A Study on Biological Data Modelling

Hideyasu Shimadzu

March 2008

To my parents

Preface

> *Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.*
> (The future of data analysis, *Annals of Mathematical Statistics* 33:1-67, 1962)
>
> — John W. Tukey

Modelling is a fundamental of Science. It is of much importance to create a model which leads people to many discoveries. A better model can be constructed in a framework which better captures the reality. The *Data modelling* starts from establishment of such a framework by honestly looking at the observed data.

Biological data modelling is somewhat specific because the variability of data remains significant even if the experiment were carefully designed. It is nothing more than the sign of alive that homoeostasis or constancy is maintained in any of biological phenomena. It is also often the case when biological data is obtained only once as a function of time. The aim of this thesis to seek for possible ways to a good biological data modelling, which will be suggested by three case studies.

The first case study is on modelling five bird count series observed monthly. Each of the series was decomposed into three components: long trend, short trend and irregular by two step smoothing. It was clearly shown that a simple linear transformation of the long trends as a whole

is a good modelling for capturing relationships between bird count series and environmental changes. It turned out that there are two bird groups, one of which increases in number as an increase of resident area and the other decreases as a decrease of farmland area. Each short trend also allows us to understand the seasonality of the behaviour of each bird.

The second case study is on modelling swimmers' speeds over the course of a male 200 m free-style race. The model is based on a dynamical model reflecting the trade–off between drag and propulsion in swimming. It does not only fit well the data but also provide a good description of the swimming strategies of each swimmer from phase to phase in the race. An individual factor measuring how much faster or slower the individual swims relative to the average swimming speed is estimated. This factor is, as expected, closely related to the final outcome of the race.

The third case study is on modelling membrane potential of a neuron. A simple but powerful input and output system has been created by noting that each nerve cell system has two different type of synapses; chemical and electrical ones. Three phase model has been introduced for the input as well as for the spikes, which is a simplified Hodgkin–Huxley model but with an extra phase, pre–activation phase. Spike occurrences are modelled by a point process with the intensity proportional to the derivative of the input. The model would be applicable for any other membrane potential changes of a neuron as an integrated model.

An important implication of these case studies is that the models created are not a simple extension of existing theories or models. Such models could not be obtained without careful analysis of the given data. Honest approach to the data was a key to success. As a summary, it is shown that

data–driven approach is likely to open a new horizon particularly in biological data modelling because an innovative modelling is always necessary to cope with the large variability of the data.

Acknowledgements

I would like to express my sincerest gratitude to my supervisor, Professor Ritei Shibata for his patient guidance and encouragement. This thesis could not have been accomplished without his support. His profound insights and enthusiasm for researches are exceptional. It is my privilege to have met his philosophy of *Data Science* earlier on.

I would also like to express my thank to Professors Makoto Maejima, Yuji Ohgi, Kotaro Oka, Kunio Shimizu for their comments and suggestions on the earlier version of this thesis, which are very helpful to improve it.

The researches presented in this thesis have been investigated as joint research with people in each area, without whom these would not have been completed. Especially, the bird census data was provided from Jiyu–Gakuen where I graduated, the swim race data was from the Medicine and Scientific Committee of Japan Swimming Federation and the neural action potential data was from Mr Toshinobu Shimoi (Keio University). I would like to thank these groups and people for their contributions and kind approve to analyse these data sets.

During my studies, there are many people had the misfortune of crossing my path and having to spend time for me. Dr Matthew Browne (the Commonwealth Scientific and Industrial Research Organisation Mathematical and Information Sciences, Australia) was my host during my

visit to CSIRO Cleveland laboratory, from October to November 2007. Many discussions with him was great help to get new ideas in my research. Professor Ryozo Miura (Hitotsubashi University) was an adjunct lecturer in Statistics when I was an undergraduate student. He taught me a fascination of data analysis and gave me a motivation to enter this area. Dr Peter Thomson (Statistics Research Associates Limited, New Zealand) is a member of the advisory board for the 21st Century Centre of Excellence (COE) Programme: *Integrative Mathematical Sciences* at Keio University. Many discussions with him in tea time and seminars during his annual visit to Keio University were substantially improved my researches. I am very grateful to these people for their kind support and encouragement.

Thanks all are due to everyone else who has given me many help and support during my studies at Keio University, especially the member of Shibata Laboratory including Dr Natsuhiko Kumasaka and Mr Yuki Sugaya.

Financial support, over my PhD studies, from COE Programme is gratefully acknowledged.

March 2008 *Hideyasu Shimadzu*

Contents

Chapter 1

Introduction

1.1 Biological data

Biological data is the data collected from biologicals sources, whose variability remain lage even if experiments were carefully organised. It is nothing more than the sign of alive but can be a burden of biological data modelling. Specific features of biological data indispensable in the modelling are *Homoeostasis* (*Constancy*), *Uncontrollable observation* and *Time dependency*. In the subsequent sections, we concentrate our attention into the following three biological data.

1. Bird count data (Bird);
2. Swimming race data (Human);
3. Neural membrane potential data (Neuron).

Table 1.1 summarises the specific features of those three data sets. Most important feature would be the constancy, in other words, the homogeneity of the data is retained in a group of birds, in each phase of swimming races and membrane potential changes by focusing on the constancy. This suggests that a better modelling approach is, as a first step, to construct a model only

Table 1.1: Specific features of the data dealt in this thesis.

	Bird	Human	Neuron
Homoeostasis	Environment	Race	Ion
Constancy	Group	Phase	Phase
Uncontrollable	Open system	Competition	*In vivo*
Observation	Once	Once	Once
Time dependency	Yes	Yes	Yes

for such data sets holding true for constancy since the model built might be simple and easily interpreted.

1.2 Modelling

1.2.1 Modelling process

There is no definite way of biological data modelling but there are two basic principles which any data scientist should observe, although those are quite general and applicable for any scientific data modelling. The first principle is "Be honest to the given data" and the second one is "Keep a good relation to the scientist in the field". The latter makes possible to discuss what is the target of modelling and how to approach the problem. Sometimes it results in re-sampling or re-experiments.

In the process of modelling, preliminary analysis of the data is also important. Only a good preliminary analysis can provide a successful model which fits the data well and gives good explanation. It is often the case when it results in coming back to the first stage, the data collection stage. The value of modelling is, of course, how large the impact is of the model created to the field of science. New discovery or constructive suggestion is modes most desirable as a result of the modelling.

Table 1.2: Fundamental aspects of the data.

	Bird	Human	Neuron
Observation	Count	Elapsed time	Membrane potential
Variety	Species	Individual	Cluster
Constancy	Group	Phase	Phase
Time dependency	Long, Short	Lap	Trend, Spike

1.2.2 Data and mechanism behind

In the stage of preliminary analysis, a crucial point is to grasp various aspects of the data in an appropriate manner. Table 1.2 summarises fundamental aspects of the data employed in the three case studies. Correct understanding of the aspects leads people to correct modelling, although it is not enough. As was mentioned, the constancy may suggest a framework for better modelling. Is is also necessary to consider whether it could be able to ignore the difference between individual.

Furthermore, it is also of importance to capture the mechanism behind the data for successful modelling. It can be achieved only by continuous discussion with the scientists in the underlying research field. In the case of physics, such well known structures have been given by differential equations. However, such models show sometimes different behaviour from the data observed. This suggests that the assumptions for such model may not be true and the model need to be improved.

1.2.3 Complexity and accuracy of the model

Modelling is an endless work. A well known principle of modelling is to make a good balance between the complexity and accuracy of modelling (Akaike, 1973, Konishi and Kitagawa, 1996, 2007). In other word, parsimonious model

is most desirable. But there is always room for improvement of the model created. Efforts to improve the model is necessary, but time and data are limited. A criterion for the stop of our effort would be if the model has reflected all necessary information in the data.

1.3 Smoothing technique: loess

Smoothing techniques are of use to extract structures lying behind data, especially, if any significant structure cannot be assumed. There is a useful technique called local polynomial regression proposed by Cleveland (1979), Cleveland and Devlin (1988) which is available on S–PLUS as a function *lowess* or *loess*.

Local polynomial regression assumes a smooth function $f(x)$, as an expected structure, behind data $(x_i, y_i), i = 1, 2, \ldots, n$, which are observed and locally approximate using polynomials. The weight used for weighted least squared method for estimation is

$$\sum_{i=1}^{n} w\left(\frac{|x_i - x|}{d_\delta(x)}\right)\{y_i - f_x(x_i)\}^2 \underset{f_x}{\rightarrow} \min, \tag{1.1}$$

where

$$d_\delta(x) = \max_{i; x_i \in U_\delta(x)} |x_i - x|.$$

Here $U_\delta(x)$ is the nearest neighbour of x defined by $[n\delta]$, which is the maximum integer of $n\delta$. The smoothing parameter δ means the proportion of observations in the nearest neighbour. Further, f_x is a p–degree polynomial given by

$$f_x(z) = \sum_{k=0}^{p} \beta_k(x)(z - x)^k.$$

The weight function w for weighted least squared in S–PLUS is a tori cubic weight

$$w(x) = \begin{cases} (1-x^3)^3 & (0 \le x < 1), \\ \\ 0 & (\text{otherwise}). \end{cases}$$

Introduce some matrix notations, for convenience. Put

$$\boldsymbol{X}(x) = \begin{bmatrix} 1 & (x_1 - x) & \cdots & (x_1 - x)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (x_n - x) & \cdots & (x_n - x)^p \end{bmatrix}, \ \hat{\boldsymbol{\beta}}(x) = \begin{bmatrix} \hat{\beta}_0(x) \\ \vdots \\ \hat{\beta}_p(x) \end{bmatrix},$$

$$\boldsymbol{W}(x) = \operatorname{diag}\left[w\left(\frac{x_1 - x}{d_\delta(x)}\right), \cdots, w\left(\frac{x_n - x}{d_\delta(x)}\right) \right], \ \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

$\boldsymbol{X}(x)$ is an $n \times (p+1)$ design matrix and $\boldsymbol{W}(x)$ is an $n \times n$ diagonal matrix of weights.

The solution of the least squares problem can be written as

$$\hat{\boldsymbol{\beta}}(x) = \boldsymbol{A}^{-1}(x)\, \boldsymbol{X}^T(x)\, \boldsymbol{W}(x)\, \boldsymbol{y},$$

where

$$\boldsymbol{A}(x) = \boldsymbol{X}^T(x)\, \boldsymbol{W}(x)\, \boldsymbol{X}(x).$$

$\boldsymbol{A}(x) = \{a_{m,r}(x)\}$ is the $(p+1) \times (p+1)$ symmetric matrix and its (m,r) element is written as

$$a_{m,r}(x) = \sum_{i=1}^{n} (x_i - x)^{m+r-2}\, w\left(\frac{x_i - x}{d_\delta(x)}\right). \tag{1.2}$$

The smoothed value at x is given by

$$\hat{\beta}_0(x) = \frac{1}{\det \boldsymbol{A}(x)} \sum_{k=0}^{p} \operatorname{adj}(\boldsymbol{A}(x))_{1,k+1} \sum_{i=1}^{n} (x_i - x)^k\, w\left(\frac{x_i - x}{d_\delta(x)}\right) y_i, \tag{1.3}$$

where $\operatorname{adj}(\boldsymbol{A}(x))_{1,k+1}$ is the cofactor of $a_{k+1,1}(x)$.

This technique will be used in Section 2 and 4 for extracting some structures which could not assume any mechanisms behind the data.

Chapter 2

Bird count series modelling to explore environmental changes

2.1 Introduction

Relationships between avifauna and natural environment have been attracting many researchers' interests in ecology. However, their interests have been rather biased to abundance of species, particularly in field studies conducted in Japan, see Higuchi *et al.* (1982), Anada and Fujimaki (1984), Hirano *et al.* (1985, 1989), Murai and Higuchi (1988), Kurosawa (1994), Ootaka and Nakamura (1996) and Maeda (1998). Although frequency changes of each species would be of much importance to capture the effects of human activities which are apt to cause environmental changes for birds, not so many works have been done on the number of birds observed in an area for each species but there are several papers, Hirano (1996), Komeda and Ueki (2002), Uchida *et al.* (2003), Shimadzu and Shibata (2005).

In this chapter, the data observed over 35 years at Jiyu–Gakuen in Tokyo, Japan, is used to explore relationships between the number of individuals and environmental changes due to human activities. The data has been collected monthly in a well organised way. By two step smoothing technique, each bird

count series is decomposed into three components, *long trend*, *short trend* and *irregular*. To explorer relationships between those long trends and several environment indices, the scale and location of each long trend is adjusted. As a consequence, two bird groups are popped up. One is the group of Turtle Dove (*Streptopelia orientalis*), Brown–eared Bulbul (*Hypsipetes amaurotis*) and Great Tit (*Parus major*) and the adjusted long trends all fit well to the curve of increasing residential area. Another is the group of Tree Sparrow (*Passer montanus*) and Gray Starling (*Sturnus cineraceus*) and the adjusted long trends all fit well to the curve of decreasing farmland area. It will be shown that each short trend provides significant information on the seasonal behaviour of each species.

2.2 Bird census

There have been various bird censuses conducted but not necessarily well organised. Its objective, the period or the method varies census by census. Census by a national institution is usually well organised and the data is open to public on the web. Two examples of such censuses are Common Bird Census (CBC; `http://www.bto.org/index.htm`) conducted by British Trust for Ornithology (BTO) over the whole of the United Kingdom since 1962, and Breeding Bird Survey (BBS; `http://www.mp2-pwrc.usgs.gov/bbs/`) conducted by the United States Geological Survey (USGS) and the Canadian Wildlife Service (CWS) since 1966.

On the other hand, bird censuses in Japan are usually conducted by individuals or small groups, so that the collected data are not necessarily open to public but scattered over individuals or groups in Japan. In this respect, the data collected at Jiyu–Gakuen is valuable because it is the result

of a continuous survey of the number of birds in a fixed area over 35 years and the data is open to public as is explained later.

2.3 Data

2.3.1 Natural environment of observational place

Higashi–Kurume city where Jiyu–Gakuen is located is 20 km far from the centre of Tokyo and on the centre of Musashino plateau on the loamy layer of Kanto. She covers 12.92 km^2 area lying 6.5 km east and west, 3.5 km south and north and has three rivers crossing to the east: the Kurome River, Ochiai River and Tateno River.

Jiyu–Gakuen campus is located on the southeast end of Higashi–Kurume city as is shown in Figure 2.1 and covers 100,000 m^2. Many trees and bushes are found in the campus, for example, Japanese maple (*Acer palmatum*), Japanese zelkova (*Zelkova serrata*), Ginkgo (*Ginkgo biloba*), Korean hornbeam (*Carpinus tschonoskii*), Red pine (*Pinus densiflora*), Japanese white oak (*Quercus myrsinaefolia*), and Japanese aucuba (*Aucuba japonica*). Also farm place, glass land and several ponds can be found in the campus.

2.3.2 Data collection

Once a month, the census is conducted by about 40 students of the secondly school and the number of birds for each species are recorded. Whole area of the campus is investigated before noon (about 30 min between 9 to 11 am) of a fine day with no rain and light wind.

Various types of bird census have been conducted (Bibby *et al.*, 2000). For example, *Territory mapping* used in CBC is the census to identify territory

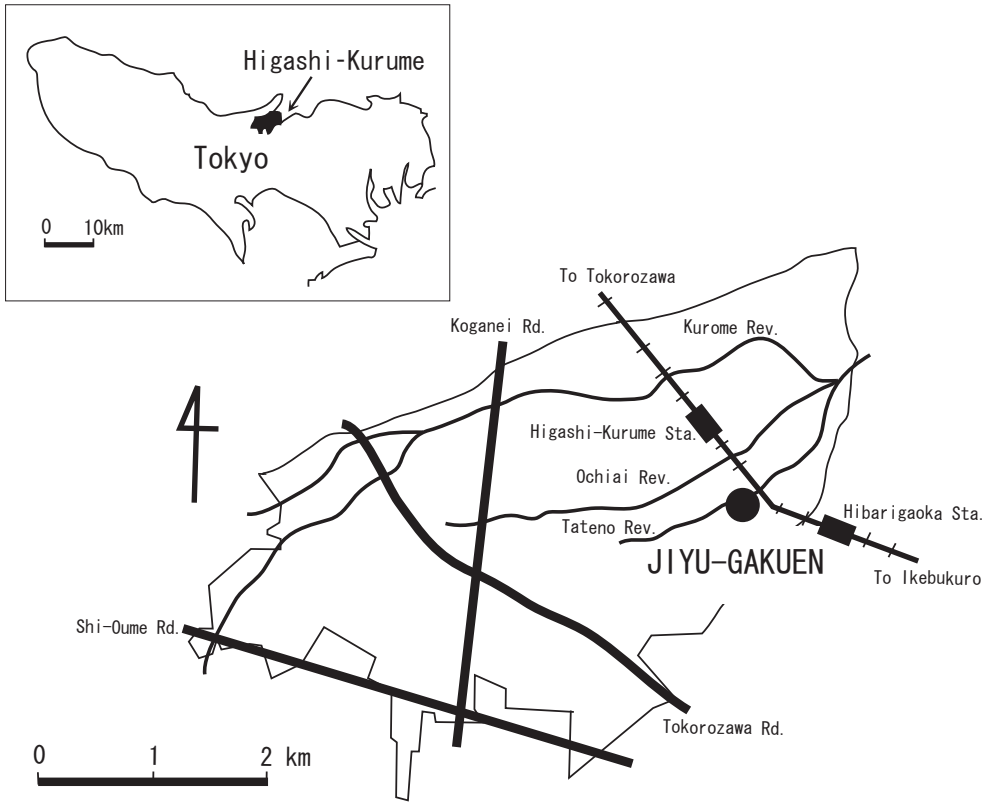

Figure 2.1: Jiyu–Gakuen in Tokyo, Japan.

of each bird and *Point counts* used in BBS or *Line transects* frequently used in Japan is the census to know the number of birds in an area. Line transects is advantageous in mountainous area. Although census has to be carefully designed depending on target species, environment conditions and quality of participants to the survey, complete sampling is adopted in this survey because of the quality of participants (Kira, 2000).

The data observed is available from Kira *et al.* (2002) or on the web

`http://www.stat.math.keio.ac.jp/DandDIII/Examples/JiyuBirdCount.dad`

which is organised along with the DandD (Data and Description) rule (Shibata, 2001, Yokouchi and Shibata, 2001).

2.3.3 Target species

More than 60 species have been observed for 32 years from 1967 to 1998, but some of them are not frequently observed as is seen on Table 2.1. In this chapter, only eight species out of 60 species are taken into consideration.

In those eight species, Azure–winged Magpie and Oriental Greenfinch are exceptional. As is seen in the count series shown in Figure 2.2, the count series of each species does not show any clear trend but rather oscillating. The reason is not the same for those two species. Oriental Greenfinch is a winter visitor of which the number largely depends on richness of foods (seeds of grass and trees) during autumn to winter. On the other hand, Azure–winged Magpie is usually moving within small bevy so that, the observed number can be large if the census were organised during their occasional visit.

Table 2.1: Observed frequency of birds for each species.

Rank	Frequency	Name	Latin name
1	383	**Tree Sparrow**	*Passer montanus*
2	377	**Rufous Turtle Dove**	*Streptopelia orientalis*
3	362	**Great Tit**	*Parus major*
4	354	**Brown–eared Bulbul**	*Hypsipetes amaurotis*
5	344	**Grey Starling**	*Sturnus cineraceus*
6	328	Azure–winged Magpie	*Cyanopica cyana*
7	305	Oriental Greenfinch	*Carduelis sinica*
8	236	**Jungle Crow**	*Corvus macrorhynchos*
9	127	Hause Swallow	*Hirundo rustica*
10	118	Duskey Thrush	*Turdus naumanni eunomus*
11	101	Black–faced Bunting	*Emberiza spodocephala*
12	89	Bull–headed Shrike	*Lanius bucephalus*
13	88	Bamboo Partridge	*Bambusicola thoracica*
14	84	White Wagtail	*Motacilla alba*
15	82	Japanese Pygmy Woodpecker	*Dendrocopos kizuki*
16	64	Bush Warbler	*Cettia diphone*
17	54	Spotbill Duck	*Anas poecilorhyncha*
18	54	Hawfinch	*Coccothraustes coccothraustes*
19	54	Japanese White–eye	*Zosterops japonicus*
20	50	Japanese Grosbeak	*Eophona personata*
21	50	Gray Wagtail	*Motacilla cinerea*
22	47	Rock Dove	*Columba livia var. domestica*
23	44	Carrion Crow	*Corvus corone*
24	40	Japanese Wagtail	*Motacilla grandis*
25	31	Daurian Redstart	*Phoenicurus auroreus*
26	24	Japanese Green Woodpecker	*Picus awokera*
27	15	Thick–billed Shrike	*Lanius tigrinus*
28	14	Japanese Lesser Sparrow Hawk	*Accipiter gularis*
29	13	Little Egret	*Egretta garzetta*
30	10	Siberian Meadow Bunting	*Emberiza cioides*
31	10	Rustic Bunting	*Emberiza rustica*
32	10	Coal Tit	*Parus ater*
33	9	Common Cuckoo	*Cuculus canorus*
34	8	Hause Martin	*Delichon urbica*
35	8	Pale Thrush	*Turdus pallidus*
36	7	Oriental Cuckoo	*Cuculus saturatus*
37	7	Brown Hawk Owl	*Ninox scutulata*
38	6	Narcissus Flycatcher	*Ficedula narcissina*
39	6	Jay	*Garrulus glandarius*
40	6	Crowned Willow Warbler	*Phylloscopus coronatus*
41	4	Goshawk	*Accipiter gentilis*
42	4	Skylark	*Alauda arvensis*
43	3	Indian Rose–necked Parakeet	*Psittacula krameri manillensis*
44	2	White–rumped Swift	*Apus pacificus*
45	2	Black Kite	*Milvus migrans*
46	2	Varied Tit	*Parus varius*

47	2	Ashy Minivet	*Pericrocotus divaricatus*
48	2	Siberian Bluechat	*Tarsiger cyanurus*
49	2	Brown Thrush	*Turdus chrysolaus*
50	2	Naumann's Thrush	*Turdus naumanni naumanni*
51	1	Teal	*Anas crecca*
52	1	Indian Tree Pipit	*Anthus hodgsoni*
53	1	Jungle Nightjar	*Caprimulgus indicus*
54	1	Blue–and–white Flycatcher	*Cyanoptila cyanomelana*
55	1	Red–breasted Flycatcher	*Ficedula parva*
56	1	Black–headed Gull	*Larus ridibundus*
57	1	Grey–spotted Flycatcher	*Muscicapa griseisticta*
58	1	Night Heron	*Nycticorax nycticorax*
59	1	Honey Buzzard	*Pernis apivorus*
60	1	White's Ground Thrush	*Zoothera dauma*

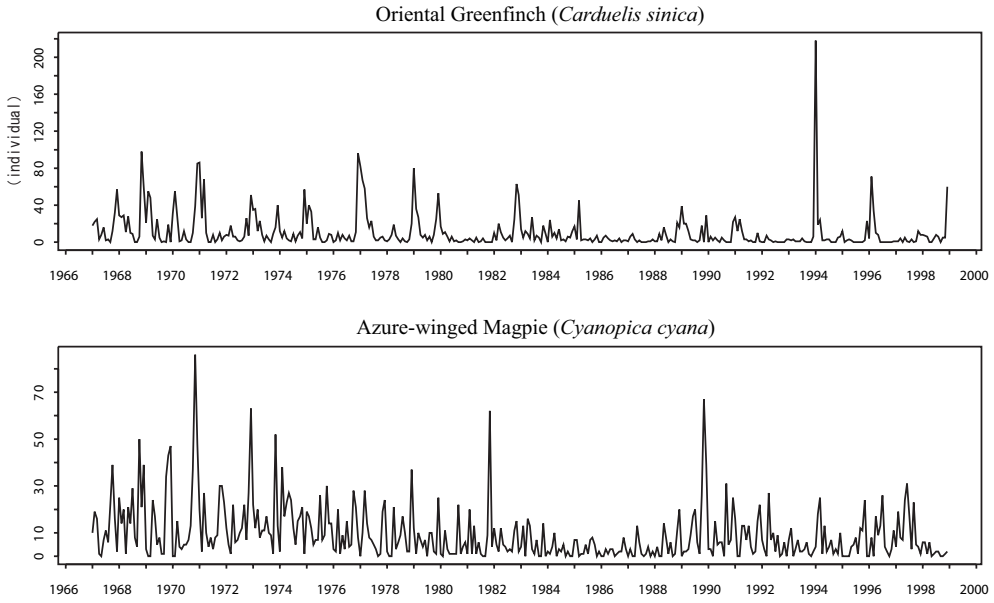

Figure 2.2: Count series of Oriental Greenfinch and Azure–winged Magpie.

The target species are now six species; Turtle Dove, Brown–eared Bulbul, Great Tit, Tree Sparrow, Gray Starling and Jungle Crow, which are indicated by bold font on Table 2.1. As a consequence, those are the species categorised into *resident bird* being observed over year in this area. It is naturally expected that their number of such a bird is strongly related with environmental changes.

2.4 Time series decomposition through loess

In time series analysis, seasonal adjustment has been widely used to extract seasonal movements. The expectation is to find a seasonal structure lying behind the data. Particularly in economics, considered are yearly, monthly or weekly seasonalities. However, in terms of birds, weekly or monthly seasonality would not be reasonable. Even if there were yearly seasonality,

it would not be so exact as in economical time series. Therefore, two step smoothing technique will be employed in place of seasonal adjustment, to decompose the original time series into three components.

There are two typical smoothing techniques:

- Spline smoothing;
- Local polynomial regression.

Spline smoothing fits a piecewise polynomial function to the given data, where the pieces are specified by the given knots. It is assumed that the derivatives of the function are continuous up to an order. On the other hand, local polynomial regression provides the smoothed value by fitting a polynomial by weighted regression in a neighbourhood of each target point. Kernel smoothing is a variant of local polynomial regression where the order of polynomials is zero. An implementation of the local polynomial regression is loess by Cleveland (Cleveland, 1979, Cleveland and Devlin, 1988) which is available on S–PLUS. There are good examples and detailed discussions on polynomial regression modelling in Chambers and Hastie (1992) or Fan and Gijbels (1996). Several works have been done for bird count series to find relation to environmental conditions by using such a local polynomial regression technique. For example, James *et al.* (1996) applied the technique to BBS data for 26 years from 1966 to 1992 and analysed 26 species of American Warblers observed in the central America. Their main result is that the number of birds largely depends on the altitude of observational place, but also they found that the deterioration of food environment by air pollution as a cause of possible.

Here it is assumed that the original series $Z_i(t)$ of species i can be

decomposed into two components and noise as

$$Z_i(t) = L_i(t) + S_i(t) + I_i(t) \ \ (t = 1, \ldots, 384).$$

where $L_i(t)$ is long trend which behaves slow in decade–long span, $S_i(t)$ is short trend having yearly span and $I_i(t)$ is irregular. The long trend $L_i(t)$ is extracted by applying smoothing technique loess to the original series $Z_i(t)$. Further, the short trend $S_i(t)$ is extracted from $Z_i(t) - L_i(t)$ by the same way with shorter span. Such decomposition approach called *two step smoothing* was applied for a financial time series (Shibata and Miura, 1997). An example of the decomposition for Tree Sparrow is shown in Figure 2.3. The top panel is the original series $Z_i(t)$ and the bottoms are following the order, long trend $L_i(t)$, short trend $S_i(t)$ and irregular $I_i(t)$.

Without any assumption of specific cycles, local polynomial regression provides smoothing values. This is desirable aspects if it is not able to assume any structure behind the data. It is necessary to chose a smoothing parameter δ by `span` and a degree of polynomial p by `degree` in S–PLUS. There are several discussions on selection of these parameters (Fan and Gijbels, 1996). However it is important how extract a reasonable trend which can be easily interpreted. Choice of parameters will be discussed in following.

2.5 Relationships between bird count series and environmental factors

Applying such smoothing technique to each of six species listed in Section 2.3.3 to extract long trend, it was clearly shown that those species are categorised into two groups, one of which increases and another decreases. Turtle Dove, Great Tit, Brown–eared Bulbul and Jungle Crow are included

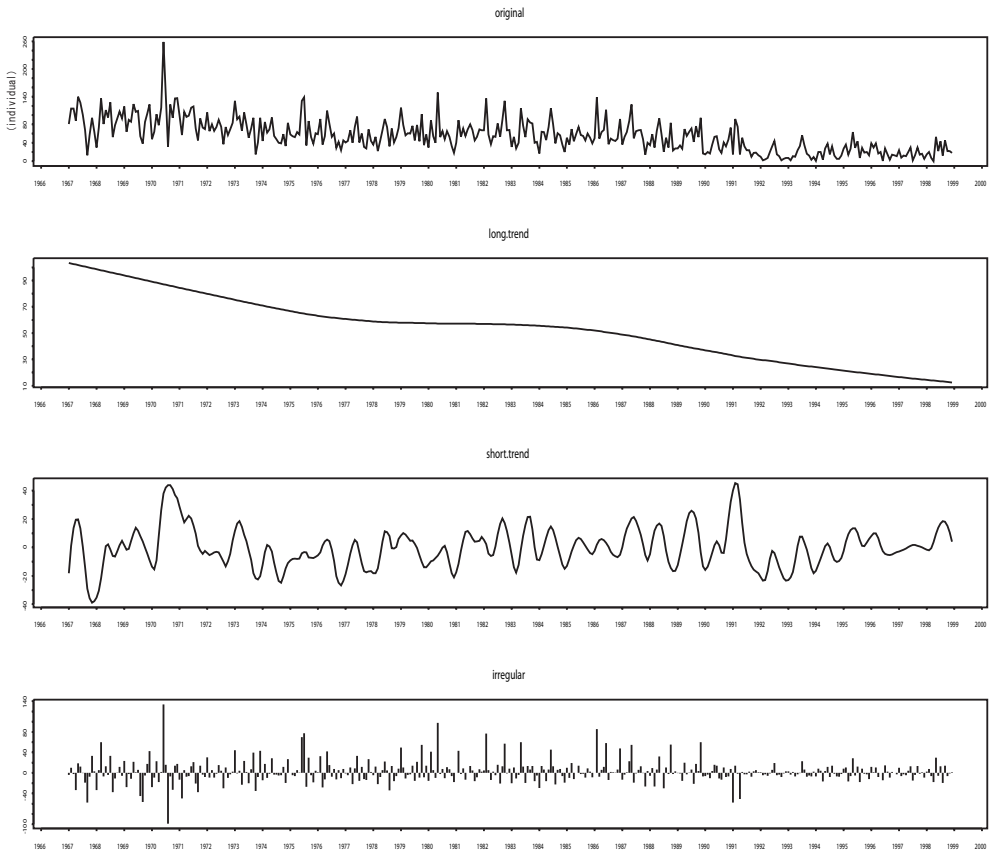

Figure 2.3: Count series decomposition of Tree Sparrow.

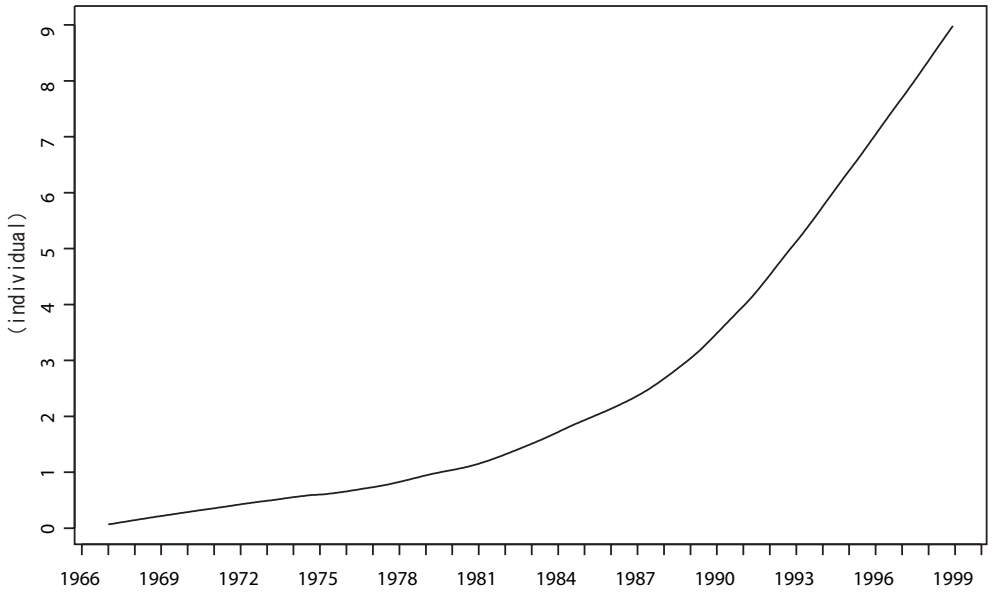

Figure 2.4: The long trend of Jungle Crow.

into the former group and Tree Sparrow and Gray Starling are included into the latter group. As to Jungle Crow, it is also increasing, but Figure 2.4 shows the different increase competitive of Figure 2.6. Such increase of Jungle Crow has been recently reported especially in urban areas over Japan. There are several causes possible, for example, the growth of trees. It would not be appropriate to discuss this species same as other so it will be dropped from the target and leave this for the future. For instance, only five species; Turtle Dove, Brown–eared Bulbul, Great Tit, Tree Sparrow and Gray Starling are analysed.

The long trends extracted from each count series are compared with environmental factors to find any clear relationships between them. Although there are, as candidates, several environmental factors, temperature, the area of classification of land, for example. It is consequently found that

four environmental factors of Higashi–Kurume city which may have strong relationships with the number of individuals:

- Resident area [km^2], *increase*;
- Farmland [km^2], *decrease*;
- Length of paved road [km], *increase*;
- Length of unpaved road [km], *decrease*.

These four factors are, as expected, related each other. However, such relations are not specific.

It would be the easiest way to overlay these two different time series having different scale. So a linear transformation is adopted for the long trend of each species to adjust their scale parameter a_i and location parameter b_i as close as possible to each environmental factor; the area of farmland $F(t)$ or the resident area $R(t)$. The parameters are defined by least squared method,

$$\begin{cases} \displaystyle\sum_{t=1}^{384} \left\{F(t) - (a_i L_i(t) + b_i)\right\}^2 \underset{a_i, b_i}{\longrightarrow} \min \quad (i = 1, 2), \\ \displaystyle\sum_{t=1}^{384} \left\{R(t) - (a_i L_i(t) + b_i)\right\}^2 \underset{a_i, b_i}{\longrightarrow} \min \quad (i = 3, 4, 5). \end{cases} \tag{2.1}$$

It is interesting to note that smoothing parameter δ for the long trend can be re–estimated so as to minimise the squared error simultaneously with the location and scale parameters. Such re–estimated smoothing parameters will be given in the next section.

2.5.1 Long trend

The estimated parameters (δ, a_i, b_i) which minimise the least squared error (2.1) are estimated with the degree of polynomial $p = 1$. The parameters

Table 2.2: Estimated smoothing parameter (`span`).

	Resident area	Farmland	Paved road	Unpaved road
Tree Sparrow	-	15 yr	-	14 yr
Gray Starling	-	29 yr	-	27 yr
Turtle Dove	32 yr	-	32 yr	-
Brown–eared Bulbul	16 yr	-	25 yr	-
Great Tit	19 yr	-	29 yr	-

Table 2.3: Estimated parameters adjust to resident area and farmland.

		Resident area		Farmland	
		a_i	b_i	a_i	b_i
1	Tree Sparrow	-	-	0.038	1.215
2	Gray Starling	-	-	0.116	1.167
3	Turtle Dove	0.478	-0.638	-	-
4	Brown–eared Bulbul	0.461	-0.028	-	-
5	Great Tit	0.876	-0.417	-	-

are shown in Table 2.2, 2.3, and 2.4. Attention to the increase of decrease of each time series, it is easily understand which environmental factor should be related with the bird count series. So then the parameters of those 3 species, Turtle Dove, Great Tit and Brown–eared Bulbul showing a increase trend are adjusted as close as possible to the resident area or the length of paved road in Higashi–Kurume city. The other hand, the rest of species; Tree Sparrow and Gray Starling are adjusted to the area of farmland and the length of unpaved road. Un estimated parameters are indicated by "-".

Figure 2.6 and 2.5 show the transformed long trends using parameters (Table 2.3, 2.4) and environment factors. These figures show that the long trends all are quite similar with environment changes, which means that the estimated parameters are significant. This show that the smoothing

Table 2.4: Estimated parameters adjust to the length of paved or unpaved road.

		Paved road		Unpaved road	
		a_i	b_i	a_i	b_i
1	Tree Sparrow	-	-	1467.282	−32246.85
2	Gray Starling	-	-	4613.627	−36988.11
3	Turtle Dove	18405.29	−114019.4	-	-
4	Brown–eared Bulbul	17448.08	−86633.47	-	-
5	Great Tit	31743.47	−91209.35	-	-

parameter chosen as $p = 1$ was enough, as well.

The estimated parameters for the three species; Turtle Dove, Brown–eared Bulbul and Great Tit in Table 2.4 take larger value than those of Table 2.3. This is because of the discontinuous behaviour of the length of paved road in Higashi–Kurume city. However, as is shown in Figure 2.6 and 2.5, their long trends show quite similar behaviour with referred environment changes even those having quite different properties. There is lying behind that the scenario of urbanisation would be considered. In fact, such rapid increase of resident area and decrease of farmland led by the high economic growth period in Japan. It is clearly shown that the resident area was lager than the farmland even since 1974.

These consideration naturally lead that the essential environmental factors for bird may be the change of resident area and farmland. Such close relationships between birds and the length of paved or unpaved road would be coincidence. It is not sure whether such significant relationship can be found from the observations in other place. This result implies the highly adaptability of birds to environment.

The two groups in Figure 2.5 can be well explained from the view

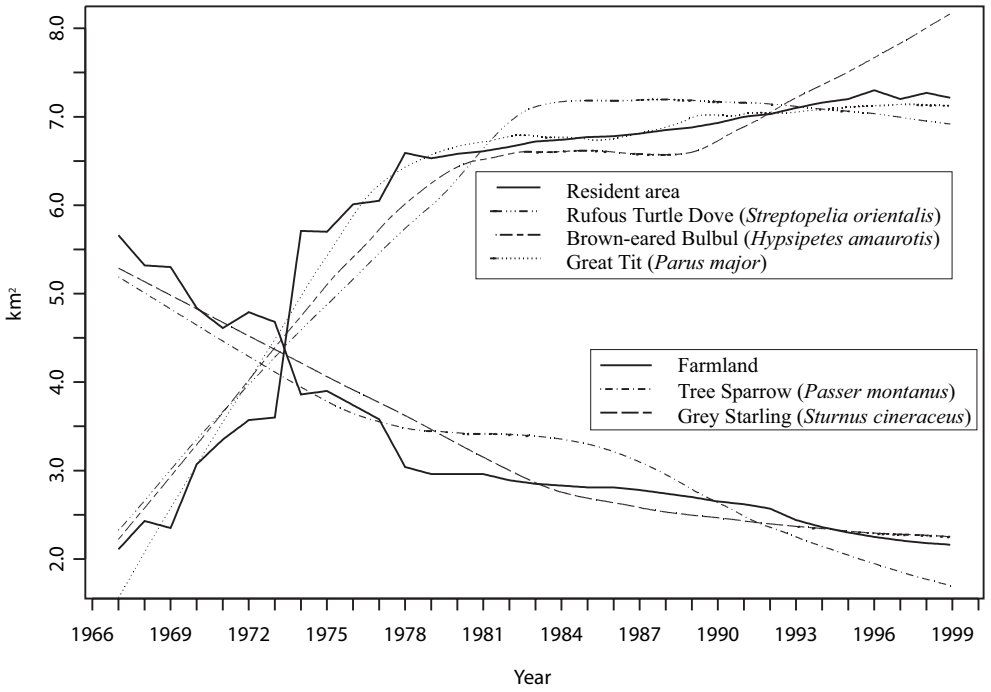

Figure 2.5: The long trends and changes of resident area and farmland in Higashi–Kurume city.

Table 2.5: Ecological characteristics of each species.

	Feeding	Nesting
Tree Sparrow	Open land	Open forest
Gray Starling	Open land	Open forest
Turtle Dove	Open forest	Forest
Brown–eared Bulbul	Forest	Forest
Great Tit	Forest	Forest

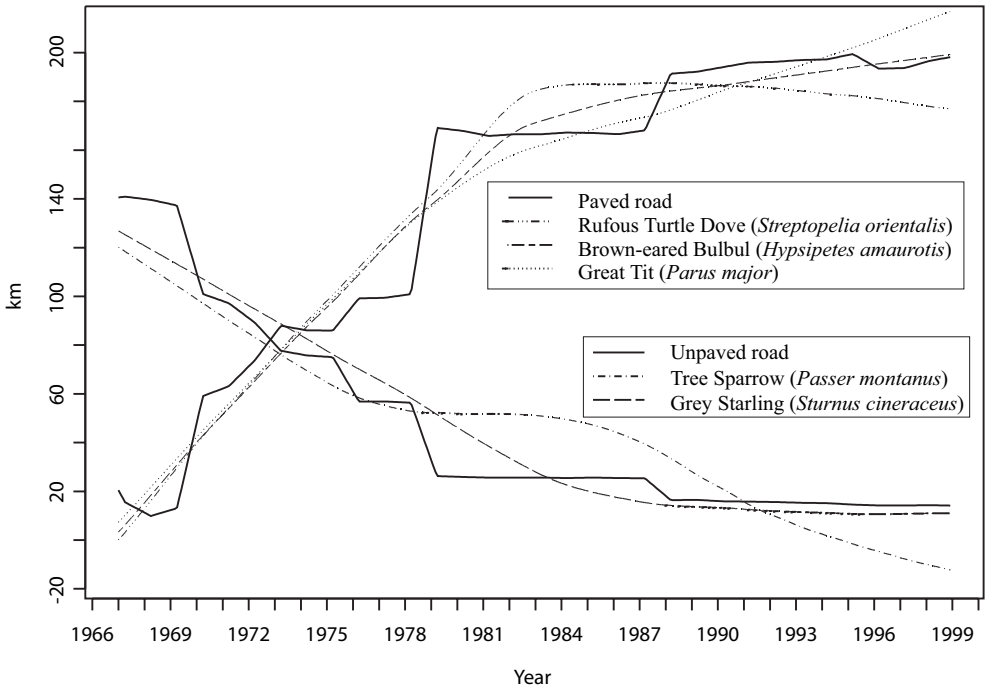

Figure 2.6: The long trends and changes of the length of paved or unpaved road in Higashi–Kurume city.

point of ecological theory, the preference of environmental conditions of each species. Such preference of environmental characteristics of each species are summarised in Table 2.5. The increase of resident area leads the increase of garden plants and shade trees which make small green island. As a consequence, Turtle Dove, Brown–eared Bulbul and Great Tit which can adopt by themselves are increase. On the other hand, Tree Sparrow and Gray Starling which feed on farmland decreased. That is, the decrease of Tree Sparrow and Gray Starling were largely depending on the condition of feeding place.

2.5.2 Short trend

On decomposition of bird count series, *middle trend* was possible to take into consideration. However, it was not significant because of its low variation in value. Therefore a short rend is derived from each original series $Z_i(t)$ by extracting long trend $L_i(t)$ like $Z_i(t) - L_i(t)$ with $p = 2$.

Figure 2.7 shows the difference due to the choice of the degree of polynomials. Estimated short trend when $p = 1$ or $p = 2$ are shown on the top and bottom panel of Figure 2.7, respectively. It is clearly shown that there is unnatural behaviour in the top panel ($p = 1$) which cannot follow the original. On the other hand, the case ($p = 2$) seems to work well.

Smoothing parameter δ was chosen as one year for Tree Sparrow, Gray Starling, Turtle Dove and Great Tit but half year only for Brown–eared Bulbul because of their wandering.

The estimated short trends are shown in Figure 2.8 and their seasonality can be found in Figure 2.9. There are three groups recognised by their behaviour. The first is Brown–eared Bulbul, the second includes Turtle Dove

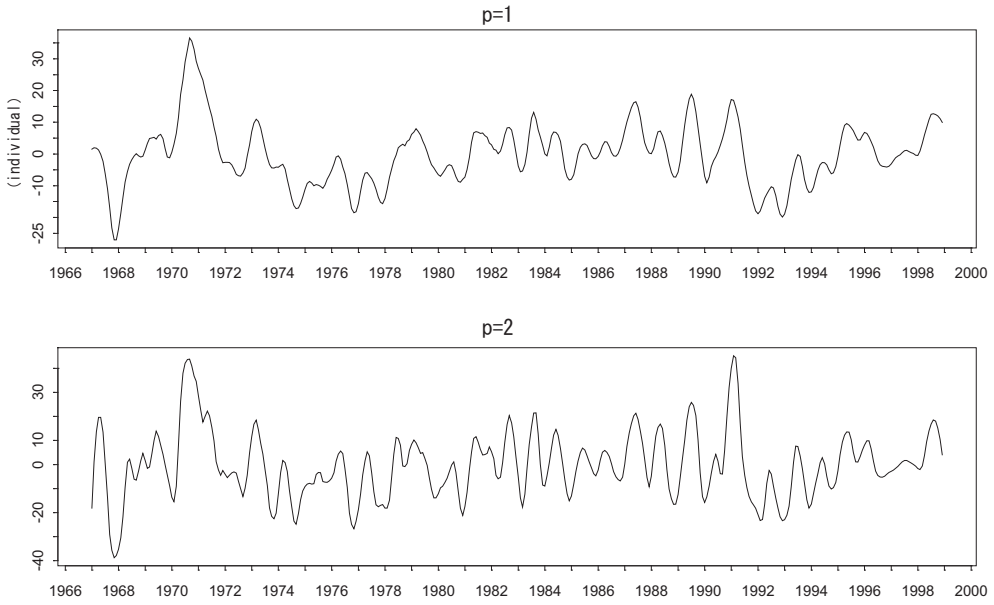

Figure 2.7: Differences in short trends due to the difference of polynomial degrees (Tree Sparrow).

and Great Tit, and the third includes Tree Sparrow and Gray Starling. Figure 2.9 shows their significant difference between these three groups which is due to their seasonality.

The short rend of Brown–eared Bulbul shows definitely different behaviour from others. This is because that Brown–eared Bulbul had been a winter wandering species but now has been resident species ever since 1973. Such phenomena has been widely known over Japan. However, the some of individuals is still wandering so two groups were found in their seasonality. As to the second group including Turtle Dove and Great Tit, it is also shown that the increase in winter especially from November to February. This is because that they are also resident species but some on them are still wandering and visiting the observation place in winter. Great Tit was also winter species in old days. These species show increase in winter but this implies that seasonal

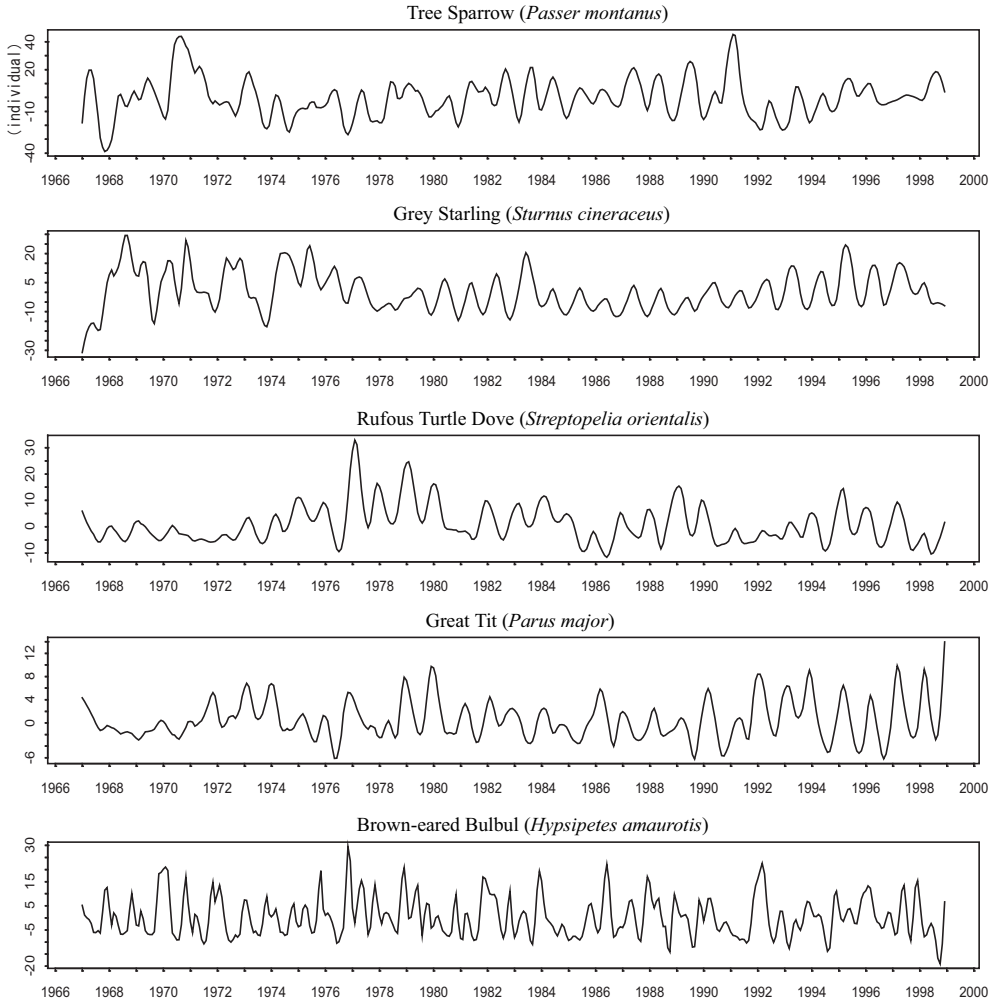

Figure 2.8: Estimated short trends of each species.

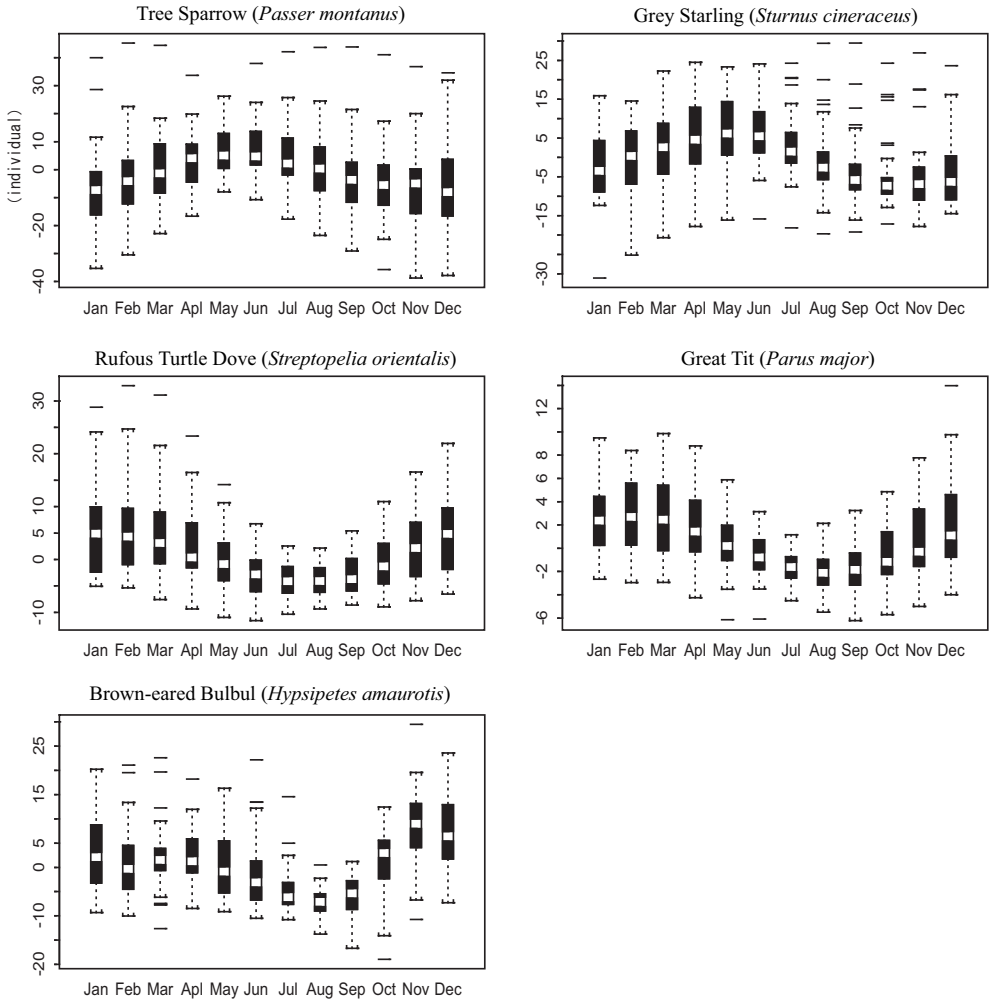

Figure 2.9: Seasonal movement of short trends.

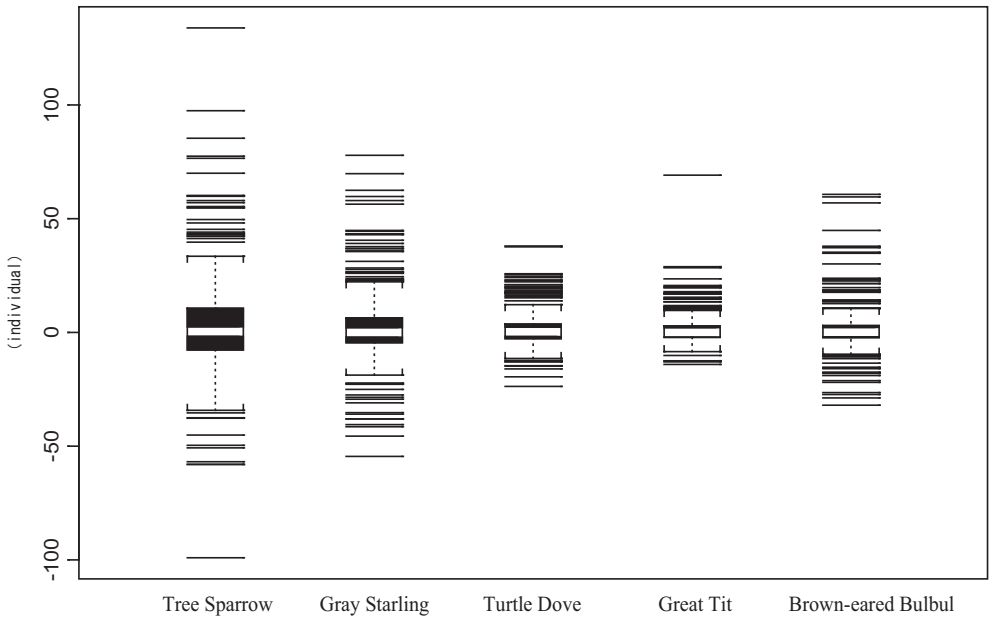

Figure 2.10: Boxplot of irregular series of each species.

effects are greater than migration effects. On the other hand, Tree Sparrow and Gray Starling show their increase in the period from April to June. This is because of migration in high possibility. After the increase, there is a decrease due to dispersion of young birds.

It is interesting to note here that the similarity between groups recognised by the long trend and short trend. This implies that Tree Sparrow and Gray Starling which show significant increase in migration are decrease in long range. On the other hand, Brown–eared Bulbul, Turtle Dove and Great Tit which show significant increase in winter wandering increase in long range.

2.5.3 Irregular series

Figure 2.10 shows the box plots of irregular series $I_i(t)$. This shows that irregulars are symmetrically distributed but having heavy tail rather than

Table 2.6: Correlation between irregulars

Tree Sparrow	Gray Starling	Brown–eared Bulbul	Turtle Dove	Great Tit
1				
0.144	1			
0.075	0.028	1		
0.212	0.130	0.236	1	
0.119	0.101	0.060	0.127	1

the normal distribution. There is no significant regular behaviour that is supported by low correlation coefficients.

This kind of aspects of irregular series are expected as being reflected the interaction between these species. This outcomes are quite natural because these species analysed here are independent in terms of the food chain theory in ecology.

These results show that the decomposition of count series through smoothing technique work quite well.

2.6 Summary

The five bird count series observed on a monthly basis from 1967 to 1998 at Jiyu–Gakuen, Higashi–Kurume city in Tokyo are simultaneously analysed. Each count series is decomposed into three components, long trend, short trend and irregular by two step loess smoothing. This decomposition explains well the relationship between the bird count and some of environmental changes. By selecting appropriate locations and scales as well as the smoothing parameters so as to minimise the residual sum of squares, it is shown that each five long trend very similarly moves with one of environmental factors. Turtle Dove, Brown–eared Bulbul and Great Tit

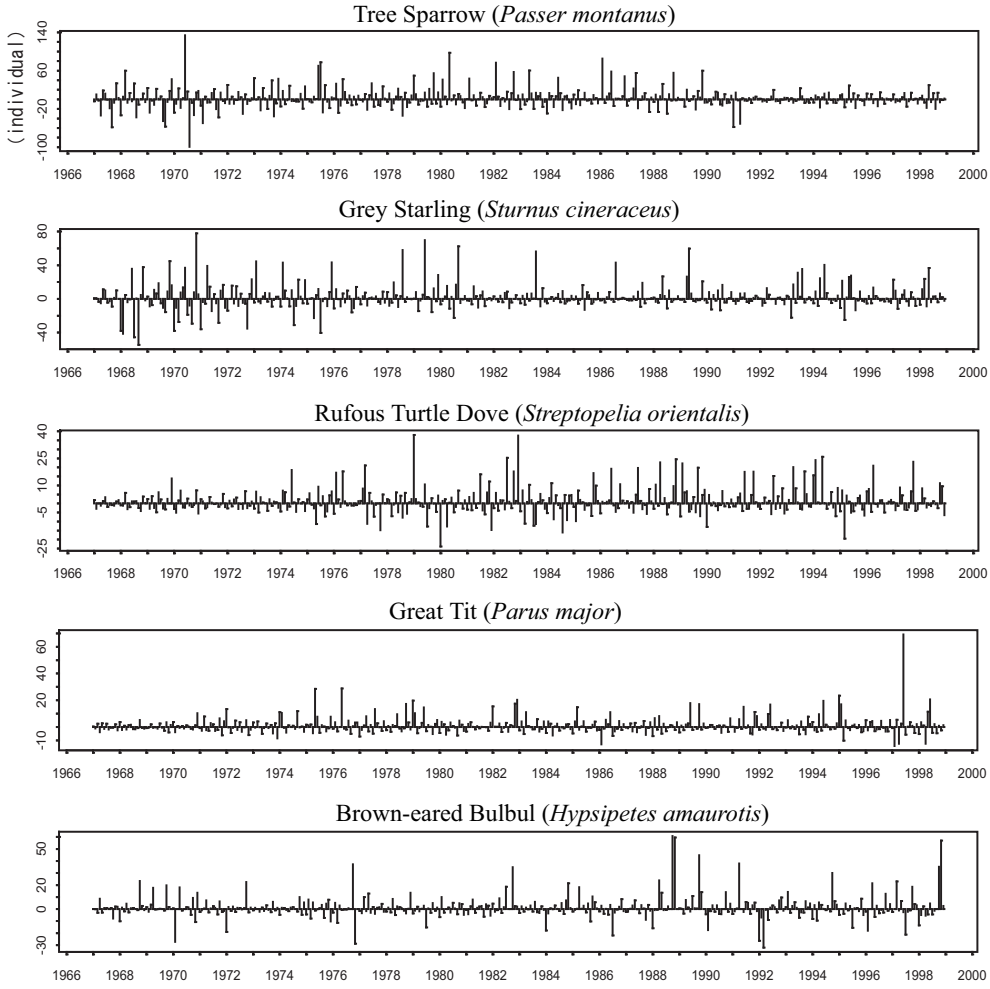

Figure 2.11: Irregular series of each species.

increased its number to link with the enlargement of resident area. Tree Sparrow and Gray Starling gradually decreased its number to link with the decrease of farmland. Such two bird groups are well described by their environment preference. Variation of each short trend can be explained by the effects of breeding season or winter wandering. The fact that each irregular series has no significant trend and very low correlation coefficients suggests a success of our decomposition.

Chapter 3

Swimmers' speeds modelling over the course of a race

3.1 Introduction

There have been many attempts to model aspects of swimming from various points of view including biomechanics, physiology and race analysis. However, there appear to have been few that model swimming speed over the course of a race. In this chapter, a new model of swimming speed and its variation over the race is proposed. This model is fitted to elapsed times at several points along the side of a pool. The model provides a good description of the strategies adopted by each swimmer over the course of the race. As a consequence, it should be of use to trainers, national selectors and those interested in the biomechanics of swimming.

The observations used here are elapsed times observed at 21 check points in the 34 preliminary male 200 m freestyle race held in the 2004 Japan Swimming Championships. A suitable dynamical model is fitted that includes a parameter describing the individual effect of each swimmer. Since a swimmer's strategy may change from location to location in a lap, each lap is split into three phases, the first, middle and last. Similarities between

phases over laps were used, although the first and last laps need special attention, leading to a more parsimonious model with reduced number of parameters. This is an important consideration here since the number of observations is limited. Section 3.3 demonstrates how to accomplish this task.

The proposed model builds directly on the deterministic models of Amar (1920), Karpovich (1933), Kolmogorov and Duplishcheva (1992) and Takagi *et al.* (1999). It extends these models by accounting for propulsion and setting them in a suitable stochastic framework. Related work from the view point of race analysis includes Arellano *et al.* (1994), Chengalur and Brown (1992), Craig and Pendergast (1979), Craig *et al.* (1985), Ikuta *et al.* (1998), Kjendlie *et al.* (2004), Matsui *et al.* (1997), Okuno *et al.* (2003) and Shimadzu *et al.* (2007). These papers focus mainly on the swimming speed in the middle phase, which is decomposed into a product of the stroke length (m/cycle) and stroke ratio (cycle/min). Relations between these two factors are mainly discussed from a largely empirical point of view. It would be natural to concentrate on such aspects if swimming in the middle phase of the race were the key to winning. However, the proposed model shows that swimming strategies in other phases are equally important for a good outcome.

The approach adopted here is to model all phases of the race to allow a better understanding of individual strategies for each phase and their impact on the race as a whole. In this way, an overall integrated strategy for improvement of swimming performance can be developed for the entire race.

3.2 Data

The 2004 Japan Swimming Championships was not only one of the major swimming competitions in Japan, but also part of the selection procedure for the Athens Olympic games. For the male 200 m freestyle race, only 34 qualified swimmers holding a record faster than 1:50.8 were invited. The race was recorded on video tapes by the Medicine and Scientific Committee of Japan Swimming Federation, with the aim of using them for scientific research. The purpose and the design of the video recording were clearly explained by the committee to team managers prior to the race and they gave their informed consent. The authors are allowed to use these video tapes from the committee with the proviso that the privacy and dignity of the swimmers should be protected.

The race was recorded on video tapes by five video cameras (60 frames per second) placed parallel to the swimming direction. To minimise perspective bias, each camera focused on just one of the intervals: 5–7.5, 10–15, 20–30, 35–40, 42.5–45 m. Based on the time stamp which was accurate to within 5 milliseconds on each frame, elapsed times were measured when a swimmer's head reached each one of 21 check points: 0, 15, 20, 30, 45, 50, 57.5, 70, 80, 95, 100, 107.5, 120, 130, 145, 150, 157.5, 170, 180, 195 and 200 m, with the exception of the ends of the pool where elapsed times were measured when a swimmer touched the wall. The second check point in the first lap was placed at 15 m instead of 7.5 m since it is hard to identify the location of each swimmer for 15 m after a dive. More details of the data collection can be found in Matsui *et al.* (1997).

3.3 Model

3.3.1 An underpinning deterministic model

A well known model for swimming speed $v(t)$ at time t is given by the differential equation,

$$\frac{dv(t)}{dt} = -\alpha v(t)^2,$$

which was proposed by Amar (1920). There have been several experiments to estimate the value of the drag parameter $\alpha > 0$ in relation to the body characteristics of each swimmer (Karpovich, 1933, Kolmogorov and Duplishcheva, 1992, Takagi *et al.*, 1999). However, this model ignores the effect of propulsion which needs to be taken into account in swimming. A more general model is

$$\frac{dv(t)}{dt} = -\alpha v(t)^2 + \beta, \tag{3.1}$$

where $\beta \geq 0$ measures the propulsion generated by the swimmer's stroke action. The solution of the differential equation (3.1) can be explicitly written as

$$v(t) = \begin{cases} \dfrac{1}{\alpha t + \frac{1}{v_0}} & (\beta = 0) \\ \dfrac{2\sqrt{\kappa}}{1 - c_1 e^{-2\alpha\sqrt{\kappa}t}} - \sqrt{\kappa} & (\beta > 0) \end{cases} \qquad (t \geq 0),$$

where v_0 is the initial speed, $\kappa = \beta/\alpha$ and $c_1 = (v_0 - \sqrt{\kappa}) / (v_0 + \sqrt{\kappa})$. The model is continuous in terms of β so that

$$\begin{aligned} \lim_{\beta \to +0} v(t) &= \lim_{\delta \to +0} \frac{2\delta}{1 - c_1(\delta) e^{-2\alpha t\delta}} - \delta \\ &= \lim_{\delta \to +0} \frac{2e^{2\alpha t\delta}}{2\alpha t c_1(\delta) - dc_1(\delta)/d\delta} \\ &= \frac{1}{\alpha t + \frac{1}{v_0}}, \end{aligned}$$

where $\delta = \sqrt{\kappa}$ and $dc_1(\delta)/d\delta = -2v_0/{v_0}^2$.

It is better to write the speed as a function of distance x rather than time t since our observed elapsed times are measured in terms of distance. To derive $v(x)$ from $v(t)$, it is enough to consider the case when $\beta > 0$,

$$\begin{aligned} x(t) &= x_0 + \int_0^t v(s)\ ds \\ &= x_0 + 2\sqrt{\kappa}\int_0^t \left(\frac{1}{1 - c_1 e^{-2\alpha\sqrt{\kappa}s}} - \frac{1}{2}\right)\ ds \\ &= x_0 + \sqrt{\kappa}t + \frac{1}{\alpha}\log\left(\frac{1 - c_1 e^{-2\alpha\sqrt{\kappa}t}}{1 - c_1}\right) \\ &= x_0 + \frac{1}{2\alpha}\log\left(c_1\frac{v(t) + \sqrt{\kappa}}{v(t) - \sqrt{\kappa}}\right) + \frac{1}{\alpha}\log\left(\frac{2\sqrt{\kappa}}{v(t) + \sqrt{\kappa}}\frac{1}{1 - c_1}\right) \\ &= x_0 + \frac{1}{2\alpha}\log\left(\frac{4c_1\kappa}{v(t)^2 - \kappa}\frac{1}{(1 - c_1)^2}\right), \end{aligned}$$

where the formula used here is given by

$$t = \frac{1}{2\alpha\sqrt{\kappa}}\log\left(c_1\frac{v(t) + \sqrt{\kappa}}{v(t) - \sqrt{\kappa}}\right).$$

This yields

$$v(x)^2 = \frac{4c_1\kappa}{(1 - c_1)^2}e^{-2\alpha(x - x_0)} + \kappa,$$

and the swimming speed $v(x)$ at distance x is given by

$$v(x) = \begin{cases} v_0 e^{-\alpha(x - x_0)} & (\beta = 0) \\ \sqrt{c_2 e^{-2\alpha(x - x_0)} + \kappa} & (\beta > 0) \end{cases} \qquad (x \geq x_0), \tag{3.2}$$

provided that the speed at the initial distance x_0 is v_0 and $c_2 = {v_0}^2 - \kappa$. However it is not appropriate to apply this model to the whole race directly since the male 200m freestyle race consists of four laps of the pool, each of length 50m. It is clear that α or β will not stay constant over the race, nor

even in a lap so a natural approach is to split each lap into several phases within which these parameters might be expected to be constant.

For simplicity, three phases are introduced for each lap. The first phase (from 0m to x_1m) is just after a dive or turn where drag, but no stroke propulsion, are expected ($\alpha > 0, \beta = 0$). By contrast, drag and propulsion are both expected ($\alpha > 0$, $\beta > 0$) in the middle phase (from x_1 m to x_2 m) and in the last phase (from x_2 m to 50 m). It is also natural to assume that a swimmer's speed stays constant in the middle phase since every swimmer should have reached an equilibrium swimming state in this phase, so that $v(x_1) = v(x) = v(x_2)$ for $x_1 \leq x \leq x_2$, that is, $\kappa = {v_0}^2$. Such an equilibrium no longer holds true in the last phase where a swimmer should have prepared for a turn or the end of the race. Also, note that the drag parameter in the first phase can be different from that in other phases because of the dive or turn in the first phase.

Combining these assumptions, a model for swimming speed of a lap which consists of three phases is then

$$v(x;\boldsymbol{\theta}) = \begin{cases} v_0 e^{-\alpha_0 x} & (0 \leq x < x_1), \quad \text{(First phase)} \\ v(x_1) & (x_1 \leq x < x_2), \quad \text{(Middle phase)} \\ \sqrt{c_2 e^{-2\alpha(x-x_2)} + \kappa} & (x_2 \leq x < 50), \quad \text{(Last phase)} \end{cases}$$

where $\boldsymbol{\theta} = (v_0, \alpha_0, x_1, x_2, \alpha, \beta)$, $c_2 = v(x_1)^2 - \kappa$ and $\kappa = \beta/\alpha$. Note that the break points x_1 and x_2 are also parameters which can differ from lap to lap. Figure 3.1 shows a stylised picture of the swimming speed $v(x)$. Then the swimming speed over the whole race is given as

$$v_j(x) = v(x - 50(j-1); \boldsymbol{\theta}_j),$$

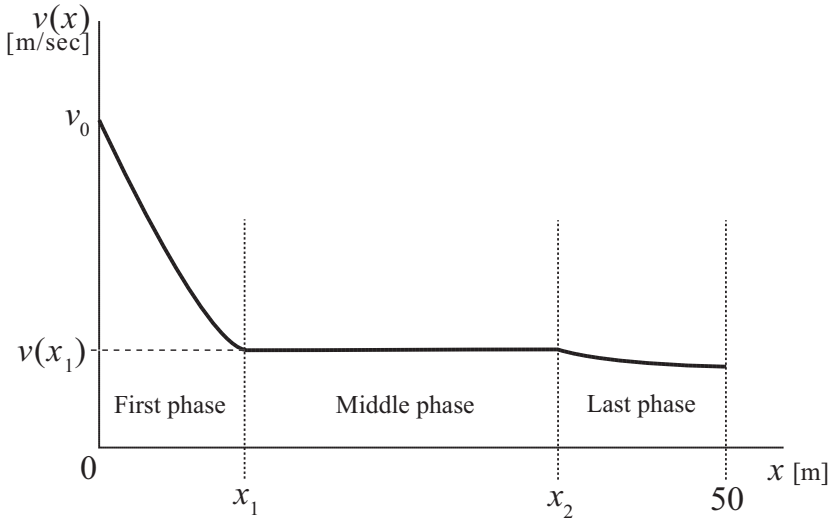

Figure 3.1: A stylised picture of swimming speed $v(x)$ over one lap of the race. The lap is divided into 3 phases, a first phase just after a dive or turn, a middle phase and a last phase just before a turn or the finish of the race.

for $50(j-1) \leq x < 50j$, $j = 1, 2, 3, 4$, where $\boldsymbol{\theta}_j = (v_{0j}, \alpha_{0j}, x_{1j}, x_{2j}, \alpha_j, \beta_j)$ is the vector of parameters for lap j. Therefore the set of parameters $\{\boldsymbol{\theta}_j; j = 1, 2, 3, 4\}$ determines a swimming speed model over the race. The estimation of such unknown parameters will be discussed in Section 3.3.2.

An important aspect of the model is the specification of an individual effect for each swimmer. We model the swimming speed of swimmer i in lap j as

$$\mu_i v_j(x), \ j = 1, 2, 3, 4,$$

where μ_i is a multiplicative factor, specific to the individual swimmer, that is assumed to be constant over the race, and $v_j(x)$ is the common swimming speed of swimmers in lap j. This multiplicative model allows for a simple understanding of a swimmer's performance relative to the common swimming speed $v_j(x)$. In particular, the values of the multiplicative factors μ_i provide

an overall measure of swimming performance that can be used to discriminate between swimmers.

3.3.2 A stochastic model for swimming speed

The observed elapsed times are not free from random fluctuations due to the swimmers as well as random errors in the observational process. If $T_{ij}(k)$ denotes the elapsed time of swimmer i at distance $x_j(k)$, where k denotes a check point in lap j, it is assumed that

$$T_{ij}(k) = \int_0^{x_j(k)} \frac{1}{\mu_i v_j(x)}\, dx + \sigma B_i\left(x_j(k)\right), \tag{3.3}$$

where $\{B_i(x); 0 \leq x < 200\}$ is standard Brownian motion representing accumulated error up to distance $x_j(k)$. Brownian motion is a continuous time process, which is widely used in various disciplines. Its basic property is that any increment $B_i(x+\Delta x) - B_i(x)$ is normal with mean zero and variance Δx and distributed independently of any other non-overlapping increment.

Thus

$$\Delta T_{ij}(k) = \frac{1}{\mu_i}\int_{x_j(k-1)}^{x_j(k)} \frac{1}{v_j(x)}\, dx + \sigma\sqrt{\Delta x_j(k)}\,\varepsilon_{ijk},$$

where $\Delta T_{ij}(k) = T_{ij}(k) - T_{ij}(k-1)$, $\Delta x_j(k) = x_j(k) - x_j(k-1)$ and the ε_{ijk} are independent standard normal random variables. The parameters of the model can now be estimated by weighted least squares

$$\sum_{i=1}^{34}\sum_{j=1}^{4}\sum_{k=1}^{5} \frac{r_{ijk}^2}{\Delta x_j(k)},$$

where

$$r_{ijk} = \Delta T_{ij}(k) - \frac{1}{\mu_i}\int_{x_j(k-1)}^{x_j(k)} \frac{1}{v_j(x)}\, dx.$$

The squared residual r_{ijk}^2 is divided by $\Delta x_j\,(k)$ because the residuals $\{r_{ijk}\}$ are expected to be independently distributed normal random variables with mean zero and variance $\sigma^2 \Delta x_j\,(k)$. The normality will be checked in Section 3.4.3.

To keep the model parsimonious, the number of parameters $\{\boldsymbol{\theta}_j; j = 1, 2, 3, 4\}$ are now reduced by assuming similarities between phases over laps. The model for the first phase is assumed to be common over laps other than the first phase in the first lap since this starts from a dive. It is also assumed that the parameters α_{0j} and x_{1j} are common over the laps other than the first ($\alpha_{02} = \alpha_{03} = \alpha_{04}$ and $x_{12} = x_{13} = x_{14}$), and that the x_{2j} are common over laps other than the last ($x_{21} = x_{22} = x_{23}$). The break point in the last lap x_{24} is different since all swimmers are focused on completing the race rather than making a turn. Furthermore, it is assumed that the drag parameter α_j for the last phase in each lap is known. For stable estimates, it is adopted that $\alpha_j = 0.428$ or $\alpha_j = 0.37$ for any lap j, that are the constants known from the results of an experiment by Toussaint *et al.* (1988) and Karpovich (1933) for a 70 kg swimmer. It will be seen that the choice of either of these values does not lead to any significant difference in the final results. These considerations reduce the total number of parameters to be estimated to 48 for the whole race.

Fortunately, it is possible to apply the above parameter estimation procedure without any numerical integration. It is enough to prove (3.4)

only when $\beta > 0$. Letting $v_0 = v(0)$, we have

$$\begin{aligned}\int_0^x \frac{1}{v(u)}\, du &= \int_{v_0}^{v(x)} \frac{1}{v}\,\frac{du}{dv}\, dv \\ &= -\int_{v_0}^{v(x)} \frac{1}{\alpha v^2 - \beta}\, dv \\ &= \frac{1}{2\alpha\sqrt{\kappa}}\left\{\log\left(\frac{v(x)+\sqrt{\kappa}}{v(x)-\sqrt{\kappa}}\right) - \log\left(\frac{v_0+\sqrt{\kappa}}{v_0-\sqrt{\kappa}}\right)\right\}.\end{aligned}$$

The integration of $1/v(x)$ for the swimming speed given in (3.2) is explicitly written as

$$\int_0^x \frac{1}{v(u)}\, du = \begin{cases} \dfrac{1}{\alpha}\left(\dfrac{1}{v(x)} - \dfrac{1}{v_0}\right) & (\beta = 0), \\ \dfrac{1}{2\alpha\sqrt{\kappa}}\left\{\log\left(\dfrac{v(x)+\sqrt{\kappa}}{v(x)-\sqrt{\kappa}}\right) - \log\left(\dfrac{v_0+\sqrt{\kappa}}{v_0-\sqrt{\kappa}}\right)\right\} & (\beta > 0). \end{cases} \tag{3.4}$$

To estimate the parameters, a program for solving nonlinear least squares, `nlminb` in S–PLUS (Chambers and Hastie, 1992), was employed.

3.4 Result

3.4.1 Common swimming speed and its parameters

The estimated parameters assuming $\alpha_j = 0.428$ and $\alpha_j = 0.37$ for the last phase of each lap are listed in Table 1 and Table 2 respectively. Note that the choice has little effect on the parameter estimation, as expected.

The estimated parameters in Table 3.1 and Table 3.2 are largely consistent with the experience of swimmers and their trainers. They also provide a good description of the characteristics of swimming in a race. Diving not only affects the initial speed $\hat{v}_{0j}$ but also the drag parameter $\hat{\alpha}_j$ and the location parameter $\hat{x}_{1j}$. As expected, $\hat{v}_{01}$ and $\hat{x}_{11}$ in the first lap are higher than the

Table 3.1: Estimated parameters ($\alpha_j = 0.428$).

Lap	$\hat{v}_{0j}$	$\hat{\alpha}_{0j}$	$\hat{x}_{1j}$	$\hat{x}_{2j}$	$\hat{\beta}_j$
$j = 1$	4.11	0.09	9.32		1.17
$j = 2$	3.06			36.74	1.09
$j = 3$	3.00	0.08	7.05		1.04
$j = 4$	2.93			45.00	1.42

Table 3.2: Estimated parameters ($\alpha_j = 0.37$).

Lap	$\hat{v}_{0j}$	$\hat{\alpha}_{0j}$	$\hat{x}_{1j}$	$\hat{x}_{2j}$	$\hat{\beta}_j$
$j = 1$	4.10	0.09	9.32		1.16
$j = 2$	3.06			36.55	1.09
$j = 3$	3.00	0.08	7.05		1.04
$j = 4$	2.92			45.00	1.43

values in other laps. It is also possible to see how swimmers exhaust their energy as the race progresses with the initial speed $\hat{v}_{0j}$ in each lap decreasing by approximately 0.07 m/s per lap. A reason why the drag parameter $\hat{\alpha}_{01}$ in the first lap is higher than other laps could be due to the impact of diving.

The effect of the finish line is also apparent with the values of $\hat{x}_{24}$ and $\hat{\beta}_4$ in the last lap being higher than those in the other laps. Since no turn is necessary at the end of the last lap, each swimmer makes their break for the finish line over the last phase of the race.

Figure 3.2 illustrates the estimated common swimming speed. The speeds in the middle phase in each lap are 1.83, 1.73 1.67 and 1.65 m/sec respectively. These values are consistent with the values reported by Matsui *et al.* (1997) and Ikuta *et al.* (1998). Any unnatural behaviour of the common

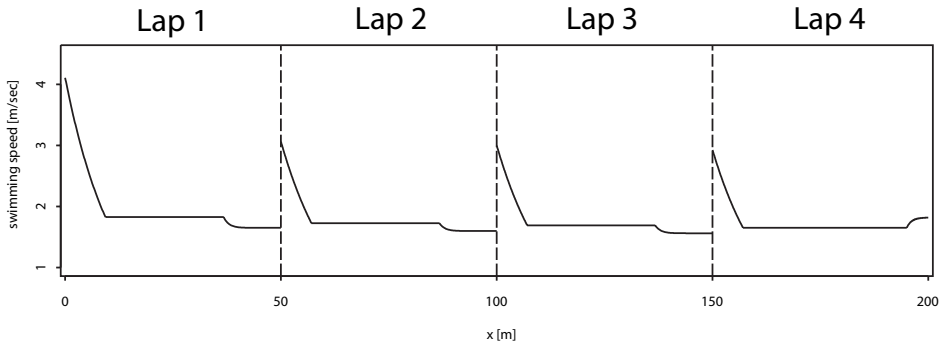

Figure 3.2: Estimated common swimming speed $\hat{v}_j(x)$ over the course of the race as a function of distance x and lap j ($j = 1, 2, 3, 4$).

swimming speed, especially around the break points, is most likely due to the assumptions made for the parsimonious parameterisation discussed in the previous section. Such behaviour could be improved if more check points were set and more data collected, particularly around phase boundaries.

3.4.2 Individual parameters

Individual effects are measured by parameter $\mu_i, i = 1, 2, \ldots, 34$. Figure 3.3 shows that, as expected, the estimated μ_i are strongly and inversely related to the final time taken to complete the race. The reason why the $\hat{\mu}_i$ are not exactly placed on the theoretical line is not only because of estimation and measurement error, but also because of the random fluctuations of effort by each swimmer in the race. The point in the top left corner of the plot corresponds to the winner of the race. The isolation of this point from the others suggests that the winner is significantly faster than the others, with his individual factor being more than 3% faster than the averaged swimming speed. By contrast, the point in the bottom right of the plot corresponds to the slowest swimmer whose factor was about 2% slower. As mentioned

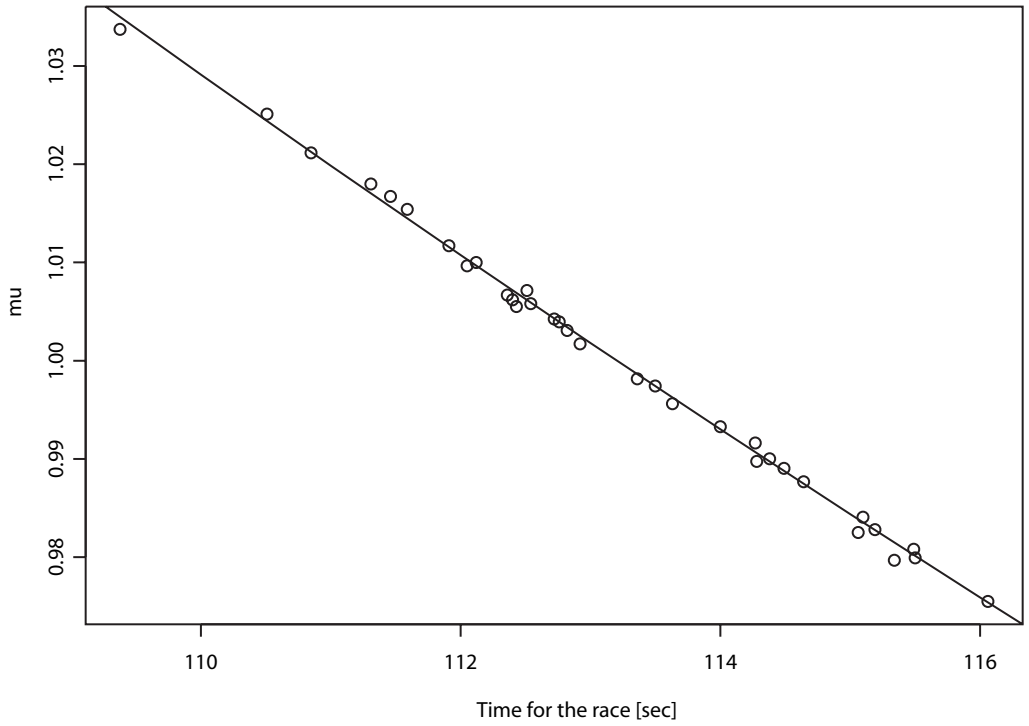

Figure 3.3: Individual fitted parameters $\hat{\mu}_i$ plotted as a function of race times together with the theoretical relationship.

previously, these factors are important as they discriminate between the swimmers.

3.4.3 Discussion

Figure 3.4 plots the standardised residuals,

$$\left\{ \hat{\varepsilon}_{ijk} = \frac{\hat{r}_{ijk}}{\hat{\sigma}\sqrt{\Delta x_j\left(k\right)}}; i = 1, 2, \ldots, 34 \right\}$$

for every lap j and check point k. The dashed horizontal lines placed at ± 3.03 indicate the 95% confidence bound for the standardised residuals of each swimmer. The bound $b = 3.03$ is calculated so that

$$\mathrm{P}\left(|E_{jk}| < b, j = 1, 2, 3, 4, k = 1, 2, \ldots, 5\right) = 0.95,$$

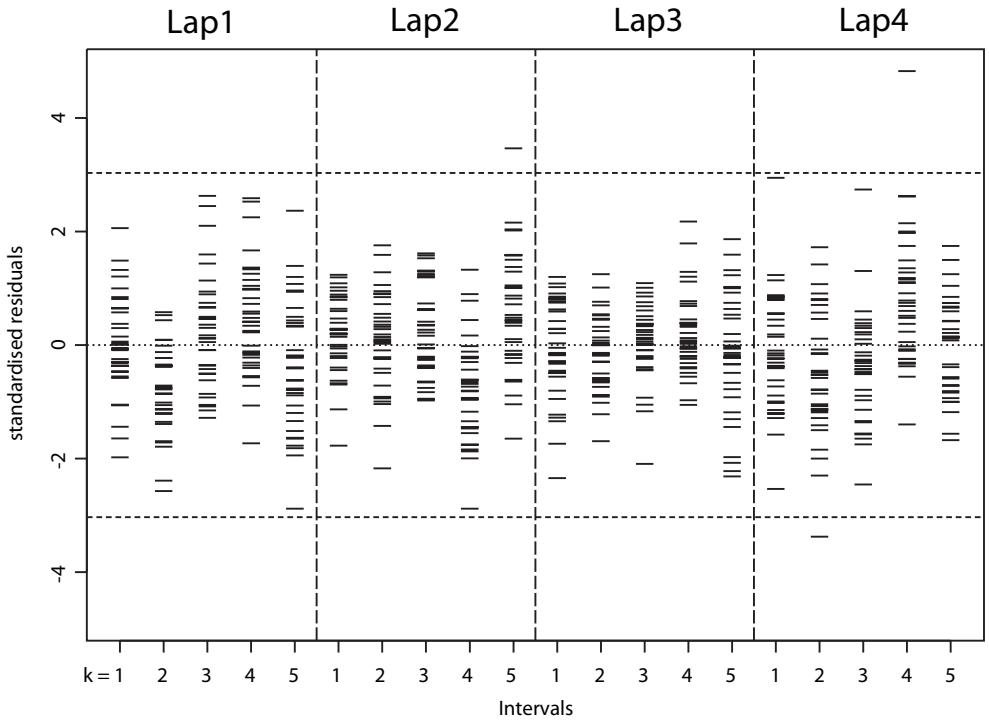

Figure 3.4: Standardised residuals with 95% confidence bounds.

where the $\{E_{jk}\}$ are independent standard normal random variables. In fact, b is the solution of $1-(1-p)^{20}=0.05/2$ where $b=\Phi^{-1}(1-p)$ and $\Phi(\cdot)$ is the standard normal distribution function.

Three standardised residuals lie outside the 95% confidence bounds. These are for swimmers ranked 13, 28 and 31 whose residual plots are shown in Figure 3.5. Residual plots such as these should be of use to swimmers and their trainers to evaluate their performance and the strategy they have adopted in a race. For example, the plot of the swimmer ranked 13 suggests that his rank would improve if he swam faster in the first and last lap. The swimmer ranked 28 has residuals that take high values before making a turn which implies a need to improve his turn technique. It is clear that

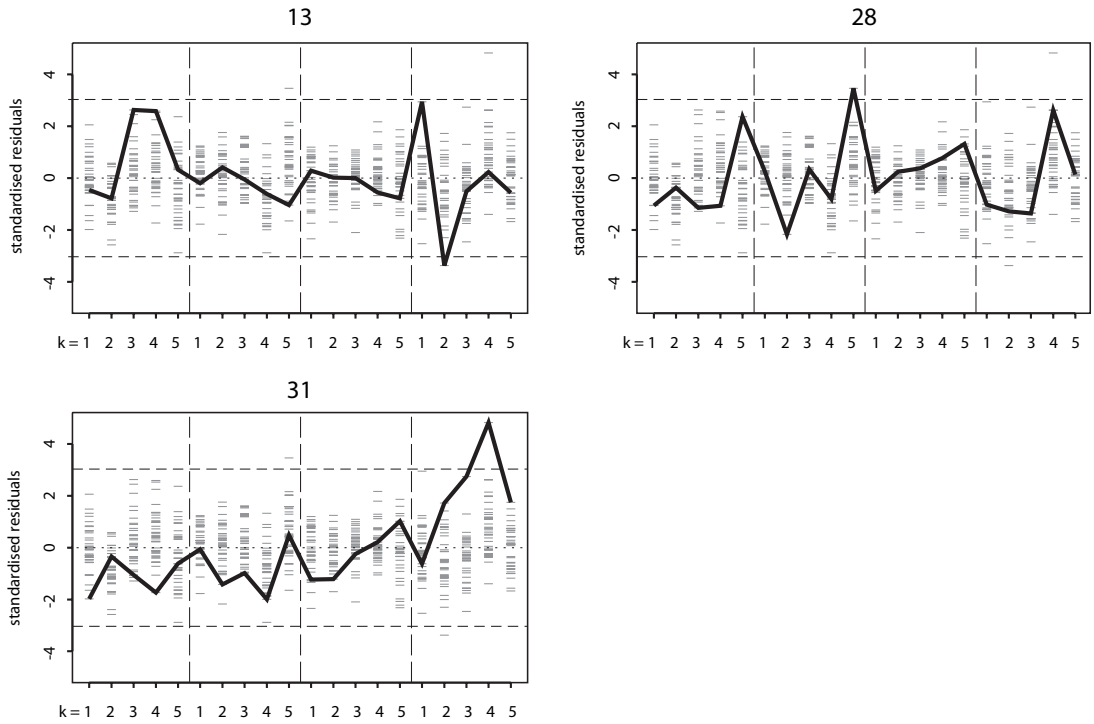

Figure 3.5: Standardised residuals for 3 outlying swimmers.

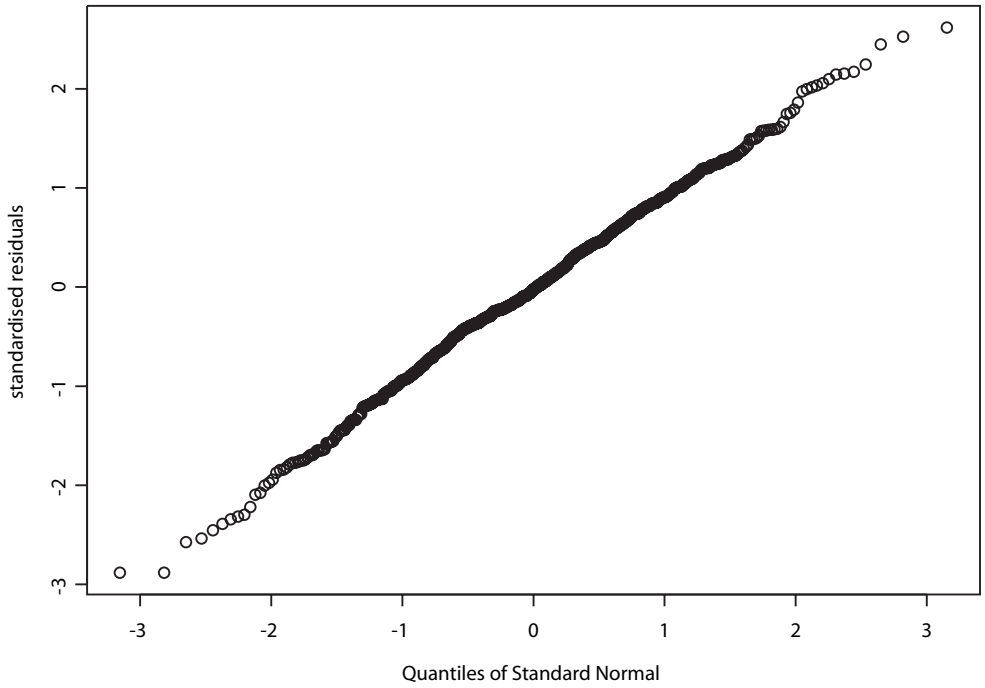

Figure 3.6: Q–Q plot of standardised residuals.

the swimmer ranked 31 started well, but exhausted his energy before the finish and failed to keep up with the other swimmers in the last lap. No doubt there are other factors that affect swimming performance (e.g. health, fitness, strategy etc) and these could be accounted for, but are left for future investigation.

With the exception of the three outlying swimmers mentioned above, the normal Q–Q plot (normal quantile–quantile plot; Chambers *et al.* (1983)) of the standardised residuals $\hat{\varepsilon}_{ijk}$ was highly linear which supports the normality of the ε_{ijk} (Figure 3.6). This indicates that the parsimonious model adopted is a good fit to the data.

3.5 Summary

A stochastic model of swimming speed over the entire course of a race has been developed. It builds on a deterministic physical model that reflects the trade–off between drag and propulsion in swimming. The model has been simplified to cope with the limited number of observations, noting similarities and dissimilarities between the four laps of the race where each lap is divided into three separate phases. The elapsed times that are observed are modelled as a function of a deterministic function of distance swum, lap of the race and phase of the lap, together with accumulated stochastic error which is modelled using Brownian motion.

The model fits the data well, is easy to understand and interpret, and also provides a good description of the swimming strategies of each swimmer from phase to phase in the race and over the race as a whole. An individual factor measuring how much faster or slower an individual swimmer performs relative to the average swimming speed of the race is simultaneously estimated in the course of fitting the model. This factor is, as expected, closely related to the final outcome of the race.

The model can be used to analyse and quantitatively evaluate the performance of individual swimmers. As a consequence, it should be of use to trainers and national selectors to improve individual swimming performances and to identify a swimmer's future potential. The model is also intended to be of interest to engineers and scientists concerned with the biomechanics of swimming and we hope that it will lead to a number of further developments.

Chapter 4

Membrane potential modelling led by an *in vivo* measurement of a single neuron

4.1 Introduction

This chapter concerns with modelling membrane potential measured from a single neuron. There have been many attempts in modelling membrane potential of neurons. A well known pioneering work is the Hodgkin–Huxley model (Hodgkin and Huxley, 1952) that is now still leading researches in neuroscience. For example, Rose and Hindmarsh (1989), Rinzel (1990), Wilson (1999) are inspired by the Hodgkin–Huxley model. However, there appears to have been few attempts to model membrane potential by *in vivo* measurement of a single neuron. In this chapter, a simple input–output system is proposed for a single neuron. It is incorporating the existence of different types of synapses that electrical and chemical, as a key to modelling. The input through electrical synapse is directly transfered to the membrane but that through chemical one is delayed and modified within the process. It is regarded that a slowly varying part of the membrane potential would be

the input through the electrical synapse and introduced a three phase model for each spike due to the input through chemical synapse. The three phase model is quite general so that it can be used as a model for the input signal. Occurrence time of each spike is modelled by an inhomogeneous Poisson process with the intensity which is proportional to a positive part of the derivative of the input. Every part of modelling is inspired by investigating every detail of the observed data and the validity is checked again with the data. This approach is time exhausting and laborious but a rewarding way of modelling the data.

4.2 Data

4.2.1 Data collection

Earthworms (*Eisenia fetida*) from the commercial supplier (Verdex Co., Ltd., Kitakyushu, Japan) were maintained in a box filled with moist soil at 4 days prior to experiments. It is used only mature earthworms whose clitellums were clearly visible and body weights were more than 300 mg. The earthworm was anesthetised in 10% ethanol for 10 min or chilled earthworm saline (125.5 mM NaCl, 10.0 mM $NaHCO_3$, 2.5 mM KCl, 2.0 mM $CaCl_2$, 1.0 mM $MgCl_2$ and 10.0 mM Tris–Buffer) for 3 min, and dissected for the isolation of segmental ganglia of the ventral nerve cord following clitellum.

Intra–cellular recording was made using a sharp glass micro–electrode filled with 100 mM potassium acetate (30–80 MΩ). Time series of intra–cellular potential were acquired with an amplifier (MEZ–8300, Nihon Koden, Tokyo, Japan), and digitised using Power Lab/8SP with Chart 5.2 (AD Instruments, Colorado Springs, CO) over 40 sec with 0.05 msec time resolution.

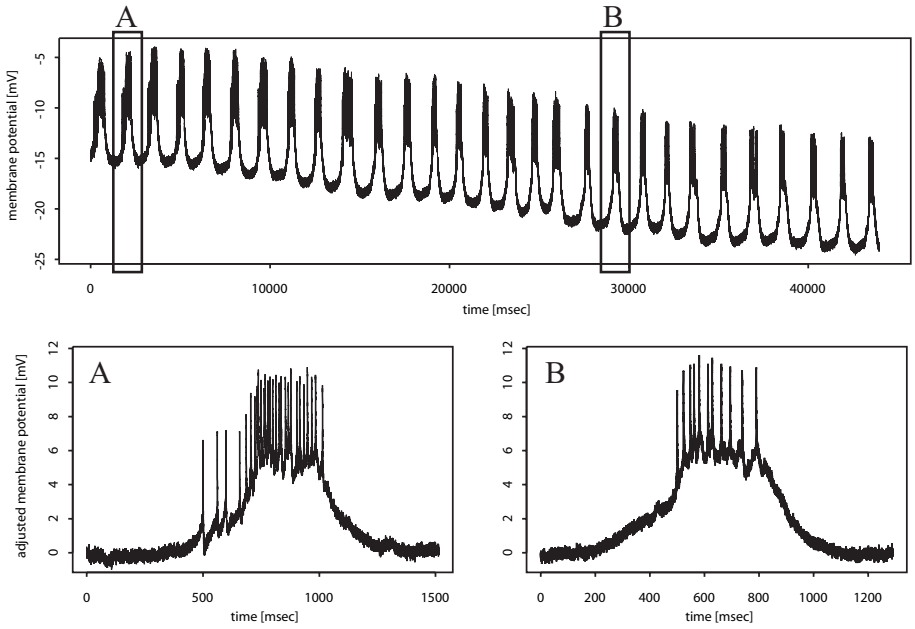

Figure 4.1: Observed time series and examples of zoomed clusters.

4.2.2 Data exploration

The whole observed time series of membrane potential is shown on the top panel of Figure 4.1. It is obvious that the series is not only oscillating but also gradually decaying. The reason for the decay can be thought in various ways, for example, the electrode used slowly slipped off or some environment factors are gradually changed. To adjust such decay, the whole original series is split into 27 clusters of which base level is adjusted as zero since the aim of this research is modelling the behaviour of the membrane potential in each clusters rather than such a global movement over the clusters. The beginning of each cluster is set at 500 msec ahead of the first spike and the end is 500 msec behind of the last spike, and applied a linear transform so that the levels at the beginning and end of each cluster to be zero. There are two examples A and B of such an adjusted clusters are shown on the bottom

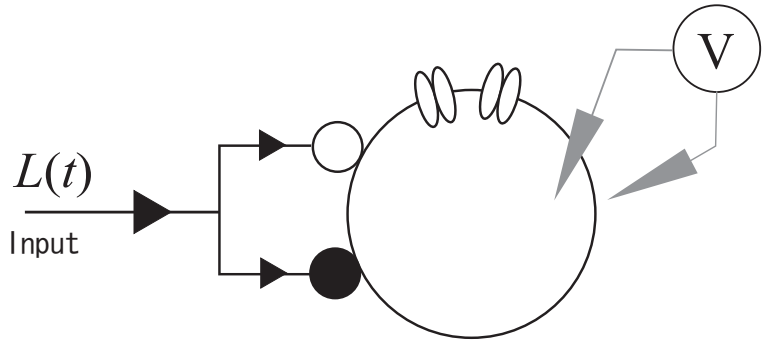

Figure 4.2: A simple input–output system for cell membrane potential.

panels of Figure 4.1. The adjusted membrane potential $V(t)$ in each cluster is modelled. Typically the time t runs from 0 to 1200 or 1500 msec in a cluster at the longest.

4.3 Model

As is seen from two examples of the clusters in Figure 4.1, there are at least two major changes on membrane potential; gradual change (trend) and instantaneous change (spike). There would be various reasons why such two types of potential changes exist. One of possibilities is that the trend directly reflects the input to the neuron and the spikes are caused by a side effect of the changes of the input. Such an observation reminds that there are two types of synaptic transmissions. One is electrical one and another is chemical one. The electrical synapse transmits the input signal without any delay or modification but the chemical synapse transmits it with a delay and modifications. The signal arriving at the chemical synapse stimulates its membrane and releases some chemical transmitters from the pre–synaptic cell into the synaptic cleft. As the transmitters diffuse over the gap the receptors of the post–synaptic cell starts an activation, a rapid increase of

the membrane potential. If the increase reaches a threshold, it results in the opening of an ion channel, for example, sodium channel. Then the opening of another channel, for example, potassium channel, follow the closing of that channel as is described by the Hodgkin–Huxley model. Figure 4.2 shows a schematic picture. The input $L(t)$ goes to the both synapses, chemical one (white circle) and electric one (filled circle). Instantaneous changes of membrane potential through chemical synapse invokes the opening and the closing of the ion channels shown on the top of big circle, the post–nerve cell.

Therefore, it would be convenient introducing three phases for the time period of each spike as is seen in Figure 4.1, post–synaptic cell activation, sodium channel opening and the closing and opening of potassium channel. It is illustrated in Figure 4.3 and will be modelled more precisely. More formally, let $s(t)$ be a function of a single spike due to the ith activation of post–synaptic cell starting at time T_j. Thus the accumulated potential changes by chemical synapse is given by

$$S(t) = \sum_{j=1}^{N} s(t - T_j),$$

where N is the number of spikes occurred within a cluster. It is worthy of note that chemical transmitting process can cause an instantaneous potential change so that the activation of post-synaptic cell is also instantaneous.

As a consequence, the membrane potential we have observed is decomposed into $V(t) = L(t) + S(t) + \xi(t)$, where $\xi(t)$ is a noise.

4.3.1 Model for spikes

The Hodgkin–Huxley model (Hodgkin and Huxley, 1952)

$$C\frac{dV_m(t)}{dt} = I(t) - g(t, V_m)(V_m(t) - v), \tag{4.1}$$

has been widely used as a model for the action potential of a membrane, where C and $g\,(t, V_m)$ are the capacitance and variable conductance of the membrane, respectively. The v is an equilibrium voltage and $I\,(t)$ is the input current. A simplified model is obtained by assuming that the conductance g is constant and there is no input current $I(t) = 0$. The solution of the differential equation (4.1) is then

$$V_m\,(t) = \alpha e^{\beta(t-t_0)} + v \qquad (t \geq t_0)\,, \tag{4.2}$$

where $\alpha = V_m\,(t_0) - v$, $\beta = -g/C < 0$. The voltage $V_m(t)$ exponentially decays to v if the initial voltage $V_m(t_0)$ is less than v and it grows to v otherwise. Roughly speaking, the Hodgkin-Huxley model works for describing the firing phenomena of a neuron by combining those two cases. Applying the model (4.2) to each case, a two phase model for a spike is given by

$$s(t) = \begin{cases} \alpha_1 e^{\beta_1(t-\tau_1)} + v_1 \quad (\tau_1 \leq t < \tau_2)\,, & \text{(Phase I)} \\ \\ \alpha_2 e^{\beta_2(t-\tau_2)} + v_2 \quad (\tau_2 \leq t)\,. & \text{(Phase II)} \end{cases} \tag{4.3}$$

Here $\beta_1 = -g_1/C$, $\beta_2 = -g_2/C$, $\alpha_1 < 0$ and $\alpha_2 > 0$.

An opening of a channel at time $t = \tau_1$ with $V_m(\tau_1) < v_1$ causes an increase of the membrane potential and results in the closing of the channel and invoking an opening of another channel before reaching the equilibrium. In this stage, at time $t = \tau_2$, the equilibrium voltage changes from v_1 to v_2 since the equilibrium voltage v is linked to the conductance g through the formula $v = gV_0/g_0$, where g_0 is the global conductance and V_0 is the battery voltage in the equivalent electric circuit model. The newly opened channel now decays the potential $V_m(\tau_2) > v_2$ toward to v_2.

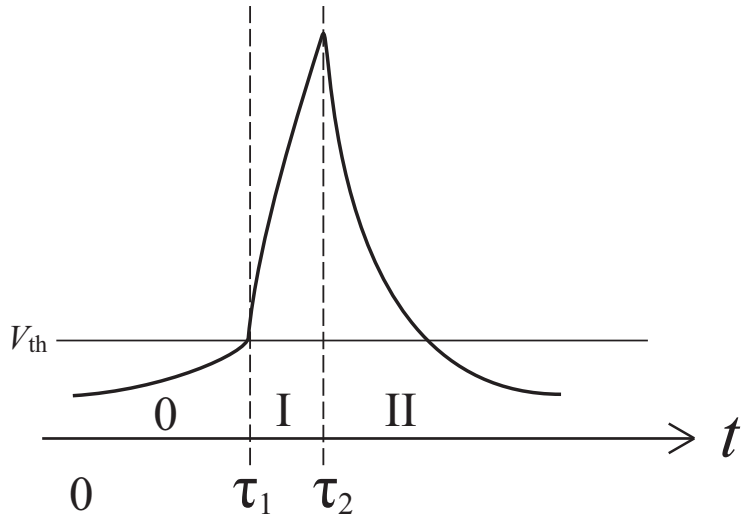

Figure 4.3: Three phases model for spikes.

However, it takes into consideration the pre–activation process in the recepoters of the post–synaptic cell, one more phase, say phase 0, is necessary before the phases I and II. Within the phase 0, it is natural to assume that the electric charge is exponentially accumurated, that is,

$$s(t) = \alpha_0 e^{\beta_0 t} + v_0 \quad (0 \le t < \tau_1)\,, \qquad \text{(Phase 0)}$$

where $\alpha_0 > 0$ and $\beta_0 > 0$. The positive β_0 is in a good contrast with the $\beta_1 < 0$ and $\beta_2 < 0$ in the phase I and II. Figure 4.3 is a schematic view of the three phase model. The phase 0 is called pre–firing phase (Guerreiro and de Araujo, 2007, Koch, 1999) where the membrane potential increases to the threshold V_{th}, then an activation of the membrane starts. The phase II after the phase I is corresponding to the refractory period.

Now consider here a model for the occurrence time T_j's. It is apparently random so that it would be natural to model it as a point process (Cox and Isham, 1980, Brillinger, 1992, Kass *et al.*, 2005). One of natural assumptions is that the intensity is proportional to the input current, that is, the derivative

of the input voltage $L(t)$ provides the intensity function,

$$\lambda\left(t;\boldsymbol{\kappa}\right)=\begin{cases}\kappa_1\left(\dfrac{dL\left(t\right)}{dt}\right)_+ & \left(t<\tau_1\right),\\[2ex] \kappa_2\left(\dfrac{dL\left(t\right)}{dt}\right)_+ & \left(\tau_1\leq t\right),\end{cases}$$

where $\boldsymbol{\kappa}=(\kappa_1,\kappa_2)$ is the vector of parameters and $(\cdot)_+$ denotes the positive part. It is worthy of note that time t runs from $-\infty$ to ∞ as same as that of $L(t)$.

4.3.2 Model for the input

Consider a model for the input signal $L(t)$. Various kind of models can be considered, but from the shape of the residual $V(t)-S(t)$, it would be a natural choice to take it as the same shape as that of spikes. This can be also thought as a dull shape of a spike raised in other neuron and transmitted to the current neuron,

$$L(t)=\begin{cases}a_0e^{b_0t}+w_0 & \left(t<t_1\right),\\[2ex] a_1e^{b_1(t-t_1)}+w_1 & \left(t_1\leq t<t_2\right),\\[2ex] a_2e^{b_2(t-t_2)}+w_2 & \left(t_2\leq t\right),\end{cases}$$

where $a_0,b_0>0$, $a_1,b_1<0$ and $a_2>0,b_2<0$. A difference from the spike $s(t)$ is that it does not start from a specific time point because the input is not instantaneous but rather slowly varying.

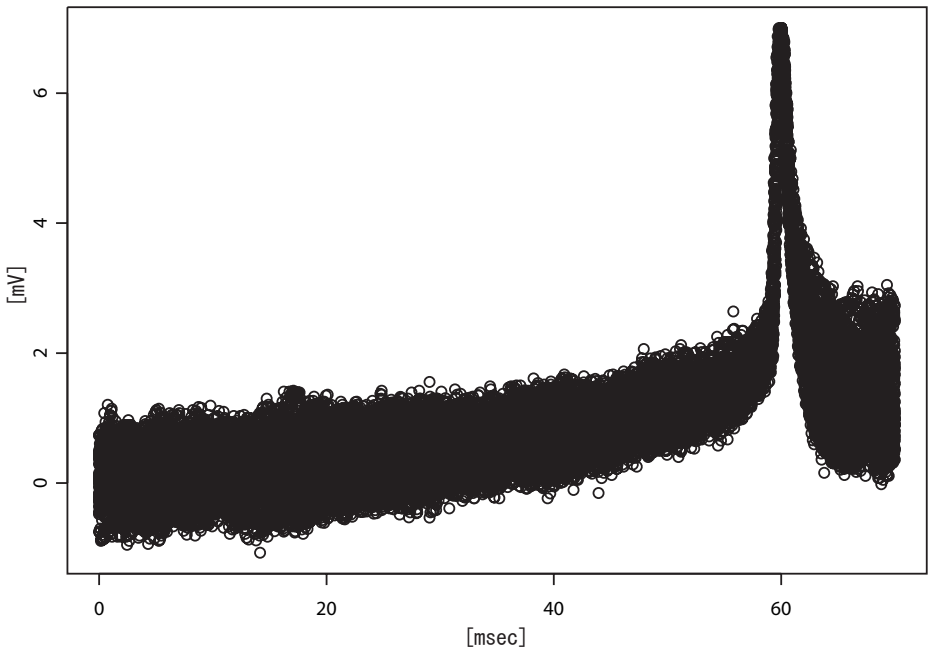

Figure 4.4: Collected spikes.

4.4 Model identification

4.4.1 Identification of the spike model

To identify $s(t)$, the first spikes in each cluster are all collected and aligned so that the peaks are at the same time position and windowed with 50 msec ahead and 10 msec behind of the peak time. The level is adjusted so that $s(0) = 0$. Such collected spikes are shown in Figure 4.4. Although the three phase model has the eleven parameters, setting the end conditions as $s(0) = s(\infty) = 0$ reduced two parameters, $v_0 = -\alpha_0$ and $v_2 = 0$. Also, it is possible to save two more parameters from the two continuity conditions at

Table 4.1: Estimated parameters of the three phase model.

$\hat{\alpha}_0$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{v}_1$	$\hat{\tau}_1$	$\hat{\tau}_2$
0.423	0.027	-2.213	-0.272	7.00	59.14	59.80

τ_1 and τ_2,

$$\begin{aligned} \alpha_1 &= \alpha_0 e^{\beta_0 \tau_1} - \alpha_0 - v_1, \\ \alpha_2 &= \alpha_1 e^{\beta_1(\tau_2 - \tau_1)} + v_1. \end{aligned}$$

Then there are seven free parameters to be estimated by non–linear optimisation program so as to minimise the residual sum of squares. The estimated parameters are shown in Table 4.1.

From this table, it is clear that the rate of the conductance change from the phase I to the phase II is $g_1/g_2 = \beta_1/\beta_2 = 8.136$ and the relative equilibrium voltage in the phase I is $v_1 = 7$ mV to that in the phase II. The shape of the estimated model for spikes is shown in Figure 4.5.

4.4.2 Identification of the input model

Based on the identified model for spikes, it is possible to produce a spike series $S(t) = \sum_j s(t - T_j)$, where the activation time T_j's used here are those directly obtained from the observation. Figure 4.6 shows an example of a simulated spike series $S(t)$.

Then the parameters of the model for $L(t)$ are estimated from the residual $V(t) - S(t)$ by a non–linear optimisation program similarly as in the estimation of parameters of the spike $s(t)$. The difference is that it is not necessary to impose the assumption $L(0) = 0$, rather it simply assumes that $L(-\infty) = L(\infty)$, that is , $w_0 = w_2$. Therefore, four parameters are reduced

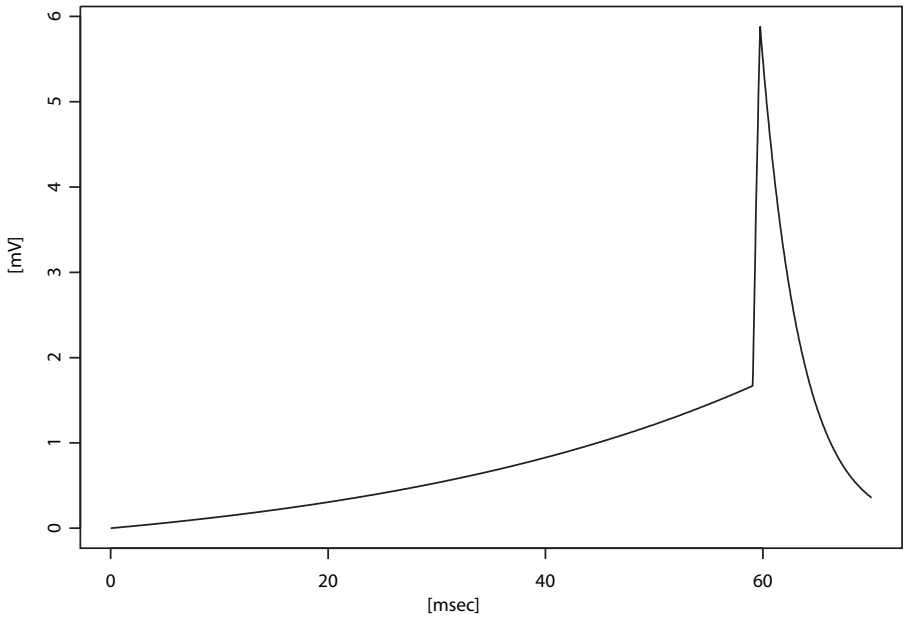

Figure 4.5: Estimated spike $s(t)$.

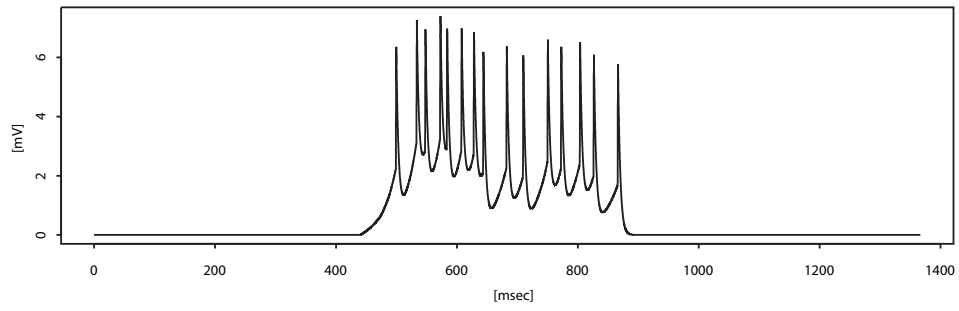

Figure 4.6: Simulated spike series $S\,(t)$.

Table 4.2: Estimated parameters of the model for $L(t)$.

Cluster	$\hat{a}_0$	$\hat{b}_0$	$\hat{b}_1$	$\hat{b}_2$	$\hat{w}_0$	$\hat{w}_1$	$\hat{t}_1$	$\hat{t}_2$
1	0.057	0.0044	-0.0014	-0.0085	-0.217	5.022	726.323	1054.364
2	0.057	0.0050	-0.0014	-0.0085	-0.217	6.022	707.323	1029.364
3	0.057	0.0050	-0.0014	-0.0085	-0.217	6.022	727.823	941.750
4	0.077	0.0050	-0.0014	-0.0085	-0.217	6.022	737.173	974.176
5	0.097	0.0050	-0.0014	-0.0085	-0.217	6.022	675.823	959.362
6	0.097	0.0050	-0.0014	-0.0085	-0.217	6.022	736.323	991.434
7	0.097	0.0050	-0.0014	-0.0085	-0.217	6.022	708.323	884.875
8	0.097	0.0050	-0.0014	-0.0085	-0.217	6.022	700.323	883.250
9	0.097	0.0050	-0.0014	-0.0085	-0.217	6.022	595.323	1061.892
10	0.157	0.0050	-0.0014	-0.0085	-0.217	6.022	664.823	900.110
11	0.157	0.0050	-0.0014	-0.0085	-0.217	6.022	645.823	869.236
12	0.157	0.0050	-0.0014	-0.0085	-0.217	6.022	656.823	814.450
13	0.157	0.0050	-0.0014	-0.0085	-0.217	6.022	657.823	814.450
14	0.157	0.0050	-0.0014	-0.0085	-0.217	6.022	657.823	777.395
15	0.157	0.0050	-0.0014	-0.0085	-0.217	6.022	617.973	928.280
16	0.157	0.0050	-0.0014	-0.0085	-0.217	6.022	615.823	853.279
17	0.127	0.0050	-0.0014	-0.0085	-0.217	6.022	689.823	912.900
18	0.207	0.0050	-0.0014	-0.0085	-0.217	6.022	611.323	771.396
19	0.207	0.0050	-0.0014	-0.0085	-0.217	6.022	615.823	833.000
20	0.207	0.0050	-0.0014	-0.0085	-0.217	6.022	642.823	800.733
21	0.207	0.0050	-0.0014	-0.0085	-0.217	6.022	626.323	698.485
22	0.167	0.0050	-0.0014	-0.0085	-0.217	6.022	661.323	929.671
23	0.187	0.0050	-0.0014	-0.0085	-0.217	6.022	676.323	820.667
24	0.187	0.0050	-0.0014	-0.0085	-0.217	6.022	611.323	941.438
25	0.187	0.0050	-0.0014	-0.0085	-0.217	6.022	658.823	876.827
26	0.187	0.0050	-0.0014	-0.0085	-0.217	6.022	643.823	833.241
27	0.187	0.0050	-0.0014	-0.0085	-0.217	6.022	623.823	786.417

as

$$
\begin{aligned}
w_0 &= w_2, \\
a_1 &= a_0 e^{b_0 t_1} + w_0 - w_1, \\
a_2 &= a_1 e^{b_1 (t_2 - t_1)} + w_1.
\end{aligned}
$$

The results are shown in Table 4.2. It is worthy of note that estimated parameters are the same except $\hat{a}_0$ and $\hat{t}_2$ over all clusters. This suggests that almost the same input $L(t)$ comes into this neuron after the cluster 10,

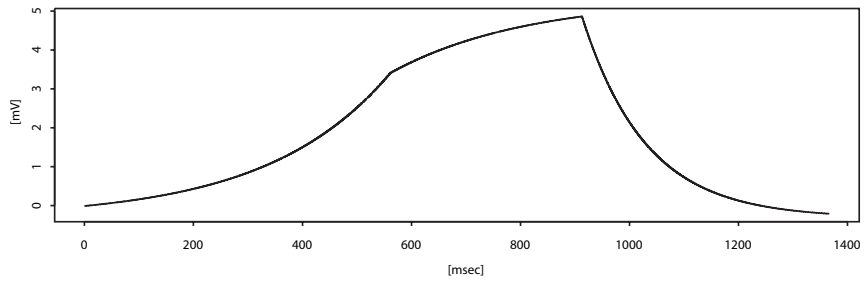

Figure 4.7: Simulated input $L(t)$.

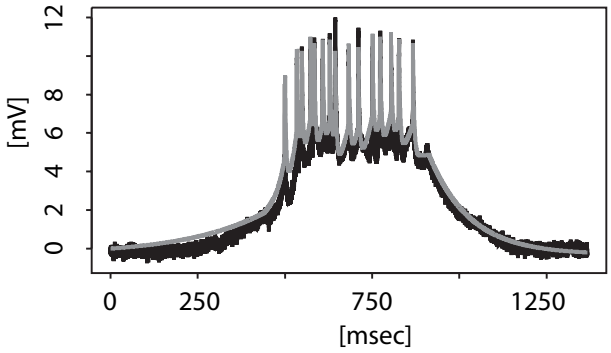

Figure 4.8: Simulated $L(t) + S(t)$ and the original time series.

although t_2 varies input to input. Figure 4.7 shows the estimated $L(t)$ for all clusters.

4.4.3 Results of simulation

An example of the results of simulation based on the identified model $V(t) = L(t) + S(t)$ for a cluster is shown in Figure 4.8. The gray line is the result of simulation and the black line is the original time series in this cluster. It is clear that the simulated data well traces the given data. Figure 4.9 is a plot of the residual and its decomposition into $\xi_L(t)$ and $\xi_S(t)$ and $\varepsilon(t)$, which are respectively the residuals related to $L(t)$ and $S(t)$, and a common residuals. The roughness of $\xi_S(t)$ is due to a constant hight of the spikes

in this simulation. It will be reduced by introduction of randomness to the hight of the spikes.

4.4.4 Identification of the intensity for the occurrence time of spikes

In the previous section, the occurrence times of spikes are obtained from the observed data. Now model it as an inhomogeneous point process as is described in the previous section. It is necessary to estimate two parameters κ_1 and κ_2, since $L(t)$ has already identified. Table 4.3 shows the result of the estimation.

Figure 4.10 shows the original time series with $L(t)$ and the intensity function $\hat{\lambda}(t)$ derived. The intensity function reflects well the occurrence time of the spikes in the top panel. To check the goodness of fit of the inhomogeneous Poisson model to the observed occurrence time $\{T_j, j = 1, 2, .., N\}$, rescaled time

$$Z_j = \hat{\Lambda}_{j+1} - \hat{\Lambda}_j \qquad (j = 1, 2, \ldots, N)\,,$$

is calculated from

$$\hat{\Lambda}_j = \int_{-\infty}^{T_j} \hat{\lambda}(t)\ dt.$$

It is possible to check the goodness of fit by a Q–Q plot for a standard exponential distribution since $Z_j \ \ (j = 1, 2, \ldots, N)$ are expected to be independently distributed as the exponential distribution as far as $T_j \ \ (j = 1, 2, \ldots, N)$ follow to the inhomogeneous Poisson process with the intensity $\hat{\lambda}(t)$. Figure 4.11 shows an example of such a Q–Q plot, which shows a good fit of the model.

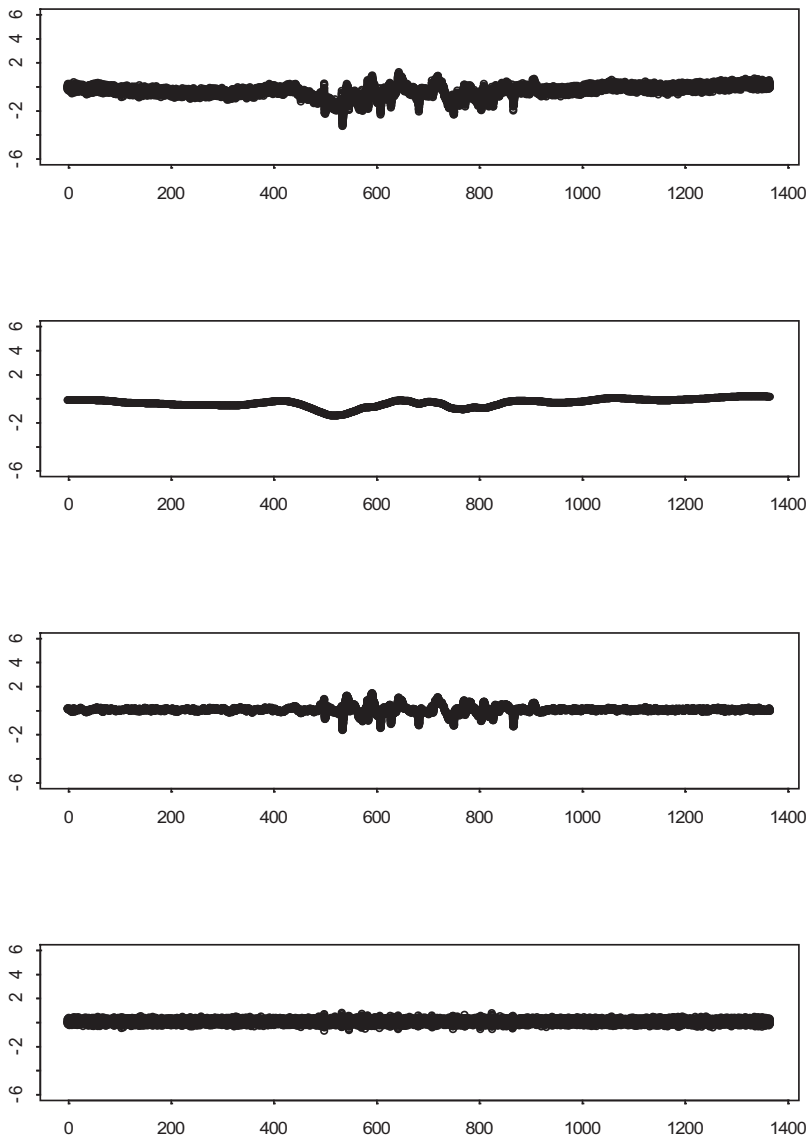

Figure 4.9: The residual and its decomposition.

Table 4.3: Estimated constants κ_1 and κ_2

Cluster	$\hat{\kappa}_1$	$\hat{\kappa}_2$
1	1.453553	11.31297
2	0.6023988	13.5511179
3	0.6023988	11.2892964
4	0.9319404	10.7045732
5	0.4659702	10.3331196
6	0.9319404	8.8725417
7	0.6414121	12.5079983
8	0.6414121	11.8017117
9	0.6414121	11.9391686
10	0.8614148	11.3205296
11	1.073636	8.49214
12	0.5368179	10.5285457
13	0.5368179	10.0272394
14	0.8052269	8.1514682
15	0.8052269	7.9091211
16	0.8052269	6.7710828
17	0.8052269	8.8311471
18	1.073636	6.301583
19	0.8052269	6.924142
20	1.073636	5.567099
21	1.073636	4.820303
22	1.073636	5.183471
23	0.8052269	6.553993
24	0.8052269	4.9054681
25	0.8052269	6.7746739
26	1.073636	5.091122
27	1.073636	4.447863

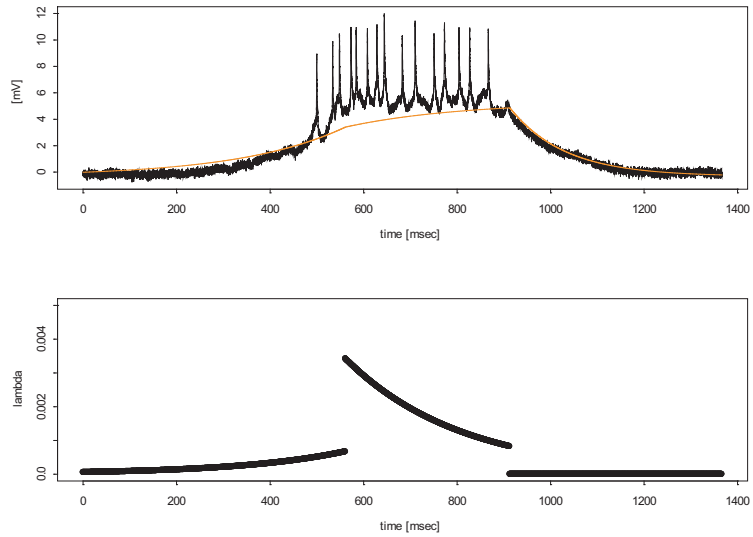

Figure 4.10: The original time series with $L(t)$ and the intensity $\hat{\lambda}(t)$ derived.

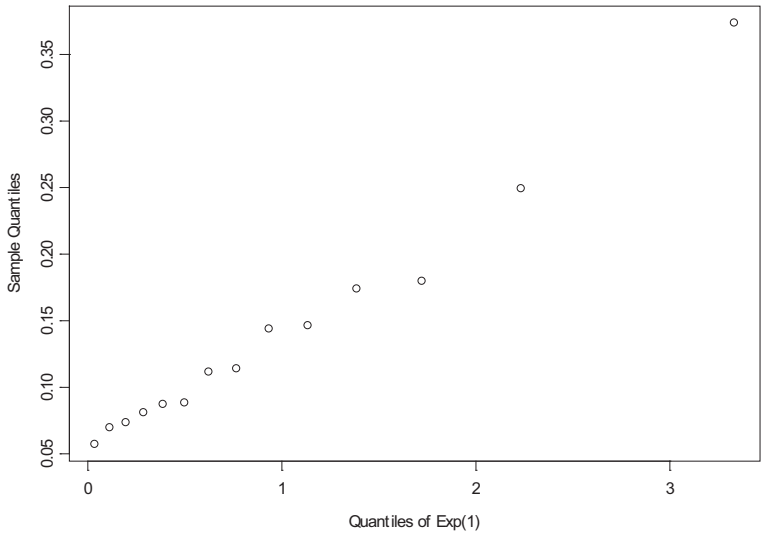

Figure 4.11: Q–Q Plot of Z_j $(j = 1, 2, \ldots, N)$ for the standard exponential distribution.

4.5 Summary

A new model for the membrane potential based on an *in vivo* observation of a single neuron is derived. The basic idea is that the input to a neuron comes through an electrical or chemical synapse. The input through the electrical synapse is directly reflected to the potential but that through the chemical synapse is delayed and modified. In this model, spikes are randomly invoked and follows to an inhomogeneous Poisson process with the intensity proportional to the derivative of the input. For the shape of a spike, it has introduced a three phase model, where the three phases correspond to the pre-activation, activation and post–activation stages. This model is quite general so that it is also possible to apply this model to the input to the neuron. The result of simulation shows good fit to the data and suggests that this model would work well not only for the underlying neuron of earth worm also for any other neurons.

Chapter 5

Conclusion

In this thesis, three case studies of biological data modelling have been demonstrated. The first case study is totally exploratory in nature so that a smoothing technique which can extract some structures behind the data was applied twice by changing the window width. As a consequence, each of five bird count series was simultaneously decomposed into three components: long trend, short trend and irregular. As a result, it was found that there are two groups whose numbers are closely related with two different environment factors. Furthermore, the variation of each short trend suggests the effects of breeding season or winter wandering. In the second case study, it was a key to introduce a physical model for swimming. A newly developed differential equation was powerful tool because of easy interpretability of the model. It describes well the trade–off between drag and propulsion in swimming but was not enough to explain observed data. The introduction of noise for swimming speed and that of individual factor as a multiplicative constant was another key. The fitted model provides us a good description of the swimming strategies of each swimmer from phase to phase in the race and over the race as a whole. In the third case study, the Hodgkin–Huxley model was extended to fit the data observed, incorporating the basic idea that

the input into a neuron comes through an electrical or chemical synapse. Furthermore, it was assumed that spikes are randomly invoked and following to an inhomogeneous Poisson process with the intensity which is proportional to the derivative of the input. This model provides a good description of the mechanism of membrane potential.

Three case studies look like independent but similar in the sense that any of the discoveries could not be achieved by a mundanely application of statistical analysis like regression or a simple modification of already existing models. The key to success is of course "Be honest to the given data" and "Keep a good relation to the scientist in the field" as was described in Chapter 1. Biological data modelling is not specific in this sense, but specific in other sense because deep understanding of the phenomena behind and good insight into the modelling are indispensable. In other words, biological data modelling can be a benchmark study in the practice of data science. My hope is that many data scientists join us for such a fascinating research experiences.

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *in* B. N. Petrov and F. Csaki (eds), *Proceedings of 2nd International Symposium on Information Theory*, Akademiai Kiado, Budapest, pp. 267–281.

Amar, J. (1920). *The Human Motor*, George Routledge & Sons, London.

Anada, S. and Fujimaki, Y. (1984). Avifauna of aguricultural land and residential area in Obihiro, eastern Hokkaido, during the breeding season, *Strix* **3**: 19–27. (In Japanese).

Arellano, R., Brown, P., Cappaert, J. and Nelson, R. (1994). Analysis of 50-, 100-, and 200-m freestyle swimmers at the 1992 Olympic games, *Journal of Applied Biomechanics* **10**: 189–199.

Bibby, C. J., Burgess, N. D., Hill, D. A. and Musoe, S. H. (2000). *Bird Census Techniques*, 2 edn, Academic Press, London.

Brillinger, D. (1992). Nerve cell spike train data analysis: a progression of technique, *Journal of the American Statisticsl Association* **87**: 260–271.

Chambers, J. and Hastie, T. (eds) (1992). *Statistical Models in S*, Wadsworth, California.

Chambers, J., Cleveland, W., Kleiner, B. and Tukey, P. (1983). *Graphical Methods for Data Analysis*, Wadsworth, California.

Chengalur, S. and Brown, P. (1992). An analysis of male and female Olympic swimmers in the 200-meter events, *Canadian Journal of Sport Sciences* **17**: 104–109.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**(368): 829–836.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting, *Journal of the American Statistical Association* **83**(403): 596–610.

Cox, D. and Isham, V. (1980). *Point Processes*, Chapman & Hall, London.

Craig, A. and Pendergast, D. (1979). Relationship of stroke rate, distance per stroke, and velocity in competitive swimming, *Medicine and Science in Sports* **11**: 278–283.

Craig, A., Skehan, P., Pawelczyk, J. and Boomer, W. (1985). Velocity, stroke rate, and distance per stroke during elite swimming competition, *Medicine and Science in Sports* **17**: 625–634.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Regression and Its Applications*, Chapman&Hall, London.

Guerreiro, A. M. G. and de Araujo, C. A. P. (2007). An extended model for a spiking neuron class, *Biological Cybernetics* **97**: 211–219.

Higuchi, H., Tsukamoto, Y., Hanawa, S. and Takeda, M. (1982). Relationship between forest areas and the number of bird species, *Strix* **1**: 70–78. (In Japanese).

Hirano, T. (1996). Changes in breeding avifauna during the past 25 years at Tomatsuriyama in Utsunomiya city, *Strix* **14**: 25–31. (In Japanese).

Hirano, T., Endo, K., Nihei, K., Kanehara, K. and Higuch, H. (1985). The relationship between the percentage of wooded area and occurrence of bird species in Utsunomiya, *Strix* **4**: 33–42. (In Japanese).

Hirano, T., Ishida, H. and Kunitomo, T. (1989). The relationship btween forest area and the number of bird species in winter, *Strix* **8**: 173–178. (In Japanese).

Hodgkin, A. L. and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve, *Journal of Physiology* **117**: 500–544.

Ikuta, Y., Okuno, K., Matsui, K., Terada, A., Honbu, Y., Ishikawa, M., Wakayoshi, K. and Nomura, T. (1998). Relationship between control of swimming velocity and stroke rate — suggestion from result of 100m and 200m free-style events —, *Japanese Journal of Sport Methodology* **12**: 1–8. (In Japanese).

James, F. C., McCulloch, C. E. and Wiedenfeld, D. A. (1996). New approaches to the analysis of population trends in land birds, *Ecology* **77**(1): 13–27.

Karpovich, P. V. (1933). Water resistance in swimming, *Research Quarterly* **4**: 21–28.

Kass, R. E., Ventura, V. and Brown, E. N. (2005). Statistical issues in the analysis of neuronal data, *Journal of Neurophysiology* **94**: 8–25.

Kira, Y. (2000). *The Natural History of Minamisawa*, Jiyu-Gakuen Shuppankyoku, Tokyo. (In Japanese).

Kira, Y., Shimadzu, H. and Yamagata, M. (2002). Records of bird census in Jiyu-Gakuen, *Jiyu Gakuen Annual Report* **6**: 161–180. (In Japanese).

Kjendlie, P., Stallman, R. K. and Gundersen, J. (2004). Adults have lower stroke rate during submaximal front crawl swimming than children, *European Journal of Applied Physiology* **91**: 649–655.

Koch, C. (1999). *Biophysics of Computation*, Oxford University Press, New York.

Kolmogorov, S. and Duplishcheva, O. (1992). Active drag, useful mechanical power output and hydrodynamic force coefficient in different swimming strokes at maximal velocity, *Journal of Biomechanics* **25**: 311–318.

Komeda, S. and Ueki, Y. (2002). Long term monitoring of migratory birds at Otayama banding station (1973-1996), *Journal of the Yamashina Institute for Ornithology* **34**: 96–111. (In Japanese).

Konishi, S. and Kitagawa, G. (1996). Generalized information criteria in model selection, *Biometrika* **83**(4): 875–890.

Konishi, S. and Kitagawa, G. (2007). *Information Criteria and Statistical Modeling*, Springer, New York.

Kurosawa, R. (1994). Bird abundance in relation to the pavement rate of Tokyo, *Strix* **13**: 155–164. (In Japanese).

Maeda, T. (1998). Bird communities and habitat relationships in a residential area of Tokyo, *Journal of the Yamashina Institute for Ornithology* **30**: 83–100.

Matsui, T., Terada, A., Tatesada, E., Honbu, Y., Ikuta, Y., Wakayoshi, K. and Nomura, T. (1997). Changes in swimming velocity and stroke variables in 5m intervals during competitive swimming race —comparisons between elite and sub-elite swimmers in 200-m freestyle race of Japanese championship—, *Japanese Journal of Sport Methodology* **11**: 87–93. (In Japanese).

Murai, H. and Higuchi, H. (1988). Factors affecting bird species diversity in Japanese forests, *Strix* **7**: 83–100. (In Japanese).

Okuno, K., Horinouchi, T., Naitou, K., Ikuta, Y., Wakayoshi, K. and Nomura, T. (2003). A study on elite swimmer's race pattern of competition swimming in individuals medley events, *Annual Report of Physical Education* **35**: 87–92. (In Japanese).

Ootaka, H. and Nakamura, M. (1996). Avifauna on the campus of Joetsu University of Education during the breeding season, *Strix* **14**: 113–124. (In Japanese).

Rinzel, J. (1990). Electrical excitability of cells, theory and experiment: review of the hodgkin–huxley foundation and an update, *Bulletin of Mathematical Biology* **52**: 5–23.

Rose, R. M. and Hindmarsh, J. L. (1989). The assembly of ionic currents in a thalamic neuron I. the three-dimensional model, *Proceedings of the Royal Society of London. Series B* **237**: 267–288.

Shibata, R. (2001). *Data Literacy*, Kyoritsu Shuppan, Tokyo. (In Japanese).

Shibata, R. and Miura, R. (1997). Decomposition of Japanese Yen interest rate data through local regression, *Financial Engineering and the Japanese Markets* **4**: 125–146.

Shimadzu, H. and Shibata, R. (2005). Analysis of bird count series by local regression to explore environmental changes, *Journal of the Japan Statistical Society J* **34**(2): 187–207. (In Japanese).

Shimadzu, H., Shibata, R. and Ohgi, Y. (2007). Modelling swimmers' speeds over the course of a race, *Journal of Biomechanics*. doi: 10.1016/j.jbiomech.2007.10.007.

Takagi, H., Shimizu, Y. and Kodan, N. (1999). A hydrodynamic study of active drag in swimming, *JSME International Journal Series B* **42**: 171–177.

Toussaint, H., de Groot, G., Savelberg, H., Vervoorn, K., Hollander, P. and van Ingen Schenau, G. (1988). Active drag relates to velocity in male and female swimmers, *Journal of Biomechanics* **21**: 435–438.

Tukey, J. W. (1962). The future of data analysis, *Annals of Mathematical Statistics* **33**(1): 1–67.

Uchida, Y., Shimadzu, H. and Sekimoto, T. (2003). The relationship between the avifauna and environmental changes at Jiyu-Gakuen in Tokyo: a statistical analysis of the bird-census data for 35 years, *Strix* **21**: 53–70. (In Japanese).

Wilson, H. (1999). Simplified dynamics of human and mammalian neocortical neurons, *Journal of Theoretical Biology* **200**: 375–388.

Yokouchi, D. and Shibata, R. (2001). InterDatabase — DandD instance as an agent on the internet—, *Proceedings of the Institute of Statistical Mathematics* **49**: 317–331. (In Japanese).