.10}$$

where R_{e0} is the constant value when the valve is fully open. The factor $\epsilon_1 \ll 1$ ensures that the exponential term remains small whenever $p_{LV} - p_e > 0$ but it increases rapidly when $p_{LV} - p_e < 0$. The constant A_1 is chosen such that the product $A_1(p_{LV} - p_e)$ rises sufficiently rapidly as the valve closes. In practice we set $\epsilon = 10^{-5}$, $A = \frac{1}{2}$. These values are simply chosen to provide the correct properties. Since the exit region is much smaller than the arterial region, we also assume $R_{e0} \ll R_a$. Another option would be to simply set a switch via a Heaviside function. However, our subsequent numerical calculations showed that this led to a poor representation of the

pressure around the dicrotic notch. Ellwein et al. [6] employ a similar, but cut-off, exponential representation for all heart valves. They do not include the factor $\epsilon_1 \ll 1$ but choose $A = 2$, which is greater than our value, and this has a similar effect. The typical behaviour of the valve is shown in Figure 4.2.

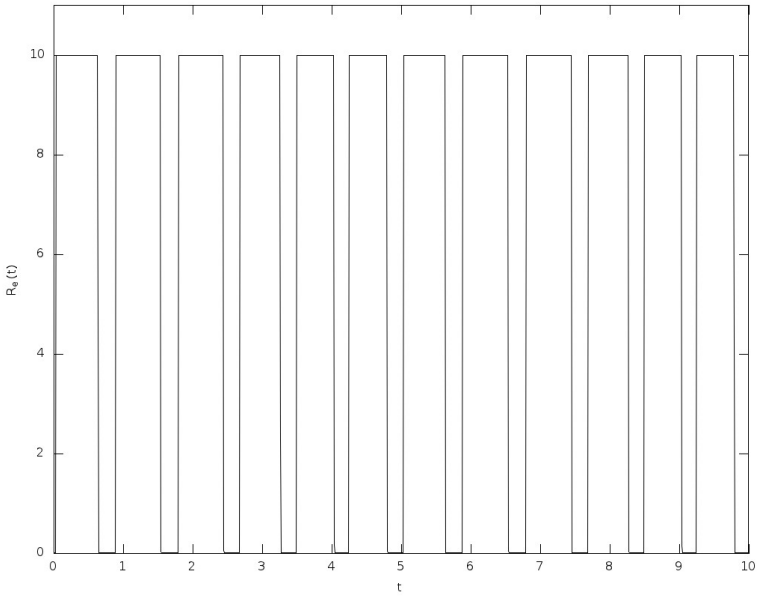


Figure 4.2: Valve resistance against time

Note that we could equally well define a valve at the entrance to the ventricle through

$$R_v = R_{v0}\left[1 + \epsilon_2(\exp(-A_2(p_v - p_{LV})))\right]. \tag{4.11}$$

However, since this region is of lesser interest, to avoid more parameter estimation we use a Heaviside function for R_v (tests confirm that this makes no noticeable difference to the results).

The modelling of the dicrotic notch is based on the assumption that it is caused when blood attempting to flow back through the valve, due to the negative pressure and inertia, impacts with the now closed valve and rebounds into the exit region. We approximate this impulse by a Gaussian, with a strength related to the pressure difference $p_e - p_{LV}$. Since pressure is a function of time, we can represent the impulse by the following function:

$$f(t) = c_4 \exp(-c_5(t - t_n - c_6\Delta t)^2/(\Delta t)^2). \tag{4.12}$$

The constants c_4, c_5, c_6 indicate the height of the pulse, the sharpness, and the position of the centre. The time t_n is when $p_e = p_{LV}$ and so it indicates

when the valve should begin closing (obviously as the pressure varies over time there are many values of t_n). The maximum value of f occurs when $t = t_n + \Delta t$; hence Δt denotes the delay in closure after the pressure drop becomes negative. It therefore controls the position of the dicrotic notch and it is an important indicator of the health of the valve. In the numerical solution we use a value of Δt from the previous cycle (and taken as some typical value for the initial condition).

The function $f(t)$ represents an input of mass to the p_e equation and so it is added to equation (4.3): to conserve mass it must be subtracted from the left ventricle —equation (4.6). In general its maximum value is much lower than the other terms in the equation and so it represents only a small contribution to the pressure. The full system to model the pressure is now given by equations (4.4), (4.5) and

$$C_e \dot{p}_e = \frac{p_{LV} - p_e}{R_e(t)} - \frac{p_e - p_a}{R_a} + f(t), \tag{4.13}$$

$$\frac{d}{dt}\left(\frac{p_{LV}}{E_{LV}(t)}\right) = \frac{p_v - p_{LV}}{R_v} - \frac{p_{LV} - p_e}{R_e(t)} - f(t), \tag{4.14}$$

with E_{LV} defined by (4.7) and R_e by (4.10).

4.4 Calculating parameter values

Key to the success of the mathematical model is the choice of parameter values. Obviously this is not a simple task. However, many of the values can be found in the literature whilst a few were obtained heuristically. Of course the term heuristics can hide a multitude of sins. Our choices are largely based on knowledge of pressure signals. For example, the constants c_4, c_5, c_6 simply come from inspection of the signal for a dicrotic notch. The beats/breath c_2 is a standard value of the order of 5 (in fact we take it to be 6 based on our experimental data). The small parameter ϵ_1 has merely to be sufficiently small so that the exponential term in equation (4.10) is negligible when the valve is closed, and A_1 sufficiently large so that the exponential term becomes significant when the valve opens.

Compliances and resistances may be calculated through a given pressure signal. For example, rearranging the volume pressure relation discussed in Section 4.2, we find

$$C = \frac{V - V_0}{p}. \tag{4.15}$$

This may be interpreted as the stroke volume to mean average pressure; see [10]. According to [1], it may be interpreted as the ratio of stroke volume

to arterial pulse pressure. Since we have data for the arterial pressure, we may calculate

$$C_a = \frac{SV}{P_{a,\,sys} - P_{a,\,dia}}. \tag{4.16}$$

We then use this to estimate $C_e = C_a$ and $C_v = 20C_a$. Similarly, the systemic resistance may be estimated from the arterial pressure signal [1] via

$$R_v = \frac{P_{a,\,mean} - P_{v,\,mean}}{SV \cdot HR}, \tag{4.17}$$

where SV is the stroke volume and HR the heart rate. Values close to those obtained from the above formulae are quoted in [4]. In the exit region (the aortic arch) we assumed that the compliance was the same as in the rest of the arterial system. However, since it is wider than other vessels, we used a lower resistance $R_{e0} = R_v$.

The full set of parameter values, together with the source for the values, are given in Table 4.1. Typical values for most model parameters may be found in [5, 12].

Parameter	Value	Units	Source	Parameter	Value	Units	Source
C_e	1.5	ml/mmHg	1	C_a	1.5	ml/mmHg	1
C_v	50	ml/mmHg	1	R_{e0}	0.016	s·mmHg/ml	1
R_a	0.06	s·mmHg/ml	1	R_c	1.2	s·mmHg/ml	1
R_v	0.016	s·mmHg/ml	1	T_{sys}	$T/3$	s	2
T_{dia}	$2T/3$	s	2	T	0.9	s	2
E_d	0.06	mmHg/ml	2	E_{s0}	3.0	mmHg/ml	2
ϵ_1	10^{-5}		3	A_1	0.5		3
R_{eMax}	10	s·mmHg/ml	3	c_1	0.1	mmHg/ml	3
c_2	6	beats/breath	3	c_3	0.01	s^{-1}	3
c_4	500		3	c_5	$4 \log 100$		3
c_6	7.5		3	ω_0	7.54		3

Table 4.1: Parameter values. Sources are: 1. Equations (4.16), (4.17); 2. References [4, 5, 12]; 3. Heuristics.

4.5 Results

In Figure 4.3 we present the pressure curves obtained from the numerical solution of the governing equations. As should be expected, the pressures decrease with distance from the heart, so the pressure curve with the highest maximum represents the left ventricle. Below this is the exit region pressure, the arterial pressure, and finally the venous pressure. Both the exit and

arterial curves exhibit a distinct dicrotic notch: the exit notch being sharp whilst in the arterial system it is more diffuse. The time axis has been shifted so that we only present pressure curves when the system has settled down, that is, when the effect of the input initial conditions has died out.

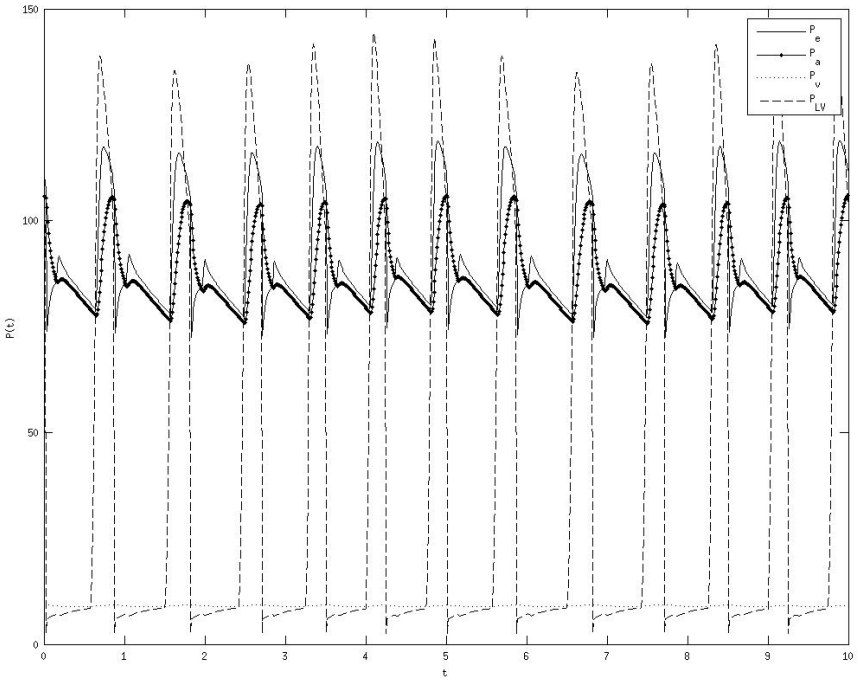


Figure 4.3: Model prediction of pressure variation with time. Moving from the highest peak to the lowest the curves are p_{LV}, p_e, p_a, p_v.

Obviously our main interest lies in the arterial pressure, since this is what is measured in practice. In Figure 4.4 we show only the arterial pressure and for comparison we also present data provided by the ICU at Hospital Vall d'Hebron, Barcelona. Clearly the agreement is excellent. Slight differences can be observed in the maximum and minimum values, but the greatest difference appears to lie in the variation of the dicrotic notch —our model misses this slightly at certain times.

4.6 Conclusion

The results presented in the previous section clearly indicate that our model can accurately reproduce a blood pressure signal with the correct choice of parameter values. The challenge now is to take a pleth and automatically

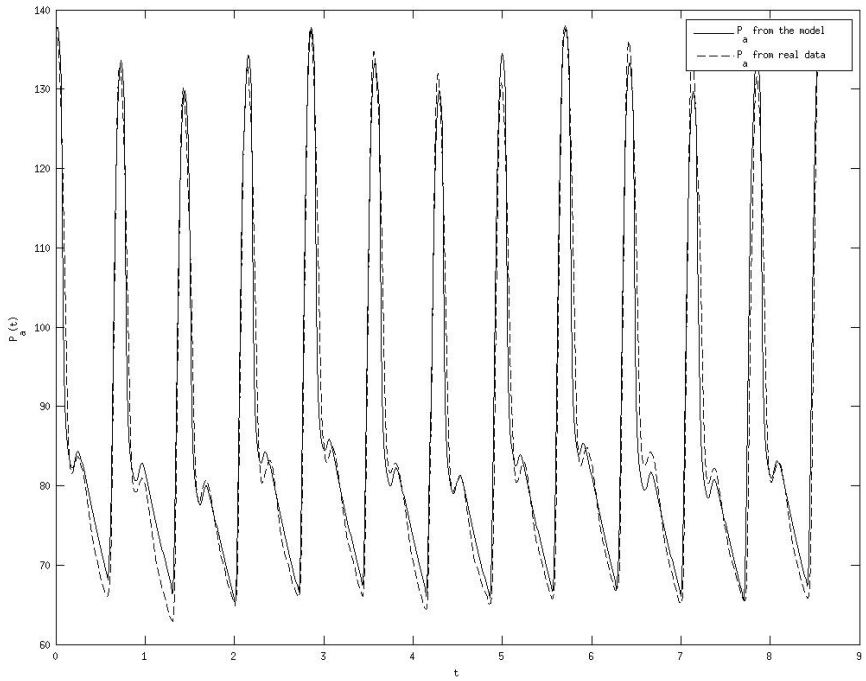


Figure 4.4: Comparison of arterial pressure from model and data from ICU at Hospital Vall d'Hebron

calculate the parameter values required for the model and, based on the output, be able to make some recommendation concerning the health of the patient. This will comprise the next stage of this investigation. In order to achieve this, we note that the vascular tree can be modelled as a non-linear time-variant channel, and there are tools that can either linearize this system in order to do the translation from catheter to pleth (i.e., Wiener filters) or just live with these non-linearities and make predictions (Kalman filter). On the positive side, we have a large amount of data both from arterial catheters and the corresponding pleth on which to validate the method.

Acknowledgements

The initial work for this paper was carried out in the Centre de Recerca Matemàtica during December 2010 at the Workshop on Mathematical Modeling of Blood Flow and the Baroreflex System. We acknowledge the assistance of other participants at that workshop, including Michelle De Decker, Francesc Font, Jonathan Low (all from the CRM) and Prof. Andrew Fowler of the University of Limerick.

Financial support was provided through a Marie Curie International Reintegration Grant *Industrial applications of moving boundary problems*, grant no. FP7-256417, Ministerio de Ciencia e Innovación grant MTM2010-17162 and the Mathematics Applications Consortium for Science and Industry (`www.macsi.ul.ie`), funded by the Science Foundation Ireland mathematics initiative grant 06/MI/005.

Bibliography

[1] J. F. Augusto, J. L. Teboul, P. Radermacher, P. Asfar, *Interpretation of blood pressure signal: physiological bases, clinical relevance and objectives during shock states*, Intensive Care Medicine **37** (2011), 411–419.

[2] L. Bernardi, F. Keller, M. Sanders, P. S. Reddy, B. Griffith, F. Meno, M. R. Pinsky, *Respiratory sinus arrhythmia in the denervated human heart*, J. Appl. Physiol. **67** (1989), 1447–1455.

[3] M. Cannesson et al., *Relation between respiratory variations in pulse oximetry plethysmographic waveform amplitude and arterial pulse pressure in ventilated patients*, Critical Care **9** (2005), R562–R568.

[4] S. J. Chapman, A. C. Fowler, R. Hinch, *An introduction to mathematical physiology*, Mathematical Institute, Oxford University, preprint, 2010.

[5] M. Dat, *Modeling cardiovascular autoregulation of the preterm infant*, Thesis, Eindhoven University of Technology.

[6] L. M. Ellwein, H. T. Tran, C. Zapata, V. Novak, M. S. Olufsen, *Sensitivity analysis and model assessment: Mathematical models for arterial blood flow and blood pressure*, Cardiovasc. Eng. **8** (2008), 94–108, DOI: 10.1007/s10558-007-9047-3.

[7] J. A. Goodwin et al., *A model for educational simulation of infant cardiovascular physiology*, Anesth. Analg. **99** (2004), 1655–1664, DOI: 10.1213/01.ANE.0000134797.52793.AF.

[8] F. Grodins, *Integrative cardiovascular physiology: a mathematical synthesis of cardiac and blood vessel hemodynamics*, Quart. Rev. Biol. **34** (1959), no. 2, 93–116.

[9] A. C. Guyton, J. E. Hall, *Textbook of Medical Physiology*, W. B. Saunders Company, 2000.

[10] J. Keener, J. Sneyd, *Mathematical Physiology*, Springer, 1998.

[11] B. Oommen et al., *Modelling time varying elastance: The meaning of "load-independence"*, Cardiovasc. Eng. **3** (2003), no. 4, 123–130.

[12] J. T. Ottesen, *Modelling of the baroreflex-feedback mechanism with time-delay*, J. Math. Biol. **36** (1997), 41–63.

[13] J. L. Palladino, J. P. Mulier, A. Noordergraaf, *Defining ventricular elastance*, in: Proceedings of the 20th Annual International Conference of the IEEE, vol. 1, Engineering in Medicine and Biology Society, Hong Kong, 1998, 383–386.

[14] T. J. Pedley, *Mathematical modelling of arterial fluid dynamics*, J. Engrg. Math. **47** (2003), 419–444.

[15] M. D. Reisner et al., *Utility of the photoplethysmogram in circulatory monitoring*, Anesthesiology **108**, no. 5, May 2008.

[16] Patent: System and apparatus for the non-invasive measurement of blood pressure, WO2010043728 (A1), ES2336997 (A1), Ribas Ripoll, V., Sabirmedical S.L.

[17] K. H. Shelley, S. Shelley, *Pulse oximeter waveform: Photoelectric plethysmography*, in: Clinical Monitoring, C. Lake, R. Hines, and C. Blitt (Eds.), W. B. Saunders Company, 2001, pp. 420–428.

[18] K. H. Shelley, *Photoplethysmography: Beyond the calculation of arterial oxygen saturation and heart rate*, Anesth. Analg. **105** (2007), S31–S36.

[19] H. Suga, *Cardiac energetics: from $E(max)$ to pressure-volume area*, Clin. Exp. Pharmacol. Physiol. **30** (2003), no. 8, 580–585.

[20] M. Ursino, *Interaction between carotid baroregulation and the pulsating heart: a mathematical model*, Am. J. Physiol. **275** (Heart Circ. Physiol. **44**) (1998), H1733–H1747.

[21] M. Ursino, M. Antonucci, E. Belardinelli, *Role of active changes in venous capacity by the carotid baroreflex: analysis with a mathematical model*, Am. J. Physiol. **267** (1994), H2531–H2546.

[22] M. Ursino, A. Fiorenzi, E. Belardinelli, *The role of pressure pulsatility in the carotid baroreflex control: a computer simulation study*, Comput. Biol. Med. **26** (1996), 297–314.