

https://dspace.org/

Deterministic dynamics in questionnares in the social sciences

Mberi Kimpolo, Charles Lebon

2011-03-28
http://hdl.handle.net/10539/9248

Downloaded from a DSpace repository. https://dspace.org/

Deterministic Dynamics in Questionnaires in the Social Sciences.

Charles Lebon Mberi Kimpolo

Supervisor: Professor David Sherwell



School of Computational and Applied Mathematics
University of the Witwatersrand
Private Bag 3, WITS 2050, Johannesburg
Republic of South Africa

A dissertation submitted to the Faculty of Science
in fulfilment of the requirements for the degree of
Doctor of Philosophy

Declaration

I declare that this thesis is my own, unaided work and it has not been submitted before for any degree or examination in any other university. It is being submitted for the degree of Doctor of Philosophy in the University of the Witwatersrand, Johannesburg, South Africa.

C.L. Mberi Kimpolo

October 5, 2010

Abstract

In this thesis a novel mathematical technique for the analysis of longitudinal surveys in the social sciences is given. This analysis maps the longitudinal data of a fixed number n of demographic variables of a single social unit into an orbit of a single point in the unit square. The $x, y-$axes denote fitness, and, significance of social variables. A finite set of $2^n \times n!$ states in the unit square is thereby defined. Data from the rural Agincourt Health and Demographic Surveillance Site survey is analysed. The data set consists of the following demographic variables: biological mother out-migration, household head is a minor and adult death. For a sample of 2669 households we record orbits for the period 1998 to 2007. Social variables relate to educational progression. The flow of household orbits is found to describe temporary in- and out-migration of biological mothers of children. The densities of the flows show that educational default is associated with out-migration. The method predicts an increase of 52 defaulting households per year for the period 2007 to 2015. This result is facilitated by visualization of orbits and by identification of appropriate dynamical models, directly from the longitudinal data. It is hoped that visualization of household fitness can better influence policy makers.

I dedicate this thesis to my father

Kimpolo Martin,

and my mother

Josephine Loubondo

whose love, care and selflessness have brought me this far

to my wife

Alix Dyna Mberi Kimpolo

and my daughter

Hattie Wright Mberi Kimpolo

Acknowledgements

My deepest reference of gratitude is first rendered to the One I owe allegiance to, my strength and my very present help in times of need, the Almighty God, the Great Provider, The Lord Jesus Christ, without whom this thesis would not have seen the light of day.

My profound gratitude goes to my supervisor, Professor David Sherwell for his guidance and encouragement. His extreme availability and continuing support towards this work will always and highly be appreciated.

I also would like to acknowledge the staff of the Demographic and Population Studies Programme: Professor Steve Tollman and Professor Kathy Kahn for their valuable inputs into the construction of this project as well as their general advice and assistance.

I wish to express my appreciation to all my fellow postgraduate students and staff at the school of Computational and Applied Mathematics, among them are Dr Amandine Robin, Mr Naval Andrianjafinandrasana, Mr Morgan Kamga Pene for their constructive suggestions, encouragements, support and useful discussions.

I would like to thank the following staff members of the Agincourt Health and Demographic Surveillance Site, University of the Witwatersrand: Dr Mark, Mr Paul Mee and Mr Afolabi for helping me in the data collection for this work.

I also would like to acknowledge with great appreciation the financial support of the following donors

- The Hewlett and William Foundation which awarded me a doctoral fellowship to support this research
- This work was also supported by the Welcome Trust [085477/Z/08/Z]
- The African Institute for Mathematical Sciences (AIMS)
- The African Mathematical Millennium and Scientific Initiative (AMMSI).

Contents

Abstract **2**

Acknowledgements **4**

List of Figures **13**

List of Tables **13**

1 Introduction **14**

1.1 Introduction 14

1.2 Review of deterministic mathematical modelling 16

1.2.1 Deterministic mathematical studies of individual-level social dynamics 16

1.2.2 Deterministic population-level dynamics models 18

1.3 Review of statistical longitudinal data analysis techniques 21

1.4 Projection 26

1.5 Motivation and objectives 26

1.5.1 Motivation 26

1.5.2 Objectives 27

1.5.3 Outline 27

2 Data Preparation **28**

2.1 Introduction 28

2.2 Description of the Agincourt District 28

2.3 The Agincourt research 30

2.4 Description of the Agincourt HDSS database 31

2.4.1 Problem of question order in design of questionnaires . . . 32
2.5 Description of the Agincourt data for the current analysis . . . 34
2.5.1 Designing questionnaire . . . 34
2.5.2 Description of data collection process . . . 34
2.5.3 Data Sampling Methods . . . 38
2.6 Conclusion . . . 45

3 General Orbit Theory of Longitudinal Data 46
3.1 Introduction . . . 46
3.2 Fitness Space . . . 47
3.3 Equally weighted questions model . . . 51
3.3.1 Distance and displacement for equally weighted questions . . . 52
3.3.2 Results for the accumulated displacements ${}_{acc}\Delta_t^k$ . . . 54
3.4 Models with Weighted Questions . . . 61
3.4.1 Results for fitness orbits for weighted questions of fixed order . . . 65
3.5 Evolutionary question order model . . . 66
3.5.1 Result for the evolutionary orbits on I_2^n . . . 70
3.5.2 Evolutionary orbits in the unit square $I_2^n \times I_n^n$ . . . 73
3.6 Dynamics of orbits in the state space $\Gamma_n = I_2^n \times I_n^n$ . . . 76
3.7 Transition matrix for social units . . . 81
3.7.1 Some properties of the transition matrix T_n . . . 85
3.8 Example of Agincourt household orbits in $\Gamma_{n=3}$ . . . 86
3.9 Social unit-level transition matrix T_n^k . . . 92
3.10 Conclusion . . . 95

4 Demographic Analysis 97
4.1 Introduction . . . 97
4.2 Population-level dynamics models . . . 97
4.2.1 Initial conditions of the social systems . . . 98
4.2.2 Population transition matrix . . . 99
4.2.3 Population density matrix . . . 100

4.2.4 Population flux vector . . . 102
4.3 Agincourt population dynamics . . . 102
4.3.1 The "fhiqo" and "riqo" starting strategies . . . 103
4.3.2 Agincourt density matrices . . . 110
4.3.3 Distribution of the number of transitions in each observation year . . . 118
4.4 Agincourt flux vector . . . 122
4.5 Initial and final states of the Agincourt population in Γ_3 . . . 125
4.6 Visualization of all orbits of Agincourt data. . . . 127
4.6.1 Orbits of the Agincourt population in Γ_3 . . . 130
4.7 Effect of household change on children's progression in school . . . 134
4.7.1 Colour coding and definition of educational default . . . 134
4.7.2 Analysis of Agincourt dynamics including education status . . . 135
4.8 Properties of Agincourt transition matrix with education status, $l_{failure} = 4$ . . . 138
4.8.1 Time series of fitness states, including educational status. . . . 146
4.8.2 Agincourt orbits in phase diagram, including educational status. . . . 147
4.9 Identification of social force in the social system . . . 156
4.10 Summary of Agincourt Demographic Results 1998 – 2007 . . . 157
4.11 General strategy for large number of questions and comment on the role of statistical methods159

5 Projecting the future of Agincourt social dynamics 164
5.1 Introduction . . . 164
5.2 Formulation of the dynamics . . . 165
5.3 Method of projection . . . 167
5.4 Detecting periodic orbits . . . 168
5.4.1 Periodic orbit for a dynamical system . . . 168
5.4.2 Agincourt period one ($\tau = 1$) orbits . . . 169
5.4.3 Agincourt period two ($\tau = 2$) orbits . . . 170
5.4.4 Agincourt period three ($\tau = 3$) orbits . . . 174
5.4.5 Agincourt period four ($\tau = 4$) orbits . . . 178
5.5 Simulation of the Agincourt population . . . 182

5.5.1 Simulation of Agincourt results . . . 183
5.5.2 Predicted Agincourt population, 2007 – 2015 . . . 187
5.5.3 Scenario planning . . . 188
5.6 Conclusion . . . 191

6 Conclusion **193**
6.1 Mathematical conclusion . . . 194
6.2 Discussion . . . 197
6.3 Future work . . . 199

Appendix **199**

A Agincourt data description **200**
A.1 Description of the variables . . . 200

Bibliography **200**

List of Figures

2.1 Distribution of the number of Agincourt households over the the observation period. 39

2.2 Distribution of answer values change for 3098 Agincourt households. 40

2.3 Distribution of the number of answers changing values per time step. 41

2.4 Distribution of the number of observed Agincourt households per year. 42

2.5 Distribution of answer value change for 2669 Agincourt households. 43

2.6 Distribution of the number of answers changing values per time step. 44

3.1 Accumulated displacement for only one switch-and-stay answer value change. 57

3.2 Accumulated displacement for only one oscillating answer value change. 58

3.3 Accumulated displacement for a uniformly random change in 8 of 26 answer values. . 58

3.4 Accumulated displacement for a uniformly random change in 26 of 26 answer values. 59

3.5 Agincourt household accumulated displacement, for many answer values changing. . 60

3.6 Agincourt household accumulated displacement, for various answer values changing. 60

3.7 Agincourt household accumulated displacement, for 3 negative changes. 61

3.8 Visualization in Σ_2^3. The trend of social unit fitness is made clear by visual orbits. . 61

3.9 Single social unit fitness orbit, for a switch-and-stay answer value change. 66

3.10 Single social unit fitness orbit, for an oscillating answer value change. 67

3.11 Fitness orbit,for a uniformly random change in 8 of 26 answer values. 68

3.12 Fitness orbit, for a uniformly random change in 26 of 26 answer values. 69

3.13 Agincourt fitness orbit, for many answer values changing. 69

3.14 Agincourt fitness orbit, for a single oscillating answer value changing. 70

3.15 Agincourt fitness orbit, for 3 answer values changing. 71

3.16 Evolutionary fitness orbit, for a switch-and-stay answer value change. 72

3.17 Evolutionary fitness orbit, for an oscillating answer value change. 72
3.18 Evolutionary fitness orbit, for uniformly random change in 8 of 26 answer values. . . 73
3.19 Evolutionary fitness orbits, for a uniform change in 26 of 26 answer values. 74
3.20 Bi-sequence space $S_3 = \Sigma_2^3 \times \Sigma_3^3$. 80
3.21 Bi-sequence space $S_3 = \Sigma_2^3 \times \Sigma_3^3$, with sequences indexed. 81
3.22 Comparison of the exponential function and the dimension of the transition matrix. 87
3.23 Agincourt evolutionary fitness-significance orbit $(e_t^{7150}, \chi_t^{7150})$. 88
3.24 Agincourt evolutionary fitness-significance orbit $(e_t^{7150}, \chi_t^{7150})$. 89
3.25 Agincourt evolutionary fitness-significance orbit $(e_t^{4325}, \chi_t^{4325})$. 90
3.26 Agincourt evolutionary fitness-significance orbit $(e_t^{4325}, \chi_t^{4325})$. 91
3.27 Agincourt evolutionary fitness-significance orbits $(e_t^{5873}, \chi_t^{5873})$. 92
3.28 Agincourt evolutionary fitness-significance orbit $(e_t^{5873}, \chi_t^{5873})$. 93
3.29 The dynamics in S_3, of the Agincourt household $k = 7150$. 94

4.1 State space S_n for $n = 2$ binary-valued questions with states numbered. 119
4.2 Agincourt distribution of the number of transitions over years. 120
4.3 Time-dependent flux over the dominant states 23 and 24. 122
4.4 Initial states of the Agincourt households in Γ_3. 126
4.5 Initial distribution of the number of Agincourt households s. 127
4.6 Final states of the Agincourt households in Γ_3. 128
4.7 Final distribution of the number of Agincourt households s. 129
4.8 Time series of the fitness component, e_t, of all the Agincourt households. 130
4.9 Time series of the significance component, χ_t, of the Agincourt households. 131
4.10 Phase diagram of the Agincourt households. 132
4.11 Orbits of the Agincourt households with time dependent (vertical axis). 133
4.12 Distribution of the number of the Agincourt household orbits over Γ_3. 133
4.13 Distribution of the Agincourt households educational status with respect to $l_{failure}$. 136
4.14 Time series of fitness component, e_t, including education status, $l_{failure} = 4$. 146
4.15 Fitness time series for defaulting Agincourt households, $l_{failure} = 4$. 147
4.16 Fitness time series for non-defaulting Agincourt households, $l_{failure} = 4$. 148

4.17 Agincourt phase diagram orbits including education status, $l_{failure} = 4$. 149
4.18 Distribution of defaulting households, s_{red}, for $l_{failure} = 4$. 150
4.19 Distribution of non-defaulting households, s_{green}, for $l_{failure} = 4$. 151
4.20 Agincourt orbits with time dependent, including education status, for $l_{failure} = 4$. . . 152
4.21 Orbits of non-defaulting Agincourt households, with time dependent (vertical axis). . 153
4.22 Orbits of defaulting Agincourt households, with time dependent (vertical axis). . . . 154
4.23 Phase diagram of defaulting Agincourt households, for $l_{failure} = 4$. 155
4.24 Phase diagram of non-defaulting Agincourt households, $l_{failure} = 4$. 155
4.25 Example of frequency distribution of changing answer value. 161
4.26 Example of frequency distribution of $f_0^j,\ j \in \{0, 1\}$ 162

5.1 Agincourt period-1 household orbits (e_t^k, χ_t^k) in Γ_3. 170
5.2 Agincourt period-1 orbits with time dependent (vertical axis). 171
5.3 Distribution of Agincourt period-1 households over states of Γ_3. 172
5.4 Agincourt period-2 household orbits (e_t^k, χ_t^k) in a sub-space of Γ_3. 172
5.5 Agincourt period-2 phase diagram in reduced space with time dependent. 173
5.6 Evolutionary fitness orbits e_t^k, for all Agincourt period-2 households. 173
5.7 Evolutionary significance orbits χ_t^k, for all Agincourt period-2 households. 174
5.8 Initial distribution of Agincourt period-2 households over states of Γ_3. 175
5.9 Agincourt period-3 household orbits (e_t^k, χ_t^k) in Γ_3. 176
5.10 Agincourt period-3 phase diagram with time dependent (vertical axis). 176
5.11 Evolutionary fitness orbits e_t^k, for the Agincourt period-3 households. 177
5.12 Evolutionary significance orbits χ_t^k, for the Agincourt period-3 households. 177
5.13 Initial distribution of Agincourt period-3 households over states of Γ_3. 178
5.14 Agincourt period-4 household orbits (e_t^k, χ_t^k) in Γ_3. 179
5.15 Agincourt period-4 phase diagram with time dependent (vertical axis). 180
5.16 Evolutionary fitness orbits e_t^k, for the Agincourt period-4 households. 180
5.17 Evolutionary significance orbits χ_t^k, for the Agincourt period-4 households. 181
5.18 Initial distribution of Agincourt period-4 households over states of Γ_3. 181

List of Tables

2.1 Annual cycle of census rounds. 33

3.1 Data of the anonymous Agincourt household $k = 7150$. 87
3.2 Data of the Agincourt household $k = 4325$. 89
3.3 Data of the Agincourt household $k = 5873$. 91
3.4 Transitions for the Agincourt household number 7150, for $1999 \leq t \leq 2006$. 93

4.1 Sorted number of transitions from the density matrix $D_3^{fhiqo}, d_{ij} \geq 100$. 115
4.2 Sorted number of transitions from the density matrix $D_3^{riqo}, d_{ij} \geq 100$. 116
4.3 Transitions of significant transitions from the density matrix D_3^{Ag}. 117
4.4 Reduced Agincourt density matrices for the indicated year, $d_{t,ij} \geq 10$. 121
4.5 Agincourt flux vector for the indicated year. 123
4.6 Distribution of Agincourt educational transitions with respect to $l_{failure}$. 137
4.7 Summary of Agincourt educational dominant transitions with respect to $l_{failure}$ . . . 138
4.8 Identification of social force 157

5.1 Distribution of the number of periodic orbits in Agincourt population 182
5.2 Agincourt average flux vector. 184
5.3 Comparison of approximate and real Agincourt number of households. 185
5.4 Comparison of average and real Agincourt number of households. 186
5.5 Comparison of simulated and Agincourt density matrices. 187
5.6 Comparison of simulated and Agincourt flux vectors for the final transition year . . 188
5.7 Distribution of simulated Agincourt population 189
5.8 Distribution of simulated Agincourt population with new 4000 households 189

5.9 Sorted number of transitions from the density matrix $D_3^{(scenario)}$. 190

5.10 Sorted number of transitions d_{ij} for previously defaulting households. 190

5.11 Fluxes of dominant transitions for the scenario 190

5.12 Scenario of Agincourt defaulting households, $2007 - 2015$. 191

5.13 Predicted number of Agincourt with additional 2000 defaulting households. 192

A.1 Description of the variables 201

Chapter 1

Introduction

1.1 Introduction

In this thesis, we use the term *social unit* to describe a social entity which is part of a larger social group or society. There are various types of social units including individual, family and household [1]. This latter will be used as the analysis level in the discussion of the data that are analysed as an application of the theory developed in this thesis.

Because of the growing complexity of social networks in modern societies [2, 3, 4, 5, 6, 7, 8], the use of deterministic models has become a topic of great interest in social sciences research [9, 10, 11, 12]. In fact, many social phenomena cannot be analysed with conventional statistical techniques [13]. For instance there are some dynamical social processes that are not equivalent to a simple sequence of time-dependent structures, which can be statistically analysed. Such "complex systems" require strong mathematical techniques which also include the description of non-equilibrium phenomena. The need of such new mathematical models is critical to improve research in the social sciences. In this thesis a new deterministic mathematical technique, to complement probabilistic methods, is presented which hopefully will add value to the scientific understanding of the social sciences.

We are interested in the very rich information of longitudinal surveys. For instance, the Agincourt Demographic Surveillance Site has yielded answers to some 200 questions asked to 14000 households over 16 years (details given below). As household conditions change, this represents some 2^{200} possible transitions of each household if answers are Yes/No. The problem is to access

this potentially huge quantity of "experimental" information.

Dynamical systems are mathematical laws that evolve a set of states in time. An example is given by angular rotation about a circle where the unit of angle is chosen so that $360^o = 1$

$$\theta_{n+1} = a\theta_n,\ 0 \leq \theta_n \leq 1 \tag{1.1}$$

where a is a fixed real number. Note that $0 \leq \theta_n \leq 1$ means that we are only interested in the decimal part of the angle θ_n. This truncation of the integer part is denoted

$$\theta_{n+1} = a\theta_n \ mod\ 1,\ n = 0, 1, 2, \dots\ . \tag{1.2}$$

Of course $\theta_1 = a\theta_0 \ mod\ 1$, $\theta_2 = a\theta_1 \ mod\ 1, \dots$ and in general

$$\theta_{n+1} = a^n\theta_0 \text{ mod } 1,\ n = 0, 1, 2, \dots\ . \tag{1.3}$$

Consider the case $a = 10$. Then the initial angle $\theta_0 = 0.31459\dots$ (digits of π) jumps to $\theta_1 = 0.14159\dots$, $\theta_2 = 0.4159\dots$ and so on. The initial angle $\theta_0 = 0.314314314\dots = 0.\overline{314}$ maps to $\theta_1 = 0.14\overline{314}$, $\theta_2 = 0.4\overline{314}$, $\theta_3 = 0.\overline{314} = \theta_0$. In the first case θ never repeats any pattern for ever (because π is an irrational number). In the second case, $\theta_{n+3} = \theta_n$, $\forall n$ and we have a period-3 pattern. We note the infinite number of initial conditions $\theta \in [0, 1)$ and the infinite number of patterns that arise from these initial conditions. Of course patterns are here just orbits on the circle.

In the Agincourt data we may imagine regular and irregular patterns as the answers to the questionnaires of a household evolve in time. Even with the richness of the data, we note that the binary answers at any moment from a long string of $0'$s and $1'$s may be regarded as a number, even a decimal number θ_n if we divide by the number of questions, and that they are all easily accommodated by the dynamical system (1.1) *if it applies*. Demographically, the situation might be more complicated because many households may be on the same pattern, *or orbit*, if (1.1) applies.

In this thesis, we attempt to define 'orbits' in some socially meaningful way, and to identify dynamical systems (if any), that might govern the orbits. This study is inspired by dynamical system theories [14, 15, 16].

1.2 Review of deterministic mathematical modelling

This section is devoted to a short review of a large and growing literature on mathematical modelling in the social sciences, [17, 18], [19] and [20] are reviews. We do this here by briefly illustrating the types of mathematical models, to better contrast our approach.

Models can be classified in two main categories including

1. Deterministic mathematical studies of individual-level social dynamics
2. Deterministic population-level dynamic models

1.2.1 Deterministic mathematical studies of individual-level social dynamics

There are few deterministic mathematical studies of social dynamics at the individual-level. Lewin [21, 22], Barber [23], Helbing and Molnar [24], Pearson and McCartney [25, 26] proposed various deterministic approaches.

Lewin [22] introduced a new approach for modeling individual behavioural changes. He argued that behavioural changes are driven by so-called *social fields* or *social forces*. The idea of identification of social forces was taken further by [23, 24]. The model developed in [24] considered individual pedestrian behaviour. The most sophisticated model of pedestrian behaviour is perhaps that of Helbing and Molnar [24]. This illustrates is an application of the ideas of Lewin [22].

Helbing and Molnar give coupled Leugevin equations

$$\frac{d\underline{\omega}_\alpha}{dt} = \underline{F}_\alpha + \quad \text{fluctuations} \tag{1.4}$$

$$\frac{d\underline{r}_\alpha}{dt} = \underline{\omega}_\alpha(t) g(\frac{v_\alpha^{max}}{||\underline{\omega}_\alpha||}) \tag{1.5}$$

with

$$g(\frac{v_\alpha^{max}}{||\underline{\omega}_\alpha||}) = \begin{cases} 1 & \text{if } ||\underline{\omega}_\alpha|| \leq v_\alpha^{max} \\ \frac{v_\alpha^{max}}{||\underline{\omega}_\alpha||} & \text{otherwise.} \end{cases} \tag{1.6}$$

Here $\underline{\omega}_\alpha$ is a preferred velocity of pedestrian α, v_α is actual velocity and $\underline{r}_\alpha$ is distance from desired destination. They construct forces $F_\alpha(t)$ in potentials with exponential rates of spatial

change. With choice of these rates they can simulate interaction of many pedestrians, for many positions $\underline{r}_\alpha(t)$, over time.

We note the individual level of study and the ad hoc form of the model. The results of simulations over many pedestrians can in principle be compared with observation of many individual pedestrians. In general, the parameters of the model are adjusted for agreement.

Pearson and McCartney [25] inspired by the methods of Barber [23] used similar social and psychological models to develop deterministic mathematical models of individual dynamics. Stochastic models have been presented in [27]. For a given set of n individuals in a network, Pearson and McCartney model the dynamic behaviour of the interactions between these individual by the following equation

$$\dot{x} = A(x)\Phi(x) \tag{1.7}$$

where x is the vector containing the nodes of the social network x^i. The matrix $A = (a_{ij})$ with a_{ij} constant or function of x. The map Φ is defined as $\Phi : x \to (f(x^1), f(x^2), \ldots, f(x^m))^T$ where f is a cubic polynomial with three real roots with negative derivatives at the two roots 0 and 1, and $m = n(n-1)$. In particular, A and f are models in six parameters, designed to introduce attraction or reaction between individuals. Again, Pearson and McCartney can simulate the passage of individual interactions through the network. Their paper [25] was fully deterministic, individuals merely distinguished by their initial conditions in a state space. We note again the ad hoc model and the necessity to adjust parameters to achieve comparison with (future) experiment.

These two models of individual behaviour contrast strongly with the individual dynamics that arise from a well-posed longitudinal data which give detailed and precise knowledge of the social unit, focused within a precisely stated and relevant purpose. It is of interest to extract the dynamical system, if it exists, directly from the individual data. We know of no such studies.

We note the use of utility functions as the determinant of individual behaviour in mathematical models in the social sciences [20, 9]. These are very high level descriptors, in a few parameters, of dynamics of individuals and are a statistical attack on the data that can allow direct access to diffusion [28, 29, 30] and other differential equations. They are then of great interest, in the context of this study. However, they are again ad hoc model, without detailed characterization of the underlying dynamics. The human variables used are few (e.g. in [31], three forces only are explained by words like 'persuasive', 'compromising' and 'avoiding' as compared to the approximately 200 questions of the Agincourt data). Also, such models make the assumption of randomness of

individual behaviour. Again we are encouraged to make direct use of 'experimental' data.

1.2.2 Deterministic population-level dynamics models

Deterministic population dynamics is widely used (reviewed in [32, 20, 33]). These methods typically probe the truth of growth rates of sub-populations, but do not probe downward to the individual. Ordinary differential or difference equation models of population dynamics are of course deterministic models used for population projection [33].

The trivial model

$$\frac{dN}{dt} = \alpha N \ , \tag{1.8}$$

for population numbering N individuals, and growth rate α, does not acknowledge any particular social force, and is remote from longitudinal data such as that of Agincourt.

The well-known logistic model [34, 35] is given by

$$\frac{dN}{dt} = r_0 N(\frac{K-N}{K}) \tag{1.9}$$

where N is the population size, r_0 is the population growth rate, and K is the capacity of the environment. The model (1.9) is an example of interaction of a population with a resource; in this case the social force is directly imposed but is very simple.

Chaotic dynamical systems [33, 36, 37, 38] are of interest as models of unpredictable population behaviour. These can model the stochastic behaviour of real phenomena perhaps capturing random fluctuations in population numbers. We have given a famous mathematical example of a chaotic system in (1.1).

Andrew [39] discussed the question of whether the behaviour of stochastic models of population dynamics agrees with equivalent chaotic deterministic dynamics [38]. Inspired by [40, 41], the methods of [39] were based on chaotic models including

1. the deterministic component of the single-species model

$$N_{t+1} = f N_t (1 + a N_t)^{-b} \tag{1.10}$$

where N_t is a population density in generation t, f is the per capita finite rate of increase, the constant a scales the density and b determines the form of the density dependence. Note if $a \ll 1$, $b = 1$, this approximates the logistic difference equation corresponding to (1.9).

2. the deterministic component of the host-parasitoid model

$$\begin{aligned} H_{t+1} &= fH_t(1+\tfrac{aP_t}{k})^{-k} \\ P_{t+1} &= cH_t[1-(1+\tfrac{aP_t}{k})^{-k}] \end{aligned} \tag{1.11}$$

where H_t and P_t, respectively host and parasitoid densities in generation t, f is the finite per capita rate of increase of the host, c is the number of parasitoid progeny produced per parasitized host, a is the area of discovery of parasitoid, and k describes the degree of aggregation over hosts of encounters with parasitoid. Similarities were found, making chaotic models of stochastic phenomena of interest in applications. We note that chaotic dynamics is a mixture of periodic orbits in many periods and of stochastic orbits. These dynamical systems warn us to search for periodic orbits in data. They also warn against the statistical modeller's assumptions of random data.

Detection of periodic orbits was discussed by Pierson and Moss [42] and So [43]. They built predicted models based on the recurrence of patterns in state space and were were successful in establishing the existence of periodic orbits in the study of the crayfish caudal.

Pawelzik and Schuster [44] developed a new method for predicting chaotic time series. This method extracted periodic orbits using time-series data of chaotic continuous dynamical system. Thus, to discover periodic orbits of many periods is to discover a property of chaotic systems. The periodic orbits are easy to simulate and they showed that these orbits can be used to construct models that could be used for projection.

Following the discussion of the above ideas, Paul and Edward [45] presented new techniques to detect periodic orbits in a dynamical system. They emphasised that the detection of periodic orbits in dynamical system *is a test of the presence of determinism* [46]. As in Pawelzik and Schuster [44], they also used experimental time series data to test their models. However, they limited their discussion to the determination of period one orbits. This is very simple and does not help to better understand the behaviour of the rest of the system. Developing their previous ideas [45], Paul and Edward [47] show that the best way of describing a dynamical system is through the detection of its periodic orbits. They prove that in a mathematical state space, periodic orbits are the equilibrium states [43]. Thus, if we are capable to detect all of the periodic orbits in this abstract dynamical space, then the systems temporal evolution can be predicted.

In this study, we are careful to detect period τ orbits.

The differential equations of diffusion [28, 29, 48, 30], or more deeply, the integral-differential

equations of kinetic theory [49], are deterministic [46] and have been used [50, 51, 52, 53, 54, 55, 56] to model population dynamics when the condition of each individual is a random walk [20]. Diffusion equations have the form

$$\frac{\partial N}{\partial t} = D\frac{\partial^2 N}{\partial x^2} \tag{1.12}$$

where t is time, x may be, for example, spatial displacement and D is the diffusion coefficient [57]. Diffusion equations are again coarse as they do not ask for experimental data for individuals, or for the forces that change individual states. D becomes a parameter to fit to observations. Further, forces on individuals need not be random, for example a new law can have lasting social impact. Helbing [20] notes their limitation to homogeneous conditions. For example they do not naturally apply to irregularly, geographically dispersed populations.

Kinetic equations [49] model simple social forces on a collective of individuals and models have been applied [20]. Helbing [20], proposes a simple master equation

$$\frac{d}{dt}P(\mathbf{X},t) = \sum_{X'(\neq X)} \left[\omega(\mathbf{X}|\mathbf{X}';t)P(\mathbf{X}',t) - \omega(\mathbf{X}'|\mathbf{X};t)P(\mathbf{X},t)\right] . \tag{1.13}$$

Here, $\mathbf{X}$ is one state of the system and the set of all states is denoted by Γ. $P(\mathbf{X},t)$ is a probability density over state $\mathbf{X}$ at time t, $\omega(\mathbf{X}|\mathbf{X}';t)$ is the transition rate from $\mathbf{X}'$ to $\mathbf{X}$ at time t. We note that the left-hand side of (1.12) and (1.13) have the same time dependence. In certain circumstances, equation (1.13) can be brought to the form of equation (1.12) by integrating over the state space. This shows that the kinetic equations contain more information of individual dynamics. Helbing has many sophisticated extensions and applications of (1.13). They are all derived from knowledge of individual dynamics, under the assumption of stochastic process [58]. Typically, the transition rates are modelled in terms of simple utility functions that capture behaviour of every individual. Again we see ad hoc modelling, all be it in a very convincing framework. The forces of interaction of individuals must also be modelled. We note for example interest in the psychological "forces" of persuasion, avoidance, compromise.

Mathematical deterministic models can be developed from two different approaches [59, 60, 61]: a discrete time approach, and continuous time approach used above. Models from the discrete time approach fall into two categories: recurrence models and matrix models. The deterministic discrete time matrix model was first introduced by Bernardelli [62], Lewis [63] and Leslie [64]. Important contributions to the study of deterministic discrete time recurrence models can be acknowledged in

the work of Dobbernack and Tietz [65]. These again involve high level modelling by rate equations.

Inspired by Bernardelli, Lewis and Leslie, Caswell was the first to introduce a detailed stage-classified demographic theory [66]. His particular focus was on stage-classified populations models that were later developed in [67]. The theory included linear and non-linear, time-invariant and time-varying, deterministic and stochastic models defined as follows. The size of the population at time t is given by the vector $n(t) = (n_i(t)), i = 1, 2, 3 \ldots, s$ where $n_i(t)$ gives the number of individuals at state i, at time t. The dynamics are specified by a $s \times s$ population projection matrix A_t, where

$$n(t+1) = A_t n(t) \tag{1.14}$$

Note that $A = (a_{ij}(t))$ where $a_{ij}(t)$ gives the rate of change of individuals from state i to stage j at time t. Thus the a_{ij} describe the vital rates which may vary through time.

Age is always regarded [68] as a basic demographic variable used to describe the state of an individual in its life cycle. Caswell used age to define the stage of the individual in his models. There are of course, other demographic variables that influence individual behaviour and also provide deeper knowledge of the individual than the age does.

All the above models use reduced information of the population and its environment compared with the information of a longitudinal data of a questionnaire as found in detailed demographic questionnaires such as in detailed surveys such as from the Agincourt Demographic Surveillance Site [69, 70].

This thesis gives a precise theory of orbits of individuals (T_3^{Ag}, Chapter Four), not as modelled with utility functions, but as exactly revealed by the best question set that can be devised to probe the reasons for change in individual state relative to purpose.

Although the core aim of this thesis is to formulate, describe and demonstrate orbit theory, the discussion of the literature review is not complete without reviewing statistical approaches commonly used for longitudinal data analysis.

1.3 Review of statistical longitudinal data analysis techniques

A large body of research methods developed for data analysis are based on statistical methods [71, 72, 73, 74, 75, 76]. The application of these techniques includes the use of both cross-sectional

and longitudinal data. We review and discuss a statistical technique for longitudinal data analysis, namely survival analysis. This type of analysis consists of a range of statistical methods developed for investigating the occurrence and time of events. The terminology used to describe this technique varies across disciplines. For example it is known as survival analysis in biostatistics, failure-time analysis (reliability theory) in engineering and event-history analysis in sociology. The central objective of this type of analysis is to measure survival time. There are various approaches used in survival analysis. Our review is focused on the following:

1. Discrete Time Event History Analysis approach [75, 77, 78, 79]. The commonly used model in this approach is the discrete-time hazard model. This is a parametric regression procedure used in survival analysis to characterize the distribution of survival time using a set of variables for a given population. The simplified discrete-time hazard model [77] is given by

$$h_{ij} = \Pr\left[T_i = j | T_i \geq j\right] \tag{1.15}$$

where h_{ij} denotes the discrete-time hazard which is the fundamental parameter of the discrete-time survival process. It defines the conditional probability that an individual i randomly selected in the population will experience the event of interest in the time period j, knowing that he or she did not experience that event in the early period to j.

Another way of describing the distribution of survival time is by using the survivor function given by

$$S(t_{ij}) = \Pr\left[T_i \geq j\right] \tag{1.16}$$

where $S(t_{ij})$ similarly defines the survival probability which is the probability that an individual i will survive past time period j. A homogeneous population is assumed in (1.16). Note that the hazard function (1.15) determines the risk associated with each time period while the survivor function (1.16) cumulates risk period by period. This can be generalized [75] to a non-homogenneous population.

The application of discrete-time event history analysis [75] requires us to know whether or when study events occur. The event can be positive (e.g. birth), negative(e.g. death) or neutral (e.g. marriage). Data in this case are recorded in such a way that if the question of study has one of the words "When" or "Whether" then an event has occurred. The determination of the dynamics requires the following: (1) A target event which is change of a state of interest. States (e.g. married/divorced) are exhaustive and mutually exclusive. (2)

Identification of the beginning of time, i.e. an initial starting point when no one under study has yet experienced the target event - everyone in study population occupies one and only one of the possible states. (3) Sensible metric for clocking time. Here time should be recorded in smallest possible units relevant to process under study.

Note that in Discrete Time Event History Analysis [75], uncensored and censored subjects (individuals experiencing events outside the study time or never experiencing events) must be simultaneously incorporated into the analysis. The latter subjects inform event non-occurrence and thus provide information about event occurrence.

2. Cox-Regression (proportional hazard) approach [80, 81, 82]

   The Cox-proportional hazards model is a semi-parametric procedure used in survival analysis to investigate the association between the survival time and a set of independent variables of interest for a given population. The general Box-Cox transformation, introduced in [80] is given by

$$Y^{(\lambda)} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda} & , \lambda \neq 0 \\ \ln Y & , \lambda = 0 \end{cases} \tag{1.17}$$

   The assumption in this model is that for each parameter λ, $Y^{(\lambda)}$ is a monotonic function of Y, where Y represents data. The model (1.17) is used to discriminate between log, linear or more general functional forms. Applying the Box-Cox transformation to variables in the linear model leads to the Box-Cox regression model given by

$$Y_i(\lambda) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots \beta_k X_{ki} \tag{1.18}$$

   which can be summarised by

$$Y_i(\lambda) = X_i\beta + \epsilon_i \tag{1.19}$$

   where X_i is the $k+1$ vector composed of the regressors and $\beta = (\beta_0, \beta_1, \cdots, \beta_k)$ are unknown regression parameters. As discussed in [82], it is not clear which variable should be of interest, the transformed $Y_i(\lambda)$ or the original Y_i.

   Note that the Cox-regression model can be regarded as a transformation of the hazard as a linear function of predictors. In this case the hazard function of the survival time is given by

$$\lambda(t, x) = \lambda_0(t) e^{\beta' x(t)} \tag{1.20}$$

   where $\lambda_0(t)$ is a baseline hazard function, $x(t)$ is a time-dependent vector of covariate values and β' is a vector of unknown regression parameters. We note that this approach uses

continuous time. The continuous-time hazard function (1.20) is a rate not a probability as in (1.15).

The method presented in [81] considers the modelling of complex data which involve covariates or risk factors. In this study [81] the general proportional hazard model is given by

$$\ln \{ -\ln S(t)\} = g(t) + \beta' z \tag{1.21}$$

where $S(t)$ is the survival function, $g(t)$ denotes the logarithm of the integrated null (or baseline) hazard and the linear predictor $\beta' z$ expresses the relative effect of the covariates z in terms of a vector of estimable parameters β'.

The discussion of [82] concentrates on several techniques that are useful for forming point and interval predictions in regression models with Box-Cox transformed variables. The techniques, including mean squared error analysis, predictive likelihood as well as stochastic simulation, take account of non-normality and parameter uncertainty in varying degrees. The authors of [82] use Monte Carlo methods to examine small-sample accuracy and find indications that uncertainty about the Box Cox transformation parameter may be relatively unimportant. For certain parameters, deterministic point predictions are biased, and plug-in prediction intervals are also biased. Stochastic simulation, as usually carried out, leads to badly biased predictions. In [82], a modification of the usual approach renders stochastic simulation predictions largely unbiased. Note also that Cox's proportional hazards model assumes that the hazard ratio is constant over time. In [82], this assumption is taken into consideration.

3. Kaplan-Meier approach [83, 84, 85] Kaplan-Meier approach is a framework which provides a method for estimating the survival curve using observed data, the actual event times. The central objective of this technique is to estimate the survival probabilities $S(d)$ given by

$$S(d) = P[D > d] \tag{1.22}$$

where D denotes the time to death. If L denotes the time that a loss occurs, then D and L are assumed to be non-negative random variables. An observation in this model consists of a bi-variable random vector (T, δ) where $T = \min(D, L)$ is the time of observation and δ indicates the nature of the observation. It is defined as $\delta = 1$ or 2 if $T = D$ or L respectively. We note that this method is an extension of the discrete-time approach described above. This technique breaks the assumption of rounding event times to construct intervals as used in the

discrete-time event approach. The Kaplan-Meier estimate of the survivor function is the same as that of the discrete-time event method. Note also that there is no Kaplan-Meier estimate of hazard.

The study presented in [83] shows that in the theory of competing risks, the non-parametric Kaplan-Meier estimator plays an important role. The authors of [85] consider the performance of the Kaplan-Meier technique relative to a more flexible parametric model. They claim that the reduction in efficiency of the Kaplan-Meier survival estimator becomes negligible fairly quickly as the number of parameters in the parametric model increases. Note that parametric estimation of the survival curve may be necessary in certain extreme cases, such as when the sample size is very small.

Apart from the approaches reviewed above, there are other methods used in survival analysis. For example the Bayesian non-parametric approach to a (right) censored data problem has been developed in [86] with particular emphasis on medical survival studies. The core aim of this study was to obtain the predictive distribution for future observations based on previous data. The authors of [86] addressed prediction and argue that it plays a central role in the real decision-making process implicit in most of the medical survival studies. The authors of [87] use regression with frailty in studies of survival. The hazard function for each individual may depend on observed risk variables but usually not all such variables are known or measurable. This unknown factor of the hazard function is usually termed the individual frailty.

In summary we note that in survival analysis, the outcome variable (response) is event time, failure time or survival time which is associated with some other independent variables of interest. Events may be discrete (for example sex, race) and continuous variables (for example age or temperature). In these approaches we have to consider censoring of observations. Note also that these techniques account for within-subject correlation, the basic assumptions have to be adhered to for the estimates to be acceptable.

These statistical methods are mathematically sophisticated and yield numbers. These may be difficult for non-specialists to interpret. In this thesis we develop a new method for longitudinal data analysis that is fundamentally different in that it develops visualisable orbits among fitness states invoking full information of the data. We will be concerned to examine its own effectiveness and will not compare these statistical methods with our theory. Our method suggests a statistical analysis and we will give a brief discussion in Section 4.11 of the statistical methods.

1.4 Projection

The longitudinal survey is, fundamentally, a more powerful instrument than a cross-sectional survey. The difference between cross-sectional and longitudinal surveys is that a longitudinal survey involves a series of measurements taken over a period of time and allows extrapolation. The discussion presented in this section will not be complete if it does not include studies which address projection. As is known, the critical test of the theory resided in projection [88]. If a theory does not predict then it fails absolutely.

In all the time-dependent, population level models above, if parameters can be chosen to give good agreement between theory and observations, it is allowed to continue the computation forward in time, and so project future population level behaviour. This is the importance of deterministic models, because these extrapolations may be used to set policy.

The deep sociological reasons why the rates are as they are, can be hard to identify. Yet in the rich information of longitudinal surveys, if questions are well asked, we might hope to find the experimental evidence for those reasons or causes [89, 14].

1.5 Motivation and objectives

1.5.1 Motivation

The current research work builds the orbits of individuals in a "fitness space", arising out of longitudinal surveys in the Social Sciences. This is a new approach. We aim to use given longitudinal data to determine deterministic orbits and embed in a dynamical system that can induce again random or periodic orbits [43] as in (1.1). This contrasts with Helbing [20] who assumes a random walk with uniform statistics for all individuals.

Orbits compare with statistical methods where there is no visualization and instead the demographer communicates through quite sophisticated "moments of distribution functions", for example. Such numbers as means and variances are perhaps hard for the layman to understand. It is easily understood that they are drastic reduction in information compared to visualized orbits.

Finally, we know of no *direct* analysis of experimental, time-dependent social data, that is, from longitudinal survey data, that induces the mathematical "laws" that take the study population

from one year to the next.

1.5.2 Objectives

The specific objectives of this research are as follows:

1. To build an orbit theory directly from empirical data to complement the theoretical ad hoc models of the above review, which contain parameters that are adjusted to the data
2. To use observed individual behaviour to complement the population level models of human behaviour
3. To have a mathematical computational system that would be able to predict future states of a social process and system
4. To implement our theory, for longitudinal data of the Agincourt Health and Demographic Surveillance Site.

1.5.3 Outline

The outline of this thesis can be summarised as follows. Attention is focussed on the data set of the Agincourt Health and Demographic Surveillance Site in Chapter Two. In Chapter Three, the models that will be used as tools for our analysis are described. In particular, a discussion based on a comparative approach between the real and simulated data of the individual-level analysis is presented. Chapter Four will then generalise the discussion of Chapter Three to the population-level. Chapter Five will concentrate on the discussion of population projection using the techniques developed in Chapter Three and Chapter Four. Finally, the general conclusion including a brief discussion of future work will be presented in Chapter Six.

Chapter 2

Data Preparation

2.1 Introduction

In this chapter emphasis is placed on preparation for analysis of Agincourt data [69, 70]. The methods developed to prepare Agincourt data can be applied to any population data. We make use of Python, Octave and Matlab programming tools to develop the software and analytic techniques for analysis of Agincourt and simulated data in this thesis.

There are various issues that researchers, involved in collecting Agincourt data, faced during the data collection process [90]. It is possible that among the data properties there are some which are complicated to understand and which can be explained if we have a better knowledge of the place they are collected. For example, the concept of poverty is differently regarded (or measured) in different environments. Thus, an individual who resides in an urban area is more likely to consider himself or herself as a poorer man or women, compared with one who lives in a rural area. In this case, the measurement of poverty will slightly differ for such different living environments. In this chapter, we provide not only the description of the data but also gives a brief description of the study site where the data were collected.

2.2 Description of the Agincourt District

The Agincourt research site constitutes a sub-district of Bushbuckridge district, Mpumalanga Province and is located in the remote, rural north-east lowveld of South Africa. It is close to the

eastern border with Mozambique. The Agincourt study site is rural, with poor infrastructure and services.

In 2001, the population of the Agincourt HDSS was about 69,000 persons residing in about 11,300 households which were distributed over 21 villages with both traditional and civic leadership [69, 70]. Almost a third of the study population was constituted of Mozambican immigrants. The study site covered $402km^2$. The area was densely populated with around 175 people living on each square kilometre. People in the Agincourt study population are largely Tsonga-speaking. One third of the Agincourt population is composed of Mozambicans who also speak Tsonga. The more detailed history and evolution of the Agincourt study population have been described elsewhere [91, 92, 69, 70].

The following is a summary of the demographic results from the decade $1992-2002$ [69, 70]. The sex ratio was about 93 for the whole population (80 for the permanent) and 96 among Mozambicans. The dependency ratio was 75 overall but 94 among Mozambicans. A net migration from the Agincourt sub-district was of 1% of the population per year. A crude natural increase of 2% per year resulted in annual population growth of 1%. Most permanent migration that occurred within Agincourt was a result of family formation and dissolution; 15% however was to nearby towns and a further 6% to cities. Temporary (labour) migration maintained strongly high rates for men (60% in men in the age group $35-54$ years) and growing proportions of adult women (from around 5% of women in the age group $15-34$ in 1997 to 19% in 2001). Close to half (43%) of the women who became temporary migrants in 1999 or 2000 had at least one child. These mothers were likely to be older, divorced or separated, with primary or tertiary education, and residing in female-headed households with high likelihood of co-residence with a grandparent or sibling. Mothers least likely to migrate were those living in a nuclear household.

The overall unemployment rate of the Agincourt population was about 40% and can be described as follows. There was a clear peak in unemployment around age 25; employment peaked at age 40 and involved 80% of men and 50% of women. Young adult women were more likely to be in formal sector employment; older married women tended to work as entrepreneurs or retailers in the informal sector. Although mining remained the main employer of migrant men, this was no longer the case for all employed men.

The average household size decreased significantly from 7 in 1993 to 6.6 individuals per household in 2000; this reflected a distributional shift to increasing numbers of smaller households.

Estimates of annual transition probabilities showed that household type changes regularly. The proportion of female-headed households increased significantly from 29% in 1992 to 33% in 2000; notably, given marked increases in AIDS-related deaths among adults, the proportion of skip-generation households remained low ($< 1\%$).

2.3 The Agincourt research

There are several research groups, with topics that use data from the Agincourt HDSS. This helps to build productive ties across scientific disciplines. These include collaborations with both African and international centres of excellence. At the local-level, we note for instance, a partnership in research between the Schools of Social Sciences, Public Health, Statistics, Economics, and Applied Mathematics at the university of the Witwatersrand. At the international-level we note collaborations (WBCA) between the Wits Demography and Population Studies Programme, the university of Boulder, the University of Colorado and the African Population and Health Research Centre (APHRC). These collaborations contribute to the development of interdisciplinary and multi-method strengths and a *growing methodological approaches*. The Agincourt HDSS is closely linked with highly productive rural and urban longitudinal research initiatives. In particular, it is supporting studies examining relationships between migration and child mortality. A study [92, 69, 70, 93] shows that temporary migration of mothers did not appear to increase the mortality risk for their children under age 5, in fact a small protective effect was found; 40 children born into Mozambican (former refugee) households have significantly worse mortality, this particular study concentrated in the $1-5$ age group. Preliminary findings from a survey [92, 69, 70, 93] of randomly sampled male labour migrants and locally resident men ($s = 857$) indicate that, while migrant men are at risk for multiple sexual partners, the highest risk was in locally employed men; lowest risk was among migrants who returned home monthly.

Children in the same household as their parents attained higher levels of schooling, as did children whose fathers were migrant workers [69, 70]. Female-headed households were not associated with lower levels of education [69, 70]. The last two results will be of interest in our analysis.

2.4 Description of the Agincourt HDSS database

The Agincourt HDSS maintains a database which consists of a relational database model that is a longitudinal representation of population data in the study site [69, 70]. The data is captured and upgraded through a computer program, Microsoft Structured Query Language (SQL) Server 2005.

The Agincourt database went through many modifications which contributed to its improvement since 1992. This improvement can be summarized as follows: The baseline census was stored in Foxpro in 1993, then converted into Microsoft Access in 1995, and followed the upgrades of the Microsoft Access software until 2001 when it was converted into SQL Server 2000. The current relational database model has been in place since 1999.

A number of tables including observations table, individuals table, residences table, locations table, memberships table and households table are designed to capture information in the database. These tables are related each other to form the relational structure of the database. Along with the basic demographic variables: birth, death and migration; there are other important demographic variables including pregnancy and marriage which are used to record the moves of an individual in or out of the database. Access to the tables is achieved through SQL commands called queries.

The Agincourt HDSS database records information at two different levels. At the individual-level, the information related to individuals is captured for all individuals who live in the study site and are members of the households in the study site. In particular, data on births, deaths, migration are collected and updated annually for all household members in the HDSS. The database also records information at the individual-level concerning the following demographic variables: cough status, child care grants (data available only for the following census modules: $2002, 2005$ and 2008), child morbidity (data available only for the census module of 2006), education status (data available only for the $1992, 1997, 2002, 2005$ and 2008 census modules and for new individuals), fatherhoods (data available only for the 2007 and 2008 census modules), father support status (data available only for the 2007 and 2008 census modules), health care utilization (data available only for the 2003 and 2006 census modules), labour status (data available only for the $2000, 2004$ and 2008 census modules) and stroke status.

The database contains a verbal autopsy table which records information used to establish the probable cause of death in areas lacking a vital registration system. At the household-level, data

collected include asset status (available only for the following 2001, 2003, 2005, 2007 census modules) and food security status.

The data collected is based on a repeated census taken on December 31^{St} of each year for which data is available. The baseline of the HDSS was established in 1992. These data are recorded as status observations at the census round immediately preceding the date of cross-section (for example name of deceased, date of death, cause of death).

The time step which can be year or month, is an important concept that we need to address here, because information for each demographic event is not recorded or upgraded on the same time step basis. We note that the date is recorded for all basic demographic events (births, deaths, and migrations) in the database.

If the date is estimated, it is then indicated in a separate field. Observations are time stamped with an observation date. This gives the date at which an interview took place, which is the date at which the data was recorded. All events and status observations can be linked to an observation date. Residences and memberships are recorded as episodes with start and end dates. As described above, a residence is the period of time an individual spends located at a specific dwelling. A membership is the period of time that an individual remains a member of a household. The events that start or end a residence or a membership are recorded. Status observations are repeated, cross-sectional measures and the dates of observation are recorded. These observations are repeated at different periodicities in the database and some have only been captured once.

Table 2.1 displays the associations that were made between census rounds (used below) and cross-section date.

2.4.1 Problem of question order in design of questionnaires

This study uses order of questions as an analytical tool. We distinguish between our analysis and the psychological effect of question order [94, 95] in a survey.

One of the most important aspects of designing a questionnaire [96] is improving the *response rate*, which requires providing the respondent with the motivation to complete the questionnaire and also give honest response [97]. The difficulties of questionnaire design are well known. In order to address these concerns, we note that respondents are sensitive to the context in which a question is asked, as well as to the particular words used to ask it. For instance, a questionnaire

Table 2.1: Annual cycle of census rounds.

Cross-section date	**Census round**
December 31st 1992	1
December 31st 1993	2
December 31st 1994	3
December 31st 1995	4
December 31st 1996	5
December 31st 1997	6
December 31st 1999	7
December 31st 2000	8
December 31st 2001	9
December 31st 2002	10
December 31st 2003	11
December 31st 2004	12
December 31st 2005	13
December 31st 2006	14
December 31st 2007	15
December 31st 2008	16

that asks straightforwardly about whether or not the respondent has tested HIV positive can be compared with a questionnaire that is prefaced by a series of attitudes of the respondent about HIV. It is important to decide in what order the questions will be asked. On the other hand, the impact of question order is often difficult to understand. There are [13, 75, 98] cases where it is showed that order of questions does not have an effect. It is noteworthy that there are almost no experimentally based general rules for ordering questions [75]. Models which are more complicated in terms of psychological or sociological interpretations have been studied using the same statistical techniques [10]. This study is the first that tries to use the question order as an important variable in the analysis of data. We must be careful to distinguish this from the design of questionnaire. We assume that questionnaires have been actually designed, that data has been cleaned and that responses are honest.

2.5 Description of the Agincourt data for the current analysis

Collecting data for orbit theory application has two main steps. We first need to design a questionnaire with respect the research topic of interest. Note that in principle, the format of data as required by orbit theory may differ from that of the database. For example, we need to construct a questionnaire whose questions have Yes/No answers. In the case of continuous data such as income, we bin incomes and then we must ask a set of questions, "is income in the i'th income bin?" Secondly, we might also have to infer Yes/No responses from multiple sources in the data.

2.5.1 Designing questionnaire

Before presenting our questionnaire, it is important to note that the four variables used in this dissertation, presented below, are all variables constructed from other information, and not the result of direct questioning. The questionnaire used in my thesis is described as follows.

$$\begin{aligned} q_0 &: \textbf{Was there a child without a biological mother in the household}? \\ q_1 &: \textbf{Was the head of the household a minor}? \\ q_2 &: \textbf{Was there an adult death in the household}? \\ q_3 &: \textbf{Was there a child not progressing well at school in the household}? \end{aligned} \tag{2.1}$$

This is a set of questions regarding the effect of household changes on children's educational outcomes, that we will present in chapter Four as a detailed demographic study. We acknowledge that this is a small subset of questions that might be asked regarding household change. In this thesis we will be concerned to develop a new method of analysis that might be applied for any number of questions. This subset will serve our purpose. We will argue that application of the method will best proceed by examination of combinations of small numbers of questions and that in this way bias in the choice of questions can ultimately be eliminated.

2.5.2 Description of data collection process

We give a detailed description of how the data related to the questionnaire (2.1) was collected. As mentioned above the data set used in this thesis has four variables. Three variables (q_0, q_1 and

q_2) have complete and available information each observation year and one (q_3) variable which is related to education information is only collected every 5 years in Agincourt HDSS [69, 70]. The first application of orbit theory presented in this thesis uses variables with a binary outcome. As noted above, this can be extended to continuous data.

Our scientific challenge in this thesis was to develop the method and from this point of view, it was reasonable to take a small number of "test questions" sufficient to get the mathematics to work. We are careful to experiment with numerical simulations as discussed in Section 3.3.2 of Chapter Three. The data set is based on a repeated cross-sections taken on December 31st of each year for which data is available. We used household information on the date of the cross-section. We used events occurring in the household over the 12 months prior to the cross-section date

For this data extraction, associations of Table (2.1) were made between census rounds (used to identify which education and residence status observation to use) and cross-section date. It is important to note that annual cycle of census rounds was only established after 1998 [91, 92, 69, 70]. As a result Agincourt data before 1998 is found not to be useful for the analysis presented in this thesis. It is not a specific selection criterion imposed by orbit theory to use data from 1998. The detailed process of using Agincourt database to answer our questions is as follows.

1. Before we describe the process of collecting data related to question q_0, we define some terminologies used in the wording of q_0. Throughout this thesis, the term "child" for the Agincourt population data will refer to a child of school going age, from age 7 to 16 who was member of the household over the period 1992 to 2007. The term "biological mother" for the Agincourt population data will refer to a the biological mother of a child of school going age. Thus, for every observation year, the answer to question q_0 is to check whether or not in each Agincourt household, among all children living in that household, there is a child without a biological mother. The absence of the biological mother from a household is measured by her number of residence months in that household. In order to capture both temporary and permanent migration of the biological mother, for each observation year we record a biological mother absent if the number of residence months of that mother is less than 12. To illustrate, suppose that household k has 3 children of school going age at the observation time t. Two children have their biological mothers in the household and one child does not have his/her biological mother in that household. In this case the answer to question q_0 for that household is Yes. We apply this process for every observation time.

In the Agincourt database the information to answer question q_0 is collected as follows. We used the residence status of biological mother. The resident status table has been combined with observation table to get the observation period. In addition to the added observation year column, it also provides the opportunity to rank the record to get rid of duplicates entries. The table generated by this process is joined to the individuals table to obtain the mother residents status. No data needs to be inferred for this question. We cannot from Agincourt data infer finer time intervals.

2. As before, let us first define some terminologies used in the wording of q_1. In this thesis, we define a "minor" as an individual of age less than 18 years. We use the Agincourt definition of "household head" as the person who is identified as a head by the older women in the household. In order to collect data related to question q_1 (household head is a minor), we used the information related to household head's age. The tables here are combined to generate the head of each household using household head relation found in the memberships tables. There were cases where two or more individuals have 'T' as household head relation value in a household for a particular. The rank function (in SQL) was employed to handle this. The rank gives the household head to the oldest person. This is decided as the oldest male rather than the oldest person will be the household head in the African context. This is added to the main table to get the household head. No data needs to be inferred for this question. Since q_0 can only be applied on an annual basis, inference on a finer time scale is of no use.

3. We define an "adult" as an individual of age 18 years and above. To answer question q_2 (adult death), we used the death table linked to individuals table. No data needs to be inferred for this question, because we cannot infer for questions q_1, and q_2.

4. To answer question q_3 (education of child), we used education status of child at school going age. This is measured as follows. We look at the total number of completed years of education (grade) of the child in each observation year where data is available. This is linked with the age of the child at that observation year to define the lag of grade behind "normal" grade. Note that in Agincourt, the average age of enrolment at school is 7 years. The education status table has been combined with the observation table to get the observation period, the grade and the age data. In addition to the added observation year column, it has a rank column to handle the duplicate records. We must recode the education column to years of

education completed. The table generated is joined later to the individuals table to obtain the educational years of the respondents. It is important to note that education data is collected only every 5 years in Agincourt HDSS [91, 92, 69, 70]. We cannot infer data for this question. Thus, let us consider the following scenarios for infering data for this question. Suppose a child's education status is grade 4 in 2002 and grade 8 in 2006, then we can assume that he or she passed every year between 2002 and 2006. In this case we could assign favourable (1) values to all the years in between. On the other hand, if we see that a child's education status is grade 4 in 2002 and grade 6 in 2006, then we can assume that this child failed to pass for two of the three unobserved years between 2002 and 2006. But we are unable to say when this happens. Note that it also becomes difficult to decide whether a child was not at school anywhere in time, or the child was at school but failed the same year of study twice. What really matters in our strategy is to clearly identify an unique observation year when change occurs for each state (e.g. educational default), which is not possible to determine in this case.

Concerning error in the data, note first that there is only a little literature relating to data quality and error rates in Demographic Surveillance Sites (DSSs) [99]. Errors vary widely. In [99] using Farafenni DSS, we find a 0.01% error after considerable data cleaning. In [91, 92, 69, 70, 99] in the Agincourt DSS, a 2% sample population is revisited but no error estimate is reported. In the absence of published errors we will give tolerance limits to our demographic conclusions.

In Chapter Four, we will define the measure of education progress for the Agincourt population data in order to answer question q_3.

We carefully select our questions according to purpose. In this thesis we consider the purpose

$$p_1 : \textbf{To investigate the effect of household change on child's progression in school} . \tag{2.2}$$

Note that q_3 directly identifies effect in purpose (2.2). Then questions q_0, q_1, q_2 are regarded as an initial set of questions that might cause educational default. Here we hypothesize that household change, with respect to questions defined in (2.1) can effect progress in school, that is that in a sub-population, household change *precedes* progress default. This introduces the possibility to analyse cause and effect in longitudinal data and suggests that questionnaires, in general, be posed in the fashion of p_1 : To investigate the effect of $A, B, C, \ldots$ on $E, F, G, \ldots$ Note that if all orbits are stopped at same moment, the state of all social units represents cross-sectional data. The pattern

of instantaneous data in a mathematical space hides causal relationships, relevant to purpose. We emphasis again this weakness of cross-sectional studies. If a cross-sectional study is not assisted by a longitudinal survey, sub-populations are identified by frequent occurrences of pairs of variables (here q_i with $q_j, i \neq j$), but they might have very different preceding states.

2.5.3 Data Sampling Methods

At this point, it is important to note that no statistical assumption is imposed on the sampling of the study population for the present analysis. However, there are some important data properties that must be taken into account in order to use the techniques that are proposed in this thesis.

Consider the population of the Agincourt households with a child at school going age (from 7 to 16) in the period between 1992 to 2007. Each household is observed with respect to the questions described in (2.1).

Figure 2.1 clearly displays the distribution of the number of households (members of the present study population) over their observation time. Thus, we find that the data collected for this study consist of 15603 households that are observed in the period between 1992 and 2007. We can also distinguish between the distribution of the number of Agincourt household with a missing value of question q_0 (2.1) from the distribution of those without a missing value for the same question. There are 6417 households without missing biological mother data. This represents about 41.13% of the study population.

Note that the first desirable criterion for the application of Orbit Theory is that there are no missing values in the data. Because of this important criteria, it is clear to see that the target population will then consist of the 6417 households which is about one third of the total study population.

On the other hand Figure 2.1 also shows the relationship between the number of households in the study population and their observation time. As this thesis presents a longitudinal study, the observation time for each household then becomes an important property to consider in the data collection process [90]. Thus, it is an additional information we use to select households from the study population. It is important to see for instance, that less than 1000 households without missing biological mother information have more than 10 observation points. Thus, the second desirable criterion for the application of Orbit Theory is to consider social units with a long observation time.

In addition to these two properties of the data above mentioned, for the specific case of the Agincourt data, another important property must be discussed in order to address the purpose (2.2). We can see as described in Table (A.1), that the information on education in the Agincourt HDSS is only captured every five years starting from 1992. Thus, in order to analyse the dynamics of an Agincourt household including the data of question q_3, which is related to education, the observation time of that household must not be less than 4 years. Thus, only Agincourt households with the observation time $l \geq 5$ must be included in the present study population.

With these strong arguments, we find that there are 3098 Agincourt households without a missing biological mother data and with the observation time greater than 5 years. Note that this population sample represents about 20% of the study population. This is the data sample that will be considered in the current analysis. We also find that for this population sample, the average observation time is $\overline{l} = 7.969$.

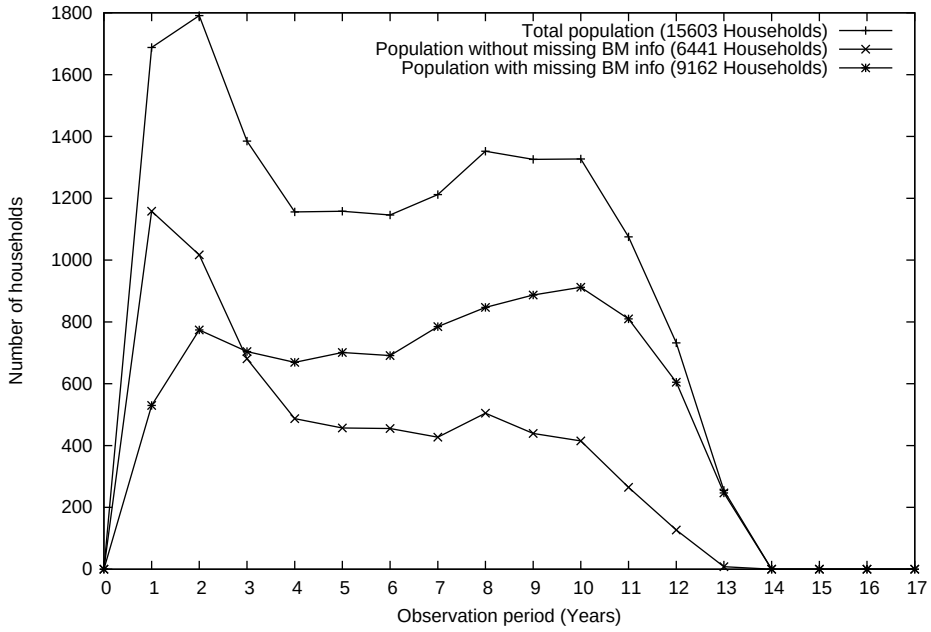

Figure 2.1: The distribution of the number of households over the length of the observation period, in the Agincourt population sample.

Figure 2.2 shows the distribution of the questions changing in our sample data. Note that the question about whether or not the household head is a minor is very stable (about 0.16% of changes in time). It shows that they are few households in Agincourt that are headed by minors.

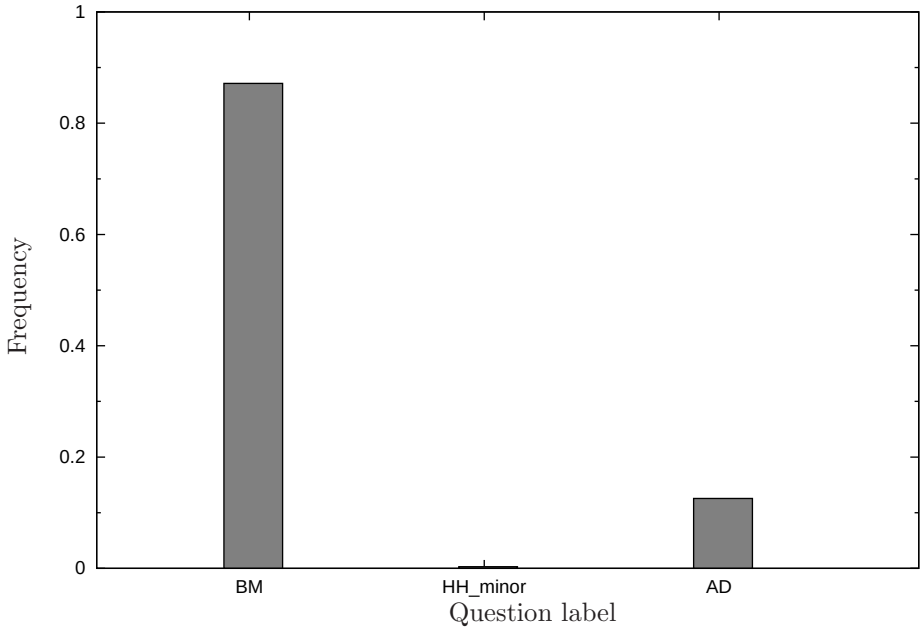

Figure 2.2: The frequency distribution of answer values changes for the Agincourt population sample of 3098 households.

The variable biological mother (BM) changes more often (about 86.72%) followed by adult mortality regime which changes for about 13.12% of the observation time.

In Figure 2.3, it is clear that 48.76% of the time nothing changes. Only one change occurs for about 47.9% of the time. The frequency of change of 2 questions is given by 3.32% of the time. The case where 3 questions change per time step is very small (about 0.02%) but it will be a useful illustration below to retain this question. Also, the average number of questions that change per time step is about $\overline{n} = 0.53$.

In Figure 2.4, we break down the distribution of Figure 2.1 for each observation year. The number of households (on the $y-$axis) is now expressed in percentage of the total population. From Figure 2.4, we can see that apparently no Agincourt household was observed in the following years: 1993, 1996 and 1997. Figure 2.4 also shows that more than 60% of the population sample has data available every year from 1998. The observation time with the appropriate data is taken from 1998 to 2007. Thus, we reduce the present population sample by only selecting the Agincourt households that meet the above criteria but now with data selected for the observation time from 1998 to 2007.

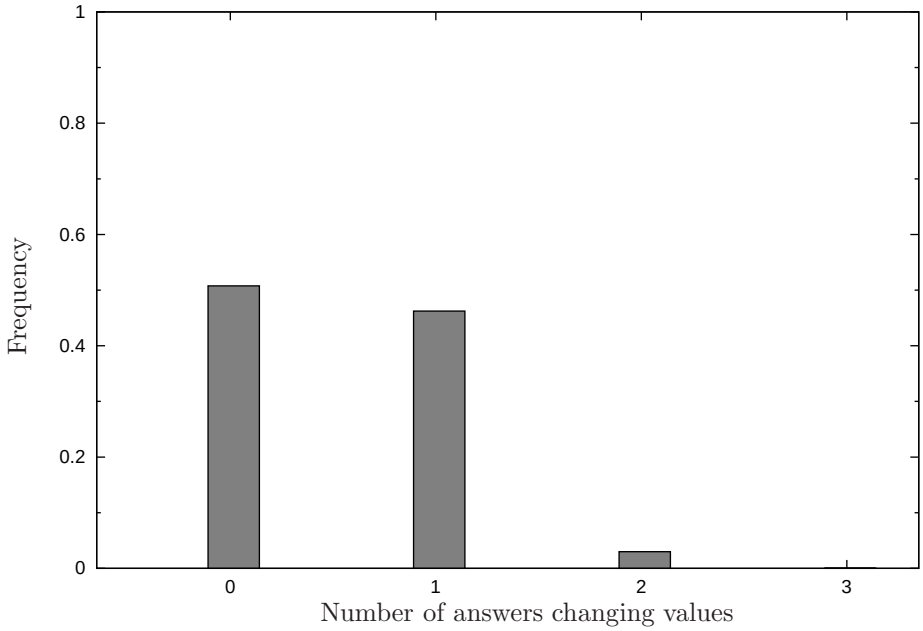

Figure 2.3: The frequency distribution of the number of answers changing values per time step, for the Agincourt population sample of Figure 2.2.

The final population sample consists of

$$s = 2669 \tag{2.3}$$

households which represents 17.10% of the study population. It is also important to redefine the distribution of questions changing answer values of Figure 2.2 and the distribution of the number of questions that change answer values per time step of Figure 2.3.

Using the same definition, but now applied to the new population sample (2.3), Figure 2.5 and Figure 2.6 respectively display the new distributions of Figure 2.2 and Figure 2.3. The population average observation time is calculated and it is given by

$$\overline{l} = 7.115 \ . \tag{2.4}$$

The new population average number of questions changing answer values does not change that much (an increase of about 10%), it is now

$$\overline{n} = 0.546 \ . \tag{2.5}$$

These are the parameters that will be used for the present analysis.

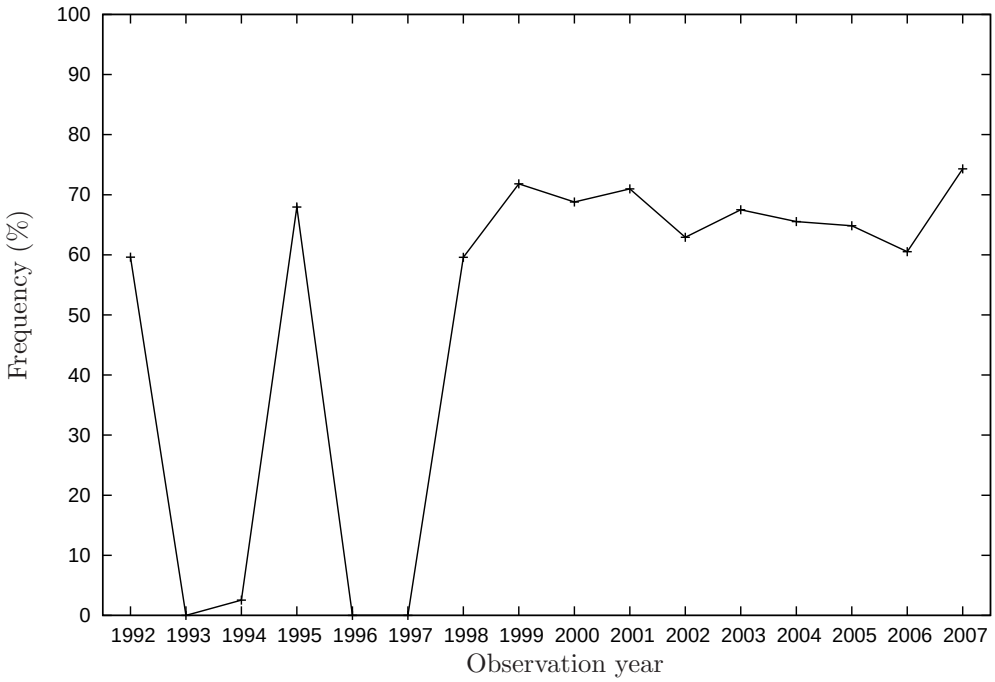

Figure 2.4: The distribution of the number of households per observation year.

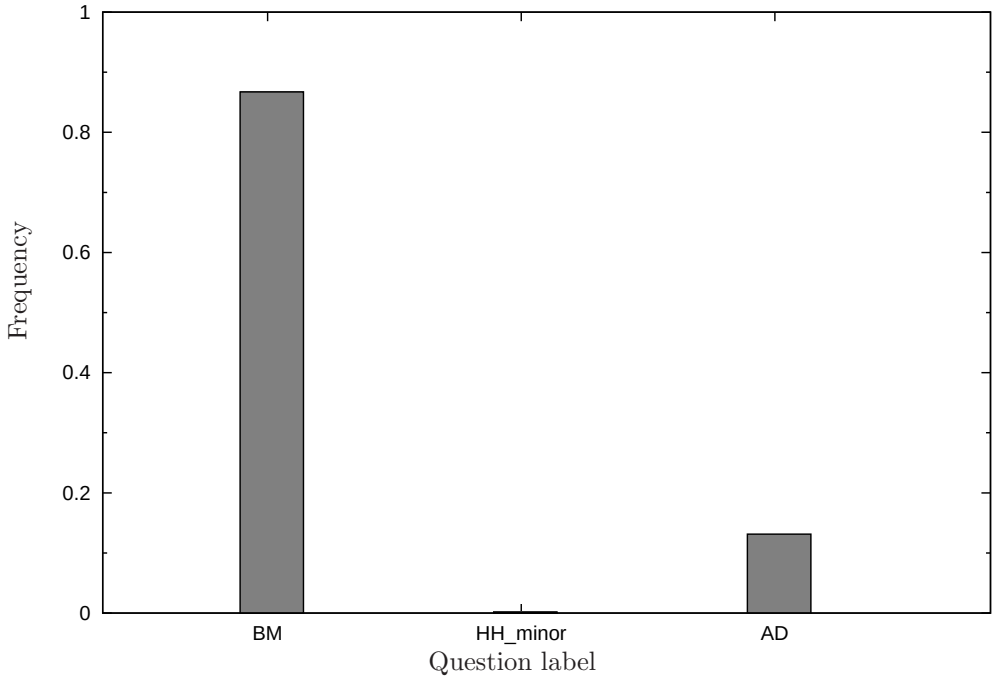

Figure 2.5: The frequency distribution of answer values change for the Agincourt population sample of 2669 households.

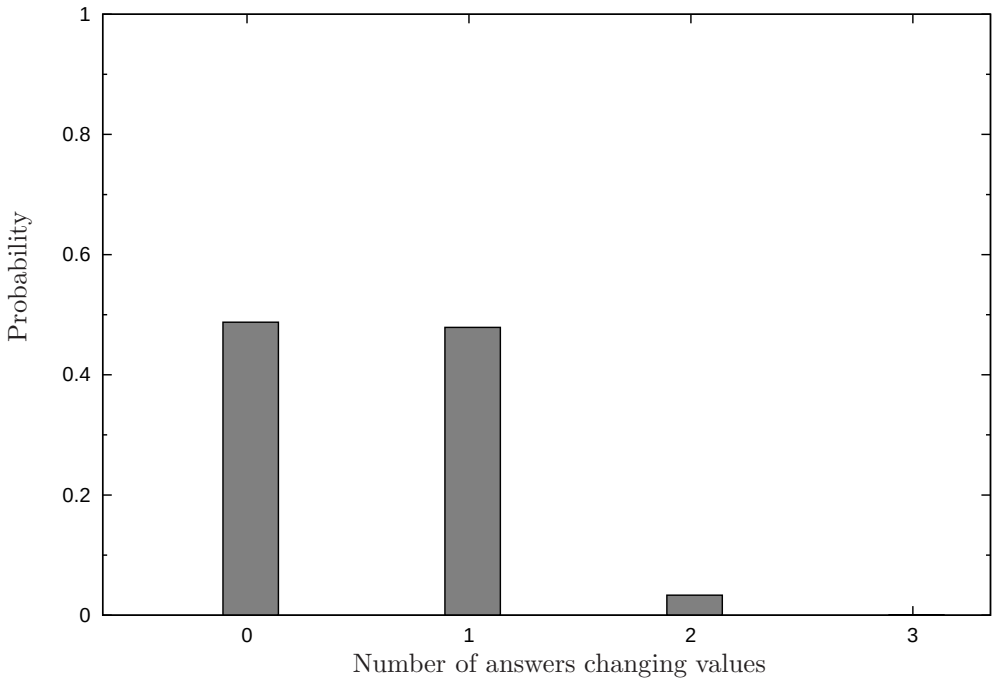

Figure 2.6: The frequency distribution of the number of answers changing values per time step, for the Agincourt population sample of Figure 2.5.

2.6 Conclusion

The objectives of this chapter were to address the data for our analysis. Thus, we clearly show that this process can be divided in two major parts. The first part of our task is the design of questionnaires and the definition of variables of interest. In order to achieve this, the study purpose must be well defined. This is simply because the study purpose directly suggests questions that might cause the effect of interest. The second part is to choose the data in such a way that they can be properly used. As in any scientific data collection [90], we understand that data must also be clean. In particular, for the purpose of the present study, we assume that households with missing values in the data are not accepted.

The discussion around these issues was particularly based on arguments that can be used to explain, for instance, some of the difficulties that researchers face when they collect data. In particular, we find that no Agincourt household was observed during 1993, 1996 and 1997. Thus, the distribution of the number of Agincourt household per observation year (see Figure 2.4) was helpful to sample the study population for the present analysis.

On the other hand, we have defined an important property of the data. This property is the order of questions (in the questionnaires) which is an additional variable that we must carefully take into account in order to use the current techniques.

Because the Agincourt data will be used in this thesis, in the discussion of this chapter we also presented a detailed description of the Agincourt study site and population. At this point, the models that constitute the present study can now be presented.

Finally, it is also important to note that throughout this thesis, any household identification number (id) that will be used represents an *anonymous* unique identifier for the household.

Chapter 3

General Orbit Theory of Longitudinal Data

3.1 Introduction

The analysis begins at the *social unit*-level just as physics begins with the understanding of individual particles by their orbits. Social units are here the households. The methods developed in this chapter will be extended to the population-level in the next chapter in order to provide a complete discussion of the present analysis.

The best example of a 'hard science' is classical, or engineering, physics. Newton's laws [100] are examples of dynamical systems, that is, systems where time is the independent variable. The information revealed by these laws is the 'orbit' of a particle, that is, the position and velocity at any moment, as affected by forces. These must agree with long sequences of measurements of the orbit variables of position and velocity.

In the social sciences, the social variables are not obvious. A longitudinal survey is regarded as a measuring devise. Questions have been chosen according to some *social purpose*. We have agreed above that we can formulate questions for *Yes* or *No* answers, that inform the purpose. If we code Yes = 1, No = 0 then we can say that an answer has a *value* 0 or 1. Thus the set of values at any moment, or the *response*, condenses as bits in a single binary string. Then, while the Newtonian particle is described by values of a few numbers, each having as many as possible digits, the social system is described by as many relevant questions as necessary with a single-digit,

1/0 answer value, which can similarly come together to form a long binary string. Both can be the basis for an orbit.

Physicists clearly define the space to visualize the movements of physical system orbits. In analogy, this study suggests methods that social scientists can use to build a new mathematical space in which the social system orbits can be visualized.

3.2 Fitness Space

Physical orbits define a *state* of the particle at each time in the space. Thus, position and velocity are measured and we say the state of the particle is known, in the space of ordinary geometrical coordinates. But we also have a sense of direction of movement (up-down, left-right, forward, backward) in the space and a sense of "how fast". In social sciences, we ask questions such that, at any moment, the state of the social unit is given, and the direction of change is also defined.

Consider the questions from (2.1) given by

$$\begin{aligned} q_2: &\quad \textbf{Was there an adult death in the household}? \\ q_3: &\quad \textbf{Was there a child not progressing well at school in the household}? \end{aligned} \tag{3.1}$$

In this example, we say that the *fitness state* of a household is determined by the set of answer values obtained by responding the questions defined in (3.1). This can change with time.

Changes of the answer values under the Yes/No answers do not guarantee us to have the dynamics with a consistent sociological content. To see this, consider the following equivalent questions in the area of public health. Assume that each questionnaire has one of the following questions

$$\begin{aligned} q_2: &\quad \textbf{Was there an adult death in the household}? \\ q_2': &\quad \textbf{Was there no adult death in the household}? \end{aligned} \tag{3.2}$$

The Yes/No responses to these questions have opposite values and can induce opposite dynamics. Either question is 'correct' and in general a questionnaire under Yes/No answers is not uniquely defined for its purpose. This means that we can imagine no absolute orbit, that is, absolute sociological truth, because measurements taken by independent sociologists will yield differing orbits, within the same given purpose of Q_t. This is in contrast to an orbit of a physical particle.

We make a fundamental assumption that, in the social sciences, the fitness state of a social unit is determined by its welfare in some social sense. In q_2, for example, adult death is less favourable. In q_3, educational default is less favourable. We code each answer value to "unfit" = 0, "fit" = 1, where "fit" represents a favourable state. This "fit" and "unfit" coding is important because independent sociologists will tend to make the same definition of fitness, at least where there is a common scientific study purpose (adult death might be a good thing where the adult is abusive).

It is clear that the "fit" and "unfit" coding is also important because it gives a sense of direction. Thus, if all answer values are zero, the individual is fully unfit. As zeros change to ones, the individual state becomes more fit. Further, it can do this more or less quickly, thus giving a sense of speed. Again we note that independent scientists will, *given the same social data and purpose*, tend to discover the same direction of movement. We say that the *fitness state* of a social unit is the current set of answer values as the above. Independent sociologists will *tend* to code in agreement. However consider the question q : **Was there death of an abusive father**? In this case it is not clear how to code fitness. We allow any coding and regard this as an hypothesis. The data will decide fitness relative to purpose. For example child progress at school may subsequently improve on the death of an abusive father. After analysis, independent sociologists will agree to code Yes = fit, in this case.

Now *change of fitness state* is the primary justification for longitudinal studies [101]. Thus, the value of q_2 may change from fit (1) to unfit (0). The value of q_3 may change from fit (1) to unfit (0). In the case adult death precedes educational default, we have a possible *cause and effect*. This is the second justification for longitudinal studies. Because adult death might cause delay in children's educational progress, we say that adult death is a (possible) *social force* [24] relevant to purpose. Each of the questions q_0, q_1, q_2 associates a social force.

Suppose that a well-posed questionnaire induces an orbit and reveals social forces and possible causes, for a particular state of a social unit, given some purpose. These qualities are directly analogous to the outcomes of Newton's Law for particles [100] which account for the advance of understanding in modern science and engineering.

Physics always seeks more and more digits to improve precision, for example, of position in geometrical space. Correspondingly, in this thesis, it is sought to allow social scientists to continually refine their questionnaires in order to fully capture the state of a social unit.

Suppose we have a population P of a finite number $s \geq 1$ of social units. Suppose that the

population is observed for a fixed length l of time period. Denote a questionnaire, administered at time t, by Q_t.

Definition 3.1. If this questionnaire, Q_t, contains $n_t \geq 1$ questions at time t then the questionnaire is defined by

$$Q_t = \{q_{t0}, q_{t1}, \ldots, q_{ti}, \ldots, q_{tn_t-1}\} \tag{3.3}$$

where q_{ti} is the *wording* of the i'th question at time t.

It is necessary to assume that the questionnaire defined in (3.3) remains unchanged for each social unit $k, k = 1, 2, 3, \ldots, s$ in the population.

Definition 3.2. The answer set given by a social unit k, at time t can then be defined as

$$A_t^k = \{a_{t0}^k, a_{t1}^k, \ldots, a_{ti}^k, \ldots, a_{tn_t-1}^k\}, \tag{3.4}$$

where $a_{ti}^k \in \{0, 1\}$ is the *answer value* to the i'th question, $q_{ti} \in Q_t$, codes so that 0 =unfit, 1 = fit, by hypothesis.

Definition 3.3. The fitness state of the k'th social unit at time t with respect to questionnaire Q_t (3.3) is defined by the set A_t^k as described in (3.4).

If we combine elements of A_t^k as a concatenated string, then directly from (3.4) we can now define the fitness state as follows.

Definition 3.4. The *binary sequence*

$$b_t^k = a_{t0}^k a_{t1}^k \ldots a_{ti}^k \ldots a_{tn_t-1}^k \tag{3.5}$$

equally captures the fitness state of the k'th social unit at time t.

As each answer has a 0/1 value, then it is clear that there are 2^{n_t} possible arrangements of answer values at time t. In a longitudinal survey, new questions may be added and, where questions are apparently irrelevant, or null, some may be deleted. Thus n_t can change with time.

Let

$$n_l = \max_t n_t, \tag{3.6}$$

which is the largest number of questions asked in Q_t to the present time l, so that there are 2^{n_l} possible states that a social unit can have. These states are points in the mathematical space of

binary finite sequences. Following the notation of [14, 102], we define a mathematical space for the present model as follows.

Definition 3.5. Let

$$\Sigma_2^{n_l} = \{b = (b_0 b_1 b_2 \dots b_i \dots b_{n_l-1}) | b_i = 0 \text{ or } 1\} \text{ , or, } \Sigma_2^{n_l} = \{0,1\}^{n_l} \text{ , } \tag{3.7}$$

be the fitness space of *finite one-sided* binary sequences of length n_l.

Note that $\Sigma_2^{n_l}$ is a sub-space of $\Sigma_2 = \{0,1\}^{\infty}$, the space of infinite one-sided binary sequence [14, 102]. The number 2 refers to the number of symbols (here because of the binary coding, we have only two symbols 0 and 1). Then the binary sequence b_t^k is a *single point* in $\Sigma_2^{n_l}$ that captures the state of a social unit k at time t. Notice that all knowledge of the social unit under purpose p is captured by this single point.

If we suppose that the questionnaire is applied to times $t = 0, 1, 2, \dots, l$, then we may define a fundamental object of this thesis as follows.

Definition 3.6. The *sequence of binary sequences*, or, sequence of points in $\Sigma_2^{n_l}$, denoted by

$$\Omega_l^k = (b_t^k)_{t=0}^l \tag{3.8}$$

defines an *orbit to time* l of the k'th social unit, in $\Sigma_2^{n_l}$.

To illustrate social changes that can be observed in $\Sigma_2^{n_l}$, consider the onset of HIV [103] that may occur at a random time to social units in a population P but once these social units are infected, the progress of this infection is perfectly predictable (while the social units live) because it is irreversible. Thus if the answer to a question q : **Have you tested HIV-positive**? will change from $0 \to 1$ in value at time t_1 and if all else goes unchanged, the *law of motion* for Ω_l^k is *deterministic* and known (it is $a_t^k = cnst.$, $t_1 \leq t \leq l$). On the other hand, opportunistic disease can be reversed, can strike again at a random time and so on, so that the *law of motion* might be random or, *non-deterministic* dynamics. In general we may expect mixed dynamics, but to understand the deterministic and random parts is presumably a contribution to social sciences. Newton's mechanics too, may involve, usefully, deterministic and non-deterministic forces. However, we make it clear that existing data is unambiguous, while a cause for each change of state may in principle be identified. Past history is always deterministic.

Social orbits are intrinsically more complicated than physical particle orbits. In physics there are just 5 known forces that can act on a particle [104] and usually just one acts at a time. In sociology, n_t questions can be independent and each will change independently. But then there will be n_t independent social forces acting on each question and thus many forces acting at any moment on the social unit. The full state of a social unit is captured by a set of responses. The full history of all this information of a social unit is captured by the orbit.

3.3 Equally weighted questions model

We make a fundamental assumption that questions have a weighting that affects welfare. We start by assuming that the order of questions is not important.

Suppose for simplicity that Q_t has n = constant, equally-weighted questions. Because of this property, Definitions 3.1, 3.2 and 3.4 then become respectively

$$Q = \{q_0, q_1, \ldots, q_i, \ldots, q_{n-1}\} , \tag{3.9}$$

$$A_t^k = \{a_{t0}^k, a_{t1}^k, \ldots, a_{ti}^k, \ldots, a_{tn-1}^k\} , \tag{3.10}$$

$$b_t^k = a_{t0}^k a_{t1}^k \ldots a_{ti}^k \ldots a_{tn-1}^k . \tag{3.11}$$

We recall that questions are asked so that the answers are "unfit/fit" and the value of an answer $a_i^k \in \{0, 1\}$ is coded so that 0 = "unfavourable" and 1 = "favourable". We suppose that questions are framed so that only this binary response is possible. Then at any time t, the state of the k'th social unit is defined by the ordered sequence given in (3.11).

Note that all information at time t of the social unit k is gathered in b_t^k. Suppose that the questionnaire has been used on a regular basis for l periods, for times $t = 0, 1, \ldots, l$. Then the complete history of the social unit k to time l is the sequence of binary sequences defined by the orbit

$$\Omega_l^k = \{b_t^k, t = 0, 1, \ldots, l\}. \tag{3.12}$$

Theorem 3.7. *The map*

$$h : \Sigma_2^n \to I = [0, 1] \subset R : b_t^k = a_{t0}^k a_{t1}^k \ldots a_{ti}^k \ldots a_{tn-1}^k \to x_t^k = 0.a_{t0}^k a_{t1}^k \ldots a_{ti}^k \ldots a_{tn-1}^k \tag{3.13}$$

is a homomorphism from sequence space Σ_2^n to the unit interval I.

This is proved in [14, 105, 106]. Thus h encodes b_t^k as a real number x_t^k, and we can uniquely switch back and forth from one to the other.

Note that there are 2^n possible points in Σ_2^n or on the unit interval that are visited by a trajectory of a social unit, that is,

$$b_t^k \in \{000\ldots0000\ldots1, 000\ldots10, 000\ldots11, \ldots, 011\ldots1\} \equiv \Sigma_2^n, \tag{3.14}$$

$$x_t^k \in \{0.00\ldots0, 0.00\ldots1, 0.00\ldots10, 0.00\ldots11, \ldots, 0.11\ldots1\} \equiv I_2^n, \tag{3.15}$$

where we add the interval 2^{-n} sequentially. So I_2^n denotes a discrete space of rational decimal numbers in base 2. Social units on the left are fully unfit, on the right are fully fit.

Definition 3.8. I_2^n is the *fitness space of real number states* of the questionnaire Q. Then we have defined an orbit on the unit interval.

$$X_l^k = \{x_t^k, t = 0, 1, \ldots, l\} \tag{3.16}$$

is the orbit of the k'th social unit on I.

The orbits Ω_l^k, X_l^k can be visualised in each space.

3.3.1 Distance and displacement for equally weighted questions

It is important to measure changes of the social system as it acts in Σ_2^n or I_2^n. If Σ_2^n or I_2^n is a metric space then it becomes simple to measure the movement of a social unit between two observation times t and t'. The importance of a homomorphism is that distances are given by the same real number in each space.

Each social unit in the population P jumps between points in Σ_2^n. Thus, following the techniques used in [14, 102] we define the distance moved by the k'th social unit in a time interval $[t, t']$ by

$$d(b_t^k, b_{t'}^k) = \frac{1}{n}\sum_{i=0}^{n-1}\left|a_{ti}^k - a_{t'i}^k\right| \geq 0 \tag{3.17}$$

where $b_t^k = a_{t0}^k a_{t1}^k \ldots a_{tn-1}^k$, and $b_{t'}^k = a_{t'0}^k a_{t'1}^k \ldots a_{t'n-1}^k \in \Sigma_2^n$ define the response sets at the two times and $a_{ti}^k, a_{t'i}^k \in \{0, 1\}$ are answer values, all of a social unit k. Then we can see that the distance is proportional to the number of differing answer values. This is intuitively satisfactory because

if none of them changes then $d(b_t^k, b_{t'}^k) = 0$ and if they all change it undergoes a maximal jump $d(b_t^k, b_{t'}^k) = 1$.

Recall that the weighting of each question is the same. If one answer value only changes, the distance remains unchanged ($d(b_t^k, b_{t'}^k) = \frac{1}{n}$) because it is independent of the position of the answer that changes value. Throughout this thesis, we will refer to this property as *non-uniqueness* of orbits. The analysis of dynamics related to this property will be referred to as *order-independent* dynamics. Then as in physics, orbits (3.8) jump in Σ_2^n, as measured by (3.17). The distance between two social units k and k' at time t is consistently defined by

$$d(b_t^k, b_t^{k'}) = \frac{1}{n} \sum_{i=0}^{n-1} |a_{ti}^k - a_{ti}^{k'}| \geq 0. \tag{3.18}$$

Then also, orbits (3.8) move relative to each other in a quantified way. The definitions (3.17) and (3.18) respectively reveal differences in the dynamics within and between the orbits as discussed by [107, 108, 109].

So far, there is no sense of direction of movement in Σ_2^n. Let us order binary sequences according to the so-called lexicographical ordering of 3.15 where the significant digits may be viewed as base 2 integers of increasing magnitude [14, 110].

Now a change $a_i : 0 \to 1$ must jump the point to the right in Σ_2^n, otherwise to the left. The new response in Σ_2^n can then be given a *fitness displacement* (as opposed to distance which is always positive), by

$$\Delta_{tt'}^k = \frac{1}{n} \sum_{i=0}^{n-1} (a_{ti}^k - a_{t'i}^k), \ t > t'. \tag{3.19}$$

This can be a positive or negative number and so the direction of movement in Σ_2^n is achieved. An orbit for the k'th social unit can apparently be visualized in I_2^n by writing down the space as in (3.15) $l + 1$ times, one below the other and then connecting states b_t^k, to b_{t+1}^k, $t = 0, 1, 2, \ldots, l - 1$. Here big jumps to the left (for example) correspond to many factors becoming unfavourable.

A relative displacement between two social units k and k' can be similarly defined by

$$\Delta_t^{kk'} = \frac{1}{n} \sum_{i=0}^{n-1} (a_{ti}^k - a_{ti}^{k'}). \tag{3.20}$$

Note that sociologists always seek to give each social change a sociological meaning. However, with the displacement $\Delta_{tt'}^k$, given in (3.19), we are unable to give a sociological explanation to each social change. To see this, consider the following example.

A fundamental objection is that while the state always has a position in (3.15), the distances (3.17) and (3.18) do not correspond to the distance jumped along (3.15). Thus if, say, $\Delta^k_{tt'} = -\frac{1}{n}$, then this does not necessarily correspond to a jump to the neighbouring sequence to the left in the lexicographical ordering. To see this, consider state $q^k_t = 10\ldots01$; then a change in value of the 1'st answer (on the left), or, the i'th answer (on the right) gives $\Delta^k_{tt'} = \frac{-1}{n}$ in either case. However in Σ^n_2 the second indeed jumps by one to its left neighbour but the first jumps left over many elements, *contradicting the meaning of distance*. There is clearly *no* re-ordering of (3.15) along a number line that will give a consistent distance or displacement for change in any one equally weighted question, if $n > 2$ (because it is not possible to place all elements adjacent to any given element). There is hence a sense of position and direction of movement in Σ^n_2, but no consistent displacement and no social relevance to its direction ($\pm$).

The non-uniqueness of orbits in the ordered space of all possible answer values, and the lack, so far, of a suitable metric space in which to visualize orbits, weakens the theory in comparison with physical dynamical systems. We must take this into account in order to enrich our theory.

To judge the usefulness of Σ^n_2 and orbits Ω^k_l, the displacements (3.19), (3.20) correspond perhaps to no more than the professional judgement of a physician of the relative state of health of the social units.

Definition 3.9. We define the *accumulated fitness displacement* by the map

$$_{acc}\Delta^k_t \equiv {}_{acc}\Delta^k_{t-1} + \Delta^k_{t-1t},\ t \geq 1 \tag{3.21}$$

where $\Delta^k_{tt'}$ is as defined in (3.19) and $_{acc}\Delta^k_t$ represents the accumulated displacement for the k'th social unit.

Note that at $t = 0$, naturally there is no initial accumulated displacement. Thus, we can assume that $_{acc}\Delta^k_0 = 0, \forall k$. Then the function $(_{acc}\Delta^k_t, t)$ can then be plotted, to visualize the accumulated favourable or unfavourable change from $t = 0$ of a social unit k.

3.3.2 Results for the accumulated displacements $_{acc}\Delta^k_t$

In this section, we present some of the results using simple illustrations and they will be followed by results from the Agincourt data. It is convenient at this point that we first describe the model parameters for all results that will be presented in this section. For simplicity, we

suppose throughout all illustrations that only one change of answer value, per time step, occurs and a constant number of questions, here $n = 26$ and $10 \leq l \leq 10000$. Note that in the present illustrations, we consider the time step being a second, a minute or a day. This can help to understand why we choose such a range for the observation period l.

The Agincourt household displacements that will be discussed in this section are randomly chosen from the study population and are anonymous.

In Figure 3.1 we give accumulated fitness displacement as defined in (3.21) for a social unit, in the case of a single favourable, switch-and-stay, deterministic answer value change (e.g. infection cure). We start the social unit at ${}_{acc}\Delta_0^k = 0$, assuming that there is no initial displacement. A social unit is now fitter in one answer value only. Perhaps an intervention has been successful or a disease naturally recedes.

In Figure 3.2 we show an example of a social unit which rests for some time before it starts a deterministic on-off-on-off-on... answer value change. We see regularly oscillating accumulated displacement. In this case, perhaps an intervention gives only temporary relief or a disease recurs.

In Figure 3.3 we have a social unit for which 8 of 26 answer values change (one at a time), with uniform probability of $\frac{8}{26}$. If the social unit starts with answers all zero valued for the changing answers (e.g. has eight pathologies as a result of some severe trauma), it is possible for ${}_{acc}\Delta_t^k$ to move favourably through at most two jumps as these pathologies are corrected (this may be deterministic [89] if the treatment is well-understood). Thereafter, the social unit wanders within the range ${}_{acc}\Delta_r^k = \pm 8/26$. If we are able to account for every jump, e.g. in terms of some preceding therapy, then the therapy is the cause of the jump, or, is the force acting on the social unit. This example shows on average a fit social unit, perhaps because many interventions are successful. ${}_{acc}\Delta_t^k$ is randomly generated and over longer times could become negative for many periods.

In Figure 3.4 we give a social unit in which all 26 answer values change (one at a time), with uniform probability of 1/26. If all answer values are initially zero, there can again be a run to a favourable state, but over long times the orbit will wander randomly over $x_r = \pm\frac{26}{26}$. As compared to *existing* longitudinal data, these figures appear to be substantially the available information of the social unit, so far as accumulated displacement is concerned. Perhaps an associated financial cost of a displacement might be useful but this must be included over and above Q_t as needed. In this example, we note the long study time. It is very rare (it has not occured here) that a run of 26 positive moves, occurs and the displacement reaches ${}_{acc}\Delta_t^k = \pm 1$.

In Figures 3.5-3.7 we give examples of accumulated displacement for three households in the Agincourt data set. In these examples (Figures 3.5-3.7) we consider the questions defined in (2.1).

We can identify particular social qualities that change. We note a tendency for household number 4325 (Figures 3.5) to fluctuate, rather than drift towards a relatively fit or unfit state. For this household, the answer to question q_0 was the only change in answer value in 12 years of observation. The present analysis clearly identifies the dynamics of this household to the movement of the biological mother (of child at school-going age) who is sometimes in, sometimes out of the household.

Figure 3.7 displays a maximum negative jump from the initial state of household 7150. In particular, all the answers to questions q_0, q_1, q_2 unfavourably change values at $t = 2001$, for this household. We also note a fluctuation about an unfit state as shown in Figure 3.6. Here we may also identify with the help of the data set the change from ${}_{acc}\Delta^{7150}_{2000} = 0$ to ${}_{acc}\Delta^{7150}_{2001} = -1$ as change from an initial ($t = 1995$) state defined by sequence 111 which remains unchanged until $t = 2000$ to sequence 000 at $t = 2001$; note 3 answers have changed value so that $\Delta^{A}_{2000-2001} = \frac{-3}{3}$; note that we identify absence of biological mother (unfavourable) from the household and that there is now a minor head (unfavourable) of the household. As a result, an adult death occurs in the household (unfavourable).

An oscillation related to one question change in answer value can be observed in Figure 3.6. For this household 5873, the answer to question q_0 was the only change in value in 9 years of the observation period.

Another social unit might show an unfavourable drift towards an unhealthy fitness state. The social unit characterized by the most zeros is in the less favourable state. The distance (3.18) and displacements (3.20) apply to two social units k and k' at time t. Performing this analysis for both social units, it may happen that a stable distance arises, indicating two possible degrees of health. The appropriate technique for analysing sub-populations in this dynamics is not geometric (Δ_t is not an orbit) but a sort. Thus for n questions for l time steps, there are 2^n possible orbits. Sort the numbers of social units that follow each orbit. This can in principle reveal the most populous orbit which becomes the *typical* orbit. It may be unrealistic to expect any two social units to be always in exactly the same fitness state. Orbits may be defined to be close if they differ by only $n_{dif} \ll n$ answer values. Then sort all social units into the population such that $d(b^k_t, b^{k'}_t) \leq \frac{n_{dif}}{n}$, for all t. This sort is just a mathematical jump in the abstract space Σ^+_n. This appears to be the

substantial demographic information, so far. It can be of obvious use.

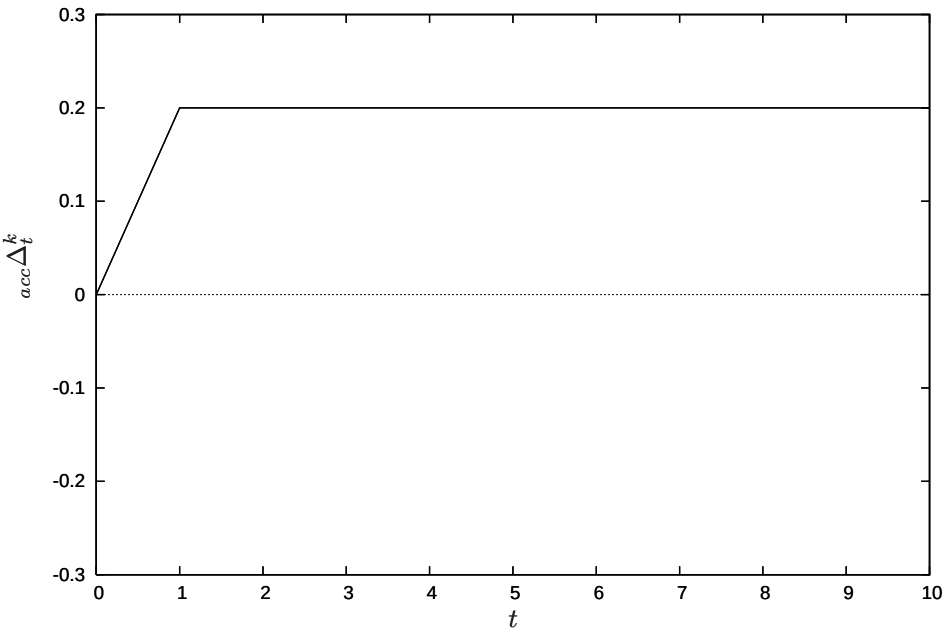

Figure 3.1: Accumulated displacement for only one switch-and-stay answer value change.

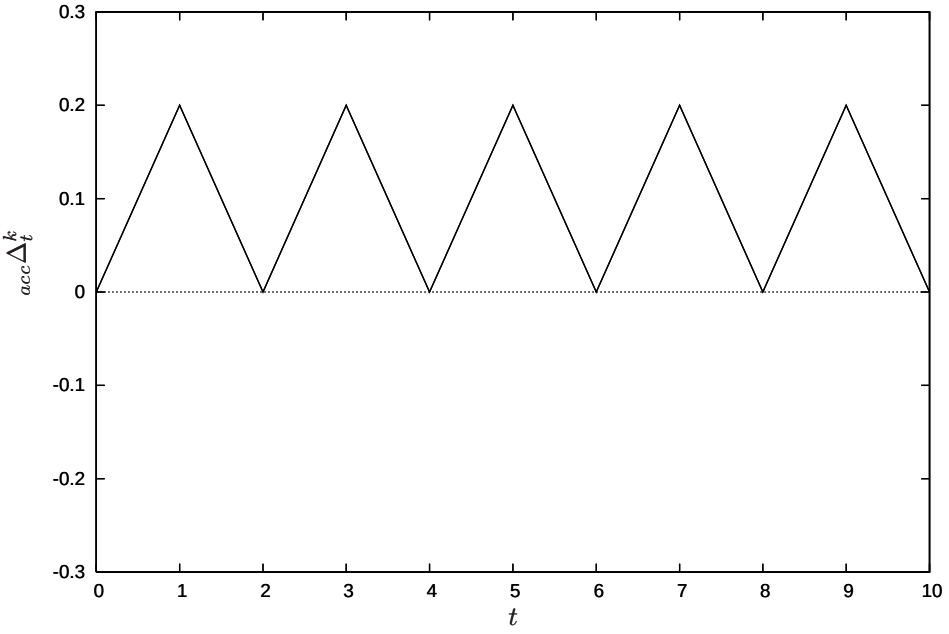

Figure 3.2: Accumulated displacement for only one oscillating answer value change.

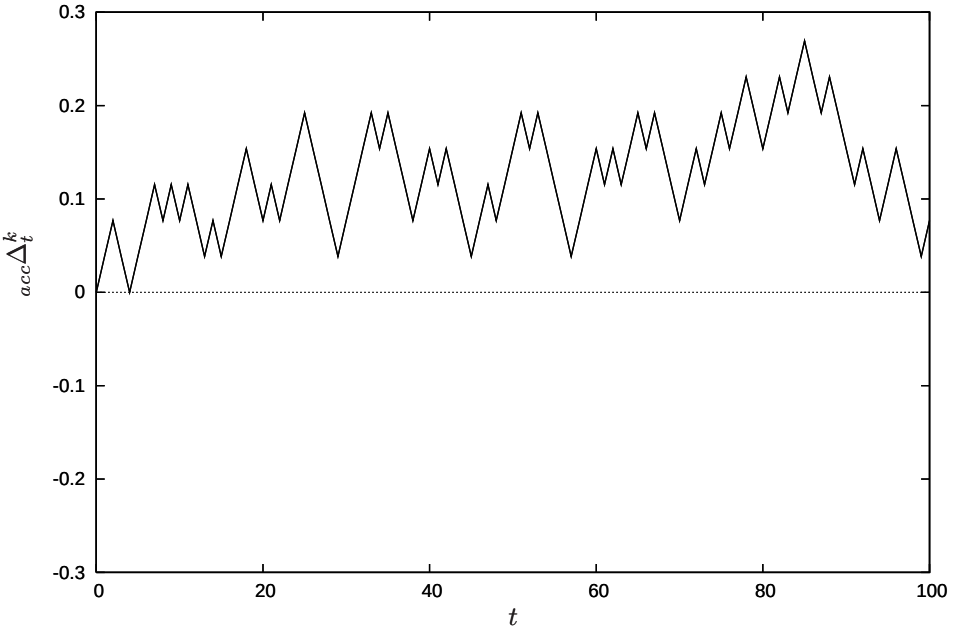

Figure 3.3: Accumulated displacement for a uniformly random change in 8 of 26 answer values. The bias to positive accumulated displacement is an accident of short derivation time.

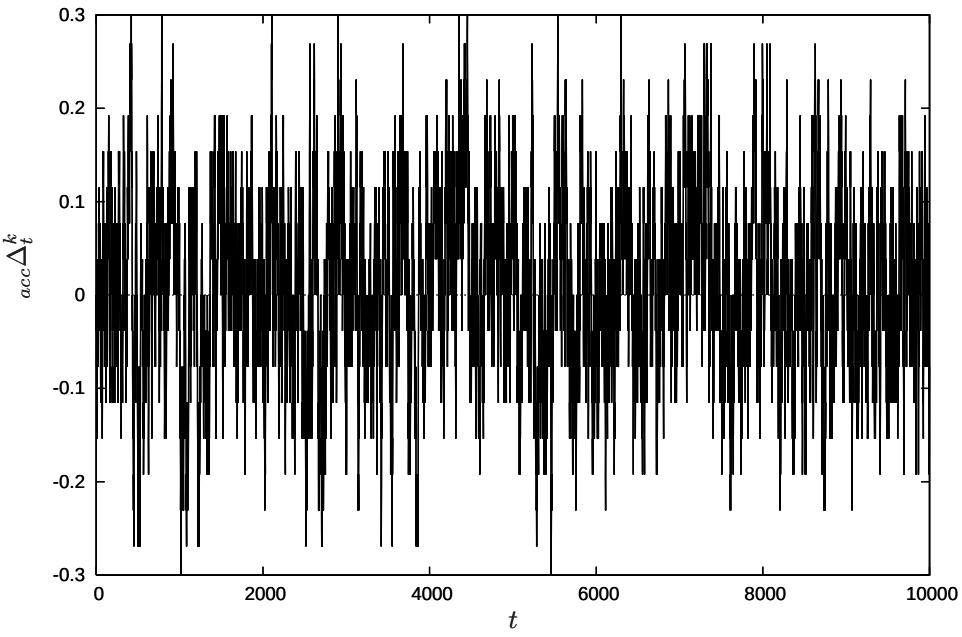

Figure 3.4: Accumulated displacement for a uniformly random change in 26 of 26 answer values. Only for very long times does random motion becomes apparent.

In all the Figures 3.1 - 3.8, the question order is as defined in (2.1) and the changing answers can be identified in these figures. The accumulated displacements for the illustrations and Agincourt households are real numbers on $[-1, 1]$ and hold equally for orbits $\Omega_l^k \in \Sigma_2^n$ or $X_l^n \in I_2^n$. In Figure 3.8 $\Sigma_2^{n=3}$ illustrated by discrete points, linearly arranged. We could instead have plotted the discrete space $I_2^{n=3}$, just the points $0.000, 0.001, \ldots, 0.111$ of the unit internal, in base 2 arithmetic, spaced 2^{-3} apart.

The fitness orbits Ω_l^k are visualized in $\Sigma_{n=3}^2$ in Figure 3.8. It will be appreciated that if we simply coded to Yes/No, different orbits will result from questions posed in the positive and negative sense. Without a uniform *convention* for Yes/No, some favourable shifts will be to the left, some to the right. Fitness is a uniform convention that captures a fundamental human characteristic. At this point it is important that we ask "is a state favourable or unfavourable?".

But it will also be appreciated that the question order is arbitrary and that independent sociologists can choose different ordering [94, 95]. This will result in different orbits and this is not satisfactory.

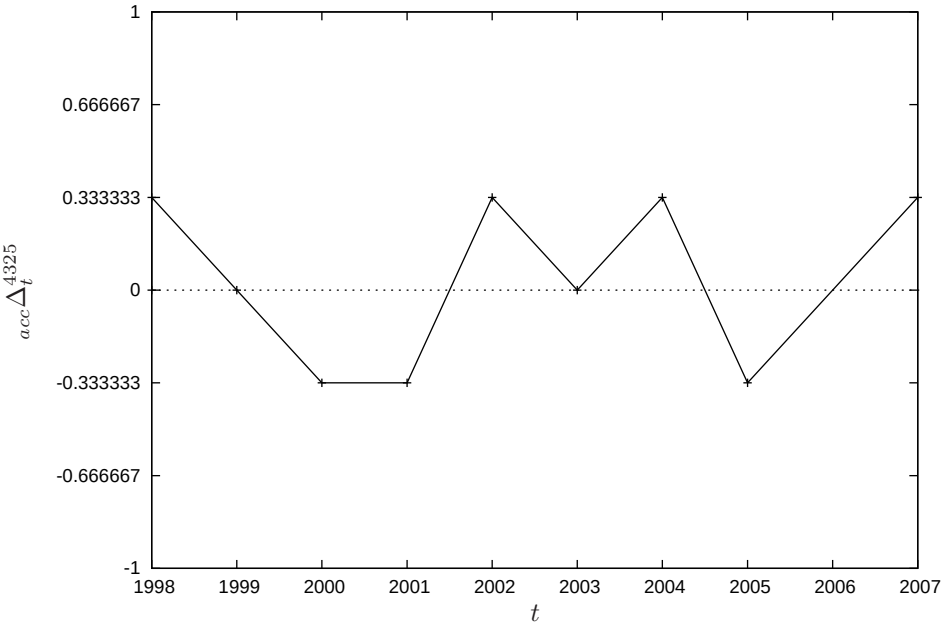

Figure 3.5: Agincourt household accumulated displacement, suggestive of random change of answer values.

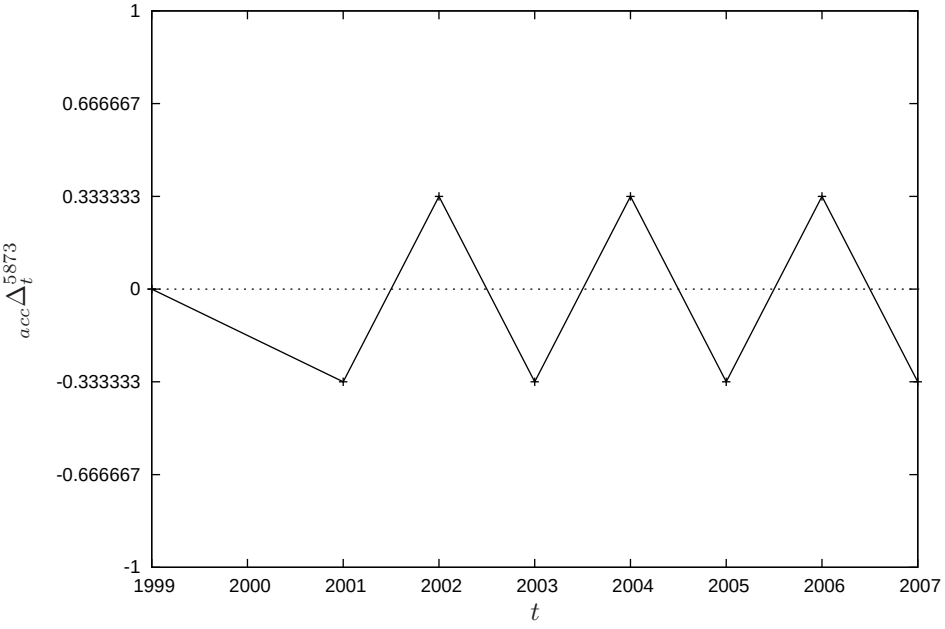

Figure 3.6: Agincourt household accumulated displacement, suggestive of simulation of Figure 3.2.

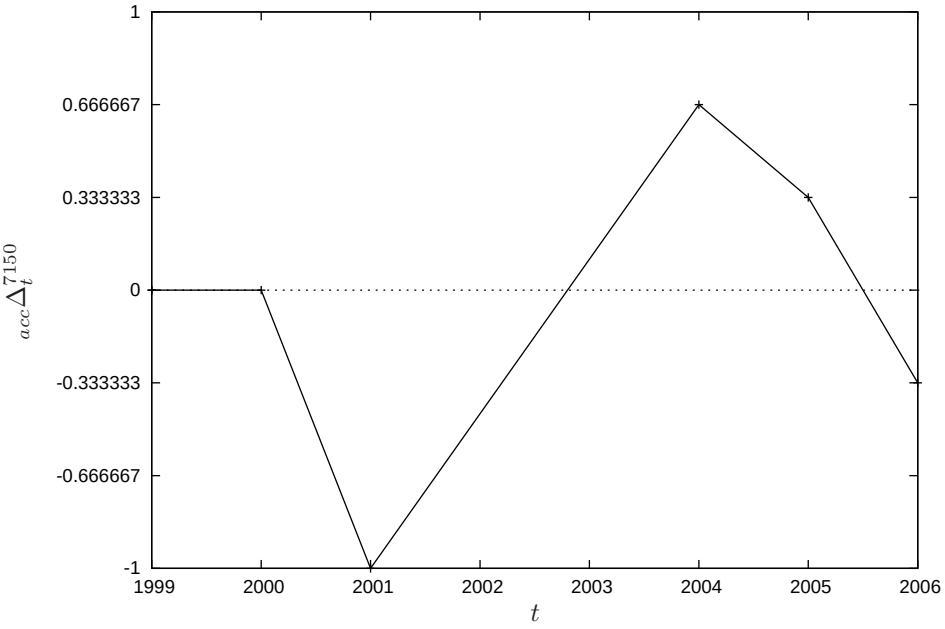

Figure 3.7: Agincourt household accumulated displacement, for 3 negative changes.

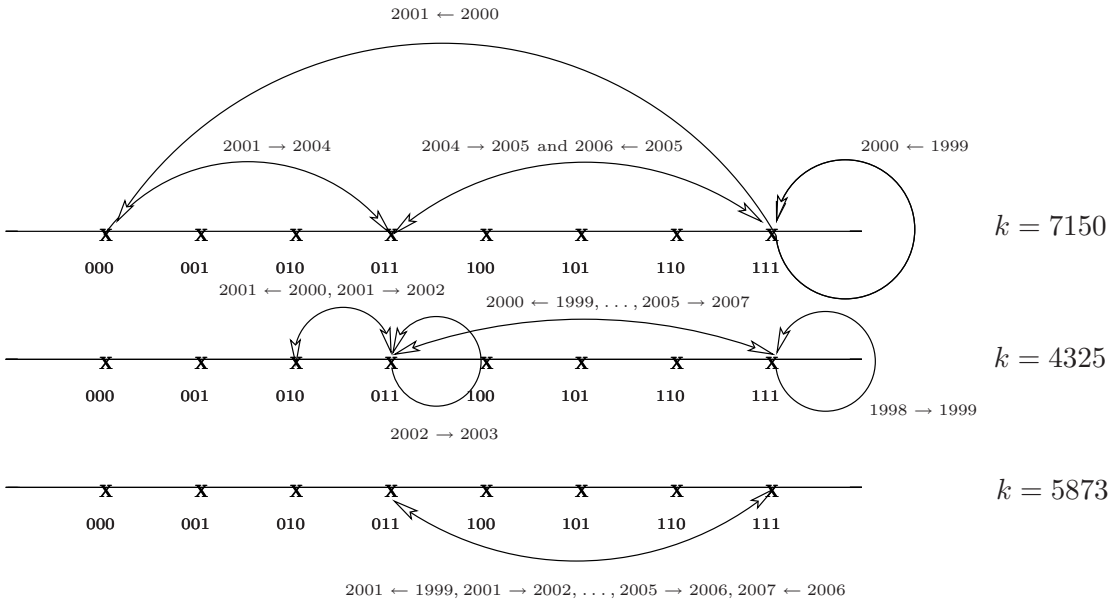

Figure 3.8: Visualization in Σ_2^3. The trend of social unit fitness is made clear by visual orbits.

3.4 Models with Weighted Questions

Visualization of the fitness state as a point on the number line with *meaningful displacement* is of great importance because it is here that orbits approach the full sense of Newtonian dynamics.

Weighting of questions may be artificially applied in order to achieve this, or, may arise naturally out of the purpose. *The assumption for the moment for weighting each question is that some good reason exists.* We allow that each household might have a different weighting (adult death may be more important in some households than absence of biological mother). We order the questions by this weighting, most important question on the left so that the order of questions varies among households. However, we begin supposing that this order remains constant for all times.

As before, suppose that we have the same socialized coding of the questionnaire Q as given in (3.9) and the associated answer set is again as given in (3.10). Again, the state b_t^k as defined in (3.11) is a concatenated binary sequence. However, we suppose a weight w_i for question i for (2.1) and order the questions by this weight, with largest weight on the left. If we attach a weight to each question $q_i \in Q$, then it becomes important to distinguish between the question number i of (2.1) and its position in Q that we denote j. It is important to make it clear that at each time step, each question q_i only has one position j in Q. This property will change the notations in (3.9) as follows.

$$Q = \{q_{i_0}, q_{i_1}, q_{i_2}, \ldots, q_{i_j}, \ldots, q_{i_{n-1}}\}\,, \tag{3.22}$$

where we have some permutations of questions q_i.

In order to achieve consistent displacements in Σ_2^n or I_2^n, we choose a particular order as follows.

Definition 3.10. The map

$$w_{i_j} \mapsto \frac{1}{2^{j+1}},\ j = 0, 1, 2, \ldots, n-1 \tag{3.23}$$

defines the weight of the i'th question which is placed in the j'th position.

It follows that we may also associate weights to the answers given to these questions by modifying the notation of (3.10). Then from (3.10) for some permutations $(i_0, i_1, i_2, \ldots, i_j, \ldots, i_{n-1})$, we have the weighted answer set given by

$${}_wA_t^k = \{a_{i_0}, a_{i_1}, a_{i_2}, \ldots, a_{i_j}, \ldots, a_{i_{n-1}}\}\,, \tag{3.24}$$

and from (3.11) we will also have the weighted concatenated sequence

$${}_wb_t^k = a_{i_0}a_{i_1}a_{i_2}\ldots a_{i_j}\ldots a_{i_{n-1}}\,, \tag{3.25}$$

which now defines the *weighted fitness* of the social unit k at time t.

The weight is coded by the order of questions. Under the homomorphism h defined in Definition 3.7 we can associate the weighted fitness ${}_wb_t^k$ to a real number x_t^k, as follows.

$$h : \Sigma_2^n \to I_2^n : {}_wb_t^k \leftrightarrow x_t^k \tag{3.26}$$

where

$$x_t^k = \sum_{i_j=0}^{n-1} w_{i_j} a_{ti_j}^k,\ a_{ti_j}^k \in \{0,1\}. \tag{3.27}$$

Another simple way of transforming the weighted fitness ${}_wb_t^k$ to a real number x_t^k is by the following coding map

$$\gamma : {}_wb_t^k \leftrightarrow x_t^k = 0.{}_wb_t^k \tag{3.28}$$

which just places a dot in front of the weighted fitness ${}_wb_t^k$ to make it a real number in base 2.

Note that from (3.23) and (3.27), it is clear to see that $0 \le x_t^k \le 1, \forall k$ and all dynamics is then visualized on the discrete unit interval $I_2^n \subset [0,1]$. Then the fitness position at time t is defined by the real number x_t^k. Given either binary sequence or real number, each particular answer $a_{ti_j}^k$ can be identified and its value determined. We refer to this property to *socialized numerical coding*. It is necessary that the order of questions is preserved by the map h or γ, and does not itself change with time.

The weighted fitness distance moved by the k'th social unit is defined by

$$d({}_wb_t^k, {}_wb_{t'}^k) = \sum_{i_j=0}^{n-1} w_{i_j} |a_{ti_j}^k - a_{t'i_j}^k| \ge 0 \tag{3.29}$$

for some permutation of question order, and the fitness displacement is similarly defined by

$$\Delta_{tt'}^k = \sum_{i,j=0}^{n-1} w_{i_j} (a_{ti_j}^k - a_{t'i_j}^k). \tag{3.30}$$

Fitness displacements (3.19), (3.20) are not consistent with jumps on an ordered space such as (3.15). We now have consistent displacements as follows. It will be seen that if $a_{t1}^k : 1 \to 0$ changes, then $d({}_wb_t^k, {}_wb_{t'}^k) = 2^{-1}$ and $\Delta_{tt'}^k = -2^{-1}$, corresponding to the geometrical distance $|x_t^k - x_{t'}^k| = 0.5$ and displacement $x_t^k - x_{t'}^k = -0.5$, on the real axis. Similarly if only $a_{tn}^k : 1 \to 0$ changes, then $d({}_wb_t^k, {}_wb_{t'}^k) = 2^{-n}$ and $x_t^k - x_{t'}^k = -2^{-n}$. These are now consistent with usual measure on $[0,1]$. The

displacement (3.30) is similarly consistent. Favourable and unfavourable moves are again achieved but now a fitness orbit $({}_wb_t^k)_{t=0}^l \in \Sigma_2^n$ can be visualized as the sequence of real numbers

$$X_l^k = (x_t^k)_{t=0}^l : x_t^k \in [0,1] \tag{3.31}$$

As x_t^k approaches 1, the state of the social unit becomes more favourable and it becomes more unfavourable if x_t^k approaches 0. In particular, big displacements are now associated with changes in heavily weighted questions. If questions are not all equally significant, the size of a displacement in $[0,1]$ may indeed approximate a sense of social consequence to the social unit. For example, it is reasonable that a Minister of Health should rank family ailments from expensive treatment to cheap treatment; Q would be designed to record the treatment history of families (say) and the orbit on $[0,1]$ would visualize that history, and reflect cost of treatment.

A change in the n'th answer value clearly gives smaller displacement than would changes in the first answer. The concept of 'nearby' must fold in the weights; thus, for orbits that are near in the sense that the states do not differ up to the n'th question, they should be within a distance $2^{-(n+1)}$ of each other in I_2^n. So the use of weighting gives us consistent visualization of orbits, the orbits are unique given some purpose and the dynamics resembles that of Newtonian particles in one-dimensional motion along $[0,1]$.

The weighting method can be modified. In Sociology the ordering of a stable invariant set may be unimportant. The binary string can be subdivided into n_{inv} invariant answers of equal weight and $n_l - n_{inv}$ variable answers of weight as before, ordered by frequency of change. Then the first n_{inv} answers all have equal weight $1 - 2^{-n_{inv}}$ and the subsequent answers have weight as in (3.27). The position on $[0,1]$ is naturally given by

$$x_t^k = \sum_{i,j=0}^{n_{inv-1}} (1 - 2^{-n_{inv}}) a_{ti_j}^k + \sum_{i,j=n_{inv}}^{n_l-1} w_{i_j} a_{ti_j}^k. \tag{3.32}$$

In a "two-weight" model we can give the first n_{inv} answers equal weight $1-2^{-n_{inv}}$, the following $n_l - n_{inv}$ answers equal weight $2^{-n_l} - 2^{-n_{inv}}$. Then from real data, if an "invariant" answer should change, simply move it into the changing segment. To move a changed answer into the invariant part would best be done with good reason, for well-posed questions. The position on $[0,1]$ is given naturally by

$$x_t^k = \sum_{i,j=0}^{n_{inv-1}} (1 - 2^{-n_{inv}}) a_{ti_j}^k + \sum_{i=n_{inv}}^{n_l-1} (2^{-n_l} - 2^{-n_{inv}}) a_{ti_j}^k, \tag{3.33}$$

where n_{inv} is the number of invariant answers. As before, $0 \leq x_t^k \leq 1$. These are examples of *piece-wise ordered* answer sets. They have obvious use in speeding up the evolutionary process, but will not be used in this thesis.

3.4.1 Results for fitness orbits for weighted questions of fixed order

In Figures 3.9-3.12 we give illustrations of fitness orbits X_l^k for some questions that are now weighted. We take the same data as Figures 3.1-3.4. We suppose an arbitrary fixed question order. Then if the social unit displaces significantly upward, a favourable trend is visualized and is explained at each time by the index of the answer that changed value. As before, only one change per social unit occurs at each time step. In these examples a social unit moves significantly by a few large displacements or many smaller displacements. It will be noted that in Figures 3.11-3.12, orbits vary within bounds set by the changing set of answers. In Figure 3.12 we illustrate a random walk on $\mathbf{I_2^3}$. Because all digits change with equal probability, the state of a social unit will sample all points of $\mathbf{I_2^3}$. Note that in Figures 3.11-3.12, 8 of 26 forces respectively act on the social unit, which wanders randomly in some domain under the influence of those forces. No single cause for change in the domain can be identified, but the forces pushing the social unit into a domain (Figure 3.11) can be identified.

Figures 3.13-3.15 display the fitness orbits of the same Agincourt households. Here the order of questions is assumed to be standard, i.e, 012. The household 5873 clearly has only one strong change in their social factors, which must now be owing to change in q_0 (i.e. absence of biological mother changes dramatically, so we can identify which answer change value). In practice, more than one answer can change value at each time step. In household 7150 where $x_{1995}^{7150} = 0.875$, (Figure 3.15), the state of the household remains favourably unchanged for 6 years, we find that just a year latter, $x_{2001}^{7150} = 0$ displays the most dramatic situation. This is related to the maximum negative change of social factors in that household captured by the accumulated displacement ${}_{acc}\Delta_{2000-2001}^{7150}$.

Note that it is possible to have a zero accumulated displacement while the fitness can vary, because while one answer changes favourably another might change unfavourably, giving a zero net displacement. We clearly have more information stored in the fitness plots than in the plots of accumulated displacements.

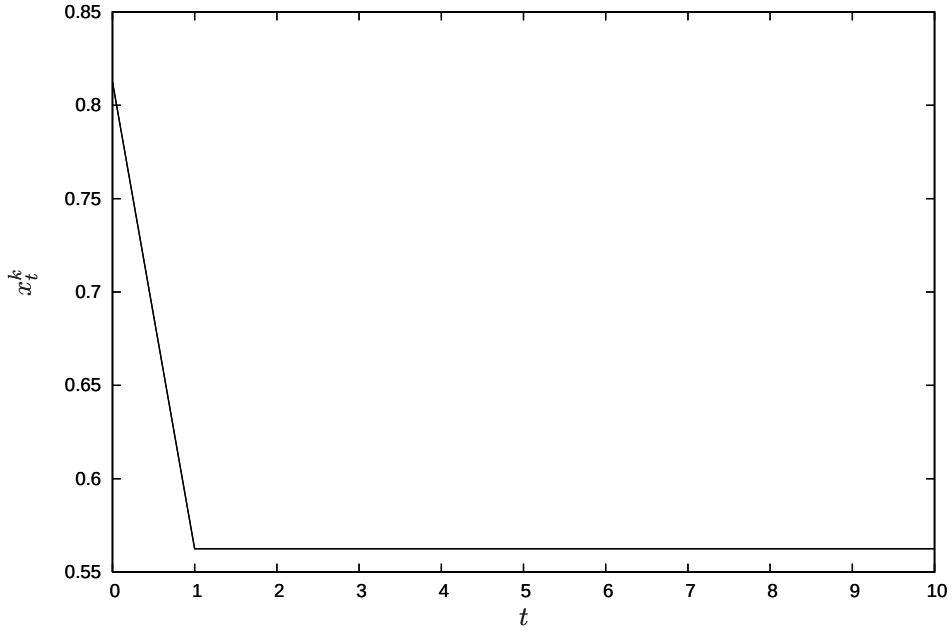

Figure 3.9: Single social unit fitness orbit, for a switch-and-stay answer value change and fixed question order.

3.5 Evolutionary question order model

Suppose that the unchanging answers are of principle interest for example, a social unit with HIV-infection. Among many questions, identify answers that have changed and in each such case, exchange each with the stable answer to the right. If this answer continues to change frequently, it will migrate to the right and the net effect of all such changes is to cause the unchanging answers to migrate to the left. Over very long times, answers will re-order themselves from left to right in increasing order of frequency of change. If there are some unchanging answers for long times, we have a social unit that is now *characterized* by the unchanging part. To think at the population-level, if there are many social units with this unchanging part, then we automatically identify a sub-population. These ideas will be explored in detail in the next chapter.

We have assumed fixed weighted question order above. The task of the sociologist, through Q_t, is now extended to deciding whether to add or subtract questions and whether to adjust the

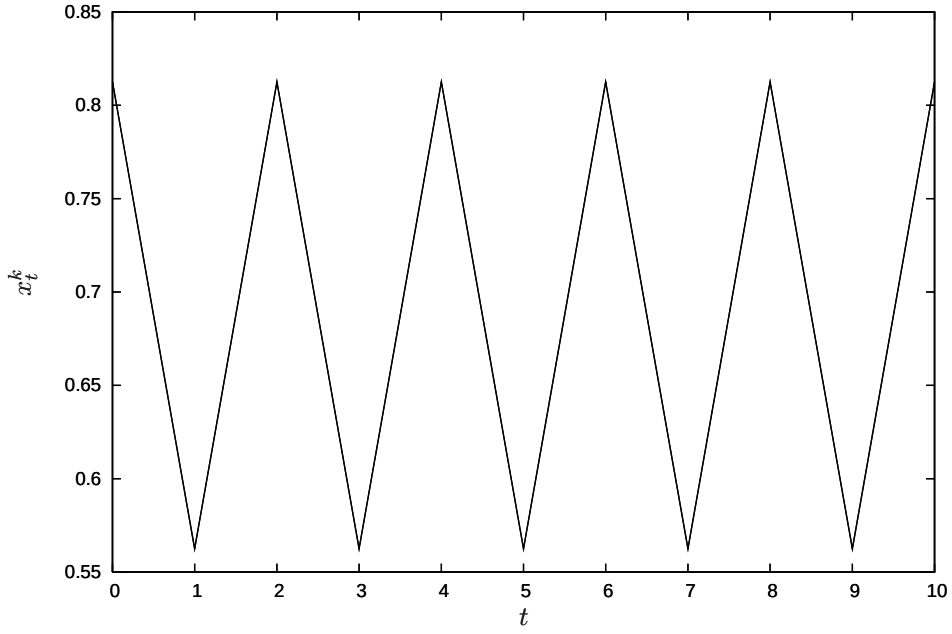

Figure 3.10: Single social unit fitness orbit, for an oscillating answer value change and fixed question order.

order/weighting of questions. The Minister of Health's ordering might or might not change with external developments. New ailments may arise and be added at some appropriate point in the question set. This may have impact and be seen as a sudden shift in the health of social units. Ailments may be dropped from Q where they do not contribute significantly to drift of a social unit on $[0, 1]$. We note that HIV/AIDS [103] did not contribute before about 1980 and may be regarded as weighted zero cost to the Minister up to that date, high cost after that date, without altering the history of Q_t. If sociologists agree that the question set is adequate, we might suppose that the social system is understood. But there will not in general be agreement among sociologists as to question order that best reflects the changes of all individuals in a population. It is a fundamental contribution of this thesis to give a rational, objective method for deciding appropriate question order.

Because a social unit migrates towards or away from a favourable state, because an unchanging part is defined, because a sub-population as in (3.33) population ("species") is defined, we refer to the *evolutionary model* [111]. The swapping of answers (or re-weighting) is contributory to understanding of the purpose of Q_t and is a new, independent tool to exhibit automatically, emergence

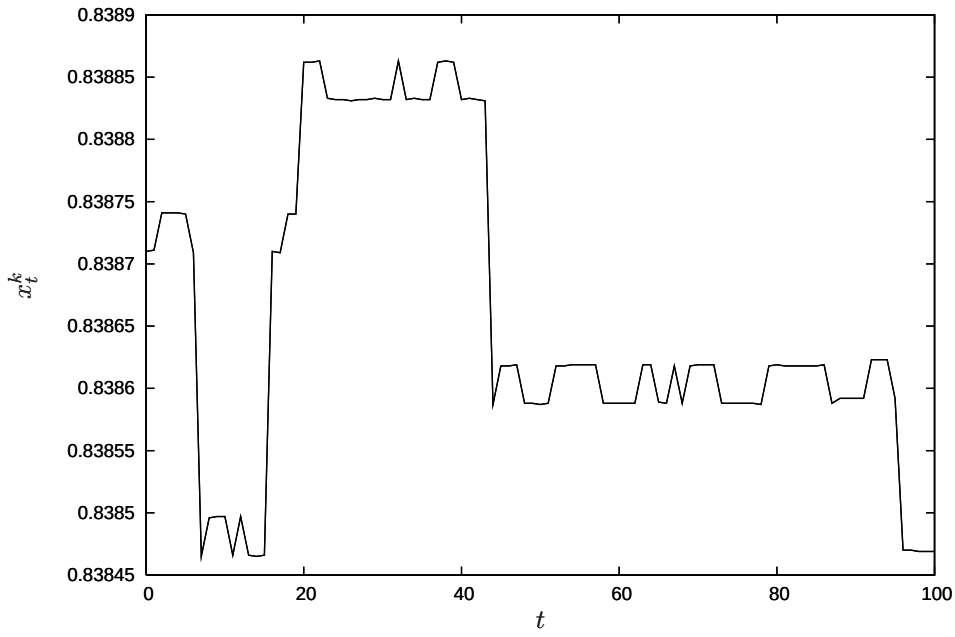

Figure 3.11: Single social unit fitness orbit, for a uniformly random change in 8 of 26 answer values and fixed question order.

of a new, objectively defined, sub-population which is conditional on the purpose of Q_t. Dynamic re-ordering is an additional deterministic function applied to the answer set per social unit. This function is informed by purpose and the given data (any such functions can be of use).

For a simple example we may reorder questions by frequency of change from the right. Or, we may have a cost function for each answer change (temporary out-migration brings in money, adult death costs money) that assigns weights. In this thesis we attach importance to the automatic identification of sub-populations and to the fundamental property of emergence of *fitness* of the sub-populations (the property of evaluation). But also, simple relative frequency of change of questions is a fundamental statistic of the data, *typical of longitudinal data*. An evolutionary orbit visualizes this relative frequency, by order of questions, or as we will now see by a "$y-$axis" that encodes question order. We will use this evolutionary model throughout this thesis.

Here the slowest varying part is of most importance and is given the largest weighting. Note now that should a long-fixed answer value $a_0 = 0$ at the first significant digit of x_t^k change value, it will move right by one space and the sequence will displace right by at least 25% of the unit interval. A change of the new, left-most factor will move q_1 back to the left with changed value

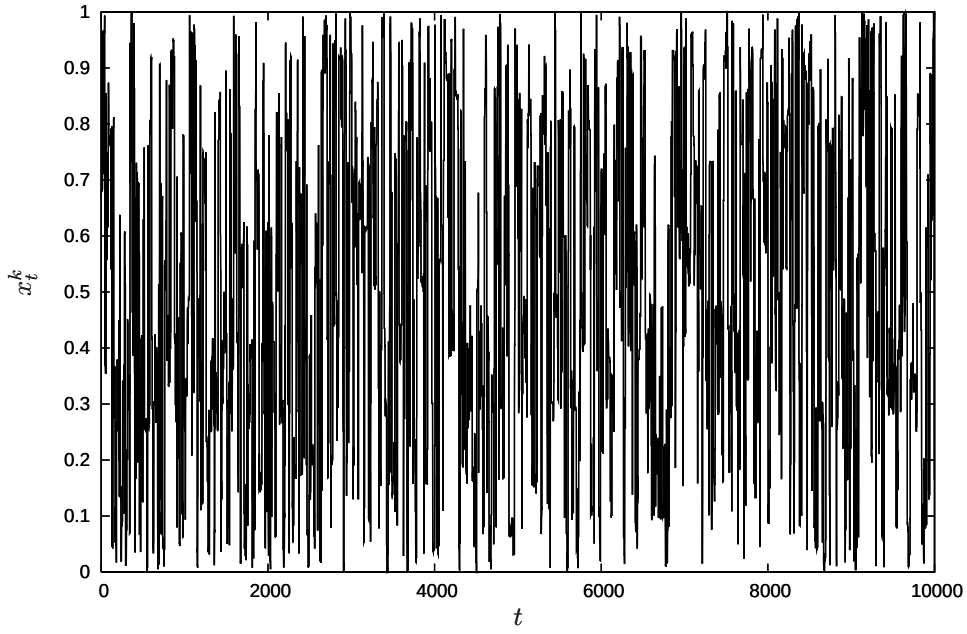

Figure 3.12: Single social unit fitness orbit, for a uniformly random change in 26 of 26 answer values and fixed question order.

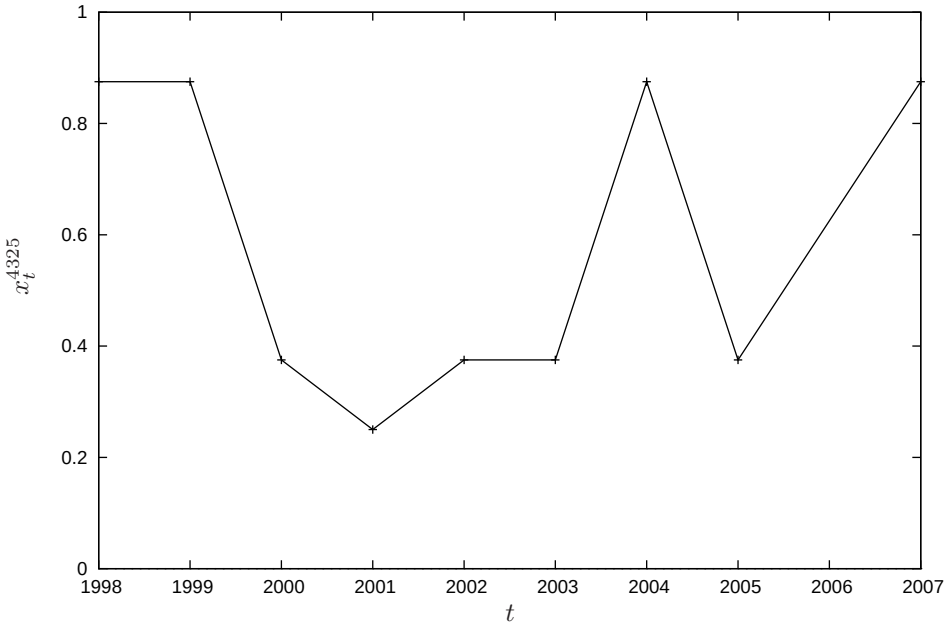

Figure 3.13: Agincourt fitness orbit, for many answer values changing. The question order is 012.

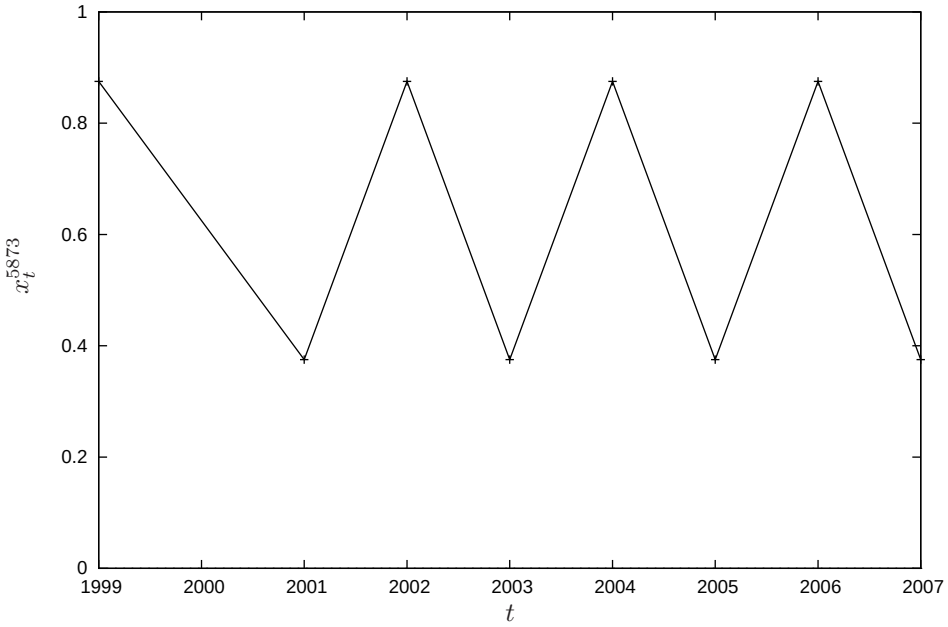

Figure 3.14: Agincourt fitness orbit, for a single oscillating answer value changing. Same ordering of questions as Figure 3.13.

$a_0 = 1$ and so give a total 50% change on the unit interval. But then an enormous displacement has taken place, making the importance of the invariant, heavily weighted factors in separating sub-populations. In contrast, only small separation arises for the rapidly-changing or low-weighted factors on the right of the binary string.

The dynamics of change of answer value per social unit may be stochastic [112] or deterministic. The game of swapping changing questions to the right is deterministic [113].

We use the properties described above to construct evolutionary orbits of social units. If change in question order is with respect to this deterministic game [113], then the orbits x_t^k defined by (3.31) now become *evolutionary fitness orbits* and will be denoted, for clarity, by

$$(e_t^k)_{t=0}^l : e_t^k \in I_2^n, \tag{3.34}$$

3.5.1 Result for the evolutionary orbits on I_2^n

Under the evolutionary model, we give the graphs of e_t^k in Figures 3.16-3.19. The technique of the previous section is used to plot position on the number line, but now the order of answers

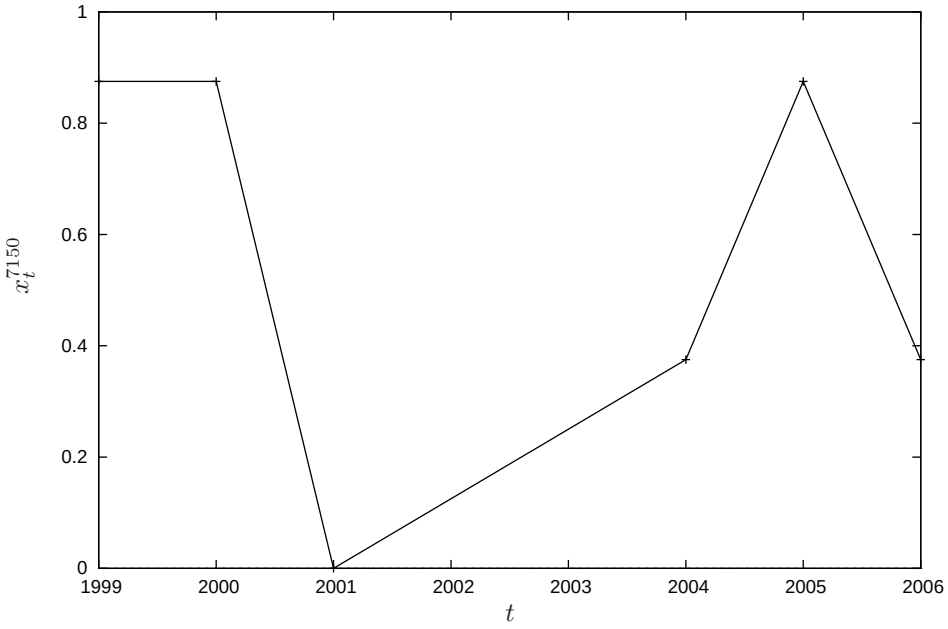

Figure 3.15: Agincourt fitness orbit, for 3 answer values changing. Same ordering of questions as Figure 3.13.

must be recorded. We also assume a standard initial question order 012. The dynamics of Figure 3.16 is unchanged from Figure 3.9 because only one switch occurs and the state is then constant where question order is reordered. In Figure 3.17 as one factor oscillates, it migrates to the right, the constant factors keep their relative order and value, thus defining a position on the e-axis while the changing factor has ever decreasing effect. The orbit settles to $e =$ constant and it is clear that among a large population, sub-populations might emerge. Of course this state has a fitness that may be attached to the sub-population. Comparing to Figure 3.10, we can clearly see changes in the dynamics because of the evolutionary game that we apply.

In Figure 3.18 there are unchanging factors and the orbit will again settle in $\mathbf{I_2^{26}}$. Note that 8 answers change values over the 26. With such significant changes, the orbit converges to a small sub-domain in $\mathbf{I_2^{26}}$ and stays forever in that domain.

Figure 3.19 applies where there are no unchanging factors and the orbit wanders over the interval $\mathbf{I_2^{26}}$ without converging.

These Figures 3.16-3.19 are not satisfactory and we do not give Agincourt examples. Although

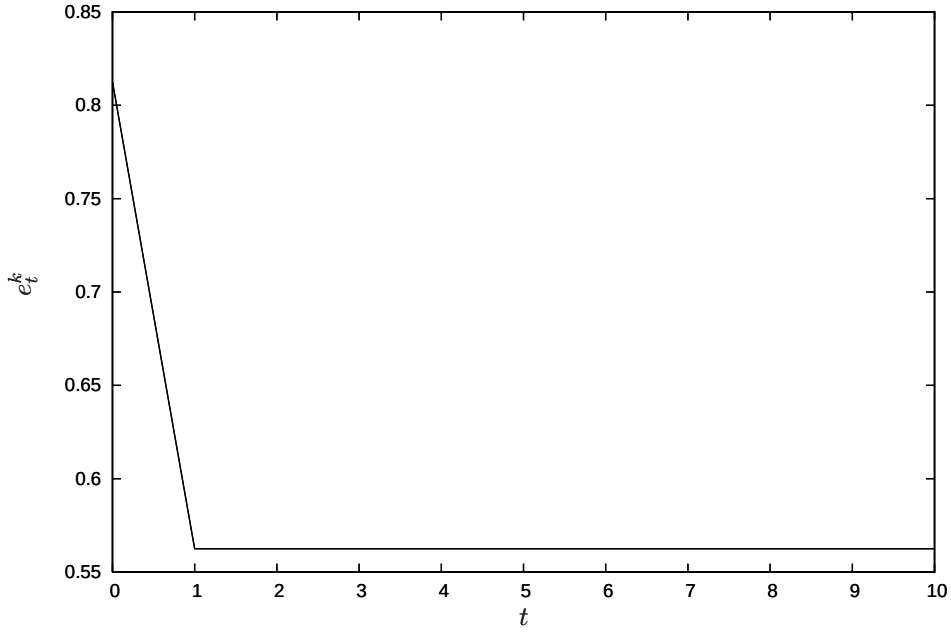

Figure 3.16: Single social unit evolutionary fitness orbit, for a switch-and-stay answer value change.

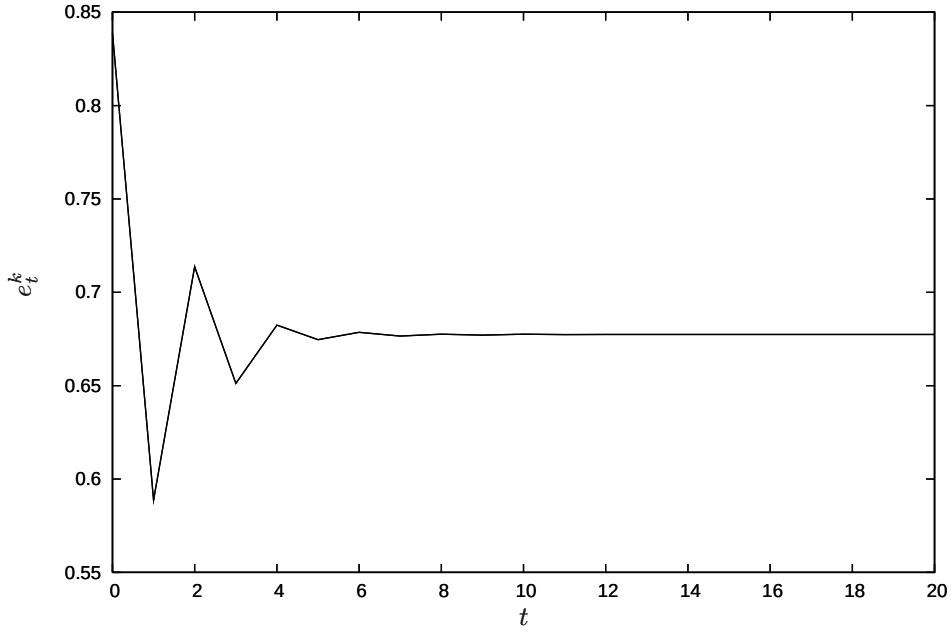

Figure 3.17: Single social unit evolutionary fitness orbit, for an oscillating answer value change.

a social unit now has a localized position on the $x-$axis, it is clear that some other social units, with completely different distribution of answers, might have the same fitness. But then two differing

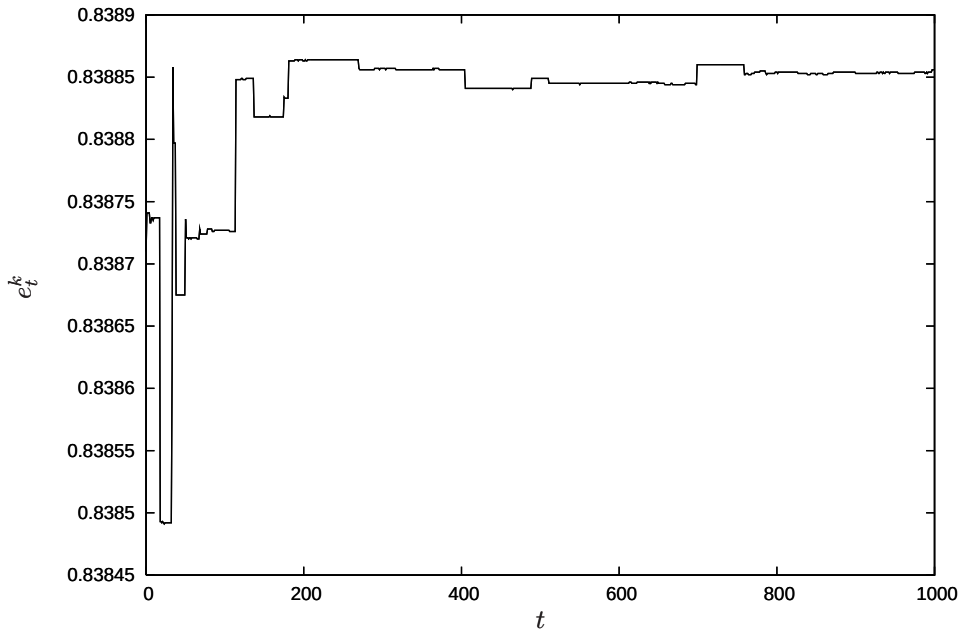

Figure 3.18: Single social unit evolutionary fitness orbit, for uniformly random change in 8 of 26 answer values.

evolutionary states (for example) are indistinguishable on that axis. Only where a sub-population has the same distribution of answers on significant digits (i.e. the evolutionary species), can we overlay orbits on the same graph, and talk naturally of the differences between like classes. The emergence of sub-populations is, for example, as for the Minister of Health above, if many orbits converge towards the same, possibly small domain.

3.5.2 Evolutionary orbits in the unit square $I_2^n \times I_n^n$

The goal of this section is to solve the problem of overlay of sub-populations as mentioned above. We have given each question an *answer value* in $\{0, 1\}$ and coded fitness to binary numbers. We will quantify a *question sequence* by *recording* its position j as described in the notation q_{i_j}. Then this is coded by the set

$$O_t^k = \{i_j | j = 0, 1, 2, \ldots, n-1\} \, . \tag{3.35}$$

This order value corresponds to the position j of the i'th question as described in the notation $q_{i_j} \in Q$.

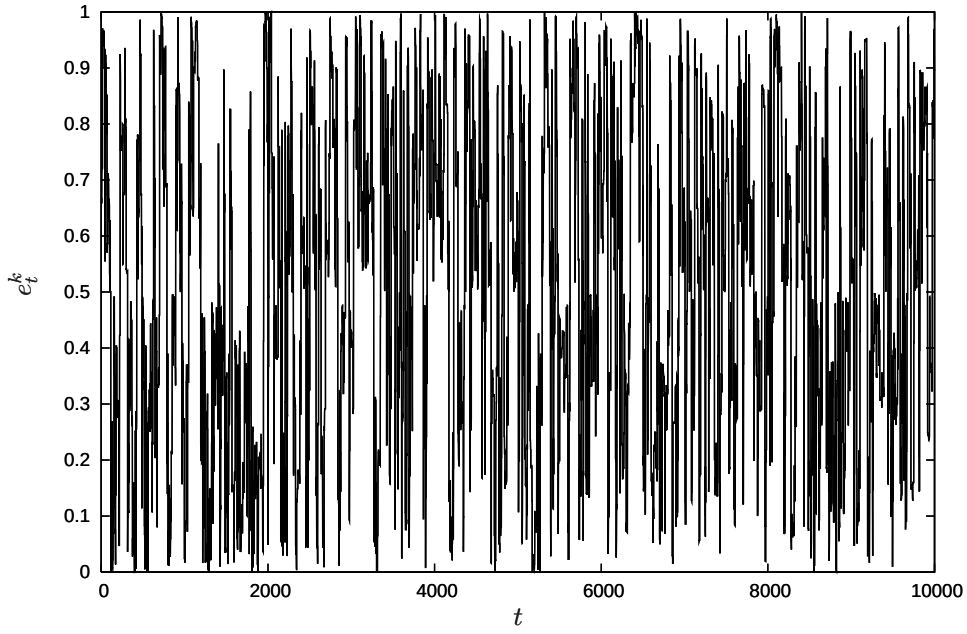

Figure 3.19: Single social unit evolutionary fitness orbits, for a uniform change in 26 of 26 answer values.

As before, if we also combine elements of O_t^k as a concatenated string, then directly from (3.35) we can define the significance as follows.

Definition 3.11. The *sequence*

$$\theta_t^k = i_0 i_1 i_2 \dots i_{n-1}, \;\; i_j \in \{0, 1, 2, \dots, n-1\} \tag{3.36}$$

captures the *significance state* of the k'th social unit at time t.

As stated above, each question order has a unique value in $\{0, 1, 2, 3, \dots, n-1\}$, thus for n questions there are $n!$ possible arrangements of question order values.

Definition 3.12. The space

$$\Sigma_n^n = \{\theta_i | i = 0, 1, 2, \dots, n-1 \text{ and } \theta_i \text{ is a permutation of symbols } 0, 1, 2, \dots, n-1\}\,, \tag{3.37}$$

is the space of *finite one-sided* sequences of length n consisting of n distinctive symbols. In this thesis, Σ_n^n is also called the *significance space of sequence states* with respect to the questionnaire Q.

Then the sequence θ_t^k is a single point in Σ_n^n that captures the significance of a social unit k at time t.

As before, let

$$w_i^o = \frac{1}{n^{i+1}},\ i \in \{0, 1, 2, \ldots, n-1\} \tag{3.38}$$

define the *order weight function*. Let

$$\chi_t^k = \sum_{i=0}^{n-1} w_i^o \theta_{ti}^k,\ \theta_{ti}^k \in \{0, 1, 2, \ldots, n-1\}\ , \chi_t^k \in I_n^n \tag{3.39}$$

where I_n^n is a discrete space consisting of $n!$ rational numbers χ_t^k. It is also clear to see that $0 \leq \chi_t^k \leq 1$. We will refer to χ_t^k as the *significance* of the answer set for the k'th social unit at time t.

Following the definitions (3.29) and (3.30), the significance distance moved by the k'th social unit is defined by

$$d(\theta_t^k, \theta_{t'}^k) = \sum_{i=0}^{n-1} w_i^o |\theta_{ti}^k - \theta_{t'i}^k| \geq 0 \tag{3.40}$$

and the significance displacement is similarly defined by

$$\Delta_{tt'}^k = \sum_{i=0}^{n-1} w_{i_j} w_i^o (\theta_{ti}^k - \theta_{t'i}^k)\ . \tag{3.41}$$

We can also derive from (3.40) and (3.41) respectively the significance distance and displacement at time t between two social units k and k'. Note that the fitness displacements in these cases may be constant so that as these new displacements grow we find questions reordering in different ways, that is, social units have different frequencies of change of answer values.

The state of social unit k is now coded by the *ordered pair*

$$(e_t^k, \chi_t^k) \in I_2^n \times I_n^n \tag{3.42}$$

in general, base 2 and base n "decimals" respectively. Then the value of an answer and the answer itself can be decided by its distance from the "decimal point".

To illustrate if $n = 3$, we code the questions

$$\{q_0, q_1, q_2\} \rightarrow \{0, 1, 2\}, \tag{3.43}$$

(just the digits of base 3 arithmetic) so that a question order sequence, say $q_1 q_2 q_0$ is coded by 120. As for the e-values above, the question order 120 is mapped to a base 3 number, i.e. 0.120. Each

coordinate may be translated (using definitions (3.27) and (3.39)) to common base 10 arithmetic for convenience of visualization, as in

$$(0.110, 0.120) = (0.75, 0.55) \tag{3.44}$$

and these can again be uniquely decoded (by going back to base 2 and base 3, using definitions (3.27) and (3.39) for question order and answer value respectively. Note that sub-populations in the $e, \chi-$plane can automatically emerge.

We again use

$$X_l^k = (e_t^k, \chi_t^k)_{t=0}^l \tag{3.45}$$

to denote the orbit of the k'th social unit in the unit square. In keeping with mathematical usage, we refer to $\Gamma_n = I_2^n \times I_n^n$ as *phase space*[14, 114]. It is at this point that we may change the question set. New questions can be added but these should be included on the right where their impact is small; then in time they must assert themselves if they are relevant. If this is the case, and it happens for many social units in a population, then a new sub-population automatically emerges.

3.6 Dynamics of orbits in the state space $\Gamma_n = I_2^n \times I_n^n$

The ideas of this thesis are strongly informed by modern dynamical systems theory [14, 110, 105, 114, 115, 16, 116]. Let

$$\zeta_t^k = (e_t^k, \chi_t^k) \tag{3.46}$$

denote the state of the k'th social unit at time t now in Γ_n. According to the dynamics established in the state space Γ_n, it is clear to see that e and χ vary with time. Define the map

$$\psi : \Gamma_n \to \Gamma_n : (e_t^k, \chi_t^k) \mapsto (e_{t+1}^k, \chi_{t+1}^k) \tag{3.47}$$

which gives the relationship between consecutive states of a social unit k over time.

According to the properties of the homomorphism γ (3.26), changes in values of the evolutionary fitness e_t^k of each social unit k are related to changes in values of its associate concatenated sequence ${}_wb_t^k$. Similarly, changes in values of the significance χ_t^k of each social unit k are attached to changes in values of its associate sequence ${}_w\theta_t^k$. So we can think in decimals or sequences.

Thus, we can associate changes under ψ (3.47) to change under a map ξ as follows:

$$(e_t, \chi_t) \overset{\psi}{\mapsto} (e_{t+1}, \chi_{t+1}) \equiv ({}_wb_t, \theta_t) \overset{\xi}{\mapsto} ({}_wb_{t+1}, \theta_{t+1}) \tag{3.48}$$

where

$$\xi : \Sigma_2^n \times \Sigma_n^n \to \Sigma_2^n \times \Sigma_n^n : ({}_wb_t, \theta_t) \mapsto ({}_wb_{t+1}, \theta_{t+1}) \; . \tag{3.49}$$

The maps ψ and ξ arise from social forces as discovered in the data and the deterministic evolutionary game.

Suppose that the order of questions is known at each time t and that we identify changes in e and χ to changes in values of b_t^k and θ_t^k for each social unit k. Let $\nu_t^k \leq n$ denote the number of questions that change answer values for the social unit k at time t. Let

$$\eta_t^k = \{i_0, i_1, \ldots, i_{\nu_t^k - 1} |\; i_j = 0, 1, 2, \ldots, n-1,\; i_j \neq i_k \; \nu_t^k \leq n\}, \tag{3.50}$$

denote the set of indexes (in increasing order) of all questions that change answer values at time t for the social unit k.

We note that $\xi = \varphi \circ \sigma$ consists of two parts, first σ (3.51) changes the answer values with respect to η_t^k for each social unit k then φ (3.54) changes the question order under the evolutionary game, also with respect to η_t^k. We write

$$\sigma : ({}_wb_t^k, \theta_t^k) \mapsto ({}_wb_t'^k, \theta_t^k) \tag{3.51}$$

so that

$${}_wb_t^k = a_{i_0}^k a_{i_1}^k a_{i_2}^k \ldots a_{i_j}^k \ldots a_{i_{n-1}}^k \mapsto {}_wb_t'^k = a'^k_{i_0} a'^k_{i_1} a'^k_{i_2} \ldots a'^k_{i_j} \ldots a'^k_{i_{n-1}} \tag{3.52}$$

where

$$a'^k_{i_j} = \begin{cases} 1 - a_{i_j}^k & \text{if } \; i_j \in \eta_t^k \\ a_{i_j}^k & \text{otherwise.} \end{cases} \tag{3.53}$$

Thus σ changes question values only. Longitudinal data merely lists answer values under a fixed answer order and σ first accepts the new data.

Then we write

$$\varphi : ({}_wb_t', \theta_t^k) \mapsto ({}_wb_{t+1}, \theta_{t+1}) \tag{3.54}$$

Here we swap changing answers, starting from the right, to the right hand end. That is

$$\varphi = \varphi_{_{0,\pi_0}} \circ \varphi_{_{1,\pi_1}} \circ \varphi_{_{2,\pi_2}} \circ \ldots \circ \varphi_{_{i,\pi_i}} \circ \ldots \circ \varphi_{_{\nu_t^k - 1, \pi_{\nu_t^k - 1}}}, \;\; i \in \eta_t^k \tag{3.55}$$

and

$$\pi_i : i \mapsto n - 1 - i \tag{3.56}$$

defines the jump of the i'th question that changes answer value, and

$$\varphi_{_{i,\pi_i}} : ({}_wb'_t, \theta^k_t) \mapsto ({}_wb^i_{t+1}, \theta^i_{t+1}) \tag{3.57}$$

$$\theta^k_t = i_0 i_1 \dots i_j \dots i_{n-1} \mapsto \theta^i_{t+1} = i_0 i_1 \dots i_{j-1} i_{j+1} \dots i_{n-1} i_j \ . \tag{3.58}$$

Thus once we have accepted the new data under σ, we apply the evolutionary game φ. If ν^k_t questions change answer values, then φ_i is applied ν^k_t-times.

To illustrate the rather complicated mathematics above of dynamics under ξ, suppose that the number of questions $n = 5$. Consider a social unit k, with the fitness state at time t given by ${}_wb^k_t = 10101$. The significance state of that social unit at time t is given by $\theta^k_t = 01234$. Suppose also that $\nu^k_t = 3$ questions change answer values for that social unit at time t and that the indexes of these questions are defined by $\eta^k_t = \{1, 2, 4\}$.

For simplicity, we write in bold with a dot on top, the questions that change answer values. Then, we have ${}_wb^k_t = 1\dot{\mathbf{0}}\dot{\mathbf{1}}0\dot{\mathbf{1}}$ and $\theta^k_t = 0\dot{\mathbf{1}}\dot{\mathbf{2}}3\dot{\mathbf{4}}$. Let us compute the fitness state ${}_wb^k_{t+1}$ and the significance state θ^k_{t+1} of the social unit k at time $t+1$ Thus, from (3.51), we have

$$\sigma(1\dot{\mathbf{0}}\dot{\mathbf{1}}0\dot{\mathbf{1}}, 0\dot{\mathbf{1}}\dot{\mathbf{2}}3\dot{\mathbf{4}}) = (11\mathbf{000}, 0\dot{\mathbf{1}}\dot{\mathbf{2}}3\dot{\mathbf{4}}) \ . \tag{3.59}$$

From the definition (3.55), we can write

$$\varphi = \varphi_{_{1,\pi_1}} \circ \varphi_{_{2,\pi_2}} \circ \varphi_{_{4,\pi_4}} \ . \tag{3.60}$$

We now use the definition (3.57), to compute $\varphi_{_{i,\pi_i}}$ for $i \in \eta^k_t = \{1, 2, 4\}$. Thus

$$\begin{aligned}
\varphi_{_{4,\pi_4}}(11\mathbf{000}, 0\dot{\mathbf{1}}\dot{\mathbf{2}}3\dot{\mathbf{4}}) &= (11\mathbf{000}, 0\dot{\mathbf{1}}\dot{\mathbf{2}}3\mathbf{4}) \ , \quad \pi_4 = 0 \\
\varphi_{_{2,\pi_2}}(11\mathbf{000}, 0\dot{\mathbf{1}}\dot{\mathbf{2}}3\mathbf{4}) &= (11\mathbf{000}, 0\dot{\mathbf{1}}3\mathbf{42}) \ , \quad \pi_2 = 2 \\
\varphi_{_{1,\pi_1}}(11\mathbf{000}, 0\dot{\mathbf{1}}3\mathbf{42}) &= (10\mathbf{001}, 03\mathbf{421}) \ , \quad \pi_1 = 3 \, .
\end{aligned} \tag{3.61}$$

Finally, the state of the social unit k at time $t+1$ is given by

$$\varphi \circ \sigma(1\dot{\mathbf{0}}\dot{\mathbf{1}}0\dot{\mathbf{1}}, 0\dot{\mathbf{1}}\dot{\mathbf{2}}3\dot{\mathbf{4}}) = (10\mathbf{001}, 03\mathbf{421}) \ . \tag{3.62}$$

We recall that under the evolutionary game (3.58), we assume that questions with changing answer values are less important compared to those that are unchanged. Thus, if the i'th question changes answer value we allow $\varphi_{_{i,\pi_i}}$ to simply shift that question to the end of the sequence. This

process redefines the weight of each question by adding more weight to unchanging questions and reducing weight to changing answers by the function π (3.56).

The elements of ${}_wb^i_{t+1}$ are equal in value to those of ${}_wb'^k_t$ but now rearranged in an order given by θ^i_{t+1}. Note that order of change of answers (from left to right) gives differing states, that is

$$\varphi_l = \varphi_{\nu_t^k-1,\pi_{\nu_t^k-1}} \circ \ldots \circ \varphi_{2,\pi_2} \circ \varphi_{1,\pi_1} \circ \varphi_{0,\pi_0} \neq \varphi_{0,\pi_0} \circ \varphi_{1,\pi_1} \circ \varphi_{2,\pi_2} \circ \ldots \circ \varphi_{\nu_t^k-1,\pi_{\nu_t^k-1}} = \varphi_r \ . \quad (3.63)$$

We refer to φ_l when we change order from the left and φ_r when answer values are changed from the right. We can now distinguish between the dynamics obtained under φ_l with that obtained under φ_l. We find that φ_r better separates sub-populations because it gives big jumps when b_t and θ_t change values compared to changes operated under φ_l.

Consider the above example where ${}_wb^k_t = 1\dot{\mathbf{0}}\dot{\mathbf{i}}0\dot{\mathbf{i}}$ and $\theta^k_t = 0\dot{\mathbf{i}}\dot{\mathbf{2}}3\dot{\mathbf{4}}$. Note that the result (3.62) is obtained using φ_r. Let us follow the definition (3.62) to compute ${}_wb^k_{t+1}$ and θ^k_{t+1} with respect to φ_l. Thus,

$$\varphi_l \circ \sigma(1\dot{\mathbf{0}}\dot{\mathbf{i}}0\dot{\mathbf{i}}, 0\dot{\mathbf{i}}\dot{\mathbf{2}}3\dot{\mathbf{4}}) = (10\mathbf{100}, 03\mathbf{124}) \ . \quad (3.64)$$

We use the definitions (3.29) and (3.40) to measure the fitness and significance distances moved by the social unit t between the two times t and $t+1$ with respect to φ_l and φ_r. These measures will help us to compare the dynamics under φ_l with the dynamics under φ_r.

$$\begin{aligned} d_r(10101, 10001) &= 0.125 \\ d_l(10101, 10100) &= 0.03125 \\ d_r(01234, 03421) &= 0.09856 \\ d_l(01234, 03124) &= 0.0896 \end{aligned} \quad (3.65)$$

This result (3.65) clearly shows that the fitness distance obtained from φ_r is 4 times the fitness distance obtained from φ_l. Also the significance distance under φ_r is slightly bigger than the significance distance under φ_l. As a result, the evolutionary process operated under φ_r is preferable in terms of identification of sub-populations.

Because of the above properties, the results presented in this thesis are obtained with the dynamics operated under φ_r. Thus, we can write

$$\xi = \varphi \circ \sigma \quad (3.66)$$

If there are n questions, then there are 2^n possible responses because each answer $a_i \in \{0, 1\}$. Thus any state ζ_t can then map to 2^n new fitness values in ζ_{t+1}. The map φ (3.54) is entirely owing to

our design. The order of questions can take any one of $n!$ values (if the first question is any one of n, once this is decided the second question is any one of $n-1$ questions and so on). Then there exist $n! \times 2^n$ possible states for a social unit and the state space of geometric points

$$\Gamma_n = \{\zeta = (e, \chi)\} \ . \tag{3.67}$$

or sequences of answer values and orderings

$$S_n = \{({}_wb, \theta)\} \tag{3.68}$$

consist of many states. It follows that $S_n = \Sigma_2^n \times \Sigma_2^n$.

The space S_n is represented by Figure 3.20. Since S_n is a discrete space, we can number elements of S_n as represented in Figure 3.21 in the case of $n = 3$ questions. The arrow illustrates a transition from state 48 defined by $({}_wb, \theta) = (111, 012)$ to state 37 defined by $({}_wb, \theta) = (100, 021)$. This transition indicates that questions 1 and 2 have changed values. Thus, we note that $\xi(111, 012) = (100, 021)$.

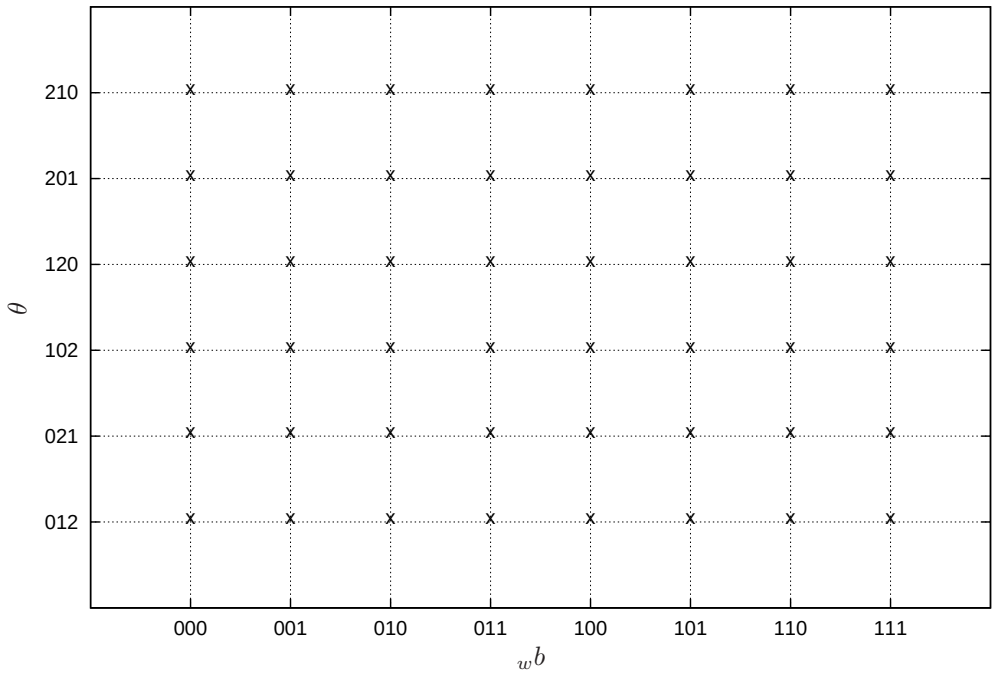

Figure 3.20: Bi-sequence space $S_n = \Sigma_2^n \times \Sigma_n^n$ for $n = 3$ questions.

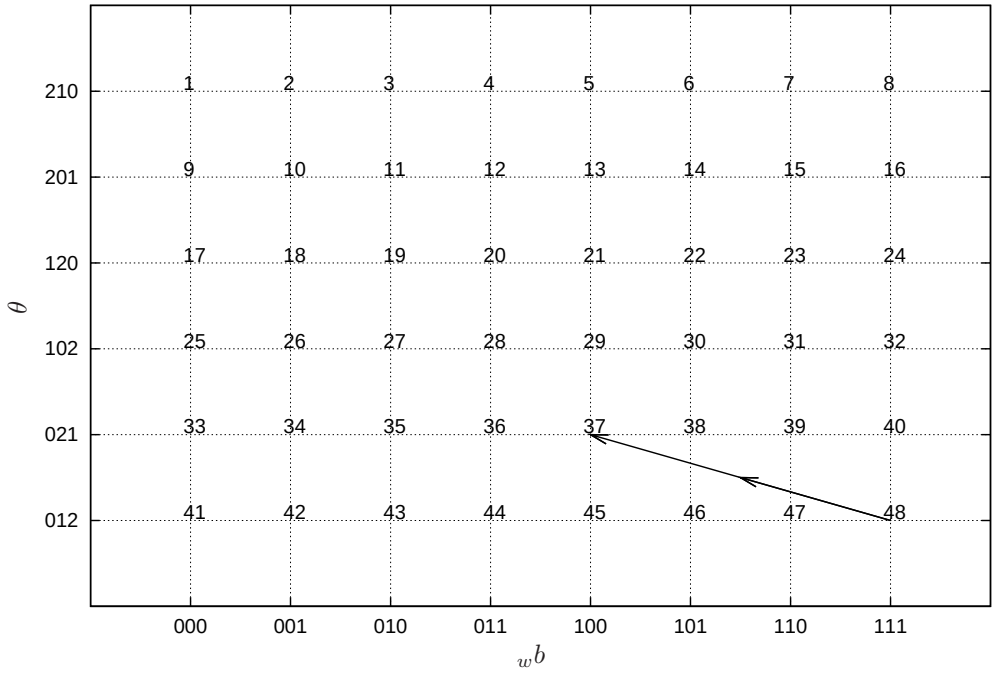

Figure 3.21: Bi-sequence space $S_n = \Sigma_2^n \times \Sigma_n^n$ for $n = 3$ questions, with sequences indexed.

3.7 Transition matrix for social units

With the dynamics of the map ψ (3.47), we can refer each move $\zeta_t^k \mapsto \zeta_{t+1}^k$ to a transition $i \mapsto j$ as in Figure 3.21. Thus, given a fixed number of questions n, a state space Γ_n combined with the process ψ can be associated to the matrix of allowed transitions for a social unit k, that is referred to as the transition matrix and we denote it by T_n^k. Thus $T_n^k \in M_{d\times d}$ where $d = 2^n \times n!$ The elements of T_n^k are defined by

$$T_n^k = [t_{ij}^k] \; i, j = 1, 2, 3, \ldots d \, , \tag{3.69}$$

where

$$t_{ij}^k = \begin{cases} 1 & \text{if} \quad i \to j \\ 0 & \text{if} \quad i \nrightarrow j \end{cases} \tag{3.70}$$

is the Kronecker delta. There are transitions that are not allowed in Γ_n due to the properties of the process ψ. Let T_n denote the theoretical transition matrix for n questions, which displays all allowed transitions in Γ_n under the process ψ. Each row of T_n adds to 2^n which represents the total number of possible binary sequence arrangements if there are n questions.

As an example, suppose that $n = 3$ questions, the theoretical transition matrix T_3 is given by (3.78). Note that each row of T_3 has $2^3 = 8$ non-zero entries. We can also note that there are many transitions that are not allowed. The total number of transitions that are allowed is given by

$$\sum_{i,j=1}^{48} t_{ij} = 384. \tag{3.71}$$

The description of $\xi = \varphi \circ \sigma$ is very complicated. T_n captures all possible state changes of a social unit that might occur in a very simple (but possibly very large) matrix of zeros and ones. That is, for a single time step, a social unit in any state i maps to state j by

$$S_{t+1} = T_n S_t \tag{3.72}$$

where

$$S_t = \begin{pmatrix} 1 \\ 2 \\ 3 \\ \vdots \\ 2^n \times n! \end{pmatrix} \tag{3.73}$$

represents a vector of all possible states. T_n captures every possible transition that a social unit can make. A transition matrix T_3 does not vary with time but records all possible transitions. T_n is known as adjacency matrix in communication theory [117].

Real data will not in general, show every possible transition. For the Agincourt data, for $n = 3$ questions, we can determine $T_3^{Ag_k}$ by methodically listing all transitions that actually occur in the k'th household. It is useful to define a deficiency matrix T_n^D which identifies the possible transitions that do not occur at Agincourt.

$$T_n^D = T_n - T_n^{Ag_k} \tag{3.74}$$

The definition (3.72) is a theoretical discrete dynamical system [16, 118, 106]. But $S_{t+1} = T_3^{Ag} S_t$ is the real Agincourt dynamical system. The real data defines a map

$$S_{t+1} = T_3^{Ag} S_t \tag{3.75}$$

from the set of states (3.73) to itself.

The maps (3.72) and (3.75) are well known in dynamical systems theory as sub-shifts of finite

type [119, 120, 121, 122]. As a simple example, the system

$$S_{t+1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} S_t, \; S = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \tag{3.76}$$

allows transitions $1 \to 1, 1 \to 2, 2 \to 2$ but not $2 \to 1$. This means that in a long time sequence of state changes, 1 can never follow 2. Thus we have the following example sequences

$$\begin{gathered} 11111111\ldots \\ 11112222\ldots \\ 12222\ldots\ldots \end{gathered} \tag{3.77}$$

In T_3 (3.78), we generate sequences in d symbols $(1, 2, 3, \ldots, 48)$ instead but it can never happen that state 24 follows state 32. It is of interest that Agincourt data might not allow certain transitions.

Each of the sequences of state symbols is an orbit in $S_n = \Sigma_2^n \times \Sigma_n^n$. In Agincourt dynamics, when the deficiency matrix is non-zero, many possible transitions do not occur. The sequences that do occur are the set of orbits of the social unit that are characteristic of the purpose. They hold complete information of state and change of a social unit at Agincourt, and are the foundation for our demographic analysis. Here we remain interested in particular orbits.

Note the very general nature of this discussion. T_n is the matrix of all possible transitions under any purpose whatsoever in n questions. T_n^{Ag} is an example of an experimental transition matrix, say T_n^{Exp}. Then this may have additional zeros, will depend on purpose, the questions asked and may vary between demographic surveillance sites. But even T_n^{Exp} completely characterises all possible transitions of that particular site. The mathematical study of all possible transitions, and of the orbits or trajectories in infinitely long sequences of transitions, of any social unit, is known as dynamical systems. T_n^{Exp} will be the basis of population projection addressed in Chapter Six.

$$\mathbf{T_3} = \begin{smallmatrix}
1&1&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1\\
1&1&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1\\
0&0&1&1&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0\\
0&0&1&1&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0\\
0&0&0&0&1&1&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0\\
0&0&0&0&1&1&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0\\
0&0&0&0&0&0&1&1&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0\\
0&0&0&0&0&0&1&1&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0\\
0&1&0&1&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&1&0&1&0&0&0\\
0&1&0&1&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&1&0&1&0&0&0\\
1&0&1&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&1\\
1&0&1&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&1\\
0&0&0&0&0&1&0&1&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&1&0&1&0&0&0&0\\
0&0&0&0&0&1&0&1&0&0&0&0&1&1&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&1&0&1&0&0&0&0\\
0&0&0&0&1&0&1&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0\\
0&0&0&0&1&0&1&0&0&0&0&0&0&1&1&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0\\
0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&1&1&0&0&0&0&0&0&1&0&1&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&1&1&0&0&0&0&0&0&1&0&1&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&0&1&1&0&0&0&0&1&0&1&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&0&1&1&0&0&0&0&1&0&1&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&1&1&0&0&0&0&0&1&0&1&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&1&1&0&0&0&0&0&1&0&1&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&1&1&0&0&0&0&0&1&0&1&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0&1&0&1&0&1&0&0&0&0&1&1&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&1&0&1&0&0&0&0&0&0&1&1&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&1&0&1&0&0&0&0&0&0&1&1&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&1&0&1&0&0&0&0&1&1&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&1&0&1&0&0&0&0&1&1&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&1&0&1&0&0&0&0&1&1&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&1&0&1&0&0&0&0&1&1&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0\\
0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&1&1&0&0&0&0&0&1&0&1&0&0&0&0&0\\
0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&1&1&0&0&0&0&0&1&0&1&0&0&0&0&0\\
0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&1&1&0&0&0&0&1&0&1&0&0&0&0&0\\
0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&1&1&0&0&0&0&1&0&1&0&0&0&0&0\\
1&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&1&0&1&0&1\\
1&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&1&0&1&0&1\\
0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&1&0&1&0&1&0\\
0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&1&0&1&0&1&0\\
0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&1&0&1&0&0\\
0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&1&0&1&0&0\\
0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&1&0&1&0&0\\
0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0&1&0&1&0&0\\
0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&1&1&0&0\\
0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&1&1&0&0\\
1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&1&1\\
1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&0&0&0&0&0&1&0&1&0&0&0&0&0&0&1&1
\end{smallmatrix} \tag{3.78}$$

3.7.1 Some properties of the transition matrix T_n

Properties of T_n can reveal useful information that can be used to interpret the dynamics of the social system of interest.

Our general process $\xi = \varphi \circ \sigma$ produces transitions that are reversible and not reversible. For example for $n = 3$, the transition $23 \rightarrow 24$ is reversible and a transition $1 \rightarrow 10$ is not reversible. To reverse any transition ($i \rightarrow j$ followed by $j \rightarrow i$) requires that T_n has an inverse ($i = {T_n}^{-1} j = {T_n}^{-1}(T_n i)$). In the general case this cannot be done, for $n \geq 2$, T_n is not *invertible* (singular) and so

$$\text{Det } (T_n) = 0 \, . \tag{3.79}$$

Of course we find this to be true for (3.78). Singularity is of interest because at least one transition $i \rightarrow j$ but $j \nrightarrow i$; an orbit such as there is enhanced possibility of cause.

It may in principle happen in experiment that no irreversible transitions occur, in which case Det $(T_n) \neq 0$. The result (3.79) reveals important properties of the matrix T_3. First, it shows that T_3 is not invertible. Second, it proves that the system under investigation, for which this matrix is associated to, is not *linearly independent* [123, 46]. We can also use this result to interpret the dynamics of any social system of 3 questions.

The symmetry of the associated matrix of the dynamical system is another important property that is useful to understand the system behaviour. In general, under ξ, for $T_n = (t_{ij})$,

$$t_{ij} \neq t_{ji} \, . \tag{3.80}$$

Thus T_n is *not symmetric*. Of course this follows if T_n is irreversible, or if Det $(T_n) \neq 0$. Symmetry merely reflects irreversibility.

It may happen that no change occurs to a social unit, in state i, that $i \rightarrow i \rightarrow \ldots i$ and we say that it *idles*. The trace of the theoretical transition matrix T_n is given by

$$\text{Trace } (T_n) = 2^n \times n!, \tag{3.81}$$

which displays the *total number of idling states*. In experiment, in rapidly changing environment, all social units might never idle once, in which case

$$0 \leq \ Trace \ (T_n^{Exp}) \leq 2^n \times n! \, . \tag{3.82}$$

Even if sometimes it is difficult to classify questions in terms of reasons for asking them, we suggest to establish the relationship between the number of questions in the questionnaires and the purposes of designing these questionnaires. As an example, consider the population growth as a study purpose. Thus from the properties of the well-known balancing equation in demography [124, 125, 32], we can see that if a questionnaire must be designed to address this particular study purpose, it will only consist of three questions. Each of these questions is related respectively to each of the three basic demographic variables: birth, death and migration. This clearly suggests that it is possible to find a fixed number of relevant questions to design a questionnaire that can address a well-defined research question.

Note that the information size is represented by the number of questions n, in the questionnaire. Another way to understand why the application of the present theory requires a constraint on the information size is to consider the relationship between the number of questions n and the dimension of the transition matrix d. Note that for n questions, we have a transition matrix of dimension $d = 2^n \times n!$

To see this clearly, we compare the exponential function and the dimension function d. Thus, Figure 3.22 shows both the exponential function and the dimension d of the transition matrix as a function of the number of questions n. We can clearly see how the function d (dimension of the matrix) increases faster than the exponential function. Such a comparison can be used to carefully decide the reasonable number of questions that is needed to address a specific study purpose, as mentioned in chapter Two. Note for example that for $n = 8$, we have a $d = 645120$ and the transition matrix $T_8 \in M_{645120\times645120}$ which can only be handled by a very sophisticated computer (with a large memory). Thus, it makes it computationally difficult to handle and mathematically difficult to analyse.

3.8 Example of Agincourt household orbits in $\Gamma_{n=3}$

In this section, we give examples of household orbits in the state space Γ_3 for the Agincourt data. We can identify particular social qualities that change for each example that will be presented. We consider three demographic variables related to questions as defined in (2.1).

Consider the Agincourt household $k = 7150$ as presented in Figures 3.23-3.24. The data is as follows Table 3.1.

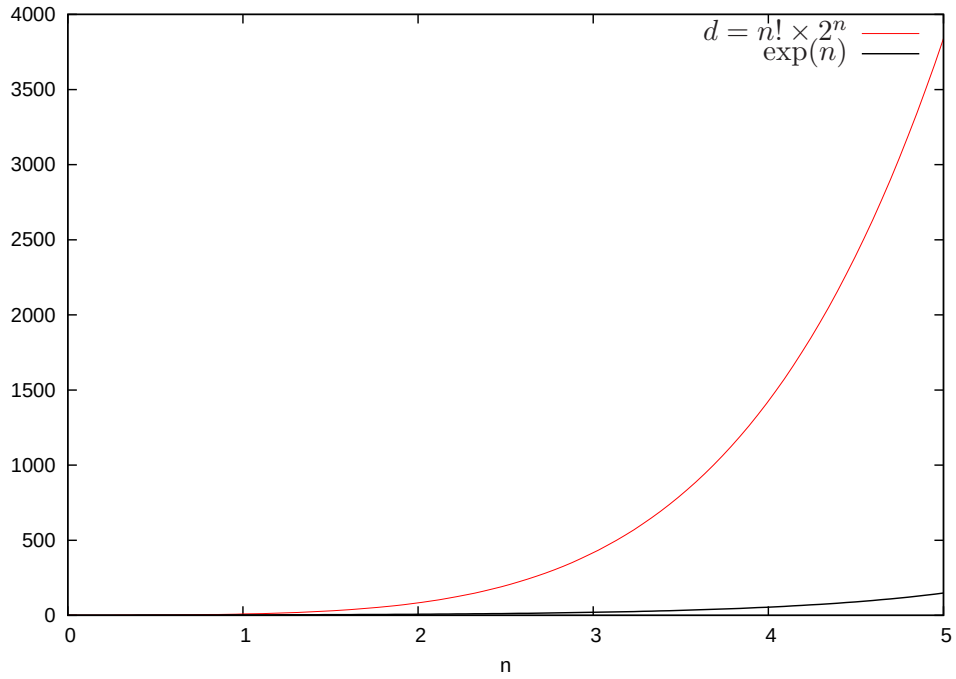

Figure 3.22: Comparison of the exponential function and the distribution of the dimension of the transition matrix d, over the number of questions n.

Table 3.1: Data of the anonymous Agincourt household $k = 7150$.

t	q_0	q_1	q_2	e	χ
1999	1	1	1	0.875	0.555
2000	1	1	1	0.875	0.555
2001	0	0	0	0.000	0.259
2004	0	1	1	0.375	0.185
2005	1	1	1	0.875	0.555
2006	0	1	1	0.750	0.555

The evolutionary orbit in Γ_3 for this household is shown in Figure 3.23 and as a time series over Γ_3 in Figure 3.24. Note that we have arbitrarily begun with question order $\chi = 0.555$.

Here we may identify with the help of the data set the change from $x^{7150}_{2000} = 0.875$ to $x^{7150}_{2001} = 0$ as change from a certain fitness state defined by ${}_w b^{7150}_{2000} = 111$ to the fitness state defined by

${}_w b_{2001}^{7150} = 000$. Note that 3 answers have changed values so that the accumulated displacement $\Delta_{2000,2001}^{7150} = -1$. This maximum negative jump illustrates the most dramatic situation that we can have in the population under these three questions. Note also that we identify an adult death (unfavourable) which in this case has a direct impact on the absence of the biological mother (unfavourable). These two unfavourable events are followed by the household being headed by a minor (unfavourable).

Three years later (at time $t = 2004$) from that dramatic year ($t = 2001$), we observe two positive changes in the characteristics of that Agincourt household. In particular, although the biological mother is still absent from the household, the household is now headed by an adult (favourable). No adult death occurs (favourable) that year. In 2005, the households gets back to its initial favourable fitness state, which is characterised here by the return of the biological mother. Unfortunately, just a year later (at $t = 2006$), the biological mother has to migrate out of the household again.

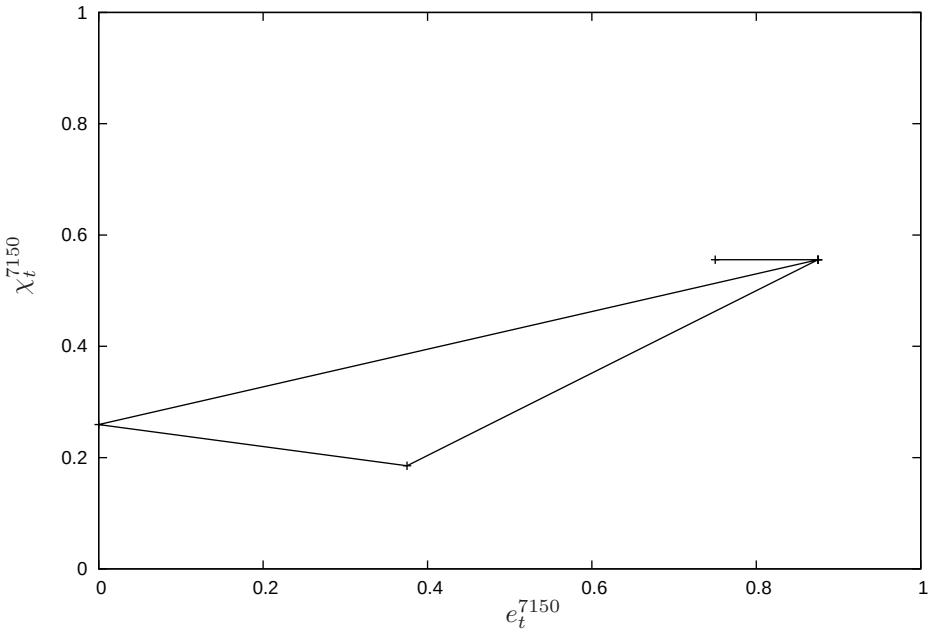

Figure 3.23: Agincourt evolutionary fitness-significance orbit $(e_t^{7150}, \chi_t^{7150})$, for 3 negative questions changing answer values in Γ_3.

In the following example, we give another social dynamics that we observe in the Agincourt population. Consider the Agincourt household $k = 4325$, the data is as follows Table 3.2.

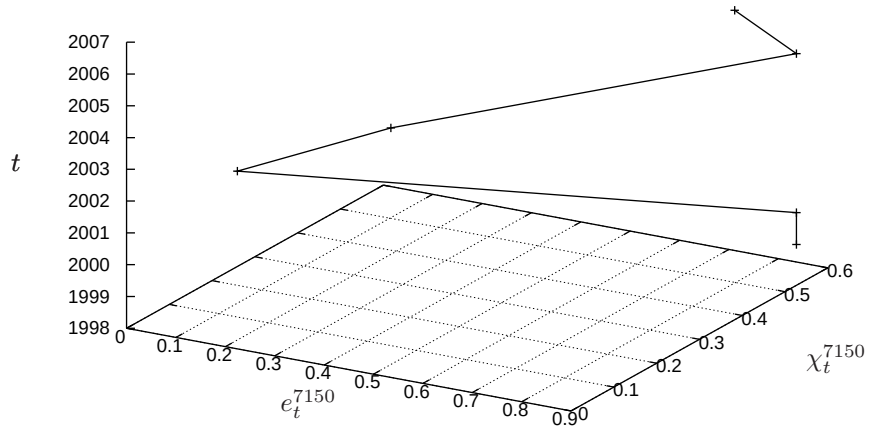

Figure 3.24: Agincourt evolutionary fitness-significance orbit $(e_t^{7150}, \chi_t^{7150})$, for many questions changing answer values, in 3−dimensional space.

Table 3.2: Data of the Agincourt household $k = 4325$.

t	q_0	q_1	q_2	e	χ
1998	1	1	1	0.875	0.555
1999	1	1	1	0.875	0.555
2000	0	1	1	0.75	0.555
2001	0	1	0	0.5	0.407
2002	0	1	1	0.625	0.407
2003	0	1	1	0.625	0.407
2004	1	1	1	0.875	0.555
2005	0	1	1	0.75	0.555
2007	1	1	1	0.875	0.555

As before, the evolutionary orbit in Γ_3 of this household is shown in Figure 3.25 and as a time series over Γ_3 in Figure 3.26. Once again we begin with the arbitrary order $\chi = 0.555$.

From Table 3.2, we can clearly see that the fitness state of this household was favourable for the first two observation years ($t = 1998$ and $t = 1999$). In 2000, the state of this household begins to change unfavourably. In particular, the biological mother out-migrates. This change is seen as an unfavourable event for this household because it is suddenly followed by an adult death that occurs in 2001. The rest of the observation time for this household is characterised by a single social dynamics which here is clearly identified to in- and out-migration of the biological mother.

One important property of the dynamics of this household is that all social changes are concentrated in a parallelogram loop as displayed in Figure 3.25.

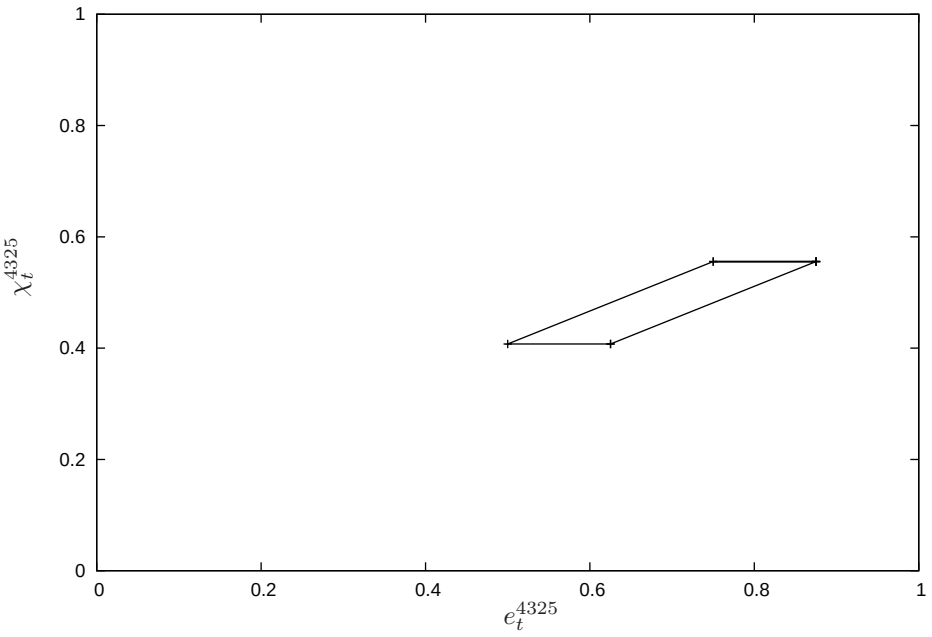

Figure 3.25: Agincourt evolutionary fitness-significance orbit, for various questions changing answer values in Γ_3.

The last example of the household-level fitness-significance state in Γ_3 is as follows. Consider the Agincourt household $k = 5873$, the data is as follows Table 3.3.

Similarly, we present in Figure 3.27, the evolutionary orbit in Γ_3, and as a time series over Γ_3 in Figure 3.28. As before, we consider the same arbitrary question order $\chi = 0.555$.

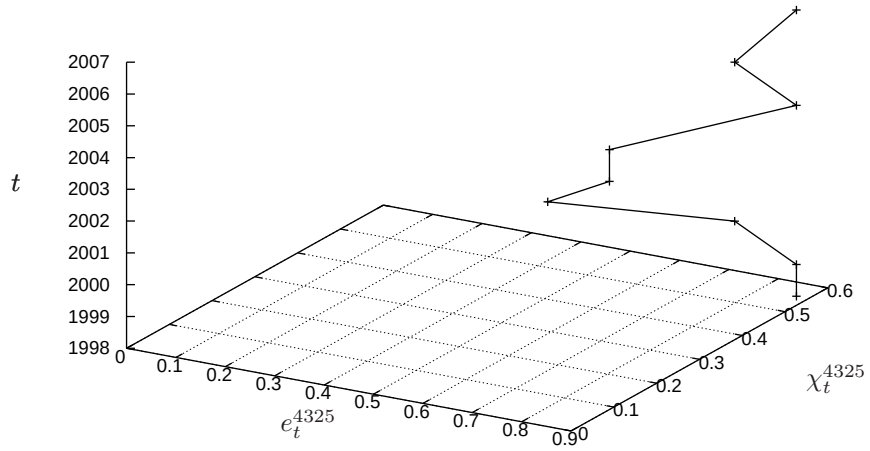

Figure 3.26: Agincourt evolutionary fitness-significance orbit, for various questions changing answer values, in 3−dimensional space.

Table 3.3: Data of the Agincourt household $k = 5873$.

t	q_0	q_1	q_2	e	χ
1999	1	1	1	0.875	0.555
2001	0	1	1	0.750	0.555
2002	1	1	1	0.875	0.555
2003	0	1	1	0.750	0.555
2004	1	1	1	0.875	0.555
2005	0	1	1	0.750	0.555
2006	1	1	1	0.875	0.555
2007	0	1	1	0.750	0.555

As previously stated, here only one question changes answer value each observation year. From Table 3.3, it is clearly shown that the single answer changing value is associated to question q_0. Directly from column 4 of the same table, we can also see that the significance value is unchanged.

To explain this phenomenon, note that because of the arbitrary initial question order $\chi = 0.120$, the changing question q_0 is initially given the weight $\omega_0 = \frac{1}{8}$, which means that question q_0 is initially less significant than other questions. With the evolutionary process over this household, we expect that no significance χ change will occur for this specific case. As a result, the dynamics of this Agincourt household is now reduced to two states of Γ_3 as shown in Figure 3.27. In this case, we expect the dynamics of this household to be related to an oscillation as clearly displayed in Figure 3.28. It is obvious to identify this social dynamics to the movement of the biological mother who sometimes out-migrates and sometimes in-migrates into the household.

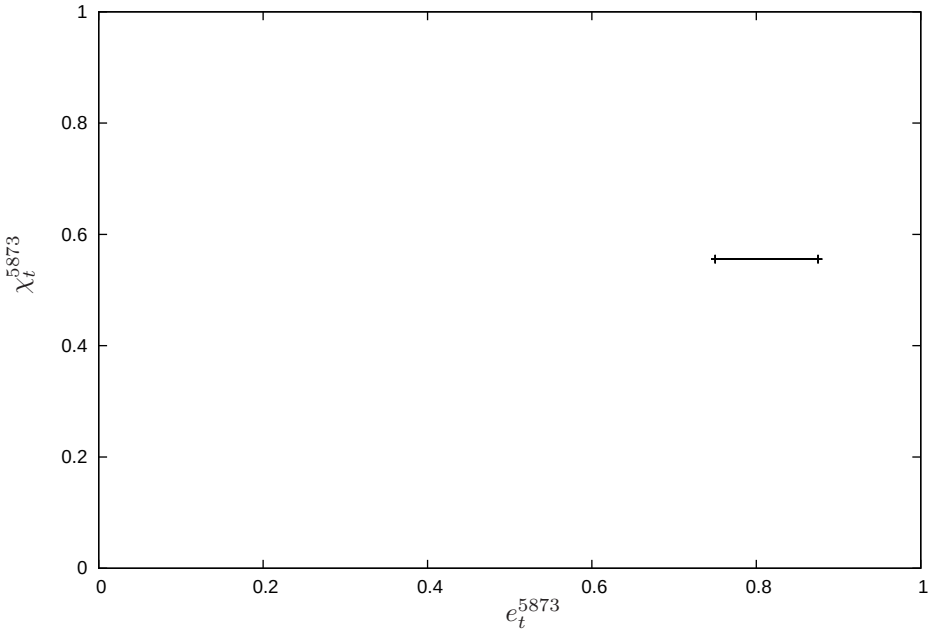

Figure 3.27: Agincourt evolutionary fitness-significance orbits $(e_t^{5873}, \chi_t^{5873})$ Γ_3, for a single oscillating answer value.

3.9 Social unit-level transition matrix T_n^k

In general T_n^k merely exhibits the transitions of the k'th social unit. In principle this may be constant for all time, but in experiment we have a finite time series and must be careful to note the periods of observation. Because this chapter is essentially restricted to the social unit-level of

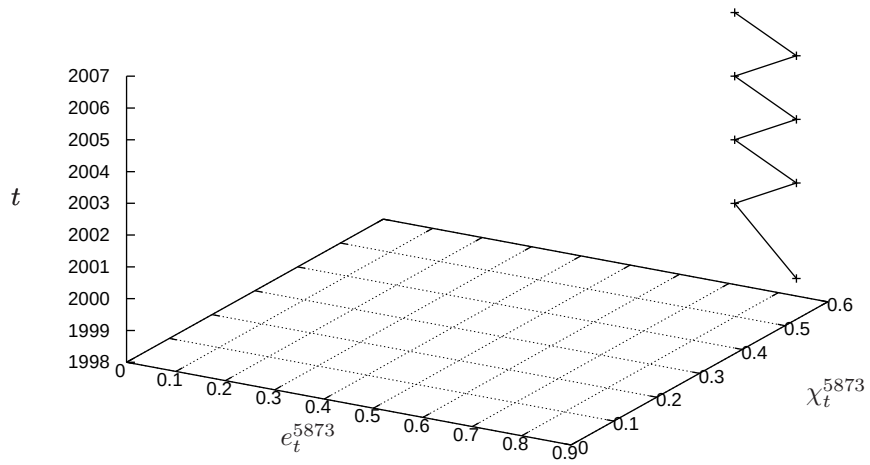

Figure 3.28: Agincourt evolutionary fitness-significance orbit $(e_t^{5873}, \chi_t^{5873})$, for a single oscillating answer value, in 3−dimensional space.

analysis, it is interesting to present an example of the transition matrix for an Agincourt social unit.

Consider data for the household number 7150 as presented in Table 3.1 to discuss its transition matrix. The transition of this Agincourt household is as follows : $24 \to 24 \to 33 \to 44 \to 24 \to 23$ which can be represented in the following table

Table 3.4: Transitions for the Agincourt household number 7150, for $1999 \leq t \leq 2006$.

State index	23	24	33	44
23	0	0	0	0
24	1	1	1	0
33	0	0	0	1
44	0	1	0	0

Table 3.4 displays the indexes of the 4 states of Γ_3 among which the social dynamics of the

household number 7150 are concentrated. Thus T_3^{7150}, for the period $1999-2006$, is a matrix of binary values.

Note that $Det(T_3^{7150}) = 0$ which means that there are irreversible transitions. Here, we note (Figure 3.29) that the household goes through 3 irreversible transitions $24 \rightarrow 33$, $33 \rightarrow 44$, $44 \rightarrow 24$ and the reversible transition $24 \rightarrow 23 \rightarrow 24$ Also, $Trace(T_3^{7150}) = 1$, which means that only one state idles in this household.

If we use Table 3.4, we can associate the dynamics of the Agincourt household number 7150 as shown in Figure 3.23 to the transitions of $\Sigma_2^3 \times \Sigma_3^3$ (see Figure 3.21). This is simply placing the dynamics of Figure 3.23 in Γ_3 into the space S_3 to obtain Figure 3.29.

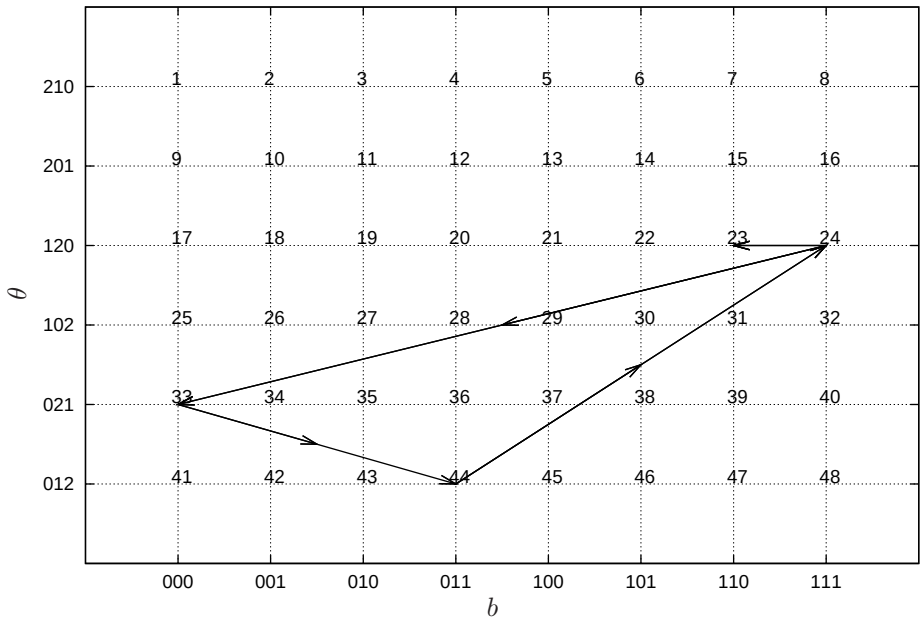

Figure 3.29: The dynamics in S_3, of the Agincourt household $k = 7150$ of Figure 3.23.

Note that the household number 7150 idles for the period $1999-2000$ in a favourable state $(111, 120)$ of symbol 24 which corresponds to no adult death in the household (favourable), the household is headed by an adult (favourable) and also the presence of biological mother in the household (favourable). In 2001, this household made a negative jump to state 33 and the household is now found in a very unfavourable state $(000, 021)$ which is related to the absence of the biological mother which possibly leads to an adult death in the household. As a result, the household is now headed by a minor.

The transition from state 33 to state 44 displays an improvement of the household status. Here we notice that there is a change of the household head from a minor to an adult. Although the biological mother is still absent from the household, there is no adult death in the household.

We also note that the level of significance of the questions is given by the values of θ. If we compare change in significance level between the transitions $44 \rightarrow 24$ and $33 \rightarrow 44$, we note that the presence of the biological mother has been given more weight compared with other questions. Thus, we can see the transition from $44 \rightarrow 24$ related to the return of the biological mother, is captured by an important jump compared to the improvement captures by a transition $33 \rightarrow 44$ in which two questions change answer values.

It is important to see that this example helps us to understand that the size of jump is not only related the number of questions that change values, but more importantly also to the weight of the questions that change answer values.

3.10 Conclusion

In this chapter, the central goal was to introduce Orbit Theory of a social unit. We have achieved an orbit comparable to those of the hard sciences at least within purpose. Thus they have a meaningful sense of direction and magnitude of jump.

New mathematical spaces Γ_n and S_n to visualize social dynamics have been defined. These spaces are related to the given number of questions $n \geq 1$ in the questionnaire and also on the coding of the answers of these questions. In order to monitor transitions that a social unit can make in Γ_n or S_n, we have clearly defined a transition matrix which contains all information of transitions.

Now, given a questionnaire Q, of a fixed number n of questions with binary answers, we have achieved the analysis of social dynamics for a single social unit. We have discussed the behaviour of a single social unit using some illustrations and also Agincourt data. In particular, for the Agincourt data, the household is the unit level of the present analysis. The techniques presented in this chapter allow us to determine properties of a household by the use of its orbits. Fitness and significance changes are easily visualised and a reason for change may be identified.

Note that we were not challenged to decide the weight to give to each question for a single

social unit. Thus, a choice of the arbitrary question order that we suggested was convenient for this specific level of analysis. Thus, we do *not* say that q_{i_0} is *twice* as important than q_{i_1}. We only say that it is *more* important, because it is in position zero.

As stated above, in physics, every change in physical systems is associated to a physical force. In this chapter, we may identify a social force that leads to change in the dynamics of a social unit. For example of no adult death in 2000 but adult death in 2001, then for this household, "no death" is the force of change. This is a demographic study and it is important to extend the discussion of the ideas of this chapter to the population-level. It is only with such a level of analysis that we may attempt to examine patterns of social unit orbits and see whether it will be possible to attach numerical significance to a specific social phenomenon. If the orbits of a significant number of social units display the same (similar) behaviour, then we may identify the social phenomenon leading that sub-population with a social force. The properties of such a social force are very important to understand cause and effect relationship which in turn can lead us to address the concept of causality, in demography.

Chapter 4

Demographic Analysis

4.1 Introduction

So far, orbits discussed for both Agincourt data and illustrative examples were limited to a single social unit. Of course sociologists and demographers are concerned with behaviour of populations. In this chapter, we focus on this important aspect and present the analysis of orbits at the population-level using Agincourt data.

Population study in the social sciences is concerned with everything that influences or can be influenced by population size, distribution, processes, structure, or characteristics [126, 124, 125, 127]. In this chapter, we consider the $(e, \chi)-$plane in real number space Γ_n or sequence space S_n and use it as the space of visualization of the population orbits. We use the techniques developed in the previous chapter, per household, and apply them at the population-level in order to provide an analysis of the dynamics. Here, particular attention will be placed on the determination of the structure or characteristics of the population using patterns of social unit orbits at any one time and in time series. Identification of sub-populations will be of interest as will the dynamics of these.

4.2 Population-level dynamics models

Before we discuss the general orbit theory at the population-level, it is important that we first discuss the parameters of the models. Note that because the methods presented here constitute an extension of those discussed in the previous chapter, the properties of certain parameters of the

data presented in the previous chapter can be assumed to be the same. Thus, it is reasonable to keep the number of questions n, in the questionnaire, the length of the observation time of each social unit l_k and the total number s of social units in the population fixed.

4.2.1 Initial conditions of the social systems

In order to analyse the dynamics of the population in the $(e, \chi)-$plane, the state $\zeta_t^k = (e_t^k, \chi_t^k)$ of each social unit k, must first be determined, for each time step t. In the previous chapter we made one assumption only (granted purpose and question set) to determine the state of a social unit. This was to initialise all example orbits from Agincourt in an arbitrary significance level, for example $\chi_0^k = 0.185 \in I_3^3$ or $\theta_0^k = 012 \in \Sigma_3^3$. The orbit arising from this is unique, but we have not yet developed a rational strategy for initial question order.

There are several scenarios we could imagine. For example, all social units can start at the same significance level, or we can randomly choose the initial significance value for each social unit. Another scenario would be to fix the initial significance value for each social unit according to the specific frequency of questions, changing answer values of that social unit.

In this thesis, we will use the following two strategies. First, we suppose that the initial question order is randomly chosen in S_n for each social unit. We refer to this case as the "riqo" (random initial question order) scenario. This may be justified in the case where most orbits are random and will thus visit all states in the space in a random order.

Second, the initial question order of each social unit is fixed according to the specific frequency of answer value change of that social unit taken from longitudinal data. We refer to this case as the "fhiqo" (frequency of household initial question order) scenario. This is natural in the sense that the operation φ (3.54) sorts states by frequency. These orbits start in an "already" sorted state, with known fitness and they will thus tend to converge more rapidly to any sub-space than will a "riqo" orbit. When frequency of answer value change is uniform for a social unit, it is reasonable to adopt the population average distribution of frequencies. Demographic time series are relatively short and when orbits are indeed random, and where there is Brownian motion [128, 129, 130, 48, 51, 52], convergence of an orbit to stable question order is typically slow ($\propto\sqrt{t}$). Note that in statistics we hypothesize a population level order and then are careful not to bias outcome in favour of that hypothesis. Here, the present strategy to change question order by frequency is a *deterministic*

action to expose sub-populations, and is not a hypothesis. We may then bias initial conditions to force out sub-populations.

A third strategy might be to start all social unit orbits with question order corresponding to the population frequency, and with their known fitness. In this case all orbits start at the same significance value χ_0. There is no advantage in doing this when the detailed knowledge is available and we reject this scenario.

4.2.2 Population transition matrix

Using the definition (3.47) for each social unit, it is possible to determine the jumps each social unit in the population can make in Γ_n. If each possible jump in Γ_n is associated to a transition as described in Figure 3.21, then this leads us to find all possible transitions of the population. Thus, the associated transition matrix can be determined using the definitions (3.69) and (3.70).

Let T_n^P denote the transition matrix for a population P, then we recall definitions (3.69) and (3.70), that is

$$T_n^P = [t_{ij}^P] \quad i,j = 1,2,3,\dots n! \times 2^n \tag{4.1}$$

where

$$t_{ij}^P = \begin{cases} 1 \text{ if } \exists\ k \in \{1,2,3,\dots,s\} \mid i \to j \\ 0 \text{ if } i \nrightarrow j\ . \end{cases} \tag{4.2}$$

We can use (4.1) and (4.2) to define the transition matrix for the population as a function of time. Let $T_{t,n}^P$ denote the transition matrix related to a questionnaire of n questions, for the population P consisting of s social units, at time t. Similarly, we define $T_{t,n}^P$ as follows:

$$T_{t,n}^P = [t_{t,ij}^P] \quad i,j = 1,2,3,\dots n! \times 2^n\ , \tag{4.3}$$

where, *at time t*

$$t_{t,ij}^P = \begin{cases} 1 \text{ if } \exists\ k \in \{1,2,3,\dots,s\} \mid i \to j \\ 0 \text{ if } i \nrightarrow j\ . \end{cases} \tag{4.4}$$

Thus, $T_{t,n}^P$ gives important properties of the social dynamics of the population P over time. For instance, it can help to capture when and what new social changes have been observed in the population P and also what other social dynamics are removed from the population. To illustrate, suppose that at time t, for a certain transition $i \to j$, we have $t_{t,ij}^P = 0$. This means that the social change related to the transition $i \to j$ does not exist in the population at time t. If after a certain

period of time $\tau \geq t$, surprisingly we find that $t^P_{\tau,ij} = 1$, then we can use τ as a reference in time that this particular social change started to be observed in the population. Another example could be the identification of an observation time, say, $t_{dramatic}$ where most of the negative social changes began in the population.

The properties of the time-dependent transition matrix $T^P_{t,n}$ can help to check the validity of the data. Suppose that data for a population have been collected for each time $t, 0 \leq t \leq l$. Compute the transition matrix $T_n = (t_{ij})$ of all theoretically possible transitions. This is given by (3.78). If there exists $t^P_{ij} = 1$ when $t_{ij} = 0$, there is an error in the data. If there exists $t^P_{ij} = 0$ when $t_{ij} = 1$, it is not an error but signals that a transition does not take place in P.

It is important to note that, in general, $T^P_n \neq T^{P'}_n$, $P \neq P'$. Thus if P is the population of households where child education is favourable, P' is the population where it is not favourable, comparison is of obvious interest.

According to the definitions (4.1), (4.2), (4.3) and (4.4) of the transition matrix, we are able to know whether or not (and when) a social change related to a certain transition $i \rightarrow j$ happens in the population. However, the transition matrix does not provide us with the information of how significant a certain social change is in the population, which is very useful to sociologists in order to advise policy-makers.

4.2.3 Population density matrix

There are several properties of Γ_n or S_n that can be of particular interest to sociologists. For example, they might be concerned with exploring the information about how many social units are in each state of interest in Γ_n. Suppose that a specific state $\zeta_{HIV} = (e_{HIV}, \chi_{HIV})$ is identified [103] as a new (very unfavourable) state in the population, perhaps a case of HIV infection newly found in an HIV-free population. If intervention strategies have to be developed in order to reduce the spread of this infection, then public health researchers would like first to investigate the dynamics of the spread of HIV infection in the population before implementing any of the intervention strategies. Thus, their particular attention would be more focused on counting the social units that are in HIV-state at each time t and also investigating how these figures change over time. With this kind of information they can better inform policy-makers on certain recommendations or interventions leading to assist the population with regard to this kind of infection.

Suppose a specific transition $(i_{TB}) \rightarrow (j_{HIV})$ is investigated in order to examine whether or not TB infection leads to HIV-infection, then public health researchers would also need to determine the number of social units making this specific transition $i \rightarrow j$ at each observation time t, in order to control the dynamics of TB infection in the population and understand why is it assumed to be responsible of HIV infection. This is possible causal information under some purpose.

In general, suppose again that we investigate the dynamics of some demographic events in a given population P. Assume that each social unit k is observed for a period of time l_k. If a specific transition $i \rightarrow j$ is of great interest to policy-makers, then the answers to the following questions would give some insights in the understanding of the dynamics of the demographic events under consideration.

1. How many social units in the population made the transition $i \rightarrow j$ at a given time t, $0 \leq t \leq l_{given}$?

2. What are the transitions that happen the most in the population?

Define the density matrix for the entire population up to time t as follows:

$$D_n^P = [d_{ij}^P] \quad i, j = 1, 2, 3, \ldots n! \times 2^n \tag{4.5}$$

where

$$d_{ij}^P = \begin{cases} \sum_{k=1}^{s} \sum_{t'=0}^{t} t_{t',ij}^k & \text{if } \ i \rightarrow j \\ 0 & \text{if } \ i \nrightarrow j \ . \end{cases} \tag{4.6}$$

Define the density matrix for each observation time $t' = 1, 2, \ldots, t$.

$$D_{t',n}^P = [d_{t',ij}^P] \quad i, j = 1, 2, 3, \ldots n! \times 2^n \tag{4.7}$$

where

$$d_{t',ij}^P = \begin{cases} \sum_{k=1}^{s} t_{t',ij}^k & \text{if } \ i \rightarrow j \\ 0 & i \nrightarrow j \end{cases} \tag{4.8}$$

The properties of T_n^P might be distinct from T_n and give valuable information regarding events that do not occur in P. There is no theoretical density matrix D_n. Note that D_n^P, $D_{t,n}^P$ are unique to P and can be very different for $P \neq P'$; thus migration might be absent in an urban community, present in a rural community [131, 132, 133, 134].

4.2.4 Population flux vector

The elements $d_{t,ij}$ of $D^P_{t,n}$ do not directly measure how the density of any state changes over time. To achieve this we use the definitions (4.7) and (4.8) to define a measure which captures that change. The flux is a net flow [135] into state i, per unit time and is

$$\delta_{t,i} = \sum_{j=1, j\neq i}^{d} (d_{t,ji} - d_{t,ij}),\ t \geq 1. \tag{4.9}$$

Note that $\delta_{t,i}$ can either be positive or negative. If $\delta_{t,i} > 0$, it means that there are increasing social units in state i at time t. If $\delta_{t,i} < 0$, it means there is a decrease in the number. If $\delta_{t,i} = 0$, then in-flow equals out-flow. This indicates that state i is a steady state [135, 136].

In physics net quantity flowing into a state per unit time is called flux. We define the population flux vector at time t as follows.

$$\underline{f}^P_t = (\delta_{t,i}) \quad i = 1, 2, 3 \ldots, n! \times 2^n,\ t \geq 1. \tag{4.10}$$

4.3 Agincourt population dynamics

In this section we use the Agincourt data as described in the second chapter to discuss the population dynamics. The discussion of this section is organized as follows. First, we describe the properties of the Agincourt transition matrix T^{Ag}_3, density matrices $D^{Ag}_3, D^{Ag}_{t,3}$ and the flux $\underline{f}^{Ag}_{t,3}$. Second, we present and discuss not one but all orbits of the Agincourt population. Finally, we provide a detailed demographic analysis of the Agincourt data with respect to changes in household characteristics and how they affect educational progression of children.

As stated in (2.1), for the analysis of the Agincourt data, the number of questions is fixed to $n = 3$. The total number of social units in the present studied population sample is fixed to $s = 2669$ which represents the number of households with an observation time $l \geq 5$ in the final selected studied population sample (see Section 2.5.3). The average length of observation time for the Agincourt data is $\overline{l} = 7.115$ (years).

We refer to Figure 2.2 for the frequency of value change for an answer to each question. This shows that Agincourt dynamics is dominated by absence of biological mother (about 86.72%), adult death (about 13.12%) with minor heads of households (about 0.16%). Figure 2.3 displays the distribution of the number of answers that change value at each time step. From this distribution,

the average number of questions that change per time step is calculated and we found that it is about 0.53. The number of answer changes per time step is dominated by no change (idle), (about 48.76%) and one change (about 47.9%).

4.3.1 The "fhiqo" and "riqo" starting strategies

We present in (4.14) and (4.15) the transition matrices for the Agincourt population respectively for the "fhiqo" and "riqo" scenarios. To better understand the discussion of the differences between the properties of the two scenarios, we present the difference $T_{diff}^{riqo,fhiqo} = T^{Ag(riqo)} - T^{Ag(fhiqo)}$ in (4.16).

Although the properties of the transition matrices for any given population data depend only on its own characteristics, it is possible to interpret them with respect to the properties of the theoretical transition matrix as described in (3.79), (3.81). Thus, it is important to also discuss the properties of the difference $T_{diff}^{Theo,fhiqo} = T_3 - T^{Ag(fhiqo)}$ as given in (4.17) and $T_{diff}^{Theo,riqo} = T_3 - T^{Ag(riqo)}$ as given in (4.18).

We examine the properties of the Agincourt transition matrices to time t to derive some social properties of the Agincourt population as proposed in (3.74). Directly from (4.14) and (4.15), we find the following results:

$$\sum_{i,j=1}^{48} (t_{ij}^{Ag(fhiqo)}) = 58, \quad \sum_{i,j=1}^{48} (t_{ij}^{Ag(riqo)}) = 93 \quad \sum_{i,j=1}^{d} (t_{ij}) = 384. \tag{4.11}$$

$$\text{Det } (T_3^{Ag(fhiqo)}) = \text{ Det } (T_3^{Ag(riqo)}) = \text{ Det } (T_3) = 0 \, . \tag{4.12}$$

$$\text{Trace } (T_3^{Ag(fhiqo)}) = 11, \quad \text{Trace } (T_3^{Ag(riqo)}) = 17, \quad \text{Trace } (T_3) = 48 \, . \tag{4.13}$$

As stated in the previous chapter, again we understand that for the Agincourt population not all transitions are allowed. Thus, the result (4.11) gives the total number of transitions that happen in the Agincourt population, with respect to each of the two mentioned scenarios. Note that the total number of transitions under the "fhiqo" scenario is about 15.10% of the total number of possible transitions (3.71). Under the "riqo" scenario we have about 24.21% of the total number of transitions. As expected, we can see that the "riqo" scenario gives more transitions (about 9%)

compared to the "fhiqo" scenario. This is an evidence that "fhiqo" gives the expected fast convergence and we can immediately attach importance to this strategy. Given the short observation time ($\bar{l} = 7.115$), this is useful.

To judge "riqo", we must first discuss randomness of orbits. It is clear that a random process [128] would visit all possible states (or make all possible transitions as in theoretical T_3) under some probability distribution. Because households do not do this, the dynamics is *not* random in Γ_3 or S_3. We should not then, select the "riqo" starting strategy, for the Agincourt data. Note that dynamics may still be random within transitions defined by T_3^{Ag}.

The result (4.12) is the same as that of the theoretical transition matrix (3.79) and shows that regardless of the initial order that we give to the questions, the dynamical system concerning the Agincourt data related to the present questions is not invertible. As a result, there are non-reversible transitions in Agincourt data. Such "one-way" change clearly relates to causality.

Note that if each state idles, then for $n = 3$ questions we have 48 (3.81) idling states. The total number of idling states for each scenario is given in (4.13) for the present Agincourt population data. We find that under the "riqo" scenario, we have 35.41% of states that idle while the "fhiqo" scenario presents about 23% of idling states. Under the "riqo" scenario, each social unit is given a random initial question order, which in turn implies a random initial state and it has more chance to start in an idle state.

We observe that there are 3 transitions ($19 \to 14$, $20 \to 38$ and $33 \to 44$) that occur under the "fhiqo" which are not observed under the "riqo". Note that these three transitions relate to the rare case (about 0.16%) of a change in question 1 (household head is a minor). Thus, we discover that the "fhiqo" scenario is not only good in convergence but has an additional advantage of revealing transitions that are lost by randomly sampling initial states.

Finally the results (4.17) and (4.18) are useful for checking errors in the data. Note that for every single transition $i \to j$, we must always have $T_{diff_{ij}}^{Theo,fhiqo} \geq 0$ and $T_{diff_{ij}}^{Theo,riqo} \geq 0$ otherwise there is an error in the data. No such error was found in the Agincourt data set.

$$\mathbf{T_3^{Ag(fhiqo)}} = \begin{smallmatrix}
0&0&0&

$$\mathbf{T_3^{Ag(riqo)}} = \begin{matrix}
0\,0\,0\,

$$\mathbf{T}^{\mathbf{riqo,fhiqo}}_{\mathbf{diff}} = \begin{matrix}
0\,0\,0\,

$$\mathbf{T}_{\mathbf{diff}}^{\mathbf{Theo,fhiqo}} = \begin{array}{c}
1\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1 \\
1\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1 \\
0\,0\,1\,1\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0 \\
0\,0\,1\,1\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0 \\
0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0 \\
0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0 \\
0\,1\,0\,1\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0 \\
0\,1\,0\,1\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0 \\
1\,0\,1\,0\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1 \\
1\,0\,1\,0\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1 \\
0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0 \\
0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0 \\
0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,1\,1\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,1\,1\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,1\,0\,1\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,1\,0\,1\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,1\,0\,1

$$\mathbf{T}^{\mathbf{Theo,riqo}}_{\mathbf{diff}} = \begin{matrix}
1\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1 \\
1\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1 \\
0\,0\,1\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,1\,1\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,1\,0\,1\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0 \\
0\,1\,0\,1\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0 \\
1\,0\,1\,0\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1 \\
1\,0\,1\,0\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0 \\
0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,1\,1\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,1\,1\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,1\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,1\,0\,1\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0 \\
0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,1

4.3.2 Agincourt density matrices

In this section, we present and discuss the properties of the Agincourt density matrices. We first discuss the properties of the density matrix for the overall observation time. After that we will examine the dynamics in the density matrices for each observation year from 1998 to 2006. We recall that the density matrix of year t is calculated from the transitions occuring between year t and year $t+1$. Thus, with this definition and considering the range of the Agincourt observation time, it is convenient that no Agincourt density matrix is defined at $t = 2007$.

Based on the present two scenarios, we present in (4.21) and (4.22) respectively the Agincourt population density matrices for "fhiqo" and "riqo" scenarios. To facilitate the discussion, we extract the dominant transitions for all years of study in Tables 4.1 and 4.2, for the indicated scenario. It is clear that these two density matrices are not symmetric. Of course this happens because T_3^{Ag} is not symmetric. But now we see that even where $d_{ij}, d_{ji} \neq 0$, density can be different. For example, $d_{23,24} < d_{24,23}$ for all years of study. It is also clear that the Agincourt population behaviour with respect to the present questions is not totally random that is, random over Γ_n or S_n. There must be a deterministic process leading the social dynamics in the Agincourt population to these sub-spaces. The dynamics within a sub-space might be random [128] or regular.

The analysis of the density matrix is quite complex compared to that of the transition matrix where entries are binary numbers. This is where we can measure the impact of each social change. Note that these two density matrices, (4.22) for "riqo" and (4.21) for "fhiqo" are computed for the whole observation time. Thus, they contain the information about the population dynamics for the observation time l.

Before comparing the two density matrices, it is important to verify that the total number of transitions in "riqo" and in "fhiqo" must be as follows

$$\sum_{i,j=1}^{48} {d_{ij}}^{Ag(riqo)} = \sum_{i,j=1}^{48} {d_{ij}}^{Ag(fhiqo)} = 16321 \tag{4.19}$$

because each scenario under consideration does not delete any transition that happens in the population.

Thus, any difference that we will mention in the discussion, will concern the distribution of the number of such transitions in each element of the density matrix. A quick look at the entries of each density matrix shows the difference in the distribution of numbers for each matrix. We see

that there are more non-idling transitions happening in the Agincourt population under the "riqo" scenario compared to the "fhiqo" (see (4.16)). This is consistent with our earlier discovery that where initial conditions are fixed by frequency of change of questions, there is rapid convergence of orbits.

More importantly, the four transitions

$$\begin{array}{ccc} 24 \to 24 & \longleftrightarrow & 23 \to 23 \\ (111, 210) & & (110, 210) \end{array} \tag{4.20}$$

are dominant for the two scenarios. Thus, regardless of the scenario, we are able to capture the social phenomena that lead to significant change within the population. We see that Agincourt households that get into a favourable state 24 are more likely to stay in that state for a long time as compared to unfavourable state 23. We understand from Figure 3.21 that this is the case of the advantaged, fully fit Agincourt sub-population.

On the other hand, we observe that state 23 corresponds to Agincourt households with absence of the biological mother. What happens in these households is that once a biological mother leaves her household, she takes time to return into the same household. As a result, we have a significant number of transitions $23 \to 23$.

Now let us examine some of the important transitions that differ between the two scenarios. The following are some of the transitions that occur in the "riqo" but do not happen under the "fhiqo" scenario. They are $7 \to 7 \leftrightarrow 8 \to 8$, $48 \to 48 \to 23$, $40 \to 40 \to 7$, $16 \to 16 \to 7$, $36 \to 8$ and $44 \to 24$.

First of all note that we have already explained (using the result (4.13)) why there are more idling states in the "riqo" than in the "fhiqo" scenario. Now apart from the idling transitions, in all the other above transitions, with the help of S_3 (see Figure 3.20) we can clearly see that only question q_0, which is related to the absence of the biological mother, is changing answer value. This has already been identified as the dominant social changes that is observed in the Agincourt population.

Secondly, if we look at the fitness value of each of the states of the above transitions, it is possible to associate each of them to the states 23 and 24 which are again the common dominant transitions as mentioned above. For example the states 7 and 23 have the same fitness value, the states $8, 16, 24, 40$ and 48 have the same fitness value.

Finally, we can clearly see that these transitions are built up of states that are defined by randomly selecting significance value. Thus, we can conclude that Agincourt household orbits that normally should start at $\theta = 120$ were given by the "riqo" scenario a different significance value, for example $\theta = 210, 021, 012$.

In order to explore the information from the density matrix D_3^{riqo} or D_3^{fhiqo}, much attention of demographers will be focused on the significant numbers. *We now assume a transition significance level,* $d_{ij} \geq 100$, *in the remainder of this thesis.*

It is important to identify the social forces that drive significant transitions. However, it can be quite challenging to go through every element d_{ij} of the density matrix in order to investigate these properties. Thus, we generate the following tables (4.1) and (4.2) respectively from D_3^{fhiqo} and D_3^{riqo}. Through each of these table, we have in a reduced form, all information that is necessary to identify the significant social phenomena leading social changes within the Agincourt population. These tables display the sub-region of S_3 where the attention of demographers must be focused.

$$\mathbf{D_3^{fhiqo}} = \begin{matrix}
0\,0\,0\,

$$
\mathbf{D}_{\mathbf{3}}^{\mathbf{riqo}} =
\begin{matrix}
0\,0\,0\,

Table 4.1: Sorted number of transitions from the density matrix $D_3^{fhiqo}, d_{ij} \geq 100$.

state i	**state** j	d_{ij}
24	24	4336
24	23	3650
23	23	3164
23	24	3163
32	32	253
32	23	194
24	31	178
31	23	139
23	31	132
23	29	129
30	30	127
31	32	126
24	29	125
30	24	116
29	30	109

Table 4.1 and Table 4.2 display by decreasing order the transitions that occur. Using each state index given in columns 1 and 2 of each table, and with the help of the state space S_3 (see Figure 3.20), we can identify the fitness and significance associated to each transition. For example 24 indexes the state $(111, 120) \in S_3$. This state is one of the above-mentioned favourable states in S_3. It corresponds to the situation of households being headed by adults, in which all children from age 6 to 17 years live with their respective biological mothers. No adult death occurs in the Agincourt households which are at this state.

Tables 4.1 and 4.2 show that regardless of the scenario, the idling state 24 or $(111, 120)$ is the most common in the Agincourt population. Thus, we find that most of the Agincourt households, once they get into this particular condition, stay there for the rest of their observation time, which is a positive social change that happens in the Agincourt population.

Note that for all the results discussed so far, the population dynamics under the "fhiqo" scenario

Table 4.2: Sorted number of transitions from the density matrix $D_3^{riqo}, d_{ij} \geq 100$.

state i	**state** j	d_{ij}
23	23	1517
23	24	1512
24	24	1477
7	7	1396
24	23	1365
8	8	1329
7	8	1302
8	7	1234
48	48	516
32	32	503
40	40	427
16	16	381
32	23	366
48	23	329
16	7	291
40	7	287
30	30	178
30	24	158
36	8	118
31	23	117
44	24	114
31	32	111

converges faster compared to the "riqo" scenario and that the dynamics is nonetheless qualitatively similar. *It is convenient to suggest that we use the "fhiqo" scenario for the rest of the analysis in this thesis. We now drop the "fhiqo" suffix and refer to* D_3^{Ag}.

Now apart from the transitions associated to idling states, we can clearly see from Table 4.1

that the transitions $24 \to 23$ (about 22.36%) and $23 \to 24$ (about 19.37%) are the most significant social changes to households. This might be associated to period 2 household orbits, which correspond to the situation of an on-off-on-. . . move of biological mothers from the households. We note that this type of move does not have a negative impact on other household characteristics under consideration. In fact, the household heads are constantly adults (favourable) and there is no adult death (favourable) in the households regardless of the biological mothers being absent or present in the households. We find that this type of move of biological mothers is not related to negative demographic dynamics for the households, *before we consider educational default*.

Table 4.1 helps to reduce the space of visualization of the population dynamics from the whole state space Γ_3 to an automatically identified region of Γ_3, defined by $0.5 \leq e \leq 1$ and $0.4 \leq \chi \leq 0.5$. In this region, we only have only a few significant transitions to investigate. The Agincourt population is then automatically divided into sub-populations with identified fitness and significance as we expected in designing the reordering of questions. The sub-population of great interest consists of households jumping between 6 states in Γ_3. These states are respectively indexed by $23, 24, 29, 30, 31$ and 32. Thus, all significant dynamics of the Agincourt population can be summarized in Table 4.3 in which the X entries represent insignificant (≤ 100) number of transitions.

Table 4.3: Transitions of Table 4.1: significant transitions from the density matrix D_3^{Ag}. Note lack of symmetry.

State index	23	24	29	30	31	32
23	3164	3163	129	X	132	X
24	3650	4336	125	X	178	X
29	X	X	X	109	X	X
30	X	116	X	127	X	X
31	139	X	X	X	X	126
32	194	X	X	X	X	253

From Table 4.3, we can see that the states 29 and 31 do not significantly idle. We have decided to ignore these infrequent transitions. Consider the transition $31 \to 31$ or $(110, 102) \to (110, 102)$ where $d_{31,31} = 7$. The idling state is related to households headed by adults, with the presence of the biological mother, but at the same time with an adult death. It means that in the Agincourt

households with the above characteristics, adult death does not significantly occur every year. Now what transition can a household make from state 29? Only transition $29 \rightarrow 30$ (no adult death in the following year) significantly happens. Note the reverse $30 \rightarrow 29$ (adult death in next year as well) does not happen so often ($d_{30,29} = 5$). Basically, all significant transitions involve biological mother.

We conclude again that the behaviour of the Agincourt population is not random with respect to the demographic variables (2.1) because orbits do not wander over the whole space Γ_3. As shown in the above tables, the dynamics of the Agincourt population sample is concentrated in a sub-space of Γ_3. In fact, this is a direct consequence of the distribution of questions changing values that we discussed in chapter 2. We can clearly see that according to the distribution given in Figure 2.2, answers given to question 1, related to the household head is a minor, are almost always unchanged. We previously argued that in such a case, this question can be deleted. As a result, the dynamics of this population is now reduced in the space $S_2 \subset S_3$ that we show in Figure 4.1. Reducing complexity is an advantage of our method. In principle we could choose more questions in Q (not available to us) look for clustering in Γ_n or S_n, and then analyse sub-populations as in Figure 4.1.

Both Table 4.3 and Figure 4.1 identify typical orbits in the Agincourt population. They are defined by transitions between the following 6 states $23, 24, 29, 30, 31, 32$ of Γ_3 or S_3. We clearly see these typical orbits are dominated by oscillations between states 23 and 24. There are only a few excursions to states $29, 30, 31$ and 24.

Definition 4.1. The approximate transition matrix $T^{Approx.} \in M_{48\times 48}$ describes only the dominant transitions of Table 4.3. Similarly $D^{Approx.} \in M_{48\times 48}$.

4.3.3 Distribution of the number of transitions in each observation year

Figure 4.2 displays the distribution of the total number of transitions that happen in each year of study. The dynamics is related to the results (4.11) that we discussed before. The total number of transitions shows no dramatic changes. For each observation year the total number of transitions that occur in the Agincourt population varies between 27 to 38. This means that the number of transitions that occur in each observation year is less than 66% of the total number (58) of transitions observed in the Agincourt population.

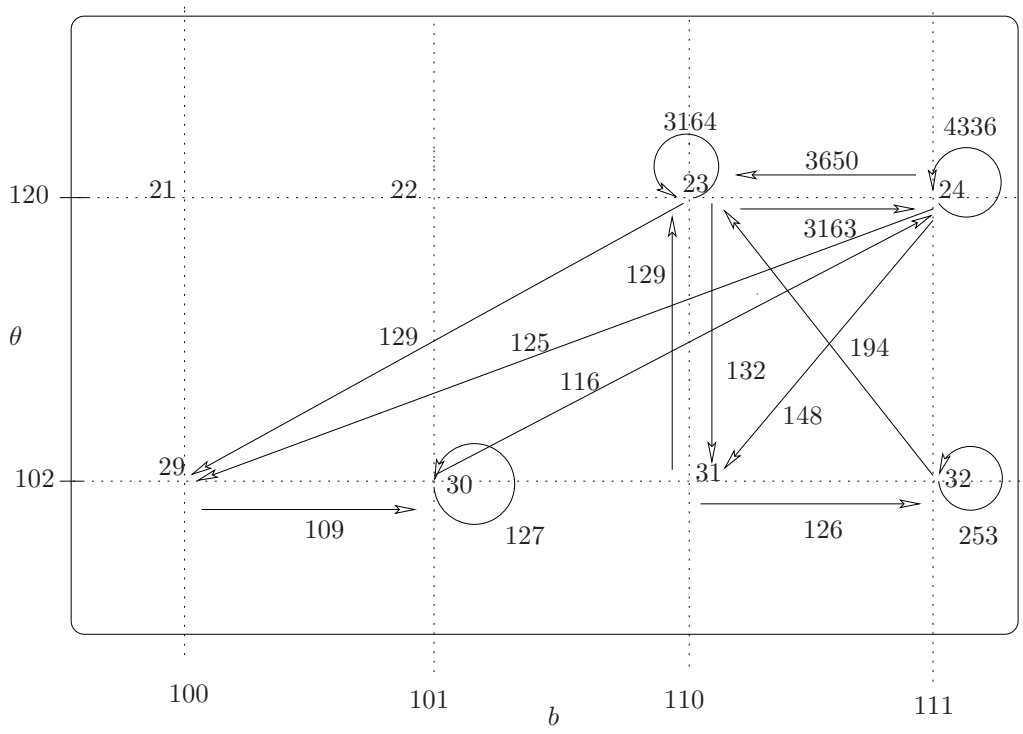

Figure 4.1: State space S_n for $n = 2$ binary-valued questions with states numbered and number of transitions. In this space, every household is headed by an adult.

As before, we summarise the dominant transitions in Table 4.4 for each year of study. Table 4.2 and Table 4.1 give dominant transitions added up over all years, $d_{ij} \geq 100$, here we give dominant transitions in each year, $d_{t,ij} \geq 10$. The exchanges $24 \to 24 \leftrightarrow 23 \to 23$ are always present. To recall, these are changes related to the absence of the biological mother. The new non-idling state change is typically $22 \to 32$, in 1998 and 1999. It is a demographic change related to adult death. Apart from this new transition, the distribution in each observation year is comparable to the overall distribution of Table 4.3 and Figure 4.1.

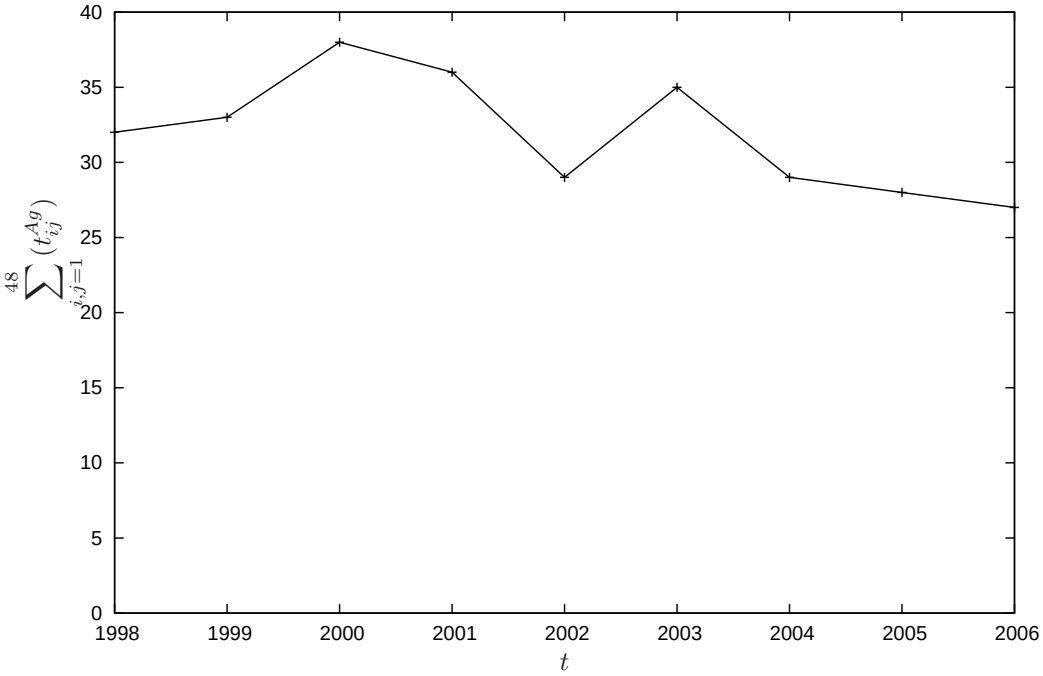

Figure 4.2: Distribution of the number of transitions in the Agincourt population over years 1998 to 2007.

Table 4.4: Reduced Agincourt density matrices for the indicated year, $d_{t,ij} \geq 10$.

i	j	$d^{Ag}_{1998,ij}$
24	24	725
24	23	365
23	24	219
23	23	133
32	32	33
22	32	22
24	31	17
32	23	16
30	24	16

i	j	$d^{Ag}_{1999,ij}$
24	24	738
24	23	555
23	24	315
23	23	233
32	32	43
32	23	33
24	31	30
30	24	25
22	32	13
23	31	10

i	j	$d^{Ag}_{2000,ij}$
24	24	560
24	23	451
23	24	402
23	23	359
32	32	38
24	31	26
32	23	23
31	23	17
24	29	17
31	32	17
23	31	15
30	30	13
30	24	11

i	j	$d^{Ag}_{2001,ij}$
24	24	564
24	23	447
23	24	409
23	23	394
32	32	36
32	23	29
31	32	23
24	31	20
31	23	16
30	30	16
23	29	15
23	31	15
24	29	13
29	24	12
29	30	10

i	j	$d^{Ag}_{2002,ij}$
24	23	415
23	23	392
24	24	351
23	24	342
23	31	23
32	32	23
32	23	22
24	29	21
31	32	18
31	23	16
24	31	14
23	29	13
30	30	13
30	24	12
29	30	11
29	24	10

i	j	$d^{Ag}_{2003,ij}$
23	23	457
24	23	433
23	24	406
24	24	404
32	32	26
32	23	24
24	31	24
23	29	23
31	23	19
29	30	19
23	31	18
31	32	17
29	24	16
24	29	16
30	30	10

i	j	$d^{Ag}_{2004,ij}$
23	23	464
24	23	406
23	24	387
24	24	374
23	29	27
31	23	24
32	32	23
32	23	22
29	30	20
29	24	19
31	32	19
30	30	17
24	31	17
23	31	13
24	29	10

i	j	$d^{Ag}_{2005,ij}$
23	23	432
23	24	382
24	24	355
24	23	344
30	30	24
23	31	22
32	23	21
32	32	21
31	23	20
24	31	19
23	29	18
24	29	18
29	24	17
30	24	17
29	30	17

i	j	$d^{Ag}_{2006,ij}$
23	24	301
23	23	300
24	24	265
24	23	234
30	30	19
31	23	18
29	30	17
24	29	16
31	32	16
29	24	13
23	29	13
23	31	13
24	31	11
30	24	10
32	32	10

4.4 Agincourt flux vector

We use the definition (4.10) to discuss the Agincourt flux vector for each year of study. We present the results in Table 4.5. Consider the dominant states 23 and 24. We note immediately the flow into 23, and out of 24 for all years $1998-2004$. There is a slow decay in the magnitude of these flows. In $2005, 2006$ the flows reverse. Since the fluxes are small we have not attached importance to this phenomenon. In Figure 4.3 we plot $|\delta_{t,23}|$, $|\delta_{t,24}|$ and clearly see that the dominant fluxes are *balanced* between 23 and 24. Thus while the magnitude of flux may change, it changes approximately equally, but in opposite direction. This balance is a constant feature of the dominant Agincourt dynamics.

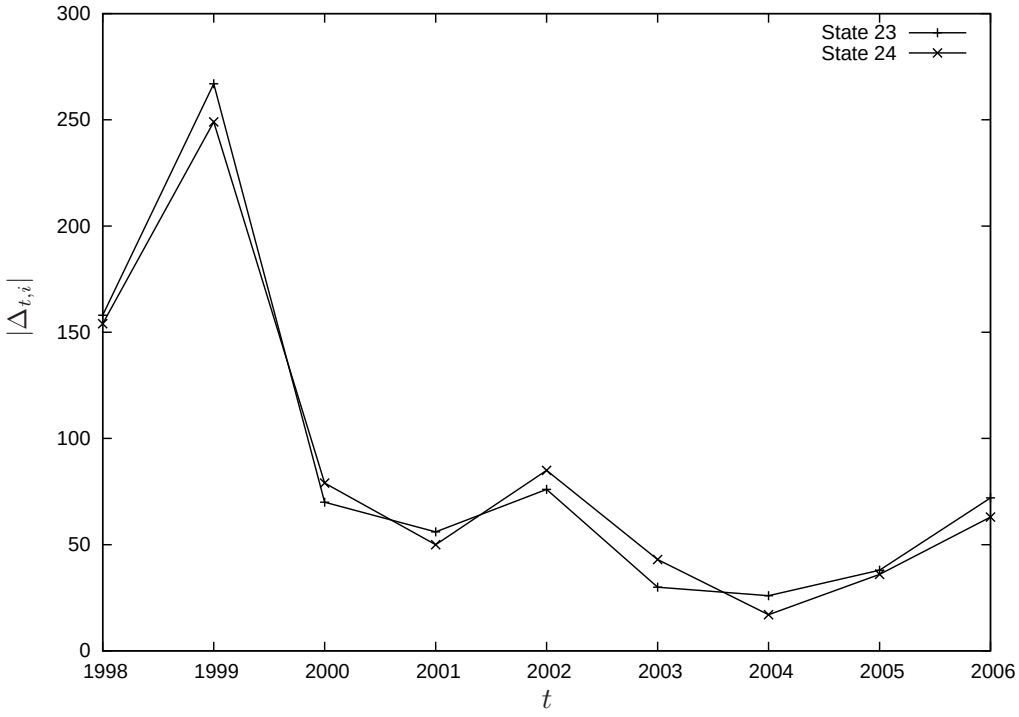

Figure 4.3: Time-dependent flux over the dominant states 23 and 24.

Table 4.5: Agincourt flux vector for the indicated year.

i	$\delta_{1998,i}$
1	0
2	0
3	0
4	0
5	0
6	0
7	1
8	0
9	0
10	0
11	0
12	0
13	0
14	1
15	0
16	-2
17	0
18	0
19	-1
20	0
21	-17
22	-24
23	158
24	-154
25	0
26	0
27	0
28	0
29	16
30	-5
31	21
32	12
33	0
34	0
35	0
36	0
37	0
38	0
39	0
40	0
41	0
42	0
43	0
44	-1
45	0
46	0
47	0
48	-5

i	$\delta_{1999,i}$
1	0
2	0
3	0
4	0
5	0
6	0
7	1
8	-1
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	0
17	0
18	0
19	0
20	-1
21	-4
22	-14
23	267
24	-249
25	0
26	0
27	0
28	0
29	3
30	-12
31	23
32	-10
33	0
34	0
35	0
36	0
37	0
38	1
39	0
40	0
41	0
42	0
43	0
44	0
45	0
46	0
47	1
48	-5

i	$\delta_{2000,i}$
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	1
9	0
10	0
11	0
12	0
13	0
14	-1
15	0
16	0
17	0
18	0
19	0
20	0
21	0
22	-5
23	70
24	-79
25	0
26	0
27	0
28	0
29	13
30	-3
31	7
32	0
33	2
34	0
35	0
36	0
37	0
38	-1
39	0
40	0
41	0
42	0
43	0
44	-3
45	0
46	0
47	-2
48	1

i	$\delta_{2001,i}$
1	0
2	0
3	0
4	0
5	0
6	0
7	-1
8	1
9	0
10	0
11	0
12	0
13	1
14	0
15	0
16	0
17	0
18	0
19	0
20	0
21	1
22	-3
23	56
24	-50
25	0
26	0
27	0
28	0
29	5
30	5
31	-3
32	-6
33	-2
34	0
35	0
36	0
37	0
38	0
39	0
40	0
41	0
42	0
43	0
44	0
45	0
46	0
47	1
48	-5

Agincourt flux vector for the indicated year (continued).

i	$\delta_{2002,i}$
1	0
2	0
3	0
4	0
5	0
6	0
7	1
8	-1
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	0
17	0
18	0
19	0
20	0
21	1
22	0
23	76
24	-85
25	0
26	0
27	0
28	0
29	13
30	-2
31	5
32	-6
33	0
34	0
35	1
36	0
37	0
38	0
39	0
40	0
41	0
42	0
43	0
44	-1
45	0
46	0
47	-1
48	-1

i	$\delta_{2003,i}$
1	0
2	0
3	0
4	0
5	0
6	0
7	2
8	-2
9	0
10	0
11	0
12	0
13	-1
14	0
15	0
16	0
17	0
18	0
19	0
20	0
21	-1
22	1
23	30
24	-43
25	0
26	0
27	0
28	0
29	7
30	8
31	7
32	-8
33	1
34	0
35	-1
36	1
37	0
38	0
39	0
40	0
41	0
42	0
43	0
44	0
45	0
46	0
47	0
48	-1

i	$\delta_{2004,i}$
1	0
2	0
3	0
4	0
5	0
6	0
7	2
8	-1
9	0
10	0
11	0
12	0
13	0
14	0
15	0
16	0
17	0
18	0
19	0
20	0
21	-2
22	-2
23	26
24	-17
25	0
26	0
27	0
28	0
29	-2
30	15
31	-12
32	-4
33	0
34	0
35	0
36	-1
37	0
38	0
39	0
40	0
41	0
42	0
43	0
44	-1
45	0
46	0
47	0
48	-1

i	$\delta_{2005,i}$
1	0
2	0
3	0
4	0
5	0
6	0
7	-3
8	3
9	0
10	0
11	0
12	0
13	1
14	0
15	0
16	0
17	0
18	0
19	0
20	0
21	2
22	4
23	-38
24	36
25	0
26	0
27	0
28	0
29	1
30	-3
31	12
32	-14
33	-1
34	0
35	0
36	0
37	0
38	0
39	0
40	0
41	0
42	0
43	0
44	0
45	0
46	0
47	0
48	0

i	$\delta_{2006,i}$
1	0
2	0
3	0
4	0
5	0
6	0
7	-1
8	1
9	0
10	0
11	0
12	0
13	1
14	0
15	0
16	0
17	0
18	0
19	0
20	0
21	2
22	-1
23	-72
24	63
25	0
26	0
27	0
28	0
29	-1
30	7
31	-10
32	11
33	0
34	0
35	0
36	0
37	0
38	0
39	0
40	0
41	0
42	0
43	0
44	0
45	0
46	0
47	0
48	0

4.5 Initial and final states of the Agincourt population in Γ_3

One way to measure the improvement in social changes of this population is to compare the initial and final states of that population. This comparison can help us to understand the direction of movement of the population which in turn can be associated to social phenomena.

In Figure 4.4 and Figure 4.5, we present the distribution of the initial states of the Agincourt population sample in Γ_3. In particular, Figure 4.4 displays the phase diagram of the initial states of 1998 while Figure 4.5 shows the distribution of the number of Agincourt households initially in each state. This initial state has been chosen by the "fhiqo" strategy.

From the above figures, we clearly see that there are two dominant initial states in this population. The first dominant state is 24, $(0.875, 0.555) \in \Gamma_3$ which defines Agincourt households headed by adult in which all children (age between 6 and 17) live with their biological mothers. In these Agincourt households no adult death is registered. This state defines about 67.55% of the Agincourt population. The second state is 23, $(0.75, 0.555) \in \Gamma_3$, with the same significance value is related to the Agincourt households with at least one child without a biological mother. This state defines about 23.34% of the population. About 9.11% of the population is distributed over the rest of states in Γ_3.

On the other hand, we identify the significance χ that dominates in the Agincourt population. Because we start each household at a significance level related to the frequency of questions that change answer values in that household (fhiqo), we automatically observe that the significance level of the Agincourt population is initially given by $\overline{\chi}_{Ag} = 0.555$ as displayed in Figure 4.5. This corresponds to question order $\theta = 120$.

Similarly, Figure 4.6 and Figure 4.7 show the final distribution of the Agincourt population in 2007. We find that the Agincourt population ends its observation time in the same two dominant states of Figure 4.5. This is not surprising because we have chosen initial conditions by the "fhiqo" strategy. Then what is of interest in these diagrams is the way the number of households redistribute over the dominant states. Now, we have 45.89% of the population at the above first dominant state and 44.81% at the second dominant state.

Now if we compare the initial and final distributions for this population, we can measure social changes that happen in the Agincourt population for the given observation time. First, we observe that most of changes happen between the two dominant states as mentioned above. There is

a decrease of about 21.66%, in the number of households that were initially at state $(e, \chi) = (0.875, 0.555)$, and an increase of about 21.47% in the number of households that were initially at state $(e, \chi) = (0.75, 0.555)$. This shows a balance in the number of households moving between the two states in both directions. This is consistent with the fluxes of Table 4.5 and with the decrease in the magnitude of fluxes from 2000 towards we expect that this balance of numbers is now stable. It is clear that the important change is the transition $(111, 120) \rightarrow (110, 120)$. Thus, from this transition, we can easily identify the social change as related to the absence of the biological mother. The leading demographic phenomenon for this sub-population, involves about 26.57%, or more than a quarter, of the Agincourt population.

These two distributions only give us an idea about how the population moves from its initial position to its final position. However, they do not inform us about what happens between the initial and the final observation times. In order to include the analysis of the dynamics between the extreme observation times, we need a more detailed visualisation of orbits. Thus, in the next sections, we will present the time series and the phase diagram of the states of the population for the whole observation time.

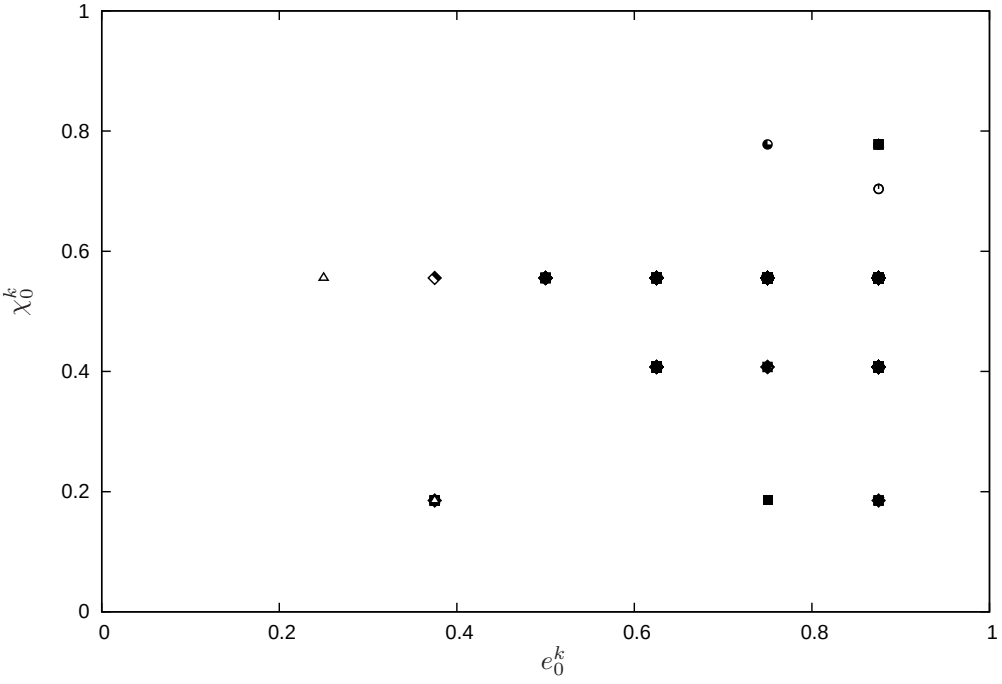

Figure 4.4: Initial states of the Agincourt households in Γ_3.

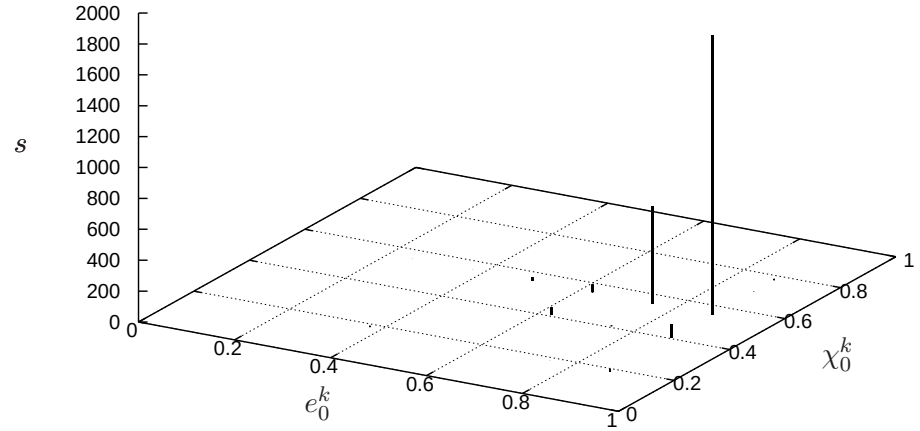

Figure 4.5: Initial distribution of the number of Agincourt households s, over states of Figure 4.4.

4.6 Visualization of all orbits of Agincourt data.

Figure 4.8 displays the time series of the fitness component of the states of the Agincourt population while in Figure 4.9 we show the time series of the significance component of the states. All 2669 orbits of our studied population are plotted.

Figure 4.8 shows that most of the orbits are concentrated in the fitness region defined by $0.75 \leq e \leq 0.875$. Although, this region is fitter, it is difficult to identify the questions that change answer values in that region. Thus, we need to look at the significance values to indicate questions that change answer values. Figure 4.9 shows a concentration of orbits in the significance region defined by $0.407 \leq \chi \leq 0.555$. Many orbits obviously overlap and these figures do not give a sensitive measure of number of households but this conclusion is supported by our knowledge of the dominant transitions involving states 23 and 24. In this figure, we have the knowledge of the questions that change answer values. This region of Γ_3 is defined by $\theta \in \{102, 120\}$. We identify that dynamics with adult household heads. Here, Figure 4.9 displays all 6 possible significance values that we can have with $n = 3$ questions.

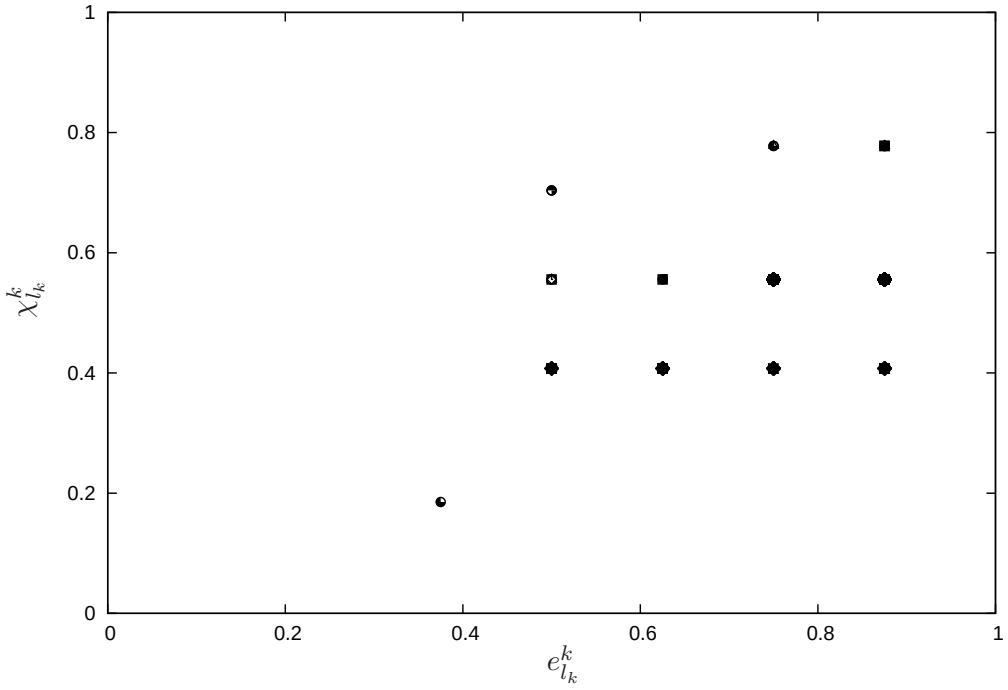

Figure 4.6: Final states of the Agincourt households in Γ_3.

We notice a few unfavourable fitness jumps to $e = 0$, in 2001 and 2005 which reveal the few cases of Agincourt households headed by minors, with an absence of biological mothers and where an adult death is registered at these observation years. Note also that among the 8 possible fitness states that we can have with $n = 3$ questions, Figure 4.8 displays only 7 of them. With the help of diagram of the space S_3 (see Figure 3.20), we identify that the missing fitness state is given by $e = 0.125$ and related to the binary sequence $b = 001$. In order to explore some valuable information regarding this missing state, we must first attach a significance level to this state which in turn will help to associate questions to their answer values. There are 6 possible significance values that we can attach to this binary sequence, to define a state in S_3. It is convenient to start by examining the Agincourt population significance value given by $\theta = 120$. Thus, if we attach this significance level to that binary sequence, we will then define the state $({}_wb, \theta) = (001, 120)$. This state defines households headed by minors, with an adult death and the presence of the biological mother. This is a situation that is not possible to find in the Agincourt population. In fact, even if there is an adult death that occurs in the household, we assume that if the biological mother is in the household and that there is no other adult that can head the household, that biological mother automatically becomes the head of that household.

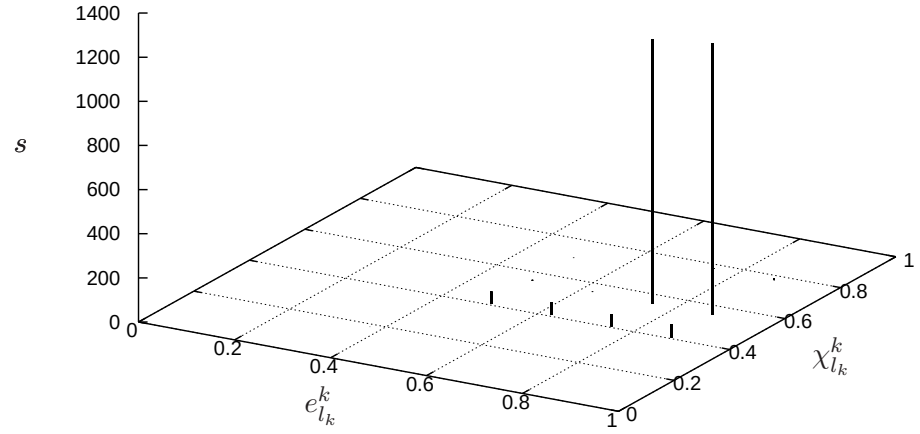

Figure 4.7: Final distribution of the number of Agincourt households s, over states of Figure 4.6.

We have identified dominant dynamics as being in the sub-space of Figure 4.1. We have defined in 4.1 approximate Agincourt transition matrix $T^{Approx.}$, defined in turn by Table 4.3. It is this dynamics which we define to be of *demographic population-level, importance*.

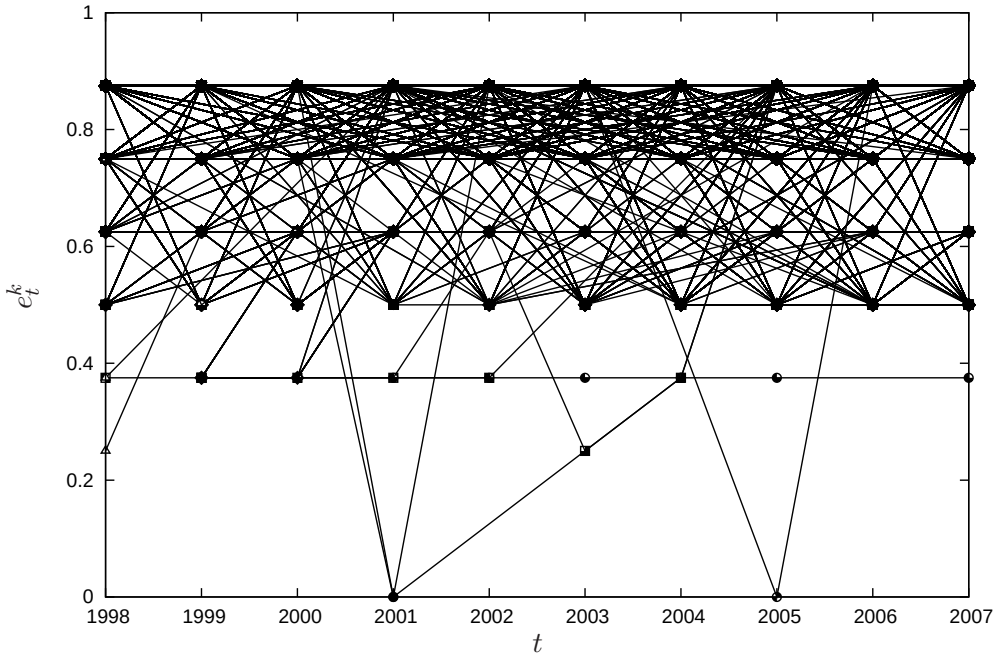

Figure 4.8: Time series of the fitness component, e_t, of all the Agincourt households, with children of school-going age.

4.6.1 Orbits of the Agincourt population in Γ_3

In the previous section we clearly showed that the time series for both the fitness and significance components give useful information to identify for example an observation time when a specific social change occurs. However, it is difficult to interpret the dynamics of fitness and significance orbits separately. In particular, we have seen that with any pair of consecutive fitness values, we are unable to identify what questions change answer values between the two times. Thus, the one-dimensional space of visualisation presents limitations. To overcome these, we now visualize orbits in Γ_3.

Figure 4.10 shows the orbits of the whole population of 2669 Agincourt households in Γ_3, for all observation years from 1998 to 2007. Now that we have both the fitness and the significance values together, their combination helps to link change in orbits to change in answer values with now the knowledge of the changing questions. We identify in Figure 4.10 that $\{(e, \chi) \mid 0.555 \leq e \leq 0.875, \quad 0.407 \leq \chi \leq 0.555\} \subset \Gamma_3$ is the sub-space of Γ_3 where most of the Agincourt population dynamics happen. Figure 4.10 is not sensitive to numbers of households in each transition and

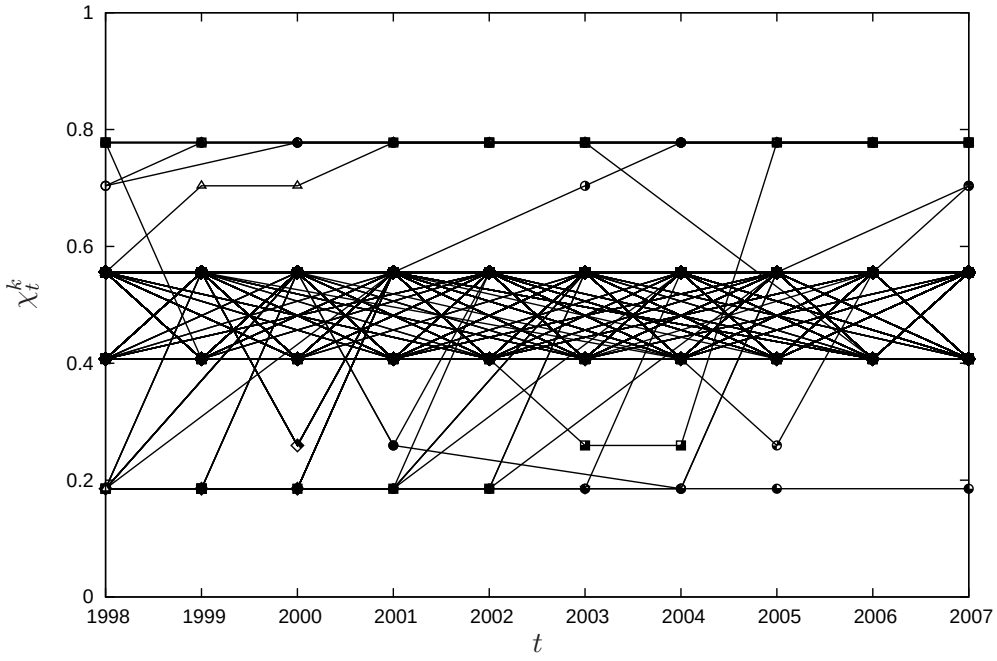

Figure 4.9: Time series of the significance component, χ_t, of the Agincourt households, with children of school-going age.

again we use earlier experience. Note that this sub-space is a Cartesian product of the fitness and significance spaces that we previously identified in Figure 4.8 and Figure 4.9 respectively. In this region changes are related to questions q_0 and q_2.

In Figure 4.11 we present the same orbits of Figure 4.10 now viewed in time (vertical axis). We see that the same above sub-space of Γ_3 is identified, where most Agincourt dynamics occur. But now, we have a clear view of these changes over time. In particular, the two points $(0, 0.259, 2001)$ and $(0, 0.259, 2005)$ of Figure 4.11 show the two years where the associated unfavourable jump occurs. Note emergence of the usual two dominant states. Recall that the first state 23, $(e, \chi) = (0.75, 0.555)$ reveals Agincourt households with absence of the biological mother and the second state $(e, \chi) = (0.875, 0.555)$ characterises Agincourt households with fitter states.

The dynamics outside that sub-space reveal some rare changes that we mentioned above. For example unfavourable jumps to $(e, \chi) = (0, 0.259)$ related to Agincourt households headed by minors, with absence of biological mother and an adult death.

In Figure 4.12 we display the overall distribution of the number of visits that households made

over the states of Γ_3. It is clear that the above-mentioned dominant states 23 and 24 are identified. As previously noticed from Table 4.1, we see that over our observation period $1998 - 2007$, about 50.11% of the visits involves state 24 and about 40% involves state 23. Less than 10% of the visits were related to other states of Γ_3.

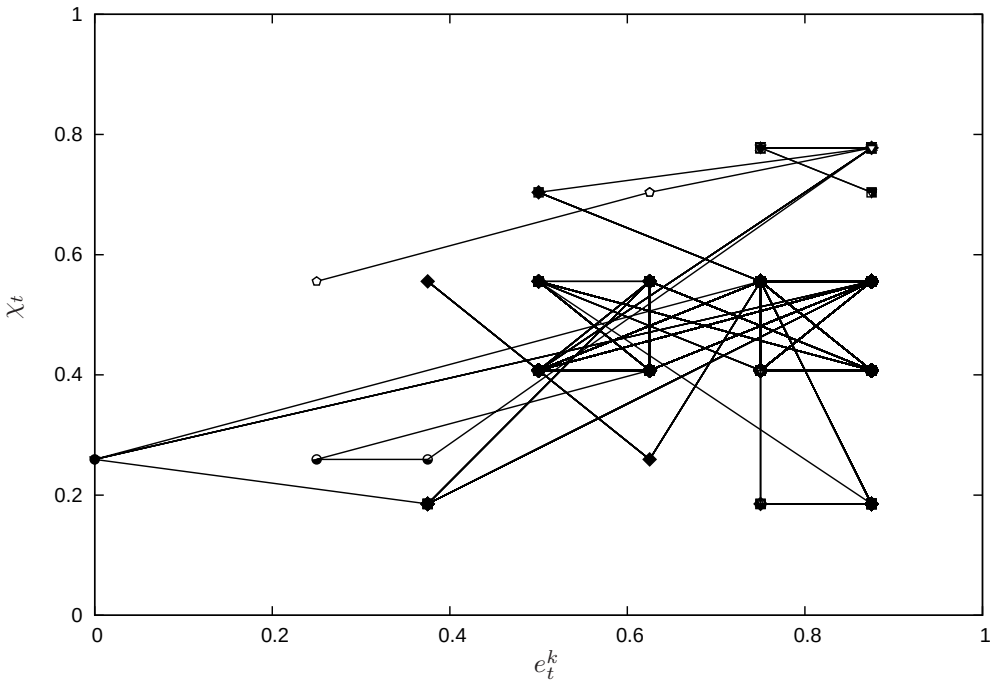

Figure 4.10: Phase diagram of the Agincourt households.

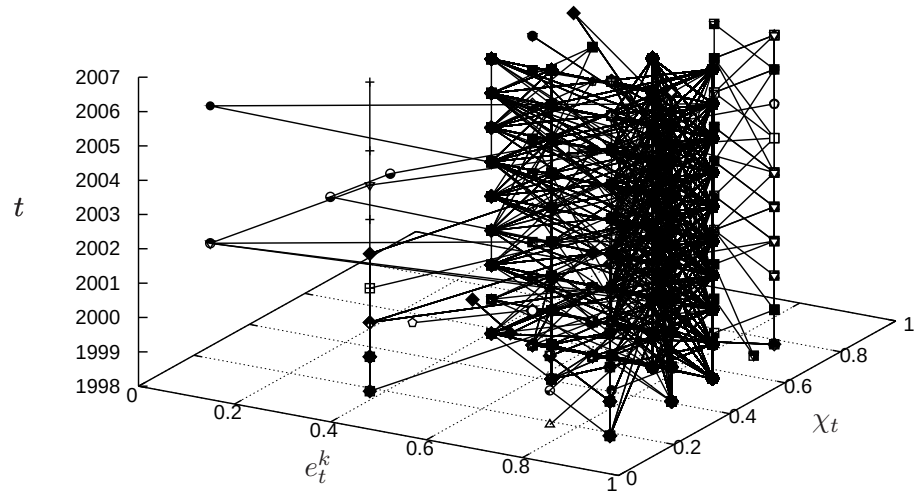

Figure 4.11: Orbits of the Agincourt households with time dependent (vertical axis).

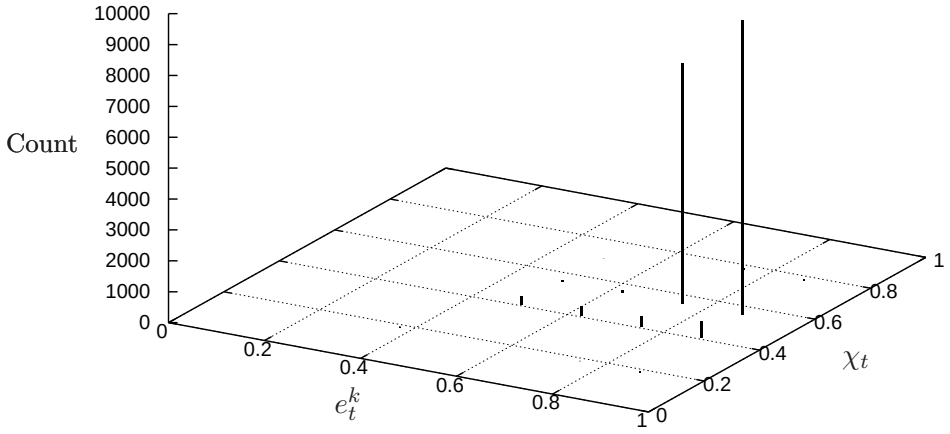

Figure 4.12: Distribution of the number of the Agincourt household orbits over Γ_3.

4.7 Effect of household change on children's progression in school

In Chapter Two, we showed that the Agincourt data collected for this study has some limitations. The information on the education in Agincourt is only available for the following years: 1992, 1997, 2002 and 2006. Note that this study does not deal with missing values in the data. Because of these Agincourt data issues, we cannot add a fourth question to the three questions that define household characteristics, for an Agincourt household, in order to include the education status of the household. The education status of the household is simply measured by asking whether or not there is a child in the household who repeats a year of study too often.

4.7.1 Colour coding and definition of educational default

We associate the colour of an orbit with defaulting or non-defaulting education in households. Red associates to orbits of households with a defaulting child and green to orbits of households without a defaulting child anywhere during its whole observation time.

As stated in Chapter Two, in order to measure the progress in education of children in the present study population, we must first define parameters that will be used to measure that progress. In the present data, we have for each observation year that the data is available, the following information for each child: ID, age and the total number of completed years of education. Let $l_{failure}$ define the maximum number of education years that a child fails during his/her school life. Thus, we associate the education progress to these 3 parameters.

Now, if we define a defaulting child as a child who fails $l_{failure}$ times during his/her school life, then we must be careful in choosing the value of $l_{failure}$ in relation with the study observation time. For the present analysis, the observation time has been sampled to the period $1998 - 2007$ and the average observation time is $\overline{l} = 7.115$.

Recall that for each household, for each child in that household, the data we collect gives the information about the age of the child and the number of complete years of education of that child. Let a denote the age of each child. Then we can associate the number of education years function to the age of the child by the function

$$y : a \mapsto y(a) \ . \tag{4.23}$$

Note that we cannot measure whether or not a child of age $a < (7 + l_{failure})$ has failed any

particular grade $l_{failure}$ times. We simply assume that a child at age $a < (7 + l_{failure})$ whatever the associated education value $y(a)$ could be, is not a defaulting child.

Now for other ages $a \geq (7 + l_{failure})$, we define a defaulting child as follows:

$$\begin{aligned} &\text{If} \quad y(7 + l_{failure} + k) > (l_{failure} - 1) + k, \quad \text{return: child is non-defaulting} \\ &\text{If} \quad y(7 + l_{failure} + k) \leq (l_{failure} - 1) + k, \quad \text{return: child is defaulting} \end{aligned} \tag{4.24}$$

where k varies with respect to the age range. For example for the Agincourt data, the age range is from 7 to 16 years. Thus, because we assume that $l_{failure} \geq 2$, we will have k such that $l_{failure} - 2 \leq k \leq 7$.

Directly from (4.24), we define a defaulting household as a household with at least one defaulting child who is found at least once during its observation years. Note that the definition of a defaulting child applies each observation year in which education data is available. However the definition of a defaulting household is applicable for its whole observation time. We can now use the definition (4.24) to colour defaulting households.

To illustrate, suppose $l_{failure} = 2$ years. If we accept the above assumptions, all Agincourt household orbits with only children of age $a < 9$ will be non-defaulting households. Otherwise, the following definition applies.

$$\begin{aligned} &\text{If} \quad y(9 + k) > 1 + k, \quad \text{return: child is non-defaulting} \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k = 0, 1, 2, \ldots 7 \\ &\text{If} \quad y(9 + k) \leq 1 + k, \quad \text{return: child is defaulting} \end{aligned} \tag{4.25}$$

4.7.2 Analysis of Agincourt dynamics including education status

We start by presenting the distribution of the educational status of the Agincourt households with respect to the value of $l_{failure}$. Thus Figure 4.13 shows how the number of Agincourt defaulting and non-defaulting households changes as we vary $l_{failure}$.

We find that about 73.25% of the Agincourt households have at least one child that repeats twice in the period between 1998 to 2007, and about 26.75% of the Agincourt households did not have a child that repeats twice in the same observation period. These figures show that on average the Agincourt population is significantly defaulting with respect to $l_{failure} = 2$. We find that most of the Agincourt population is defaulting at this level. Because the majority of the population

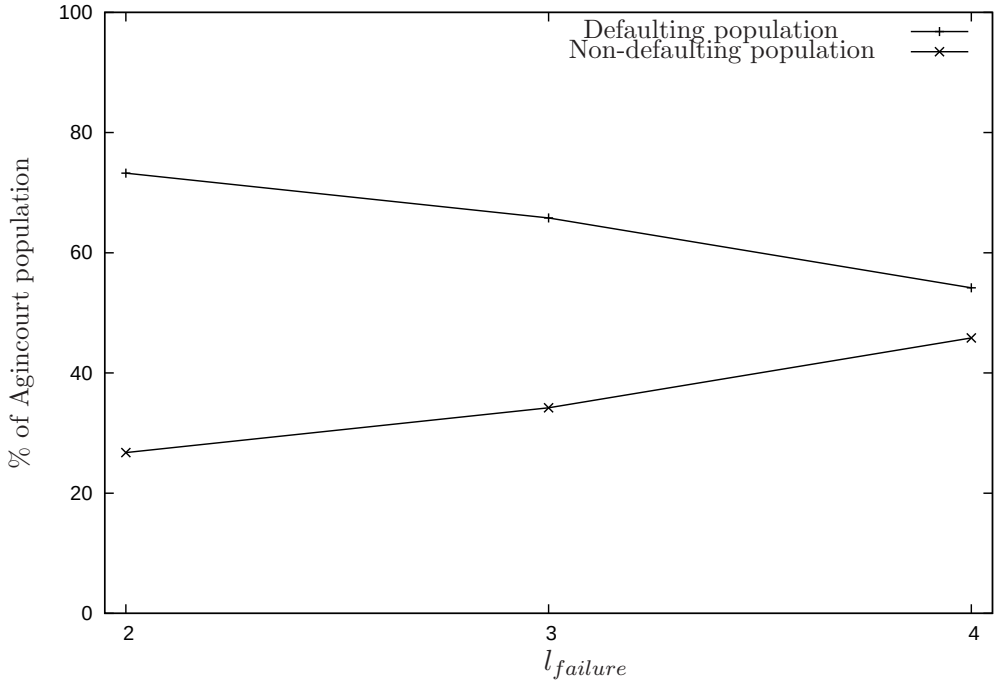

Figure 4.13: Distribution of the Agincourt households educational status with respect to $l_{failure}$.

defaults, we must suspect that this is owing to failure of the school system and it thus becomes important to increase $l_{failure}$.

With $l_{failure} = 3$, we find 65.79% of Agincourt defaulting population against 34.21% non-defaulting population and we must again suspect the school system. If $l_{failure} = 4$ we have a balance in the two populations. We now have 54.17% of defaulting population and 45.82% of non-defaulting population.

We have clearly understood by the properties of the transition and density matrices, the relationship between the Agincourt household states. The results related to the analysis presented in the previous sections will be extracted here to better achieve the analysis of the effect of change in household states on the education status. We give density matrices in (4.33) and (4.34).

We present in Table 4.6 the relationship of the Agincourt educational distribution including the dominant transitions with respect to the value of $l_{failure}$. The stared entries denote the dominant transitions. We note dominant idling in state 23 and transition $24 \rightarrow 23$ in the case of defaulting households. This contrasts with dominant idling in state 24 and transition $23 \rightarrow 24$ in progressing

households.

Finally we summarise in Table 4.7 both the distribution of Figure 4.13 and the properties of Table 4.6.

Table 4.6: Distribution of Agincourt educational transitions with respect to $l_{failure}$. Note that $\overline{Ed}$ refers to defaulting households.

Education measure	**Dominant transitions**	d_{ij}^{Ag}		$d_{ij}^{\overline{Ed}}$		d_{ij}^{Ed}	
	$i \rightarrow j$	#	%	#	%	#	%
	24 → 24	4336	26.56	3119	71.93	1217	**28.06***
$\mathbf{l_{failure} = 2}$	24 → 23	3650	22.36	2793	**76.52**	857	23.48
	23 → 23	3164	19.38	2545	**80.43***	619	19.57
	23 → 24	3163	19.37	2418	76.44	745	**23.56**
	24 → 24	4336	26.56	2773	63.95	1563	**36.05***
$\mathbf{l_{failure} = 3}$	24 → 23	3650	22.36	2500	68.49	1150	**31.51**
	23 → 23	3164	19.38	23.6	**72.88***	858	27.12
	23 → 24	3163	19.37	2174	**68.73**	989	31.27
	24 → 24	4336	26.56	2256	52.03	2080	**47.97***
$\mathbf{l_{failure} = 4}$	24 → 23	3650	22.36	2082	**57.04**	1568	42.96
	23 → 23	3164	19.38	1904	**60.17***	1260	39.83
	23 → 24	3163	19.37	1796	56.78	1367	**43.22**

Finally, we decide to use $l_{failure} = 4$ because there are two significant populations (50% of total) with education progress and default that can be compared and household phenomena might be expected to emerge from the general educational failure.

We follow the analysis used in the previous section to organise the discussion. Thus, we start by presenting the transition matrices T_3^{Ed}, $T_3^{\overline{Ed}}$ and density tables.

Table 4.7: Summary of Agincourt educational dominant transitions with respect to $l_{failure}$

Education measure	**Population (%)**		**Dominant transition**	
$l_{failure}$	$\overline{Ed}$	Ed	$\overline{Ed}$	Ed
2	73.25	26.75	$23 \rightarrow 23$ $24 \rightarrow 23$	$24 \rightarrow 24$ $23 \rightarrow 24$
3	65.79	34.21	$23 \rightarrow 23$ $24 \rightarrow 23$	$24 \rightarrow 24$ $24 \rightarrow 23$
4	54.18	45.82	$23 \rightarrow 23$ $24 \rightarrow 23$	$24 \rightarrow 24$ $23 \rightarrow 24$

4.8 Properties of Agincourt transition matrix with education status, $l_{failure} = 4$

We decompose the Agincourt transition matrix T_3^{Ag} defined in (4.14) in two different transition matrices. The first is the transition matrix for defaulting households $T_3^{\overline{Ed}}$ which captures all possible transitions that Agincourt defaulting households make. The second is the transition matrix of non-defaulting households T_3^{Ed} which similarly display all possible transitions occuring in the non-defaulting Agincourt population. We present the transition matrix for non-defaulting households in (4.30) and the transition matrix for defaulting households is given in (4.31).

To better understand the properties of the difference between the two transition matrices, we also analyse the difference $T_{dif}^{Ed,\overline{Ed}} = T_3^{Ed} - T_3^{\overline{Ed}}$ which gives the difference in the transitions happening in both sub-populations. $T_{dif}^{Ed,\overline{Ed}}$ is presented in (4.32).

In the previous sections, we have defined some properties of the transition matrices including the trace, determinant that give valuable information regarding the dynamics of the population. Here, we use the same definition to determine these properties for each of the two Agincourt sub-populations.

$$\sum_{i,j=1}^{48}(t_{ij}^{Ag}) = 58, \qquad \sum_{i,j=1}^{48}(t_{ij}^{Ed}) = 44, \qquad \sum_{i,j=1}^{48}(t_{ij}^{\overline{Ed}}) = 55\ . \tag{4.26}$$

$$\text{Det}\ (T_3^{Ag}) = \ \text{Det}\ (T_3^{Ed}) = \ \text{Det}\ (T_3^{\overline{Ed}}) = 0\ . \tag{4.27}$$

$$\text{Trace}\ (T_3^{Ag}) = 11, \qquad \text{Trace}\ (T_3^{Ed}) = 10, \qquad \text{Trace}\ (T_3^{\overline{Ed}}) = 11\ . \tag{4.28}$$

The result (4.26) gives the total number of transitions that occur in both Agincourt defaulting and non-defaulting sub-populations. In particular from 58 transitions that we observe in the Agincourt population, there are 44 (about 75.86%) transitions that occur in non-defaulting Agincourt households and 55 (about 94.82%) that occur in the Agincourt defaulting population. Note that

$$\sum_{i,j=1}^{48}(t_{ij}^{Ag}) < \sum_{i,j=1}^{48}(t_{ij}^{Ed}) + \sum_{i,j=1}^{48}(t_{ij}^{\overline{Ed}}) \tag{4.29}$$

which means that there are common transitions occuring in the two sub-populations.

The result (4.27) indicates that for both sub-populations, the dynamics is not reversible. This property suggests a possibility of identifying causal effects.

The number of idling states for both sub-populations is given in (4.28). We find that almost all idling states in the Agincourt population are also idling in both sub-populations.

Now looking at the difference (4.32) between the two sub-populations, we can clearly see that there are transitions that occur in both sub-populations. There are 3 negative values in $T_{dif}^{Ed,\overline{Ed}}$ (4.32). This identifies the 3 transitions that occur in non-defaulting Agincourt households which do not happen in defaulting households. They are the following transitions $20 \rightarrow 38$, $38 \rightarrow 23$ and $48 \rightarrow 21$. In particular the transition $20 \rightarrow 38$ is related to a positive change in household head from a minor to an adult even if we have at the same time an adult death. The other two transitions define two changes happening at the same time which are absence of biological mother followed by an adult death.

On the other hand, there are 14 (about 29.16%) transitions that occur in the defaulting Agincourt households which are not observed in the non-defaulting households. They are the following $13 \rightarrow 8 \rightarrow 14 \rightarrow 14$, $19 \rightarrow 14$, $23 \rightarrow 33^{**}$, $24 \rightarrow 33^{**}$, $30 \rightarrow 35^{**}$, $33 \rightarrow 24, 33 \rightarrow 44 \rightarrow 22$, $35 \rightarrow 36 \rightarrow 8$ and $47 \rightarrow 48 \rightarrow 47$. Let us examine for instance the transitions that we denote by **.

The transition $23 \rightarrow 33$ is related to an adult death which is followed by change in household head. Although there are only few Agincourt households that are headed by minors, this result shows their education status, which is found to be defaulting. The transition $24 \rightarrow 33$ is one of the most unfavourable changes that happen in the Agincourt population. In this case the household is fully unfit. The transition $30 \rightarrow 35$ also indicates negative change to household head.

$$\mathbf{T_3^{Ed}} = \begin{bmatrix}
0&0&0&

$$\mathbf{T}_{\mathbf{3}}^{\overline{\mathbf{Ed}}} = \begin{array}{c}
0\,0\,0\,

$$
\mathbf{T}_{\mathbf{dif}}^{\mathbf{Ed},\overline{\mathbf{Ed}}} =
\begin{array}{c}
0\,0\,0\,

$$
\mathbf{D}_{\mathbf{3}}^{\mathbf{Ed}} = \begin{matrix}
0\,0\,0\,

$$\mathbf{D}_{\mathbf{3}}^{\overline{\mathbf{Ed}}} = \begin{matrix}
0\,0\,0\,

4.8.1 Time series of fitness states, including educational status.

In Figure 4.14 we show the overlapping time series of the fitness component of the Agincourt population states. As before, we see that most orbits for both red and green colours are concentrated in the sub-space of I_2^3 defined by $0.75 \leq e \leq 0.875$. Recall that the transition in this sub-space is related to the movement in and out of biological mothers.

Note also that none of the non-defaulting Agincourt household orbits crosses below the fitness line $e = 0.375$ or $b = 011$. The defaulting Agincourt household orbits reach the most unfavourable fitness level, $e = 0$ or $b = 000$ and wander almost everywhere on I_2^3.

To help better visualisation of the orbits of each sub-population, we separate the two sub-population and show the fitness time series of each in Figure 4.15 and Figure 4.16. Again for each sub-population, most orbits are in the same sub-space of I_2^3 as above mentioned.

As stated before, because with the fitness time series dynamics, we are not able to identify the questions that change answer values, it is important to present the phase diagram which combines fitness and significance states.

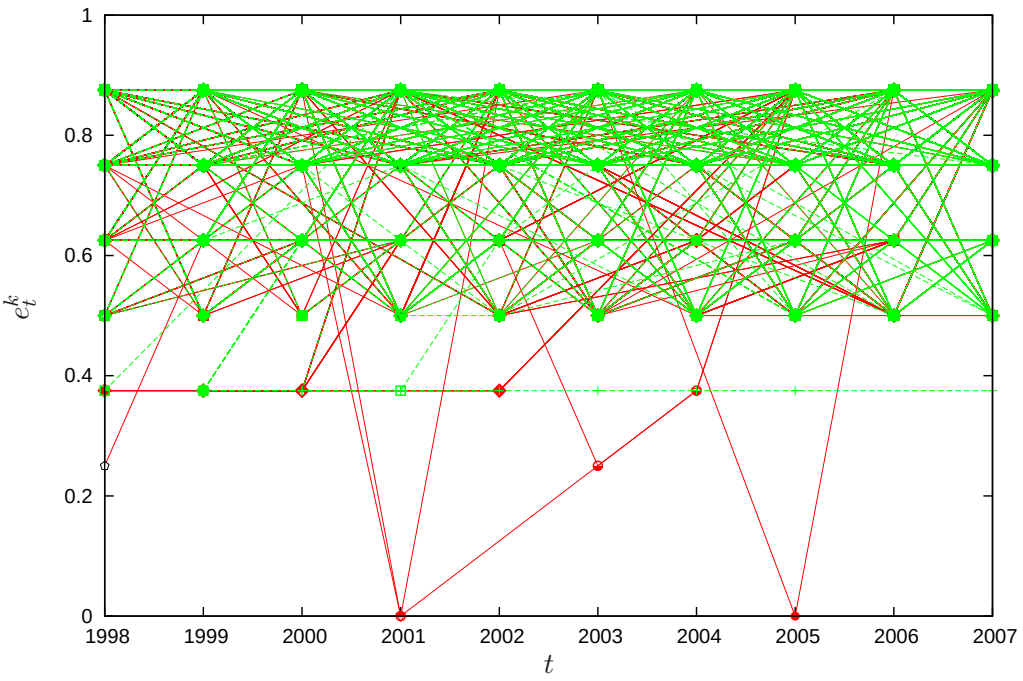

Figure 4.14: Time series of fitness component, e_t, with an educational colour coding, $l_{failure} = 4$.

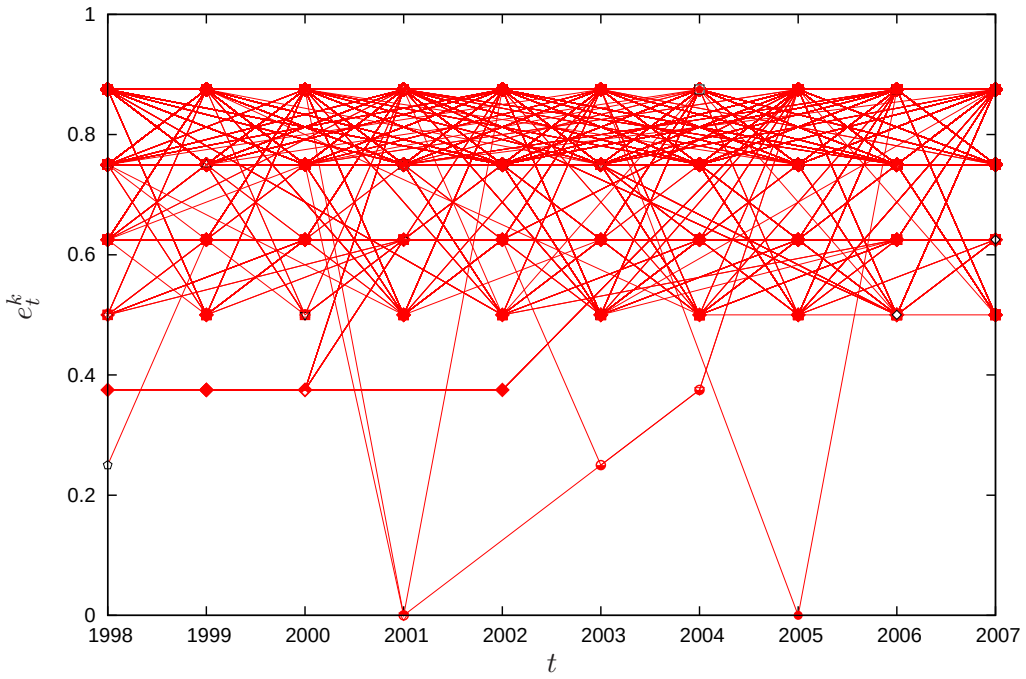

Figure 4.15: Time series of fitness component, e_t, for the defaulting Agincourt households, $l_{failure} = 4$.

4.8.2 Agincourt orbits in phase diagram, including educational status.

In Figure 4.17, we show the educational colouring orbits of the Agincourt population in Γ_3. We separate the two sub-populations to better visualise orbits of each group. Figure 4.23 gives the phase diagram of defaulting Agincourt households and Figure 4.24 for the non-defaulting Agincourt households.

Figure 4.24 shows that the non-defaulting Agincourt population is automatically divided in three sub-populations. The first, which is the dominant sub-population, is located in the sub-space of Γ_3 defined by $\{(e, \chi)|0.5 \leq e \leq 0.875,\ 0.407 \leq \chi \leq 0.55\}$.

The second sub-space is given by $\{(e, 0.185), e \in \{0.375, 0.75, 0.875\}\}$. This sub-population defined by a fixed significance value $\chi = 0.185$. The three states of households in this sub-space are the following: absence of biological mothers, an adult death and the last one corresponds to a fully fit household.

The last sub-population is defined by $\{(e, \chi)|\ e \in \{0.75, 0.875\}, \chi \in \{0.703, 0.777\}\}$. The three states in this sub-population have two social meanings. The first state corresponds to households

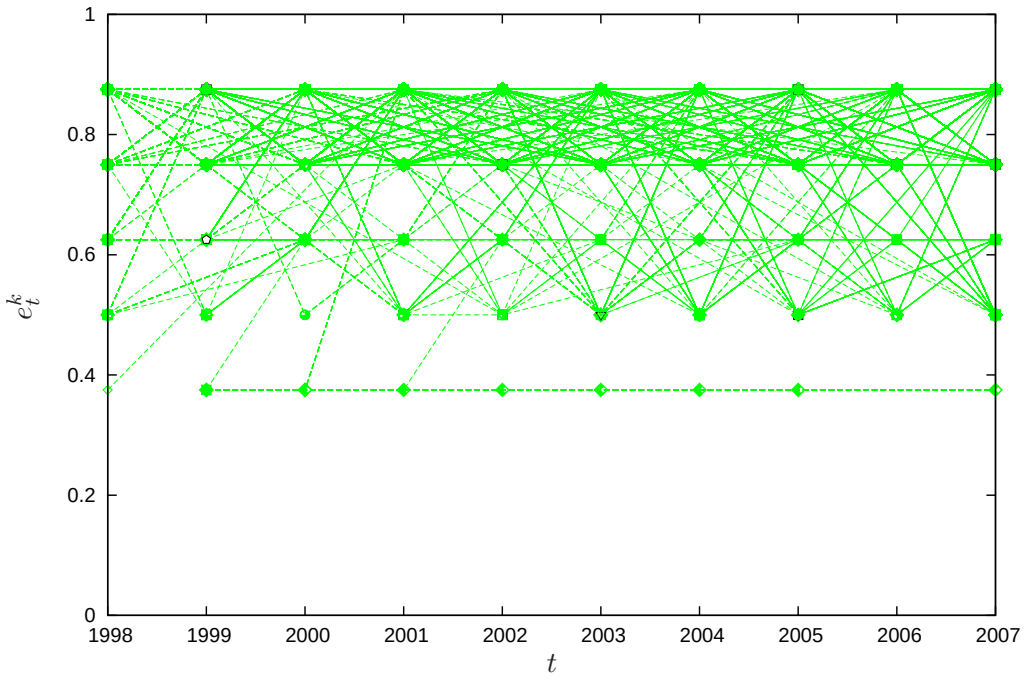

Figure 4.16: Time series of fitness component, e_t, for the non-defaulting Agincourt households, $l_{failure} = 4$.

with an absence of biological mothers and the other one corresponds to households with better conditions. Finally the maximum jump that occurs in non-defaulting Agincourt population is a positive social change $e = 0.375 \rightarrow e = 0.875$ which corresponds to a return of biological mothers (favourable).

The dynamics of defaulting Agincourt households as shown in Figure 4.23 can also be divided in different dynamic of sub-populations. Here the dynamics is not well separated. Note that the leading sub-population in this case is the same as the above first sub-population. We also observe many negative jumps to the region $e \leq 0.407$.

On the other hand, Figure 4.22 presents the 3−dimensional view of defaulting Agincourt orbits. Similarly in Figure 4.21, we show the 3−dimensional view of non-defaulting Agincourt orbits. These figures give clear view of the dynamics over time. Again for both sub-populations, most of the orbits are in the same dominant sub-space that we defined above. We observe the jumps outside that sub-space for each sub-population. In particular the stable sub-population of non-defaulting Agincourt households is now clearly shown in Figure 4.19.

The distribution of the number of Agincourt households over the states of Γ_3 is given in Figure 4.18 for the defaulting households and in Figure 4.19 for non-defaulting households. In particular, note that in both sub-populations, most of the orbits are defined by two states of Γ_3. The first state is $(0.875, 0.555)$ which defines Agincourt households in better conditions. The second is $(0.75, 0.555)$ which corresponds to Agincourt household with an absence of biological mothers. However, as before, there are more defaulting households in state $(0.75, 0.555)$ than the non-defaulting households.

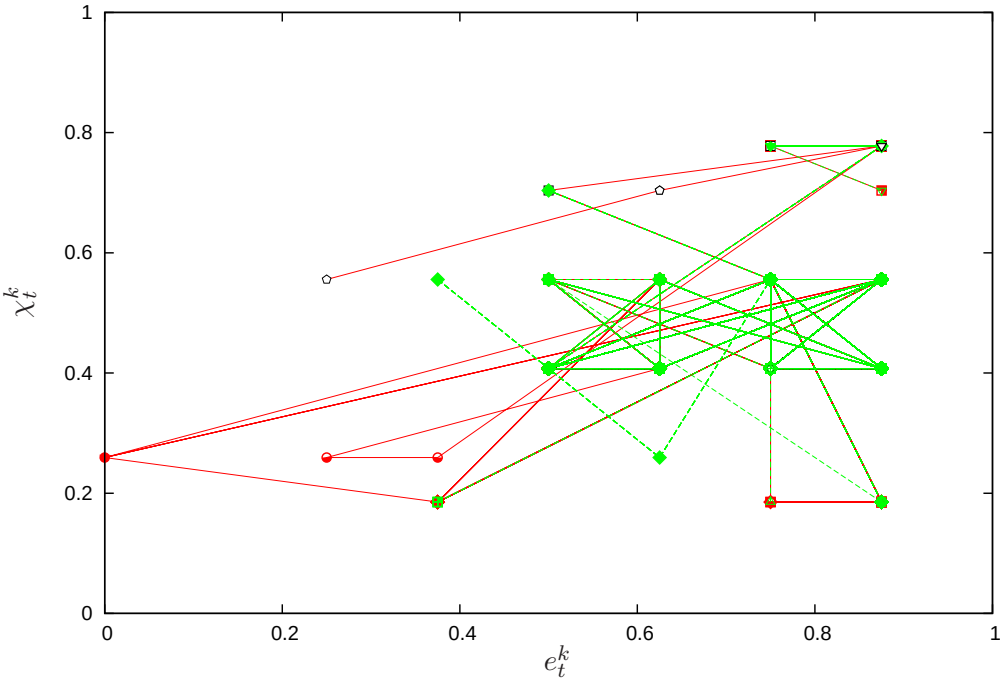

Figure 4.17: Phase diagram of the Agincourt households, with an educational colour coding, $l_{failure} = 4$.

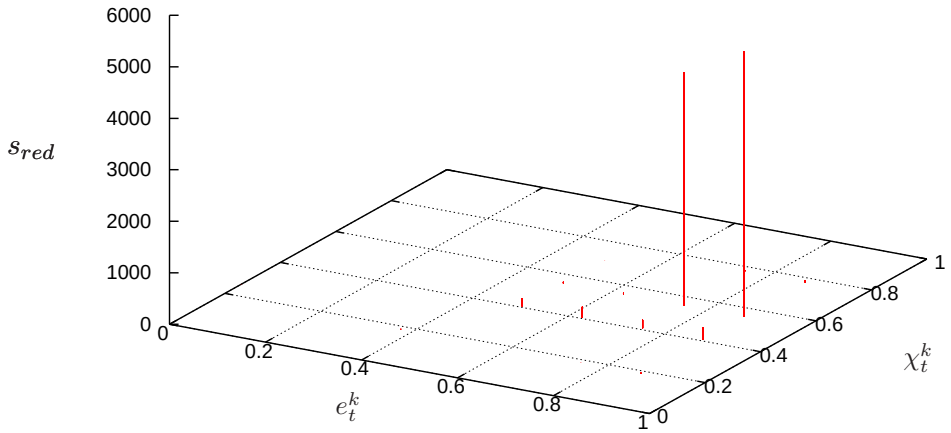

Figure 4.18: Distribution of the number of the Agincourt defaulting households, s_{red} over Γ_3, for $l_{failure} = 4$.

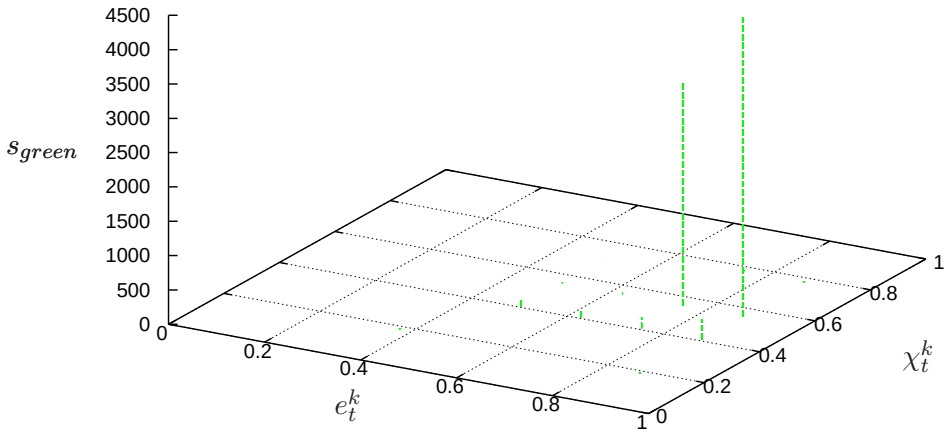

Figure 4.19: Distribution of the number of the Agincourt non-defaulting households, s_{green} over Γ_3, for $l_{failure} = 4$.

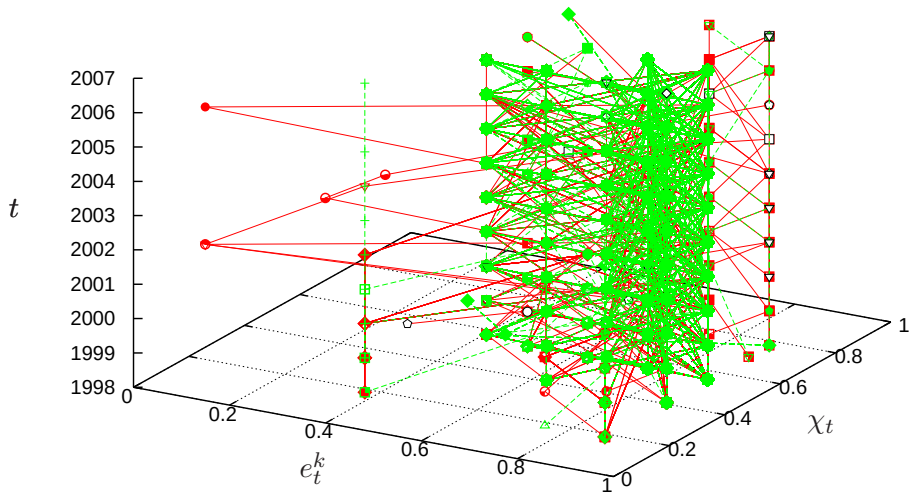

Figure 4.20: Orbits of the Agincourt households, with time dependent (vertical axis), with an educational colour coding, for $l_{failure} = 4$.

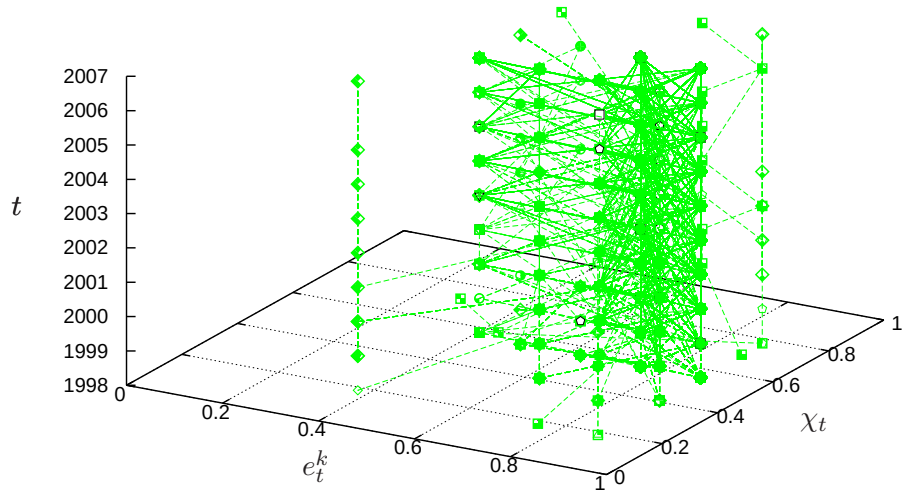

Figure 4.21: Orbits of non-defaulting Agincourt households, with time dependent (vertical axis), for $l_{failure} = 4$.

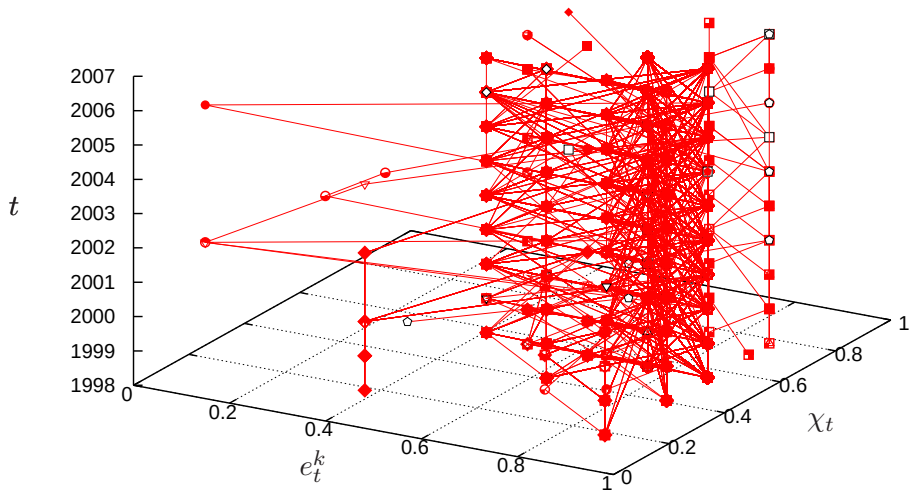

Figure 4.22: Orbits of defaulting Agincourt households, with time dependent (vertical axis), for $l_{failure} = 4$.

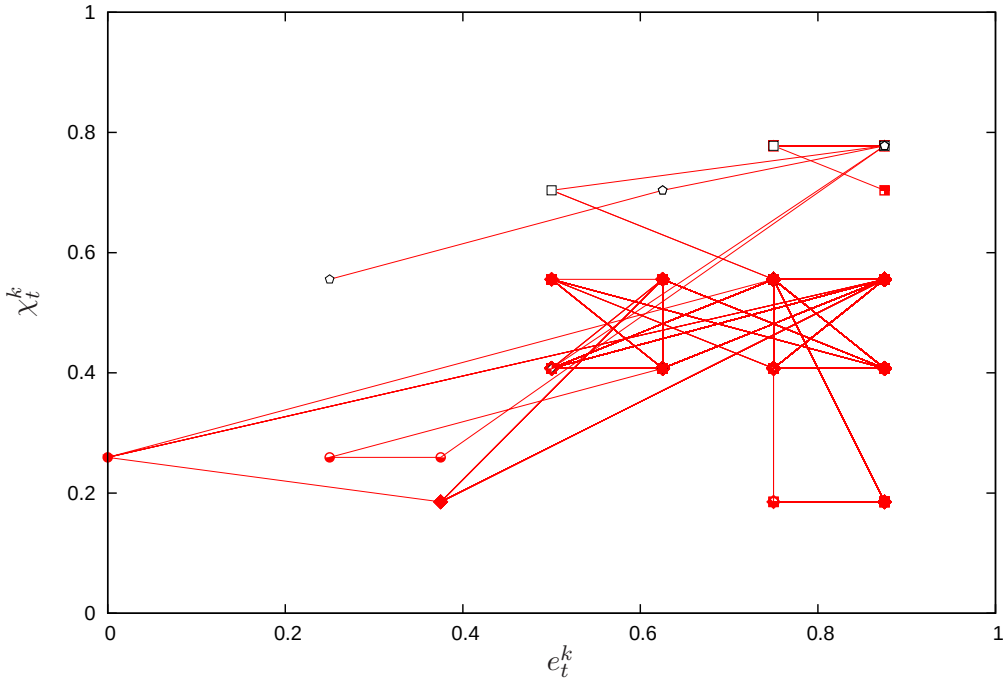

Figure 4.23: Phase diagram of defaulting Agincourt households, for $l_{failure} = 4$.

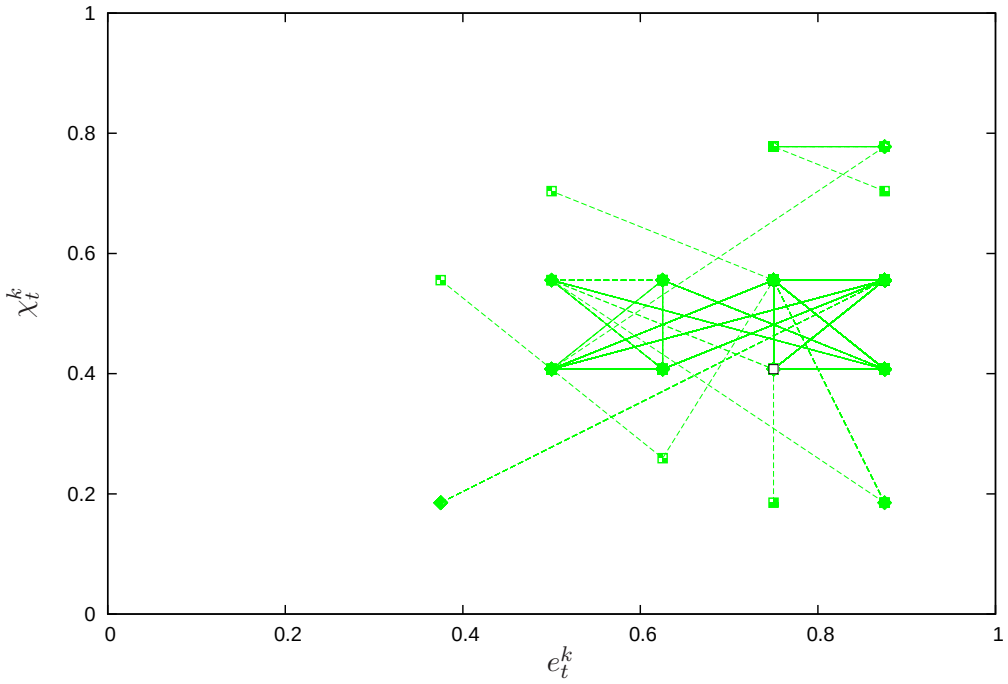

Figure 4.24: Phase diagram of non-defaulting Agincourt households, $l_{failure} = 4$.

4.9 Identification of social force in the social system

One of the central objectives of this thesis is to use the present techniques in order to identify social forces that lead the dynamics of social systems. In physics we feel justified by saying that, for example, because the force of a tennis racquet reverses the direction of motion of the tennis ball, that the racquet *causes* the reversal of the ball. We seek to identify force and cause in the present dynamical system.

Note that there are two levels of analysis that are used. They are the social unit-level and the population-level analysis. In physics, the forces on individual objects decide (for example atoms) the overall dynamics of many objects (for example the flow of water). In social dynamics if many households suffer the same individual-level force, it is natural to call this *demographic force*.

Force is sometimes obvious. For the answer to question q : does there exist HIV infection in the household?, the transition $a_t = 1 \to a_{t+1} = 0$ is owing to the "force" of HIV infection. Note that this may not be the same as $a_t = 0 \to a_{t+1} = 1$ where we must be careful to ask if sudden absence of HIV infection is owing to death of an individual, or his/her out-migration, which are completely different forces from that of infection. Thus there are various cases that we can identify

1. $a_t \to a'_{t+1} \neq a_t$ can in value be known, but possibly different forces in either direction.

   Yet suppose q : has the biological mother out-migrated? If the answer is Yes, it is now not clear what social force caused this. We must specifically ask this knowledge by better asking many questions. Perhaps

   $$\begin{aligned} q_0 &: \text{ Has the biological mother out-migrated to work?} \\ q_1 &: \text{ Has the biological mother out-migrated owing to illness?} \\ q_2 &: \text{ Has the biological mother out-migrated because of marriage?} \end{aligned} \tag{4.35}$$

   Then forces such that $a_i \to a'_{t+1} \neq a_t$ become clear as in item 1, only after asking the questions. If the answer to one of these questions is Yes, then

2. A force may not be identifiable by the question set.

3. If none of the questions of 1 is answered Yes, then an observed transition $a_t \to a'_{t+1}$ cannot be associated with a force of migration and we must ask further questions to achieve this. If we do not ask these questions here, we cannot expect to identify the forces of migration.

4. If two or more of the questions are answered Yes by various households, then multiple forces may be identified and they act at the same time.

5. In principle, a force might be identified for each of very many a (so that social dynamics is more complicated than physics).

6. However our purpose is educational progression and we can ask if mother migration, adult death or minor household heads are *forces relative to our purpose*

7. In physics we associate cause with a force. We do the same here.

From Table (4.3), the principle Agincourt transitions we have, relative to the purpose of educational default, the demographic forces

Table 4.8: Identification of social force

$24 \rightarrow 23$	Biological mother's out-migration is a demographic force and *cause* for educational default
$23 \rightarrow 24$	Biological mother's in-migration is a demographic force and *cause* for educational progression

The transitions $23 \leftrightarrow 24$ are mathematically reversible, but there is a different demographic force in each direction and different cause. The transition $32 \rightarrow 23$ is not reversible. It is the demographic force of biological mother out-migration from a fully fit household that is itself fed by in-migration from state 31. This reflects a population with different social history, owing to the timing of in- and out-migration of the biological mother.

4.10 Summary of Agincourt Demographic Results $1998 - 2007$

There were two main objectives in this chapter. The first was to extend the theory developed in chapter Three to the population-level. We have achieved the analysis at the population-level by investigating orbits of many social units at the same time. In particular, Agincourt data have been included in the discussion. We have achieved identification of sub-populations by characterising dominant fitness and significance.

We started the analysis in the chapter by considering two different scenarios related to the choice of initial question order. We found that the "fhiqo" scenario converges faster than the "riqo" scenario. As a result we select the "fhiqo" scenario.

The second objective of this chapter was to associate education status of Agincourt households in the discussion of properties of orbits. Since the present analysis does not accept variables with missing values in the data, we have to build new strategies to include the education variable which has many missing values for the Agincourt data. To solve this problem, we have coloured the orbit of each household. Through this process, we have achieved distinguishing between defaulting and non-defaulting populations.

The results in the two main parts of the discussion presented in this chapter, are in agreement. In summary, we have found the following:

1. Not all transitions with respect to $n = 3$ questions occur in the Agincourt population. We find about 23% of idling states in the Agincourt population. This suggests the presence of a stable sub-population in the Agincourt population.

2. Insignificant numbers of Agincourt households are headed by minors because q_1 was very stable in the state of adult household head. As a result the visualisation of orbits was then reduced to a sub-space of Γ_3 as presented in Figure 4.1. In particular we find that this reduced sub-space was defined by the following 6 states $23, 24, 29, 30, 31$ and 32 of Γ_3 as in Table 4.3.

3. The Agincourt population dynamics is dominated by two states 23 and 24. Most of significant social changes observed in the Agincourt populations were related to the transitions $24 \rightarrow 24 \leftrightarrow 23 \rightarrow 23$. In-and out-migration of biological mothers were identified as the major demographic events taking place in the Agincourt population with respect to our variables.

4. The education measure $l_{failure} = 4$ better separates the Agincourt population of households in educational default. At $l_{failure} = 4$, we find that there are 54.18% of Agincourt defaulting households and 45.82% of non-defaulting Agincourt households. The results at $l_{failure} = 2$ that 73.25% of households are defaulting is taken as evidence of the general failure of education at Agincourt and is rejected as a criterion for the effect of household change on educational default.

5. We find in the early years of our study that there is a net flux of households for transition

$24 \rightarrow 23$.

6. For the non-defaulting Agincourt households, the dominant transition was $24 \rightarrow 24$ which suggests that households in a fully fit state stay in these conditions for a long period of time.

7. For the defaulting Agincourt households, we find that once biological mothers out-migrate, they take long to return to their households. This was revealed by the transition $23 \rightarrow 23$ which was found to be the most dominant in this sub-population.

8. We conclude (see the results of Table 4.6) that Agincourt educational default is related to absence of biological mothers. As noted in Chapter Two, there is to our knowledge no published information about stochastic errors in the Agincourt data. However from Table 4.6 we note that for $l_{failure} = 4$, the percentage idling transitions $23 \rightarrow 23$ and $24 \rightarrow 24$ will tolerate an error of $\pm 10\%$ without altering our conclusions that absence of biological mothers is associated with educational default. We add that at $l_{failure} = 2$ level, we can only tolerate an error of 4.5% so that we have an additional criterion for choosing $l_{failure} = 4$.

4.11 General strategy for large number of questions and comment on the role of statistical methods

The application of orbit theory is not limited to a small number of questions $n = 3$ as in the analysis presented above. We have discussed simulations with $n = 26$ questions in Section 3.3.2 and Section 3.5.1 of Chapter Three. We have given simulations with $n = 26$ questions and discussed clustering (see Figure 3.18). However we have needed to use decimal notation for fitness to see the effect of change of digits on the right of the fitness sequence; this illustrates that the phase space becomes almost continuous for large n (recall that for n questions, the number of states is $d = 2^n \times n!$; to illustrate, for $n = 3, 4, 5, 6$ we have $d = 48, 384, 3840, 46080$ respectively). To deal more thoroughly with large n we may adopt deterministic strategies. It is also clear that because d increases very quickly with the number of questions, it might sometimes be useful to shorten computational time by using statistical methods.

Deterministically, we may sensibly use the following strategy: Suppose we have long time series. To say that for example, 10 variables are associated with educational default, rather than just one (as in the case of mother out-migration), does not enable us to relate a small number of possible

causes. However, suppose we eliminate one variable. We get a small state space S_{n-1}, but if the orbits in S_{n-1} have significantly the same connectivity as in S_n then that variable has no independent relevance to change of state and thus to cause. In this case we drop that variable. We proceed in this fashion to eliminate all such irrelevant variables.

To illustrate, consider the strong transitions at Agincourt as shown in Figure 4.1. Notice that we may delete question q_1 (Minor household head) without changing the connectivity of the figure. However, to remove question q_2 (Adult death) will give a set of connections in S_1 (just the two states 23 and 24 with connections as shown) that is obviously different from S_2. The same would apply if we remove question q_0. Alternatively, we may start with small numbers of questions (as in this thesis) and work through permutations to find strong associations. Indeed, it is a primary importance of our method that clustering associates demographically important transitions. In our argument significantly connected patterns can define properties of a cluster. Note that each cluster carries approximately independent transitions and each deserves detailed study. All this suggests study of permutations of a reduced number of questions.

However this deterministic method has the serious constraint that demographic time series are relatively short. At Agincourt we have an average observation time of 7 years. It is obvious that if one answer value changes at each time step that for example for 7 questions, it could take 7 years for a question on the right to migrate one place at a time to the left. Indeed if the right hand question diffuses randomly through questions order, it could take 7^2 years to reach the left hand side. For 3 questions, it could take 9 years to diffuse from left to right which is comparable to the average Agincourt observation time. For this reason also, we should choose a small number of questions to give each answer an opportunity to show its effect, and then search through permutations of small numbers of questions.

Our method does suggest a statistical approach for analysis if n is large. To illustrate, let us double the number of questions used in the above demographic analysis and assume $n = 6$ questions (or variables). In longitudinal studies, frequencies of change of variables is a fundamental property of the data. Thus, from a statistical point of view, it is easy to determine the frequency of each answer value change just by reading the data. Let f_i denote the frequency of change of answer value for question i. We conveniently model each answer frequency $f_i, i = 0, 1, \cdots, 5$ as a probability so that

$$\sum_{i=0}^{5} f_i = 1 \ . \tag{4.36}$$

To illustrate, suppose

$$\begin{cases} f_0 &= 0 \\ f_1 &= 0.15 \\ f_2 &= 0.4 \\ f_3 &= 0 \\ f_4 &= 0.35 \\ f_5 &= 0.1 \end{cases} \tag{4.37}$$

The frequency of change of each answer value of (4.37) can be visualized in Figure 4.25 where a_i labels the answer value for question i.

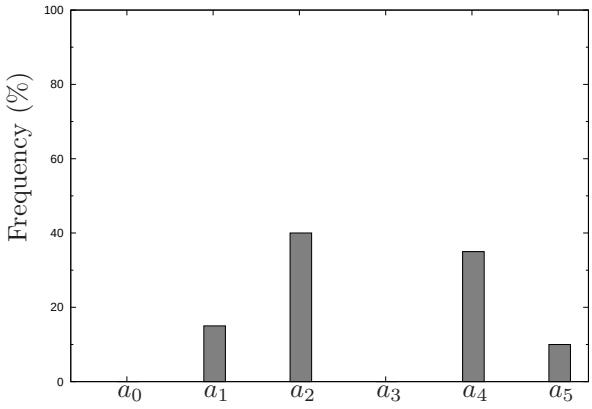

Figure 4.25: Example of frequency distribution of changing answer value.

From the distribution of Figure 4.25, our method asks us to identify the significance level of clustering. Thus, because the answer values a_0 and a_3 do not change ($f_0 = f_3 = 0$), the clusters in this case will be defined in the state space S_6 by all the significance sequences $30a_{i_1}a_{i_2}a_{i_3}a_{i_4}$ (e.g. 301245) and $03b_{i_1}b_{i_2}b_{i_3}b_{i_4}$ (e.g. 031245) where a_{i_j} and $b_{i_j} \in \{1, 2, 4, 5\}$ which here start with 30 and 03 respectively.

In general, for any number of questions n the result (4.37) is straightforward statistics of the data that can be easily determined. By including the question order in the analysis of the data, orbit theory indicates that if $i = 0, 1, 2, \cdots, n-1$ is the most slowly changing answer value, the clustering in the population under this scenario is automatically given by the significance level defined by the

sequence starting by i. Thus, this helps to identify a sub-region of S_n on the significance axis where clustering will occur, even before we plot orbits.

In order to achieve identification of clustering in the state space S_n, we must also determine the fitness region of clustering. Thus, we need additional information from the data. By reading the data, we can determine the distribution of favourable (1) and unfavourable (0) responses to each question. Let f_i^j denote the frequency of question i having an answer value $j \in \{0, 1\}$.

To illustrate, let us use the above example (4.37), suppose that $f_0^1 = 0.7$ and $f_0^0 = 0.3$ as displayed in Figure 4.26. Clearly question 0 has more favourable answers. Thus, because the clustering involves question 0, we will expect to have more clustering on the sub-region of S_n defined by $(x, y) = (1c_{i_1}c_{i_2}c_{i_3}c_{i_4}c_{i_5}, 03d_{i_1}d_{i_2}d_{i_3}d_{i_4})$, where $c_{i_j} \in \{0, 1\}$ and $d_{i_j} \in \{1, 2, 4, 5\}$, that is, on the right half of S_6.

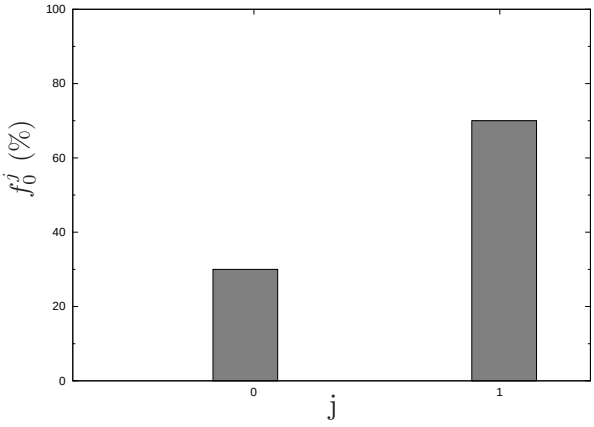

Figure 4.26: Example of frequency distribution of $f_0^j,\ j \in \{0, 1\}$

This statistics is related in a primitive way to the statistical techniques mentioned in the review (Section 1.3). The survival function underlies these techniques. The frequency distribution of Figure 4.26 may be interpreted as relative likelihood for change in each variable with f_2 the fastest and if we associate change to unfavourable status with question 2 then we have a "survival measure" of that variable. Note that for an individual where many changes occur unfavourable and favourable changes must balance and that this statistic is definitely a demographic property of the population. With reference to the simple Kaplan-Meier survival formula (1.22) applied to our data, with $d =$ 1998 (i.e. for the whole period of observation) we have

$$P(\text{Out-migration of mother}) \gg P(\text{ Adult death}) \gg P(\text{ Minor household head}) \sim 0 \qquad (4.38)$$

for defaulting and non-defaulting households with

$$P^{\overline{Ed}}(\text{ Out-migration}) > P^{Ed}(\text{ Out-migration}) \tag{4.39}$$

Note that since $f_0 = f_3 = 0$, these two clusters cannot change in time. If we drop q_0 and q_3, then q_5 will define clustering, but it can cluster itself on the left and on the right depending on f_5^j. But then clustering can be time-dependent (e.g. clustering on the right $1998 - 2002$, on the left $2003 - 2007$). It is immediately clear that orbit theory, which reveals the time-dependence of clustering in figures such as Figure 4.11, would show such a transition. Note direction of change (e.g. $23 \to 24$ and $24 \to 23$); in our simple statistical method these will not be identified. Of course application of the rigorous statistical techniques of the review above will detect a jump in survival probability at such a transition.

It should be noted that we have not been able to avoid statistical analysis altogether. Note that Figure 4.13 is a survival curve and is essential in deciding a criterion of educational progression. Thus if we define D to be delay in educational progression and d to be grades (or years) of delay in educational progression, then a "survival curve" is defined precisely by (1.22).

Yet, orbit theory presents a method of visualizing all possible states and corresponding transitions that we can have with any number of questions n in the easy-to-understand state space S_n. Full information is preserved, yet we can visualize patterns and extract demographic information.

Chapter 5

Projecting the future of Agincourt social dynamics

5.1 Introduction

In the hard sciences, the critical test of a theory resides in projection. It is here that the opportunity emerges to say that a theory fails absolutely, if it fails a projection. Suppose a purpose is understood in terms of stable Q_t. Then many individuals entering the population (e.g. those not sampled under Q_t) can be placed on the appropriate *typical orbit* at some time t by interviewing with Q_t - then if that momentary response lies on one flow only, its future is reasonably predictable by the typical orbit. Fluctuations between states $23 \leftrightarrow 24$ are clearly identified above as typical. Of course this might be the first social unit of a new flow, or in a flow missed by the choice of sample from the population. These are discovered properties in the demography and a typical orbit is obviously deeper information than that offered by statistical analysis especially if cause can be identified along the orbit. Demographic information is directly extracted from a typical orbit. If the deterministic models mentioned in the previous chapters are applicable, they present opportunities for projection. If they fail we may say unambiguously that the present theory, adapted for projection has failed.

Social systems present complex behaviours and we might suppose that human behaviours are irregular and not predictable. The existence of periodic behaviour at Agincourt will disprove this.

5.2 Formulation of the dynamics

Let $\underline{s}_{t,n}$ denote the *state row vector* at time t. Recall that with n questions, we have $d = 2^n \times n!$ possible states. Thus, we have

$$\underline{s}_t = (s_{t,i}) \quad i = 1, 2, 3 \ldots, d \tag{5.1}$$

where $s_{t,i}$ represents the index of state i at time t. Relate $\underline{s}_{t+1}$ to $\underline{s}_t$ by

$$\underline{s}_{t+1} = \underline{s}_t T_t^P \tag{5.2}$$

As stated above T_t^P is the transition matrix of the population P at time t. This is a deterministic dynamical system that captures the full set of transitions that occur at Agincourt. It gives the possible dynamics of individuals or households at Agincourt.

Let us make some iterations from the definition (5.2). For simplicity we let $T_t^P = T_t$. Thus, we have, with matrix multiplication

$$\begin{array}{rcc} \underline{s}_1 & = & \underline{s}_0 T_0 \\ \underline{s}_2 & = & \underline{s}_1 T_1 = \underline{s}_0 T_0 T_1 \\ \underline{s}_3 & = & \underline{s}_2 T_2 = \underline{s}_0 T_0 T_1 T_2 \\ \vdots & = & \vdots \\ \underline{s}_t & = & \underline{s}_{t-1} T_{t-1} = \underline{s}_0 T_0 T_1 T_2 \ldots T_{t-1} \end{array} \tag{5.3}$$

Thus, we write

$$\underline{s}_t = \underline{s}_0 T^{t-1} \qquad t \geq 1 \tag{5.4}$$

where

$$T^{t-1} = \prod_{t'=0}^{t-1} T_{t'} \ . \tag{5.5}$$

Here, we capture the detailed dynamics of P, given an initial state vector. The relation (5.4) is an example of a dynamical system [16, 137, 138, 139, 118, 120]. At this stage, $\underline{s}_t$ is just an abstract vector of state indexes. It is the 'space' in which our population moves. The relevant general way in which states at time t change in going to time $t+1$ is hidden in T_n^P, for n questions. It may be specialized, for example for 3 questions on the Agincourt population, T_3^{Ag}. To see how T_n^P itself contains information, note that if each $T_{t,n}^P = T_t$ in (5.3) and if each $T_t = T_0$ =constant, then (5.5) becomes the power matrix of T_0, that is

$$\underline{s}_t = \underline{s}_0 {T_0}^{t-1} \tag{5.6}$$

Then in [118] it is shown that while the elements of our basic transition matrix T_0 contain only $1'$s and $0'$s, the elements of ${T_0}^{t-1}$ count the number of *orbits* (not the number of households) that go from state i to state j in $t-1$ steps. This applies more generally to (5.5). For this reason, we note the important point that this map is not suitable for demographic purpose where the number of households on an orbit is of important.

As a further example, if T^{t-1} has no inverse (Det $(T^{t-1}) = 0$) then there are transitions $i \to j$, $j \nrightarrow i$, for some i, j.

Now the properties of the social transitions of the population P are computed by the characteristics of the properties defined in (5.4) and (5.5).

Longitudinal data gives us not only T^{t-1}, but also D_3 (4.21). With the knowledge of the initial density of states, d_{ij} of $D_3 = (d_{ij})$, (4.5), (4.6) and the transition matrices $T_t, t \geq 1$, we are now able to determine the number of social units at any state i, $i = 1, 2, 3, \ldots d$, at any time t as follows. We construct the density row vector

$$\underline{m_t} = (m_{t0}, m_{t1}, \ldots, m_{td}) = (m_{ti}) \tag{5.7}$$

where m_{ti} represents the number of social units at state i, at time t. Similarly to (5.2) and (5.4) we relate m_{t+1i} and m_{ti} as follows. By definition we have

$$m_{t+1i} = m_{ti} + \delta_{ti} \tag{5.8}$$

where $\delta_{t,i}$ is as defined in (4.9). It is just the net flow [135] into state i, in one time step.

Thus, from (5.8) and (4.10) we can write

$$m_{ti} = m_{0i} + \sum_{t'=0}^{t-1} \delta_{t'i}, \quad \underline{m_t} = \underline{m_0} + \underline{F_t} \tag{5.9}$$

where

$$\underline{F_t} = \sum_{t'=0}^{t-1} \underline{f_{t'}} \ = \underline{m}_{1998} + \underline{F}_{2006}; \tag{5.10}$$

The density vector of the Agincourt population in 2007 will be predicted to be more generally

$$\underline{m}^{Ag}_{2007} = \underline{m}^{Ag}_{1998} + \sum_{t'=0}^{2006} \underline{f}^{Ag}_{t'} \tag{5.11}$$

It is at this level that we have a simple model for predicting ahead. This too, is a deterministic dynamical system. It is not defined over an abstract space as in (5.2) and specifically models population level flows into and out of a state.

5.3 Method of projection

It is a tradition in the formulation of any new theory that first, models are built which constitute the basis of the theory. Now that the models have been achieved, in the following, we present the outline of the techniques that are used to include projection.

To predict ahead for Agincourt, we begin by simulating the longitudinal data of households. We use as much information of Agincourt data as possible. Then we test the simulated transition, density and flux matrices against the time matrices for agreement. Then if no dramatic changes occur in comparison to the Agincourt population, we may predict ahead by running longer simulations. This may be useful. For example, 4000 new households have been added to the Agincourt population study (in 2009), no longitudinal data is of course available, but if the new population is reasonably similar to the existing population, we can reasonably forecast the new dynamics.

Once we have confidence of similarity with full data, we can investigate scenarios. Thus suppose we make one change only to the data, say we pay biological mother to stay at home. Keeping all else unchanged, and supposing that indeed this will improve educational progression (as found in chapter Four), we can simulate forward from the present state of Agincourt, to see how the population evolves. Questions we might ask are "how many years before educational default is halved?"

In outline our method of simulation is

1. Use the average observation time, $\overline{l}$, from Agincourt data as calculated in (2.4).

2. Identify periodic orbits from Agincourt data, that complete at least two clear oscillations within $\overline{l}$. Determine their population fractions and typical orbits. If there are in excess of that expected from random sampling, the simulation of some population is just the periodic fraction on the identified orbits. This is trivial.

3. Assume that the remaining orbits are random. Determine frequency of change of an answer value (Figure 2.5) to each question and frequency of occurrence of n questions changing

answer values $n = 0, 1, 2, 3$ (Figure 2.6). Simulate household data accordingly.

4. From the approximate density of states at $t = 0(1998)$ define initial occupation numbers for simulated households, denoted $\underline{m}_0$.

5. $T_3^{Ag(Approx.)}$, $\underline{F}^{Approx.}$ are the Agincourt approximate matrix of allowed transitions and vector of fluxes. For each household as in (5.9), iterate the dynamical system $\underline{m}_{t+1} = \underline{m}_t + \underline{f}_t$.

6. Compare $D_{3,5}^{Sim}$ with $D_{3,2006}^{Approx}$, also $F_{3,4}^{Sim}$ with $F_{3,2005}^{Ag}$

7. If not in good agreement, use T_3 not T_3^{Ag}. This is useful anyway to decide if the approximation T_3^{Ag} is good. $T_{3,t}$ varies between $1998 - 2007$ and if T_3 is not useful, it might be necessary to model its time dependence.

8. When good agreement

   (a) set initial conditions on $D_{3,2006}^{Ag(Approx.)}$ to simulate forward for the existing Agincourt population. If we have first data for the 4000 new households, then we can use that to predict ahead, as well.

   (b) simulate effects of interventions, e.g., mother-grants to keep them at home. Estimate rates of change into the future.

5.4 Detecting periodic orbits

5.4.1 Periodic orbit for a dynamical system

Suppose $\zeta_t = (e_t, \chi_t)$.

Definition 5.1. A periodic orbit with period τ for the map

$$\psi = \varphi \circ \phi : \Gamma_n \to \Gamma_n : \zeta_{t+1} = \psi(\zeta_t), \qquad \zeta_t = (e_t, \chi_t) \in \Gamma_n, \qquad n \geq 1 \tag{5.12}$$

is the set of τ distinct points

$$\zeta_t = \psi^t(\zeta_0) \ , \ t = 0, \cdots, \tau - 1 \text{ with } \psi^\tau(\zeta_0) = \zeta_0 \ . \tag{5.13}$$

where ψ^τ represents the composition of ψ with itself τ times. The smallest positive value of τ for which this equality (5.13) holds is the period of the orbit.

We have coded ζ_t to a pair of integer values for each state. Then equivalently suppose $s_t = (b_t, \theta_t)$

Definition 5.2. Let $\underline{s}$ be the vector of state $(b, \theta) \in S_n$ as in (3.68). Let $s_{t+1} = s_t T_n$ where T_n is the *constant* transition matrix of the (theoretical or measured) dynamics [119, 120]. Then a periodic trajectory of period τ is one such that

$$s_{t+\tau} = s_t {T_n}^{\tau}, \ \forall \ t \ . \tag{5.14}$$

In addition to the definitions of periodic orbits given in Definition 5.1 and Definition 5.2, in this thesis, we note that all period $\tau \geq 2$ orbits are assumed not to be of period $\tau = 1$. In general, note that all period $\tau = 2k, \ k = 2, 3, \ldots$ are assumed not to be of period $\tau = k$. This is useful to distinguish for instance, between social units of period-2 orbits with those of period-4 orbits.

5.4.2 Agincourt period one ($\tau = 1$) orbits

Period one orbits ($\tau = 1$) are called fixed points [140, 44, 42, 139, 45] of a dynamical system. These are very special cases of predictable behaviour, because this type of dynamics is related to stationary household state, that is to idling states on the diagonal of T_3^{Ag}. The analysis of this particular sub-population is especially simple in S_n.

In particular, for a questionnaire consisting of $n = 3$ questions, the number of distinct period one orbits s_1 as in (4.13),

$$s_1 = \quad \text{Trace } (T_n^{Ag}). \tag{5.15}$$

Figure 5.1 and Figure 5.2 display period one orbits for the Agincourt population. We clearly see that there is only one orbit defined by state 24 or $(111, 120)$ that characterises Agincourt period one sub-population.

It is important to note how the initial question order is determined for period one households. Recall that period one orbits have stable behaviour. Thus, it is important that the initial question order that is taken from the "fhiqo" scenario, returns for the case of no change, the average population significance defined by $\overline{\theta_{Ag}} = 120$. The total number of Agincourt period one households is given in Figure 5.3. We find 4 (about 0.15% of total population) Agincourt period one households.

Recall that state defined by $(e, \chi) = (111, 120)$ represents households fully fit. Note also that the state $(111, 120)$ indexed by 24 is the most idling state in the present Agincourt population and

the associated transition $24 \rightarrow 24$ or $(111, 120) \rightarrow (111, 120)$ represents 31.28% of the transitions that we observe in the Agincourt population. It represents more than one third of the total number of the transitions that occur in the Agincourt data. Thus, it is convenient that all period one households are defined by that state.

This type of analysis is useful for projection which in this specific case is obvious. If such a sub-population is identified then, we can simulate its behaviour. As stated above, note that a period one orbit defines a sub-population. The properties of these orbits give reduced information of the associated sub-population dynamics.

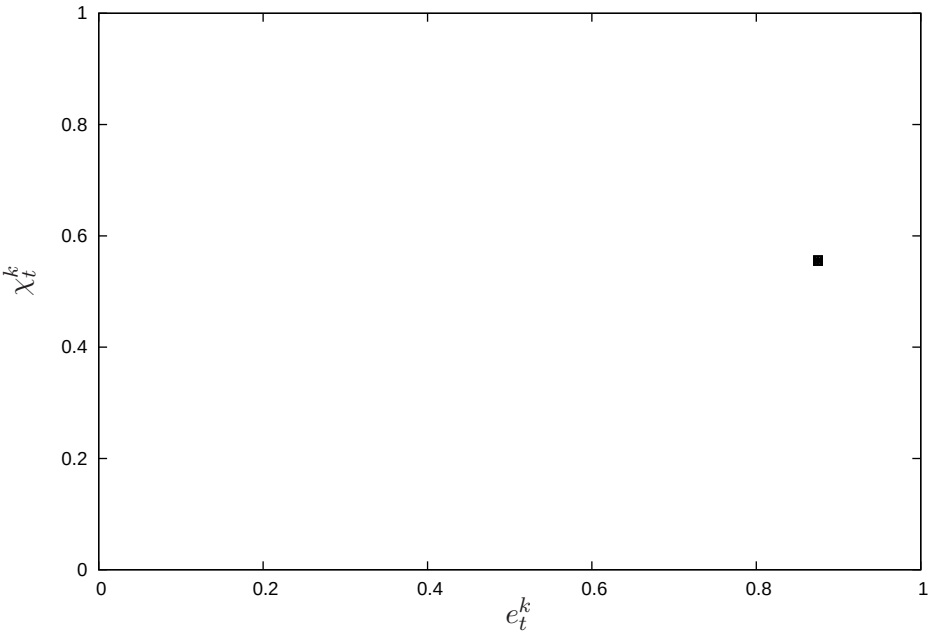

Figure 5.1: Agincourt period-1 household orbits (e_t^k, χ_t^k) in Γ_3.

5.4.3 Agincourt period two ($\tau = 2$) orbits

The dynamics of the Agincourt period two orbits are presented in Figure 5.4 and Figure 5.5. In particular, these figures clearly show that all the Agincourt period two orbits are represented by two states of Γ_3 or S_3. These two states are defined by $(b, 120)$ where the fitness component of the states is located in the region defined by $b \in \{110, 111\}$. We also see that all the Agincourt

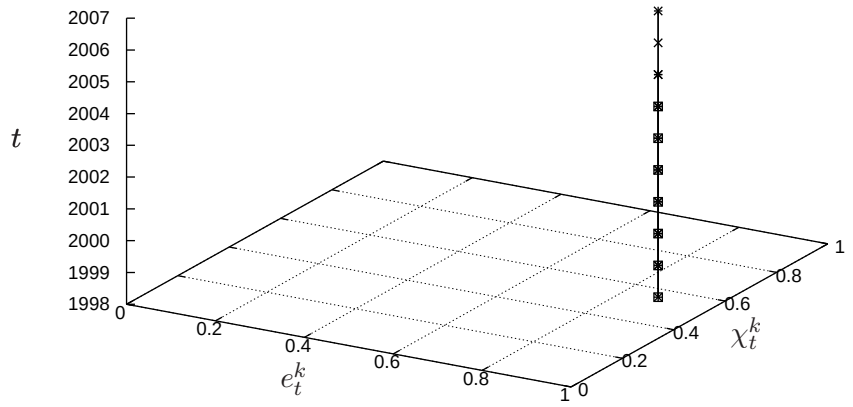

Figure 5.2: Orbits with time dependent (vertical axis), for Agincourt period-1 household orbits (e_t^k, χ_t^k).

period two orbits have the same and constant significance given by $\theta = 210$, as displayed in Figure 5.6. This significance is related to no change in answer value of question q_1. In these period two sub-populations, although in- and out-migration of biological mothers dominated, no household is headed by a minor.

In Figure 5.8, we show the initial distribution of the number of period two Agincourt households over the associated two states of Figure 5.4. In contrast to the above case, this initial distribution differs from the final distribution because the period two households do not all stay where they start. The total number of period two Agincourt households is 3. We note a slightly decrease of about 0.04%, in the number of period two households compared with the number of period one households.

As before, projection in this case, is also obvious. Because we know what the dynamics are, for each social unit in this sub-population, we can predict that if the biological mother out-migrates from the household at a given point in time t, then we know exactly that at time $t + 2$, she will return into that household, and vice versa. Again, it is also possible to identify typical orbits associated to this period two sub-population.

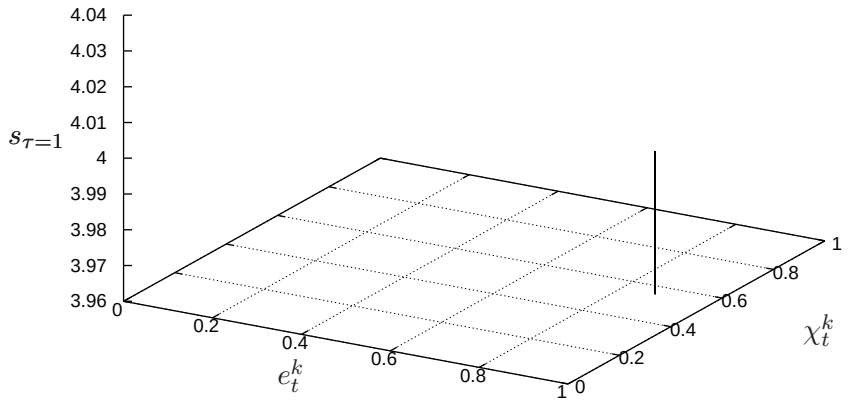

Figure 5.3: Distribution of the number of Agincourt period-1 households over the states $(e_t^k, \chi_t^k) \in \Gamma_3$.

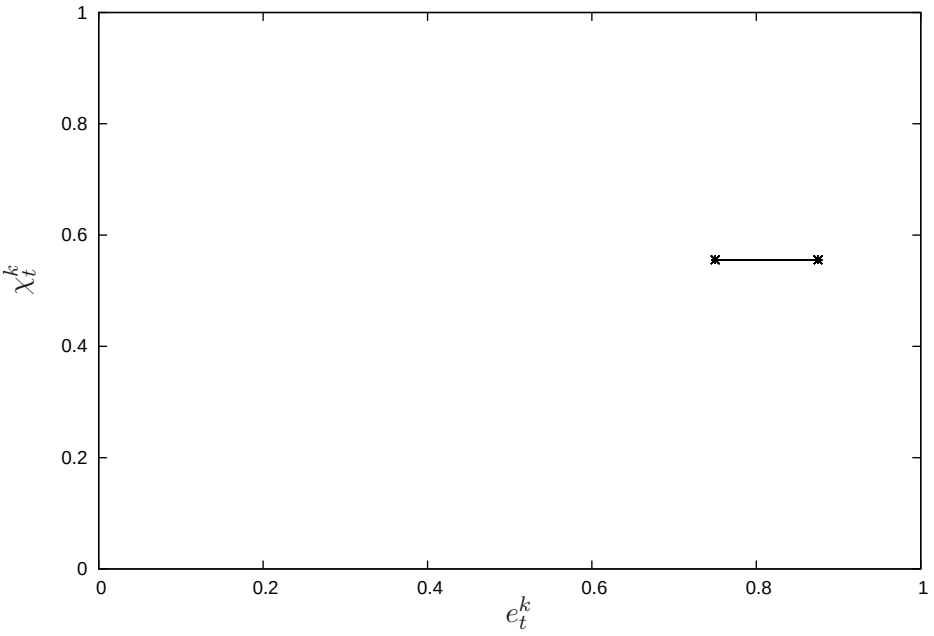

Figure 5.4: Agincourt period-2 household orbits (e_t^k, χ_t^k) in a sub-space of Γ_3.

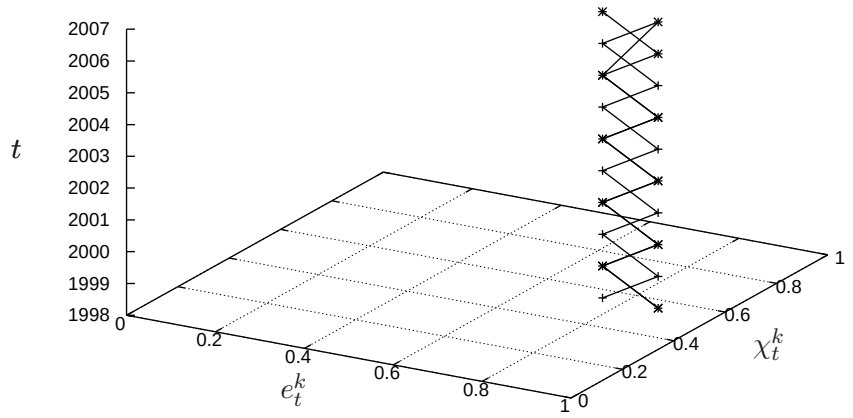

Figure 5.5: Phase diagram in reduced space, with time dependent (vertical axis), for the Agincourt period-2 household orbits (e_t^k, χ_t^k).

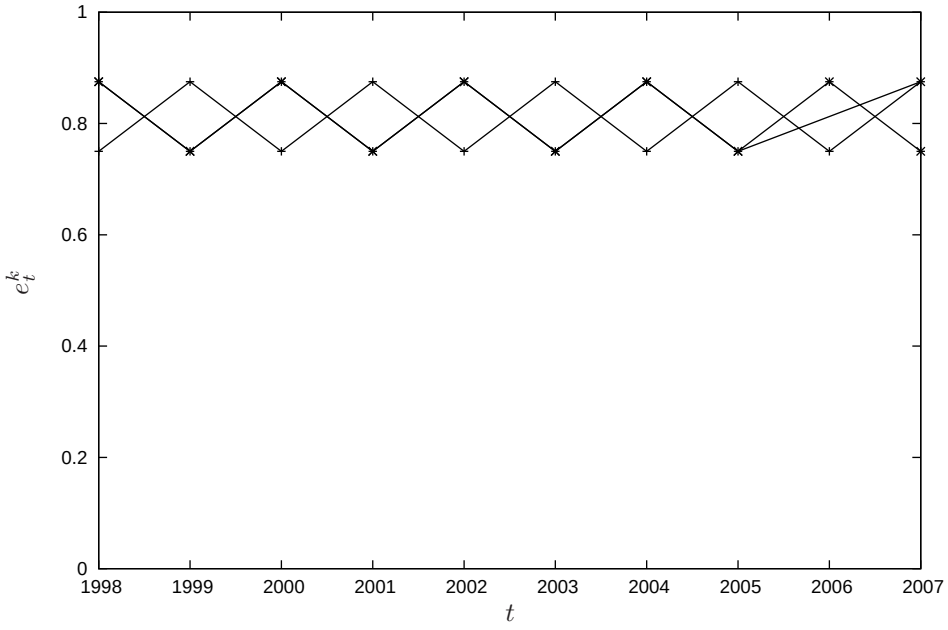

Figure 5.6: Evolutionary fitness orbits e_t^k, for all Agincourt period-2 households.

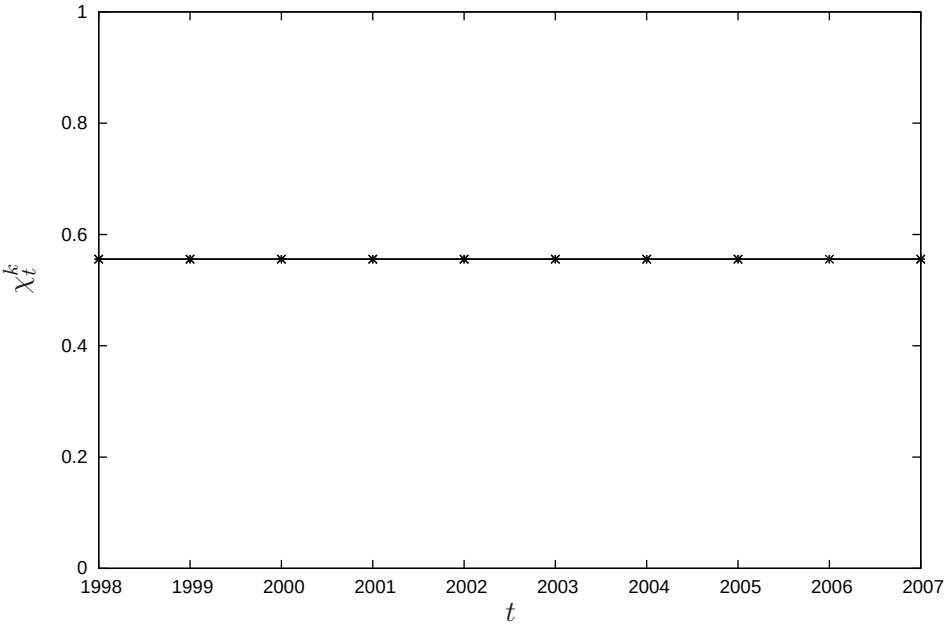

Figure 5.7: Evolutionary significance orbits χ_t^k, for all Agincourt period-2 households. Note that $\theta = 120$, since only q_0 is changing

5.4.4 Agincourt period three ($\tau = 3$) orbits

In Figures 5.9-5.12 we present the dynamics of the Agincourt period three orbits. Figure 5.9 and Figure 5.10 show that the Agincourt period three orbits are defined by the same two states $(110, 210)$ and $(111, 210) \in \Gamma_3$. The significance level $\chi = 210$, being also constant and the same (see Figure 5.12), we have the same social dynamics as above. In particular, we can see that here the dynamics are also linked to the movement of biological mother who is sometimes in and sometimes out from the household. In this case, when she is out of the household, she just delays her return for one time step, and vice versa.

Figure 5.13 displays the initial distribution of the number of period three Agincourt households over the states of Figure 5.9. We find that there are 4 Agincourt households of period three orbits starting in 1998. But now they all started at state $(111, 120)$, hence they were initially all in a favourable state. As before, we also identify the movement of the biological mother as the major social change in this sub-population. We also note a slightly increase of about 0.04% in the

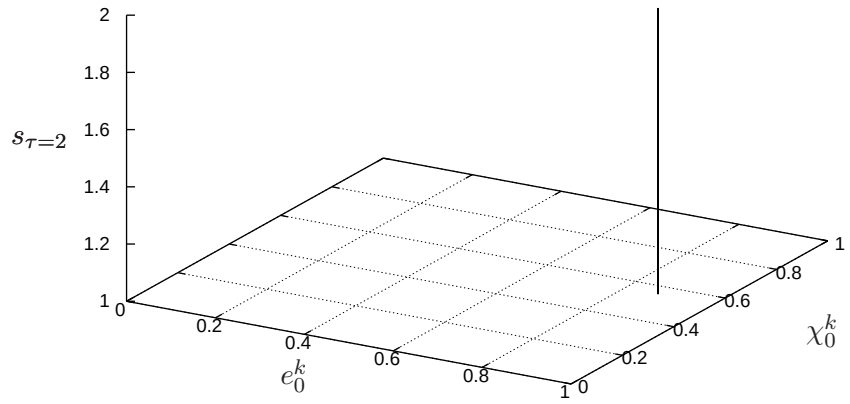

Figure 5.8: Initial distribution of the number of Agincourt period-2 households over the states $(e_t^k, \chi_t^k) \in \Gamma_3$.

number of households if we compare the number of period two households with that of period three households in the Agincourt population.

Various combinations of dynamics can lead to period three orbits. For example, a period three household orbit can be made of part of period two behaviour and another part of period one behaviour or vice versa. As before, projection of a period-3 orbit is trivial. For simulation of the whole population, the fraction of period-3 orbits is assumed together with their orbits.

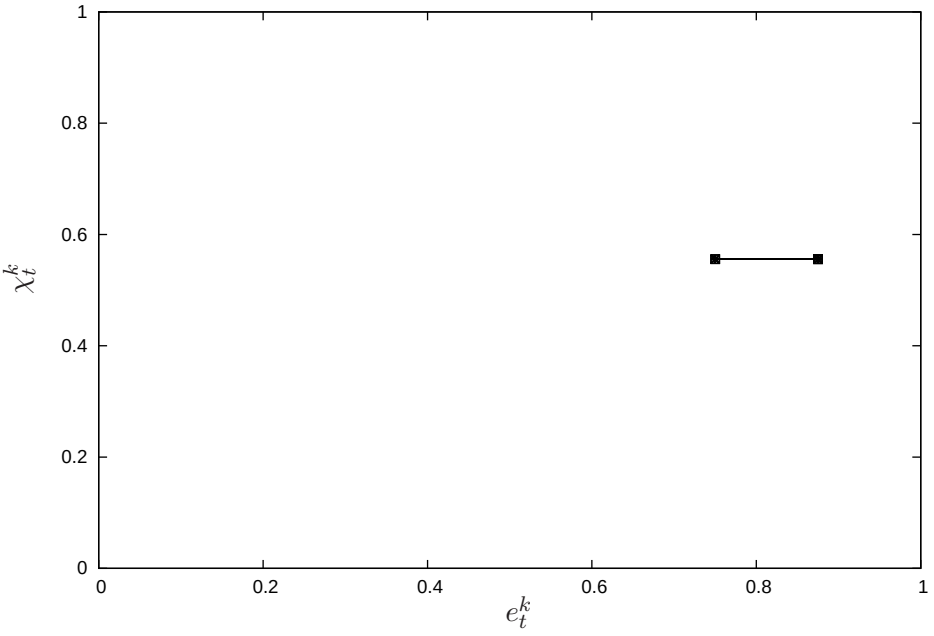

Figure 5.9: Agincourt period-3 household orbits (e_t^k, χ_t^k) in Γ_3.

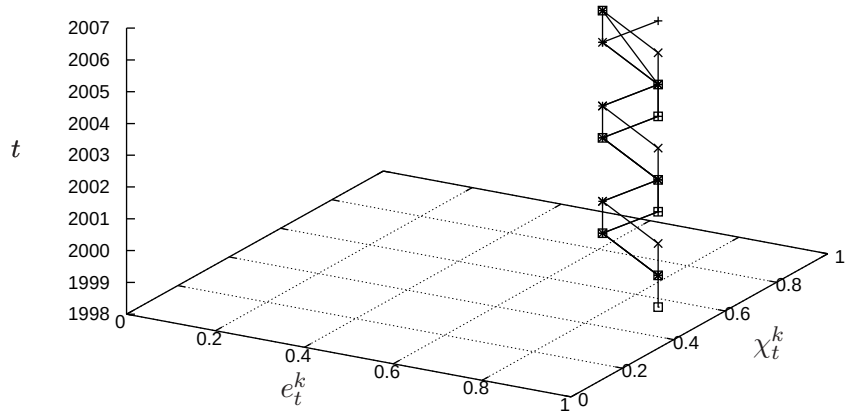

Figure 5.10: Phase diagram with time dependent (vertical axis), for the Agincourt period-3 household orbits (e_t^k, χ_t^k).

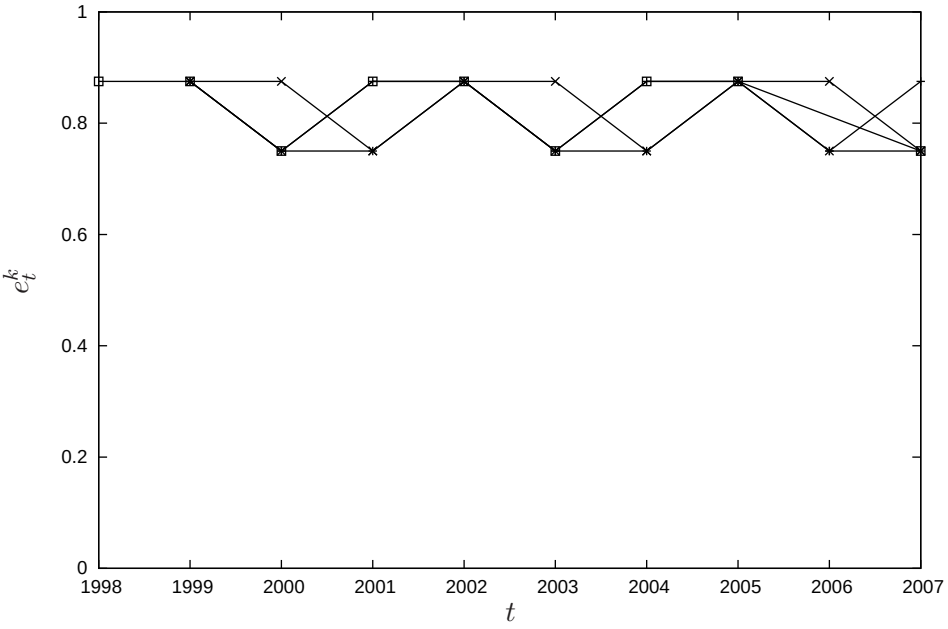

Figure 5.11: Evolutionary fitness orbits e_t^k, for the Agincourt period-3 households.

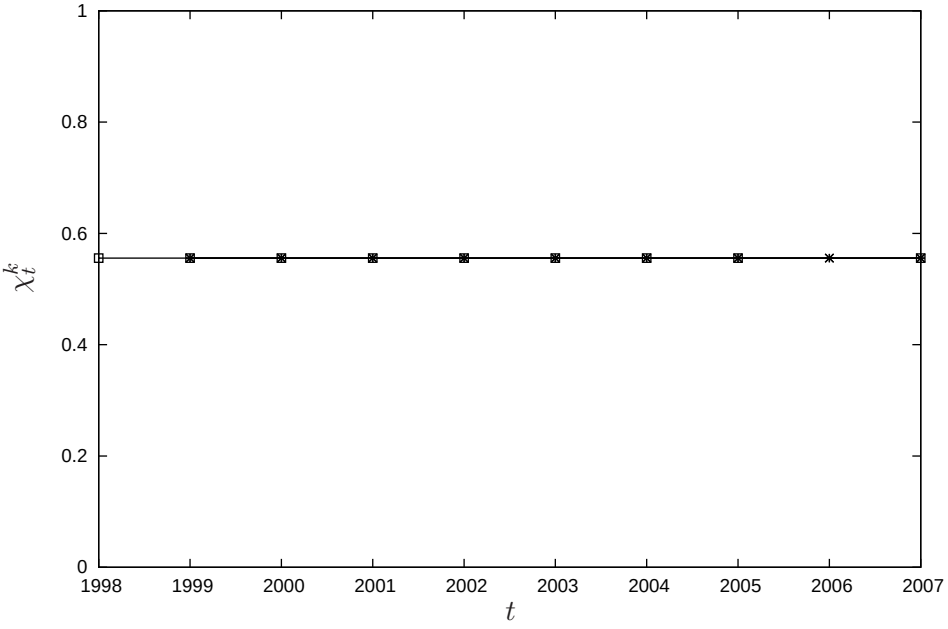

Figure 5.12: Evolutionary significance orbits χ_t^k, for the Agincourt period-3 households.

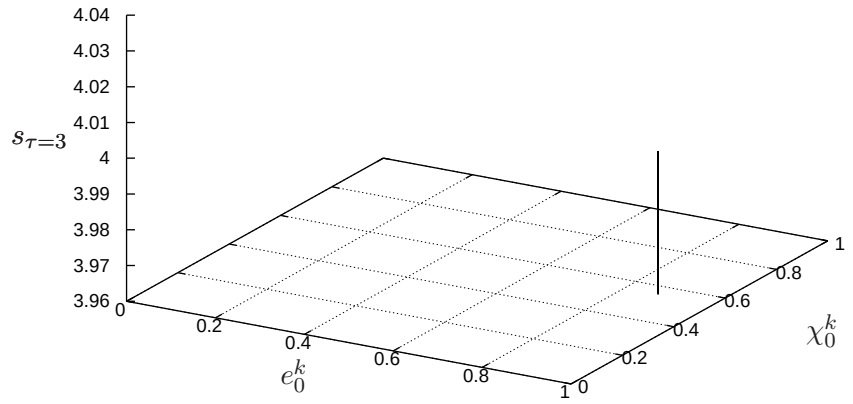

Figure 5.13: Initial distribution of the number of Agincourt period-3 households over the states $(e_t^k, \chi_t^k) \in \Gamma_3$.

5.4.5 Agincourt period four ($\tau = 4$) orbits

Following the same methods, we present the dynamics of the Agincourt period four orbits in Figures 5.14-5.17.

Similarly, Figure 5.14 and Figure 5.15 show again the dominant two states $(110, 210)$ and $(111, 210)$ of a sub-space of Γ_3. The significance level $\chi = 210$, being also constant and the same (see Figure 5.17), we have the same social dynamics as above. In particular, we can see that here the dynamics are also restricted to the movement biological mother who is sometimes in and sometimes out of the household. In this case, when she is out of the household, she just delays her return for one year, and vice versa.

In Figure 5.18 we present the initial distribution of the number of period four Agincourt households over the states of Figure 5.14. The total number of period four Agincourt households is 12. As before they almost all started at the same favourable sate $(111, 120)$. The movement of the biological mother is also the major social change in this sub-population. Now we have an increase of about 0.3% in the number of households if we compare the number of period three households

with that of period four households in the Agincourt population.

The period four dynamics can be constructed from different scenarios. As before, we can also predict the future of this type of sub-population with the knowledge of its states.

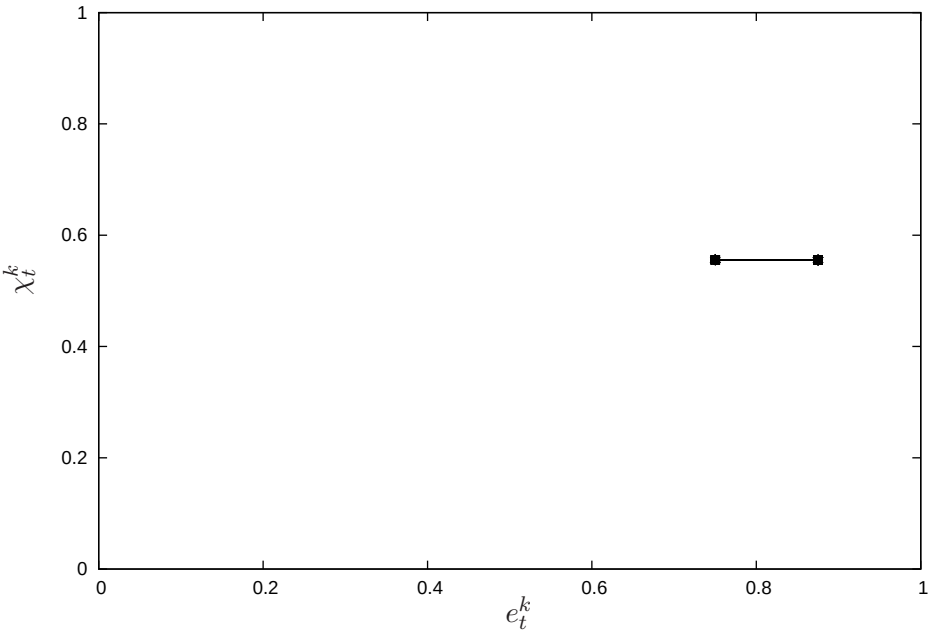

Figure 5.14: Agincourt period-4 household orbits (e_t^k, χ_t^k) in Γ_3.

Using the definition (5.13), the following table summarises the information about the distribution of the number of the Agincourt household periodic orbits.

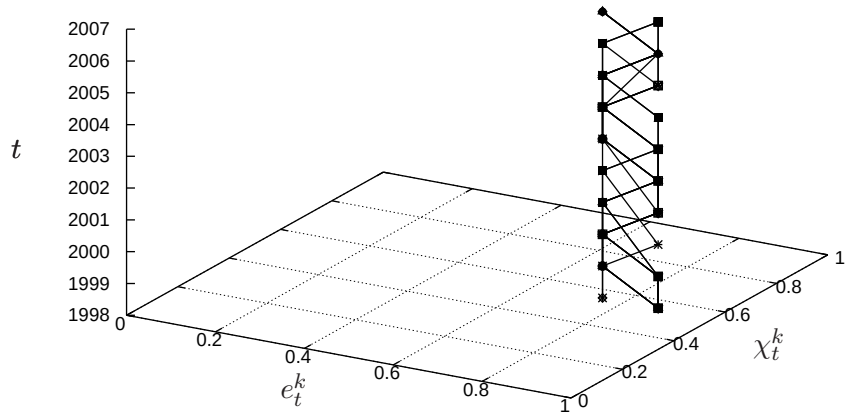

Figure 5.15: Phase diagram with time dependent (vertical axis), for all Agincourt period-4 household orbits (e_t^k, χ_t^k).

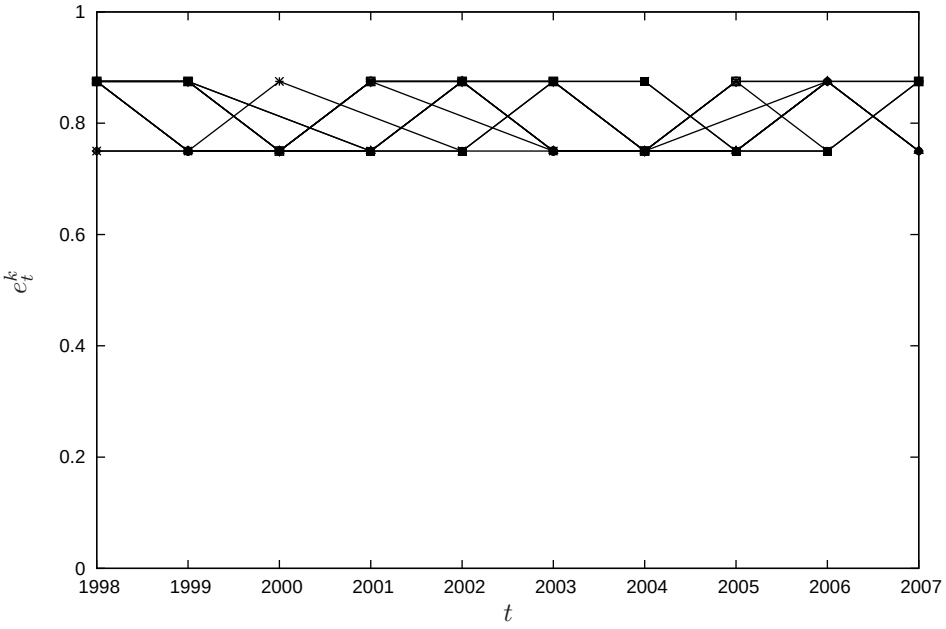

Figure 5.16: Evolutionary fitness orbits e_t^k, for the Agincourt period-4 households.

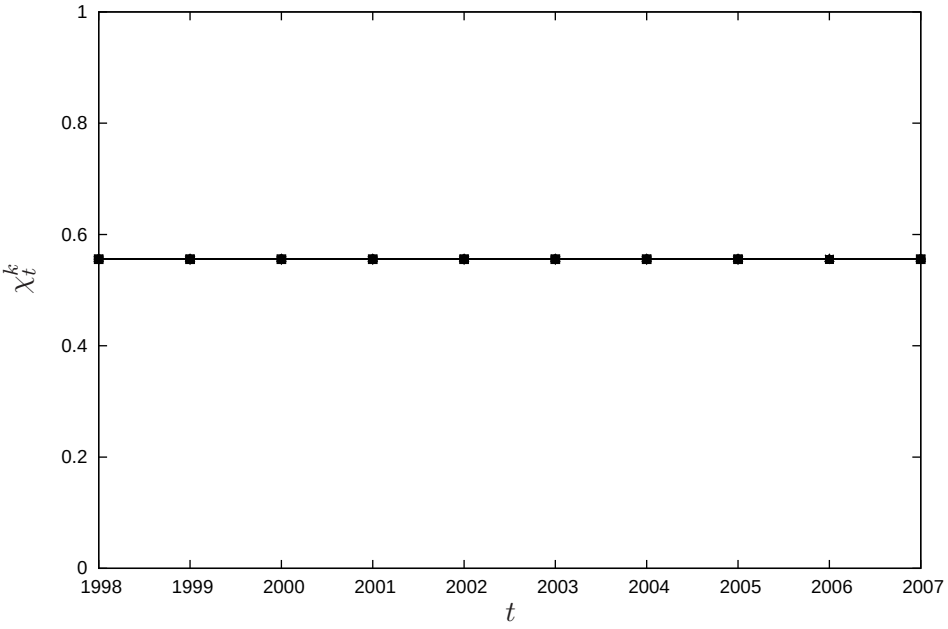

Figure 5.17: Evolutionary significance orbits χ_t^k, for the Agincourt period-4 households.

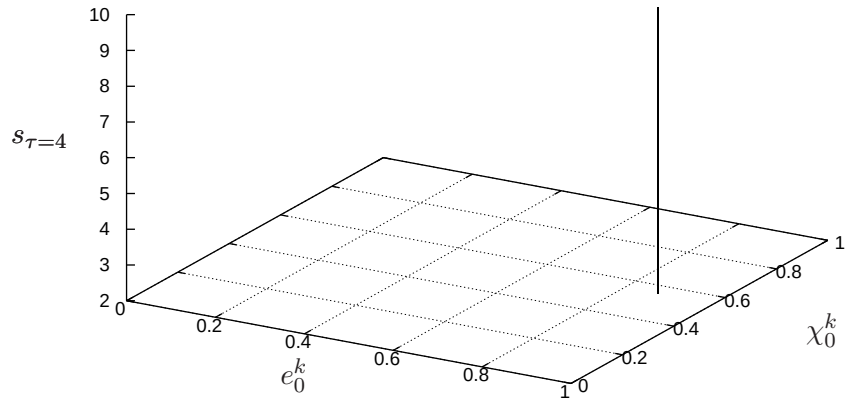

Figure 5.18: Initial distribution of the number of Agincourt period-4 households over the states $(e_t^k, \chi_t^k) \in \Gamma_3$.

Table 5.1: Distribution of the number of periodic orbits in Agincourt population

Period (τ)	Number of households	Percentage %
1	4	0.15
2	3	0.11
3	4	0.15
4	12	0.45
Total	23	0.86

The above techniques allow us to automatically separate any population P in two parts P_1 and P_2. We refer to P_1 as part of the population with a periodic behaviour and P_2 with a non-periodic behaviour. From Table 5.1, we find that less than 1% of the Agincourt population is periodic with respect to our variables of interest, which in turn defines P_1^{Ag}. Thus, we can assume that the rest (about 99%) of the Agincourt population with a non periodic behaviour, which we refer as P_2^{Ag} is stochastic. The behaviour of P_2^{Ag} can then be simulated.

As noted, periodic behaviour is not above the expected periodic orbits arising from random sampling.

5.5 Simulation of the Agincourt population

The main purpose of this section is to simulate the dynamics of the present Agincourt population. The discussion will only include the population-level analysis. The techniques presented in this section can be applied to simulate the dynamics of any population, given longitudinal data of the population. We follow the method of Section 5.3.

First of all note that in order to simulate the Agincourt population dynamics, it is important to determine the population parameters that will be used in our simulations. Thus, as stated in Section 4.3, recall that the number of questions is $n = 3$, the average observation time for the Agincourt population is calculated using the definition (2.4), we find $\overline{l} = 7.115$. Note that for the present simulations $\overline{l}$ must be an integer, thus it is convenient for the present simulations to use $\overline{l} = 7$ as the observation time for each household. The total number of Agincourt households in the present study population is $s = 2669$, we will use s as the total number of households in our

simulations then model regular and random stochastic population P_1 and P_2.

Second, note that the results of Section 5.4 show that the Agincourt population is dominated (more than 99%) by a non-periodic behaviour. The Agincourt population is divided into two parts. Here P_1^{Ag} consists of 23 households (about 0.86% of the total population) and P_2^{Ag} contains 2664 households which represent about 99.14% of the population. The periodic population P_1^{Ag} may be neglected for the simulations.

Note also that there are various levels of simulations. As stated above, in this thesis we focus on the outline of Section 5.3 that we divide in two main parts including stochastic level of simulations and scenarios planning.

5.5.1 Simulation of Agincourt results

Before we start, it is important to note for the Agincourt population P_2^{Ag}, the distributions of Figure 2.5 and Figure 2.6 do not change significantly. Thus, they may be used in the following simulations.

The results of Table 4.3 and Figure 4.1 are used to determine the approximate density matrix $D_3^{Ag(Approx.)}$ which is obtained with respect to Definition 4.1. The approximate Agincourt dynamics constitute reduced information that we will use for the simulations. The simulations should agree with D_3^{Ag}.

We use the definition (5.9) to determine the distribution of Agincourt households over states of Γ_3 for each observation year $t = 1998, 1999, \ldots, 2006$. With reference to (5.9), we obtain approximate fluxes $\underline{f}_t^{Approx}$ by deleting fluxes not in the states of Table 4.3 and Figure 4.1. The approximate and Agincourt distributions of households over dominant states are presented in Table 5.3.

We compare the results of Table 5.3, specially for the observation year 2006 to the distribution presented in Figure 4.6 and Figure 4.7. It is clear that they are in agreement. If we time average the Agincourt fluxes

$$\underline{f}^{average} = \frac{1}{9} \sum_{t=1998}^{2006} \underline{f}_t^{Approx} , \tag{5.16}$$

where $\underline{f}^{average}$ is given in Table 5.2.

We repeat the simulations and present in Table 5.4. We again note agreement. This final result

Table 5.2: Agincourt average flux vector.

i	$\delta_i^{Average}$
21	0
22	-3
23	63
24	-66
29	4
30	0
31	7
32	-5

gives the very simple and elegant model

$$\underline{m}_{t+1} = \underline{m}_t + \underline{f}^{average} \tag{5.17}$$

for predicting the future. In contrast to the map (5.2), this equation gives the number of households that flow into a given state averaged over time. This is now appropriate for demographic modelling.

Another method of simulation is as follows. From the Agincourt density matrix D_3^{Ag} we extract $D_3^{Approx.}$ which is defined only for the dominant transitions. For each row i of $D_3^{Approx.}$ we use the probability

$$p_{ij} = \frac{d_{ij}}{\sum_{i,j} d_{ij}} \ , \quad i,j = 23,24,29,30,31,32 \tag{5.18}$$

as a transition probability from state i to state j. We then start m_i households in state i and according to the probability p_{ij}, sample their transitions to dominant state j.

Note that if every social unit is observed each observation time, we will have the same number of social units

$$s_t = \sum_{i=1}^{d} m_{ti} \quad \forall t \tag{5.19}$$

We give the results in Table 5.5 for the final transition $2006 - 2007$. We note the reasonable agreement in the percentage transitions of the simulations. We do not ask for close agreement with Agincourt densities. Thus, in the simulations we use $\overline{l} = 7$, in reality households come and go and may have not completed to 2007.

Table 5.3: Comparison of approximate and real Agincourt number of households over dominant states.

i	$m_{1998i}^{Ag} = m_{1998i}^{Approx.}$
21	23
22	48
23	636
24	1826
29	0
30	32
31	2
32	58

i	m_{1999i}^{Ag}	$m_{1999i}^{Approx.}$
21	6	23
22	24	26
23	794	798
24	1672	1679
29	16	0
30	27	16
31	23	19
32	70	64

i	m_{2000i}^{Ag}	$m_{2000i}^{Approx.}$
21	2	23
22	10	13
23	1061	1061
24	1423	1434
29	19	0
30	15	-9
31	46	59
32	60	44

i	m_{2001i}^{Ag}	$m_{2001i}^{Approx.}$
21	2	23
22	5	13
23	1131	1135
24	1344	1353
29	32	17
30	12	-20
31	53	66
32	60	38

i	m_{2002i}^{Ag}	$m_{2002i}^{Approx.}$
21	3	23
22	2	13
23	1187	1188
24	1294	1294
29	37	23
30	17	-10
31	50	62
32	54	32

i	m_{2003i}^{Ag}	$m_{2003i}^{Approx.}$
21	4	23
22	2	13
23	1263	1263
24	1209	1208
29	50	36
30	15	-11
31	55	65
32	48	28

i	m_{2004i}^{Ag}	$m_{2004i}^{Approx.}$
21	3	23
22	3	13
23	1293	1292
24	1166	1157
29	57	40
30	23	8
31	62	71
32	40	21

i	m_{2005i}^{Ag}	$m_{2005i}^{Approx.}$
21	1	23
22	1	13
23	1319	1317
24	1149	1130
29	55	38
30	38	28
31	50	58
32	36	18

i	m_{2006i}^{Ag}	$m_{2006i}^{Approx.}$
21	3	23
22	5	13
23	1281	1280
24	1185	1165
29	56	40
30	35	28
31	62	79
32	22	-3

Table 5.4: Comparison of average and real Agincourt number of households over dominant states.

i	$m_{1998i}^{Ag} = m_{1998i}^{Average}$
21	23
22	48
23	636
24	1826
29	0
30	32
31	2
32	58

i	m_{1999i}^{Ag}	$m_{1999i}^{Average}$
21	6	23
22	24	44
23	794	699
24	1672	1759
29	16	4
30	27	32
31	23	9
32	70	53

i	m_{2000i}^{Ag}	$m_{2000i}^{Average}$
21	2	23
22	10	40
23	1061	762
24	1423	1693
29	19	8
30	15	32
31	46	16
32	60	48

i	m_{2001i}^{Ag}	$m_{2001i}^{Average}$
21	2	23
22	5	36
23	1131	825
24	1344	1626
29	32	13
30	12	33
31	53	24
32	60	43

i	m_{2002i}^{Ag}	$m_{2002i}^{Average}$
21	3	23
22	2	32
23	1187	888
24	1294	1560
29	37	17
30	17	33
31	50	31
32	54	38

i	m_{2003i}^{Ag}	$m_{2003i}^{Average}$
21	4	23
22	2	28
23	1263	952
24	1209	1493
29	50	21
30	15	33
31	55	39
32	48	33

i	m_{2004i}^{Ag}	$m_{2004i}^{Average}$
21	3	23
22	3	24
23	1293	1015
24	1166	1427
29	57	25
30	23	34
31	62	46
32	40	28

i	m_{2005i}^{Ag}	$m_{2005i}^{Average}$
21	1	23
22	1	20
23	1319	1078
24	1149	1360
29	55	30
30	38	34
31	50	54
32	36	23

i	m_{2006i}^{Ag}	$m_{2006i}^{Average}$
21	3	23
22	5	16
23	1281	1141
24	1185	1294
29	56	34
30	35	34
31	62	61
32	22	18

Table 5.5: Comparison of simulated and Agincourt density matrices, for the final transition year.

i	j	$d^{Ag}_{2006,ij}$	$d^{Ag}_{2006,ij}(\%)$	d^{Sim}_{5ij}	$d^{Sim}_{5ij}(\%)$
23	24	301	23.589342	537	20.119895
23	23	300	23.510972	560	20.981641
24	24	265	20.768025	616	23.079805
24	23	234	18.338558	504	18.883477
30	30	19	1.489028	48	1.798426
31	23	18	1.410658	21	0.786812
29	30	17	1.332288	50	1.873361
24	29	16	1.253918	20	0.749344
31	32	16	1.253918	24	0.899213

In Table 5.6, the simulated fluxes into dominate states are compared and we note reasonable agreement, after accounting for loss of households in 2007.

Comparing the models 5.17 and 5.18, we find 5.17 to be preferable because of its simplicity and because $\overline{l} = 7$ is short. We will use 5.17 to predict for the final transition $2006 - 2007$.

5.5.2 Predicted Agincourt population, $2007 - 2015$

In Table 5.7 we give predicted conditions over the dominant states for the whole population, over the period $2007 - 2015$. We find average increase of 52 defaulting households per annum.

In Table 5.8, we begin 4000 households, uniformly distributed over dominant states. Recall that 4000 new households have been added to the Agincourt data set. By 2015 we find emergence of states 23 and 24. The rate of emergence in state 23 of defaulting households is approximately $\underline{f}^{average}_{23} = 63$ per annum if our sample of 2669 households. The outflow from state 24 is $\underline{f}^{average}_{24} = -66$ per annum. If these averages model a steady state at Agincourt, then the new households are in danger of experiencing growing educational default at a proportional rate of approximatively 99 households per annum. Because the 4000 new households of Agincourt do not have a well-defined initial state, the rate is of use in making decisions. Thus the new households suffer a possible

Table 5.6: Comparison of simulated and Agincourt flux vectors for the final transition year

i	δ^{Ag}_{2006i}	δ^{Sim}_{5i}
21	2	0
22	-1	0
23	-72	-45
24	63	40
29	-1	-5
30	7	-6
31	-10	10
32	11	6

disservice by moving to Agincourt. Increased educational resources are necessary to accommodate defaulting children. The graphs of Figure 4.13 and this rate of default is a server criticism of the Agincourt educational system.

5.5.3 Scenario planning

The results of the previous chapter show that the Agincourt population is dominated by migration of biological mothers. In particular we find that out-migration of these mothers was the major cause of children's educational default. Thus, an important scenario we would imagine is to simulate Agincourt dynamics with no out-migration of biological mother. This means the answer value of question q_0 must be changed to 1 for all the observation years. We accordingly generate a new data set where $a_0 = 1$ for all time and then rerun orbit theory over the period $1998 - 2007$.

The new Agincourt dominant transitions are summarised in Table 5.9 and Table 5.10. As expected, the Agincourt population is now fitter than before. Table 5.9 shows the population shift to fully fit states $24, 32, 48$. Table 5.10 shows emergence of the fully fit states, for originally defaulting households. Note that unfavourable transition $24 \rightarrow 31$ have increased; this is owing to adult death and merely states that where households were in states 23 and experienced adult death, they are now in state 24.

Under this scenario, the average fluxes on the dominant sub-space are presented in Table 5.11.

Table 5.7: Distribution of simulated Agincourt population

i	m_{6i}^{Sim}	m_{7i}^{Sim}	m_{8i}^{Sim}	m_{9i}^{Sim}	m_{10i}^{Sim}	m_{11i}^{Sim}
21	5	5	5	5	5	5
22	3	0	-4	-8	-12	-16
23	1196	1259	1322	1385	1448	1512
24	1225	1158	1092	1025	959	892
29	55	59	63	68	72	76
30	57	57	57	58	58	58
31	54	61	68	76	83	91
32	60	55	50	45	40	35

Table 5.8: Distribution of simulated Agincourt population with new 4000 households

i	m_{6i}^{Sim}	m_{7i}^{Sim}	m_{8i}^{Sim}	m_{9i}^{Sim}	m_{10i}^{Sim}	m_{11i}^{Sim}
21	500	500	500	500	500	500
22	500	496	492	488	484	480
23	500	563	626	689	752	816
24	500	433	367	300	234	167
29	500	504	508	513	517	521
30	500	500	500	500	501	501
31	500	507	514	522	529	537
32	500	495	490	485	480	475

Table 5.9: Sorted number of transitions from the density matrix $D_3^{(scenario)}$.

state i	**state** j	d_{ij}
24	24	11615
32	32	1783
48	48	1659
24	31	465
31	32	436

Table 5.10: Sorted number of transitions d_{ij}^{Ag} for previously defaulting households (4.34).

state i	**state** j	$d_{ij}^{scenario}$	d_{ij}^{Ag}
24	24	6460	2256
32	32	1047	111
48	48	925	9
24	31	243	90
31	32	236	56

Table 5.11: Fluxes of dominant transitions for the scenario

state i	$f_i^{average,scenario}$
21	0
22	-14
23	0
24	-104
29	0
30	0
31	23
32	94

Table 5.12: Scenario of Agincourt defaulting households, 2007 – 2015.

i	$m_{2007i}^{Ag,\overline{Ed}}$	$m_{2008i}^{Ag,\overline{Ed}}$	$m_{2009i}^{Ag,\overline{Ed}}$	$m_{2010i}^{Ag,\overline{Ed}}$	$m_{2011i}^{Ag,\overline{Ed}}$	$m_{2012i}^{Ag,\overline{Ed}}$	$m_{2013i}^{Ag,\overline{Ed}}$	$m_{2014i}^{Ag,\overline{Ed}}$	$m_{2015i}^{Ag,\overline{Ed}}$
21	0	0	0	0	0	0	0	0	0
22	0	-14	-28	-42	-56	-70	-84	-98	-112
23	0	0	0	0	0	0	0	0	0
24	1706	1601	1496	1391	1287	1182	1077	973	868
29	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0
31	108	131	154	178	201	224	248	271	294
32	558	652	746	840	934	1029	1123	1217	1311

In Table 5.12, we predict ahead using

$$\underline{m}_{t+1} = \underline{m}_{t+1} + \underline{f}^{average,scenario}. \tag{5.20}$$

States 24 and 32 are both fully fit. We see emergence of favourable state 32. The total of favourable states declines very slowly at about 9 fully fit households per annum. Adult death remains in households and we see an increase in educationally defaulting households with adult death (state 31) at a rate of about 21 households per annum. We note that this data is taken only in the dominant sub-space and that the fully fit state 48 could become important.

In Table 5.13, we again suppose that the 4000 new households are uniformly distributed over dominant states. In addition to that, we suppose that half of the new population is educationally defaulting with respect to $l_{failure} = 4$. Behaviour is very similar to that of Table 5.12. We clearly see that a policy of keeping biological mothers at home will be a service to this community.

5.6 Conclusion

There were two main objectives in this chapter. We first developed the model (5.17) to predict forward in time from 2007.

We were careful to ensure that the number of periodic orbits did not exceed levels that could

Table 5.13: Predicted number of Agincourt with additional 2500 defaulting households, 2007 − 2015.

i	$m_{2007i}^{Ag,\overline{Ed}}$	$m_{2008i}^{Ag,\overline{Ed}}$	$m_{2009i}^{Ag,\overline{Ed}}$	$m_{2010i}^{Ag,\overline{Ed}}$	$m_{2011i}^{Ag,\overline{Ed}}$	$m_{2012i}^{Ag,\overline{Ed}}$	$m_{2013i}^{Ag,\overline{Ed}}$	$m_{2014i}^{Ag,\overline{Ed}}$	$m_{2015i}^{Ag,\overline{Ed}}$
21	250	250	250	250	250	250	250	250	250
22	250	236	222	208	194	180	166	152	138
23	250	250	250	250	250	250	250	250	250
24	250	145	40	-63	-168	-273	-377	-482	-587
29	250	250	250	250	250	250	250	250	250
30	250	250	250	250	250	250	250	250	250
31	250	273	296	319	343	366	389	413	436
32	250	344	438	532	626	721	815	909	1003

be expected from random sampling.

For the Agincourt data, we find that the the non-periodic sub-population was the significant (more than 99%) sub-population. We therefore simulated a stochastic process of sampling households data according to Agincourt statistics.

The second objective was to simulate the Agincourt population dynamics and compare the results with the real data. Again, as in any simulations the choice of the initial conditions is a central point that needs to be discussed. We decided to simulate from the Agincourt initial conditions using the average observation time $\overline{l} = 7.115 \approx 7$ and the total number of households $s = 2669$. The simulation results are in good agreement with the Agincourt real data.

We simulated ahead from 2007 Agincourt data to 2015 and found decline in educational progression of some 52 households per annum. At this point we simulated effects of interventions to adopt policy that keeps biological mothers at home. This reduced the rate of educational default to 9 households per annum.

Chapter 6

Conclusion

The aim of this thesis was to develop a new mathematical "orbit theory" for analysing longitudinal data. This theory specifically concerns the social sciences. Thus the sense of direction of motion of a social unit comes from change in its social fitness.

It is important to this work that our space of states Γ_3 or S_3 specifies the state of a social unit and is not a state at the population level as is achieved by starting analysis with the deterministic models reviewed in Section 1.2.2. However clustering induces demographic properties that may be studied by the statistical techniques of Section 1.3 or by a deterministic study of the detailed state space defined by the clusters.

Our method achieves analysis while preserving full complexity under our purpose. By this we here mean that $n = 3$ questions are hypothesised to be relevant to purpose. The value of answer to each question becomes an element of a sequences or a decimal digit with social variable identified with the position of that question. In principle we may have n independent social variables and the complete state of the household is captured by a single real number (or sequence). Contrast this with the measured position of a physical particle where additional digits merely improve the accuracy of the single physical variable of position. The orbits that we generate preserve this full information. In Chapter Four we have seen automatic identification of sub-populations in Γ_3. This too, retains full information of the longitudinal data relevant to purpose. Further, the time series of a single household records all transitions of the household within purpose.

6.1 Mathematical conclusion

We achieved the objectives of this thesis as follows.

- We have stated our purpose in (2.2)

  p_1 : **To investigate the effect of household change on child's progression in school** .

  We regard the form of p_1 as important in the contest of longitudinal data. We emphasise the word *effect*. Thus we suggest that in general, purpose is stated to extract causal relationship.

- Household change was characterized by questions as in (2.1).

- The education measure $l_{failure} = 4$ captures educational default of households.

- With our questionnaire of $n = 3$ questions, we define fitness and significance space Γ_3 or S_3 and code raw data. This may be generalized to n questions, but as discussed in Section 4.11, to seek cause from few effects seems to be advisable. Work must be done to survey permutations of questions if n is large.

- We define the maps $\psi : (e_t^k, \chi_t^k) \mapsto (e_{t+1}^k, \chi_{t+1}^k) \in \Gamma_3$ (3.47) and $\xi : ({}_wb_t, \theta_t) \mapsto ({}_wb_{t+1}, \theta_{t+1}) \in S_3$ (3.48). In these maps change in fitness value is taken directly from longitudinal data. This may be regular or stochastic. Significance is determined by reordering questions in an evolutionary sense. This is a deterministic process imposed on the data. Here we look at the social unit level (household level)

- The maps induce individual level orbits Ω_t^k, for household k in fitness-significance space Γ_3 or S_3.

- We determine the theoretical transition matrix T_3 (3.78) which captures all possible transitions under $n = 3$ binary-valued answers to questions.

- Agincourt transitions are extracted in (4.14), $T_3^{Ag} \subset T_3$. Dominant transitions are identified from the density matrix (4.21), D_3^{Ag}. Dominant fluxes $\underline{f}^{Ag}$ are identified in Table 4.5.

- Typical orbits are identified in Table 4.3 and Figure 4.1. They are revealed to dominantly be random oscillations between states 23 and 24 as biological mothers temporarily migrate. There are a few random excursions to states $29, 30, 31$ and 32 in the case of adult death.

- Our orbits in Γ_3 or S_3 clearly extract sub-populations as expected, by clustering data. This has facilitated our demographic analysis. The analysis processes full information of the longitudinal data.

- As in the previous item visualization of orbits facilitates analysis. Figures showing time-dependence of the sampled Agincourt data is given in Figure 4.11. In this figure, we see that the clustering (about states 23 and 24) is constant in time.

- Periodic orbits are extracted but their number does not exceed that expected from random sampling. It may be assumed that the dynamics hidden in Agincourt longitudinal data is stochastic.

- Educational default is identified with out-migration of biological mothers (state 23). We identified out-migration as a social force acting against educational progression or, as a possible cause of educational default under our question set (Table 4.8). We have noted that at the $l_{failure} = 4$ level this conclusion will tolerate a $\pm 10\%$ stochastic error in the data.

  We have noted above that this thesis develops a new method of analysis that best proceeds by detailed analysis of reduced numbers of questions identified by clustering. A thorough demographic study would search through permutations of questions regarding household change and search for strong associations with educational default.

- Projection: the Agincourt demographic dynamics was approximated by the map $\underline{m}_{t+1} = \underline{m}_{t+1} + \underline{f}^{Average}$ (5.17) with $\underline{m}_0 = \underline{m}_{1998}$. We find good agreement between Agincourt and approximated dynamics.

- For Agincourt we predict for the years $2007 - 2015$ that out-migration of biological mothers will cause increased educational default at the rate of 52 households per annum (Table 5.7).

- Scenarios: suppose we pay biological mothers to stay at home (BM: $a_t : 0 \to 1, \forall\ t$). We have run orbit theory for this new data. This scenario predicts that the number of default households will improve from 52 to 9 (Table 5.12).

- The *reduced* model (5.17) may be compared with those of the Introduction. Comparing (5.17) with (1.9) we see that the Agincourt community is better modelled by additive maps as opposed to multiplicative maps based on rate of change.

- Yet, at the level of individual orbits, the Agincourt transition matrix, T_3^{Ag} reveals the detailed dynamics of the population extracted directly from the data. We have used only superficial information of determinant and trace. Define for the k'th household the sequence $i_0 i_1 i_2 \cdots i_j$ of its visits to points in S_n with index i. Then the map σ_T defined by

$$i_{j+1} = \sigma_T(i_j) \tag{6.1}$$

tells us what the next state will be if we know i_j. This map is known as a sub-shift of finite type [119, 120, 121, 122] and has deep mathematics. It is directly linked to the longitudinal data. We believe this is a significant step forward in mathematical modelling [141] of processes in the social sciences. It is important to note that (6.1) gives the dynamics of all possible orbits under purpose but it does not give the number of households on any one orbit. We have chosen in this thesis to present only the high level demographic results in order to explore the usefulness of orbit theory. It is for this reason that we have constructed the density matrices and flux vectors in order to achieve (5.17) and this has required use of the detailed data. The mathematical properties of (6.1) will be explored elsewhere. However we note that T_3^{Ag} is unique to the Agincourt data and has extracted only a subset of all possible transitions as given in (3.78). All this is completely inaccessible to the deterministic models of the Introduction. It does suggest new statistical approaches as in Section 1.3. We add that because we have clearly defined states and the state space, it becomes possible to use survival analysis on each state. Statistic of the fluxes will be of great interest as well.

- Our reduced models have involved *no hypotheses* other than choice of questions and a decision on the 100−transitions cut-off. Concerning favourable/unfavourable coding. We have supposed that out-migration of biological mothers is unfavourable to educational progression and we note that our conclusion is consistent with this assumption. Minor household head is not significant, so its coding is irrelevant. Concerning adult death, we see from Figure 4.1 that there is balanced rate of death from both unfavourable state 23 and favourable state 24 to the death state 29. Because these are by far the dominant states we do not have significant evidence that adult death causes educational default and the coding is acceptable.

- Concerning existing deterministic theories as mentioned in the Introduction, we note that (5.17) and (6.1) contain no parameters. Neither of these equations is ad hoc but arises naturally from longitudinal data. Equation (5.17) can be made more precise by time-modelling

$\underline{f}^{average}$. Thus, it will be noted from Figure 4.1 that the 6 dominant states can be reached by a single orbit that might have repeated patterns (e.g. $23 \leftrightarrow 24$) and by orbits that connect all points in an endless non-repeating pattern. The map is indeed chaotic [38] and as was suggested in the discussion of the map (1.1), it is therefore able to represent the periodic and stochastic orbits that we have discovered in the data.

- We have not compared our method with statistical methods. The automatic clustering of orbit theory and our confirmation that the dynamics is stochastic, suggests statistical correlation among clustered variables [142, 143]. In this way orbit theory can have practical use as preconditioning of data for statistical analysis. We have selected our appropriate statistical method in Section 4.11. A detailed comparison of statistical (event history analysis) and orbit theoretical approaches is under way elsewhere.

6.2 Discussion

Our data set was restricted to $n = 3$ questions only. It is possible for example that socio-economic change such as income effects the ability of a child to get to school and to have good resources for success. Our strategy in the case of many questions would be to investigate automatic clustering in Γ_n, this is a space of real numbers and in principle, can handle many decimal places corresponding to social variables. Where clustering is identified, we would eliminate variables not associated with the cluster and thereby reduce the phase space as we have done above in Figure 4.1. Each cluster can then be analysed for cause and effect [142, 143]; we note in this case that we would re-frame purpose for the local analysis. The interconnection between clusters might sometimes be significant and in this case we would build a new phase space by simply renumbering states consecutively through associated clusters.

It is important in the design of Q_t that the coding to binary is unique if the notion of favourable/unfavourable holds. In this case it may be hoped that different sociologists would arrive at the same Q_t, for the same given purpose. If the number of questions differs somewhat, so that there is redundancy in one of say two questionnaires, the simplest questionnaire should be chosen. In turn, the best questionnaire will be that with the least number of questions that, by consensus, is sufficient to address purpose. If there is such general agreement, then it is reasonable to say that Q_t *stabilizes* and the purpose is *understood*. If this understanding suggests practical

actions (to favourably alter the direction of orbits) then we can say that it is *useful*.

To ensure that the minimal necessary question set is achieved, each question can be systematically deleted from Q_t to judge the effect on orbits (not necessary in our case). Then the number of accessible states in the state space Γ_n is reduced. If the deleted question has little or no effect on the common states in Γ_n that question may be deleted. For the Agincourt analysis, we have showed that change in question q_1 was very small. As a result, the dynamics of the whole Agincourt population was reduced to a Γ_2. Conversely, if it has noticeable effect it must be retained. If new questions are to be added and their frequency of change is unknown, they should be placed at insignificant digits (on the right), so that the evolutionary dynamics may automatically bring them to significant digits, with significant effect on the orbits. If there is no effect on the common states of the old and new orbits, then the new question may be deleted.

It may be that the purpose suggests additional elements to the value set, for some questions, for example the answer value $a_i \in \{0, 1, \varnothing\}$ where $\varnothing =$ not relevant. Additional values must have unambiguous social content (be coded in the same way by all sociologists). Consider q : **did you breast feed your child**? with answer value set $\{yes, no, father\}$ with a numerical coding $\{0, 1, 2\}$. In this case it is equivalent to work in base 3 numbers rather than binary numbers. In Γ_n irrelevant evolutionary orbits (i.e., those of the father) will go to a y-value in the interval $[0.2, 1.0)$, that is, where the ternary numerical coding begins with the digit 2 after the decimal point. In this case, the flow suggests a cost-saving restriction of the survey to female respondents. Note that old data can be unambiguously translated for this modified questionnaire. No new mathematical phenomena are implied in this case and in this thesis the binary case only is considered.

Questions may offer no choice of reply as in q : **tax number**? Single-valued questions are unambiguous and reveal nothing dynamical. The purpose of single-valued questions is surely as a convenient initial identification of sub-populations, relevant to purpose, by the sociologist.

'Open questions' are not directly acceptable but they may be useful. Thus q: why did you steal? may elicit many responses. However, every response that is a reason for stealing can immediately be viewed as a reply to a satisfactory question - this builds a set of good questions from each open question. Note that a questionnaire with open questions can then be translated to a satisfactory form. Other typical questions choose from many possible values. An obvious example is the value of the question q : **what is your income within ten thousand Rand**? It is clear such questions can be translated to a set of questions q_i: **is your income in the i'th income bracket**? Translation

of all responses to open questions might be very laborious.

6.3 Future work

The mathematics of the dynamical system as defined in (6.1), $i_{j+1} = \sigma_T(i_j)$ is very deep. We note that this is technically a sub-shift of finite type [119, 120, 121, 122]. As stated, it is chaotic and has advanced properties such as (Kolmogorov) entropy, that might or might not have interpretation in demography. It will be of interest to continue these mathematical investigations.

Of great importance, is an explicit comparison of statistical and orbit theoretical methods. It is there that the strengths and weaknesses of the methods can be made clear and offer guidance to future researchers.

Appendix A

Agincourt data description

A.1 Description of the variables

Table A.1: Description of the variables

Variable	Description
Year	This variable represents the year for the cross-sectional analysis for example 2001 refers to the cross-section on December 31st 2001
Household ID	This variable represents the anonymised unique identifier for the household
Anon ID	This variable represents the anonymised unique identifier for all children included in the analysis i.e. all children aged between 7 and 16 years of age members of the household between December 31st 1992 and December 31st 2008
MotherResMonths	This variable represents the number of months that the childs mother was resident as recorded as a Residence Status observation in the census round associated with the cross-section date (see Table 2.1). **Database field = ResidenceStatus.ResMonths** **Possible values** n- number of months mother was resident NULL No data recorded (missing data)
Education Recode	This variable represents the highest level of education obtained for the child derived from the Education Status observation in the census round associated with the cross-section date (as described in Table 2.1). **Database field = EducationStatus.Education** This is recoded from the raw data to represent the number of years of education for each child. This data was recorded for all household members in the following census rounds. Round 1(1992), Round 4(1997), Round 8(2002), Round 12(2006). In addition a status observation is recorded for all individuals whenever a new household is established or a new individual migrates into a household. See Table 2.1 for mapping between Education status codes and number of years of education completed **Values** n Number of years of education completed NULL No education status information recorded
HHHead IsMinor	On the cross-section date, the current household head is identified using the household head relation and membership start and end date fields in the membership table. Where two possible household heads are identified the oldest is selected. Where no household head is identified, the field is given a NULL value. The age of the household head on the cross-section date is calculated. **Possible values** "Yes" Household head aged < 18 "No" Household head aged ≥ 18 "NULL" No household head identified.
AdultDeath	Possible value = "Yes" if any deaths have occurred for individuals who were members of the household during the year 1st Jan December 31st . Note that in unusual circumstances an individual who dies may have been a member of the household during the year terminate their household membership prior to December 31st. These are included as Adult Deaths for that household and given the value "Yes".
Household	This variable represents the unique identifier for the household for all household included in the extracted data set
Household established	The variable represents the date when the household was established and it is in the format = $(mm/dd/yyyy)$
Household dissolved	This variable represents the date when the household was dissolved and it is in the format = $(mm/dd/yyyy)$

Bibliography

[1] W. Roseberry. A Unit for Social Analysis. (Book Reviews: Households). *Science*, 228:1081–1082, May 1985.

[2] Bellomo N., Bianca C. and Delitala M. Complexity Analysis and Mathematical Tools towards the Modelling of Living Systems. *Physic of Life Reviews*, 6:144–175, 2009.

[3] Ihar A. M. and Viatcheslav V. B. Mathematical Representations of the Dynamics of Social System: I. General Description. *Chaos, Solitons and Fractals*, 6:195–206, 2005.

[4] Wolfgang Weidlich. Physics and social science - the approach of synergetics. Technical Report 01, Institut fur Theoretische Physik, Universitat Stuttgart, Germany, 1991. Physics Reports (Review Section of Physics Letters) North - Holland.

[5] Peter S. Albin and Hans W. Gottinger. Structure and Complexity in Economic and Social Systems. *Mathematical Social Sciences*, 5:253–268, 1983.

[6] Kazemier B. H. and Vuysje D., editors. *The Concept and the Role of the Model in Mathematics, Natural and Social Sciences*. Dordrecht: Reidel, 1961.

[7] Saaty Thomas L. and Weyl F. J. *The Spirit and Uses of the Mathematical Sciences*. New York: McGraw-Hill, 1969.

[8] Coffman C. V. and Fix G. J., editors. *Constructive Approaches to Mathematical Models*. New York: Academic Press, 1979.

[9] Weidlich W. Sociodynamics: A systematic approach to mathematical modelling in the social sciences. *Amsterdan: Harwood Academic Publishers*, 2000.

[10] Paul F. Lazarsfeld, editor. *Mathematical Thinking in the Social Sciences*. The Free Press, A Corporation, 1954.

[11] George J. KLIR. The Role of Reconstructability Analysis in Social Science Research. *Mathematical Social Sciences*, 12:205–225, 1986.

[12] Peter D. Killworth and H. Russell Bernard. A Model of Human Group Dynamics. *Social Science Research*, 5:173–224, 1976.

[13] Mario di Bacco and Gianluca Poggi. Some remarks on statistical indices. *Metron - International Journal of Statistics*, pages 69–81, 1999. Dipartimento di Statistica, Probabilit e Statistiche Applicate - University of Rome.

[14] Robert L. Devaney. *An Introduction to Chaotic Dynamical Systems*. Allan M. Wylde, Second edition, 1989. Addison-Wesley Publishing Company.

[15] Cristopher Moore. Generalized shifts: Unpredictability and Undecidability in Dynamical Systems. *Nonlinearity 4*, January 1991.

[16] George Osipenko and Stephen Campbell. Applied Symbolic Dynamics: Attractors and Filtrations. *Discrete and Continuous Dynamical Systems*, 5(1):43–60, January 1999.

[17] Bender Edward A. *An Introduction to Mathematical Modeling*. New York: Wiley, 1978.

[18] Murray Francis J. *Applied Mathematics: An Intellectual Approach*. New York: Plenum Press, 1978.

[19] Michael Olinick. Mathematical Models in the Social and Life Sciences: A Selected Bibliography. *Mathemntieal Modelling*, 2:237–258, 1981.

[20] Helbing D. *Quantitative Sociodyanmics-Stochastic and Models of Social Interaction Processes*. Dordrecht: Kluwer Academic Publishers, 1995.

[21] K. Lewin. *Principles of Topological Psychology*. McGraw-Hill, 1936.

[22] Lewin K. Field theory in social science. *Harper and Brothers, New York*, 1951.

[23] Barder B. Neofunctionalism and the theory of the social system. *In: The Dynamics of Social Systems (P. Colomy Ed.) - London: Sage*, pages 36–55, 1992.

[24] Dirk Helbing and Peter Molnar. Social Force Model for Pedestian Dynamics. *Physical Review E*, 51(5):4282–4286, May 1995. The American Physical Society.

[25] David W. Pearson and Mark McCartney. Dynamics of social networks: A deterministic approach. *Int. J. Appl. Math. Comput. Sci.*, 12(4):545–551, 2002. amcs.

[26] Gilbert N. and Doran J., editors. *Simulating Societies: The Computer Simulation of Social Phenomena*. London:UCL Press, 1994.

[27] Sniders T. and Duijn M. Simulation for statistical inference in dynamic network models. *Berlin: Springer*, pages 493–512, 1997.

[28] Goran Peskir. On the diffusion coefficient: The einstein relation and beyond, stoch. models. Research Report No. 424 19, Dept. Theoret. Statist. Aarhus, 2001.

[29] A. Einstein. *Ü*ber die von der molekularkinetischen Theorie der Warme geforderte Bewegung von in ruhenden Flussigkeiten suspendierten Teilchen. Ann. Phys. 17 (549-560). English translation "On the motion of small particles suspended in liquids at rest required by the molecular-kinetic theory of heat" in the book 'Einstein's Miraculous Year'. *Princeton Univ. Press*, 1998 (85-98).

[30] Joseph Klafter and Igor Msokolov. Anomalous diffusion spreads its wings. *Physics World*, pages 1–4, August 2005.

[31] Dirk Helbing. A Mathematical Model for the Behavior of Individuals in a Social Field. *Journal of Mathematical Sociology 19(3)*, pages 189–219, 1994.

[32] Nathan Keyfitz and Hal Caswell. *Applied Mathematical Demography*. New York, NY : Springer, 3rd edition, 2005.

[33] May Robert McCredie. *Stability and Complexity in Model Ecosystems*. Princeton University Press, Princeton, NJ, 1974.

[34] Rahim Moineddin et al. A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, 2007.

[35] Nigel Meade. A modified Logistic Model Applied to Human Populations. *Journal of the Royal Statistical Society.*, 151(3), 1988.

[36] Stuart L. Pimn. The Complexity and Stability of ecosystems. *Nature*, 307:321–326, 1984.

[37] May Robert McCredie. Will a large complex system be stable. *Nature*, 238:413–414, 1972.

[38] Barreto Ernest and So Paul. Mechanisms for the development of unstable dimension variability and the breakdown of shadowing in coupled chaotic systems. *Phys. Rev. Lett.*, 85(12):2490–2493, Sep 2000.

[39] Andrew D. Taylor. Deterministic Stability Analysis Can Predict the Dynamics of Some Stochastic Population Models. *The Journal of Animal Ecology*, 61(2), 1992.

[40] Morrison G. and Barbosa P. Spatial heterogeneity, population 'regulation' and local extinction in simulated host-parasitoid interactions. *Oecologia*, 73, 1987.

[41] May M. Robert. Host-parasitoid systems in patchy environments: a phenomenological model. *Journal of Animal Ecology*, 47, 1978.

[42] Pierson David and Moss Frank. Detecting periodic unstable points in noisy chaotic and limit cycle attractors with applications to biology. *Phys. Rev. Lett.*, 75(11):2124–2127, Sep 1995.

[43] So P, Francis JT, Netoff TI, Gluckman BJ and Schiff SJ. Periodic orbits: a new language for neuronal dynamics. *Biophys J.*, 74(6):2776–85, Jun 1998.

[44] Pawelzik K. and Schuster H. G. Unstable periodic orbits and prediction. *Phys. Rev. A*, 43(4):1808–1812, Feb 1991.

[45] So Paul and Ott Edward. Controlling chaos using time delay coordinates via stabilization of periodic orbits. *Phys. Rev. E*, 51(4):2955–2962, Apr 1995.

[46] Michael W. Spicer. Determinism, Social Science, and Public Administration, Lessons From Isaiah Berlin. *American Review of Public Administration*, 35(3):256–269, September 2005.

[47] So Paul et al. Detecting unstable periodic orbits in chaotic experimental data. *Phys. Rev. Lett.*, 76(25):4705–4708, Jun 1996.

[48] Joseph Klafter and Igor Msokolov. Anomalous diffusion spreads its wings. *Physics World*, pages 1–4, August 2005.

[49] Bianca C. On the Modelling of Space Dynamics in the Kinetic Theory for Active Particles. *Mathematical and Computer Modelling*, 51:72–83, 2009.

[50] J. S. Coleman. *Introduction to Mathematical Sociology.* The Free Press of Glencoe, New York, 1964.

[51] A. Diekmann. The log-logistic distribution as a model for social diffusion processes. *Journal of Scientific and Industrial Research*, 51, 1992.

[52] M. Granovetter and R. Soong. Threshold models of diffusion and collective behavior. *Journal of Mathematical Sociology*, 9, 1983.

[53] R.B. Jacobsen R.L. Hamblin and J.L.L. Miller. *A Mathematical Theory of Social Change*. New York, 1973.

[54] A.M. Kennedy. The adoption and diffusion of new industrial products: A literature review. *European Journal of Marketing*, 17(3), 1977.

[55] V. Mahajan and R. A. Peterson. *Models for Innovation Diffusion*. London, 1985.

[56] G.L. Truly S.L. Fulmer W.M. Schaffer, L.F. Olsen and D.J. Graser. *Periodic and chaotic dynamics in childhood infections*. Springer Berlin, 1988. From Chemical to Biological Organisation.

[57] A. Einstein. *Ü*ber die von der molekularkinetischen Theorie der Warme geforderte Bewegung von in ruhenden Flussigkeiten suspendierten Teilchen. Ann. Phys. 17 (549-560). English translation "On the motion of small particles suspended in liquids at rest required by the molecular-kinetic theory of heat" in the book 'Einstein's Miraculous Year'. *Princeton Univ. Press*, 1998 (85-98).

[58] Traulsen Arne, Röhl Torsten and Schuster Heinz Georg. Stochastic gain in population dynamics. *Phys. Rev. Lett.*, 93(2):028701, Jul 2004.

[59] John Impagliazzo. *Deterministic Aspects of Mathematical Demography*, volume 13. Springer-Verlag Berlin Heidelberg, 1985.

[60] Lotka A.J. Relation between birth rates and death rates. *Science, N.S.*, 26:21–22, 1907.

[61] Sharpe F.R. and Lotka A.J. A problem in age distribution. *Philosophical Magazine*, 21(6):435–438, 1911.

[62] Bernardelli H. Population waves. *Journal of the Burna Research Society*, 31:1–18, 1941.

[63] Lewis E.G. On the generation and growth of a population. *Sankhya*, 6:93–96, 1942.

[64] Leslie P.H. On the use of matrices in certain population mathematics. *Biometrika*, 33:183–212, 1945.

[65] Dobbernack W. and Tietz G. Twelfth international congress of actuaries. Technical report, 1940. vol. 4, pp. 233.

[66] Caswell H. *Matrix Population Models: Construction, Analysis, and Interpretation*, chapter 5. Sinauer, Sunderland, Massachusetts, USA, 2nd edition, 2001.

[67] Caswell H. Applications of markov chains in demography. *Markov Anniversary Meeting. Boson Books, Raleigh, North Carolina, USA*, pages 319–334, 2006.

[68] Lotka A.J. Theorie analytique des associations biologiques. part ii. analyse demographique avec application particuliere a lespece humaine. actualites scientifiques et industrielles, no. 780. hermann et cie, paris, france. (published in translation as: Analytical theory of biological populations, translated by d.p. smith and h. rossert. *Plenum Press, New York, 1998*, pages 319–334, 1939.

[69] Tollman SM, Herbst K. and Garenne M. The agincourt demographic and health study, phase 1. *Johannesburg: Health Systems Development Unit, Department of Community Health, University of the Witwatersrand*, 1995.

[70] Kathleen Kahn, Stephen M. Tollman, Mark A. Collinson, Samuel J. Clark, Rhian Twine, Benjamin D. Clark, Mildred Shabangu, Francesc Xavier Gomez-Olive, Obeb Mokoena and Michel L. Garenne. Research into health, population and social transitions in rural South Africa: Data and methods of the Agincourt Health and Demographic Surveillance System. *Scandinavian Journal of Public Health*, 35(Suppl 69):8–20, 2007.

[71] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. New York: John Wiley and Sons, 1980.

[72] J. F. Lawless. *Statistical Methods and Methods for Lifetime Data*. New York: John Wiley and Sons, 1982.

[73] D. R. Cox and D. Oakes. *Analysis of Survival Data*. London: Chapman and Hall, 1984.

[74] D. Collett. *Modeling Survival Data in Medical Research*. London: Chapman and Hall, 1994.

[75] Judith D. Singer and John B. Willet. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurence*. Oxford University, 2003.

[76] Paul D. Allison. *Survival Analysis Using SAS: A Practical Guide, Second Edition.* Cary, NC: SAS Institute Inc., 2010.

[77] Judith D. Singer and John B. Willett. It's about Time: Using Discrete-Time Survival Analysis to Study Duration and the Timing of Events. *Journal of Educational Statistics*, 18(2):155–195, 1993. American Educational Research Association and American Statistical Association.

[78] Paul D. Allison. Discrete-Time Methods for the Analysis of Event Histories. *American Sociological Association*, 13:61–98, 1982. Sociological Methodology.

[79] Samuel J. Clark et al. Returning home to die: Circular labour migration and mortality in South Africa. *Scandinavian Journal of Public Health*, 35(Suppl 69):35–44, 2007.

[80] Box G.E.P. and Cox D.R. An Analysis of Transformations. *Journal of the Royal Statistical Society*, 26:211–252., 1964.

[81] A. A. Noura and K. L. Q. Read. Proportional Hazards Changepoint Models in Survival Analysis. *Journal of Business and Economic Statistics*, 39(2):241–253, 1990. Blackwell Publishing for the Royal Statistical Society.

[82] Sean Collins. Prediction Techniques for Box-Cox Regression Models. *Journal of Business and Economic Statistics*, 9(3):267–277, July 2010. American Statistical Association.

[83] E.L. Kaplan and Meier P. Non-parametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53:457–481., 1958.

[84] James B. Robertson and V. R. R. Uppuluri. A Generalized Kaplan-Meier Estimator. *The Annals of Statistics*, 12(1):366–371, March 1984. Institute of Mathematical Statistics.

[85] Paul Meier, Theodore Karrison, Rick Chappell, and Hui Xie. The Price of Kaplan-Meier. *Journal of the American Statistical Association*, 99(467):890–896, September 2004. American Statistical Association.

[86] L. Mark Berliner and Bruce M. Hill. Bayesian Nonparametric Survival Analysis. *Journal of the American Statistical Association*, 38(1):772–779, September 1988. American Statistical Association.

[87] C. A. McGilchrist and C. W. Aisbett. Regression with Frailty in Survival Analysis. *Biometrics*, 47(2):461–466, June 1991. International Biometric Society.

[88] Yaman BARLAS. Theory and Methodology: Multiple tests for validation of system dynamics type of simulation models. *European Journal of Operational Research*, 42:59–87, 1989.

[89] Helen Trottier and Pierre Philippe. Deterministic modeling of infectious diseases: Theory and methods. *The Internet Journal of Infectious Diseases*, 1(2):1–6, 2001.

[90] Neil Gilbert and Harry Specht. *Planning for Social Welfare*, chapter Probem Analysis: Data Collection Techniques, pages 311–316. Prentice-Hall, Inc., Englewood Cliffs, N.J. 07632, 1977.

[91] Tollman SM. The Agincourt field site: Evolution and current status. *S Afr Med J*, 89:853–8, 1999.

[92] Tollman SM, Herbst K, Garenne M, Gear JSS and Kahn K. The Agincourt Demographic and Health Study: Site description, baseline findings and implications. *S Afr Med J*, 89:858–64, 1999.

[93] Stephen M. Tollman Mark A. Collinson and Kathleen Kahn. Migration, settlement change and health in post-apartheid south africa: Triangulating health and demographic surveillance with national census data. *Scandinavian Journal of Public Health*, 35(Suppl 69):77–84, 2007.

[94] Howard Schuman and Stanley Pressser. *Questions and Answers in Attitude Surveys*, chapter Questions Order and Response Order. Quantitative Studies in Social Relations. Academic Press, Inc., 1981.

[95] James Alan Fox and Paul E. Tracy. *Randomized Response: A Method for Sensitive Surveys*. Sage Publications, Inc., 1986.

[96] Foddy William H. *Constructing Questions for Interviews and Questionnaires : Theory and Practice in Social Research*. Cambridge University Press, 1993.

[97] Wakeling J Bagnall G Colthart I Illing J Kergon C Morrow G Spencer J van Zwanenberg T. Burford B, Hesketh A. Asking the right questions and getting meaningful responses: 12 tips on developing and administering a questionnaire survey for healthcare professionals. *Med Teach.*, 31(3):207–11, Mar 2009.

[98] Jean M. Converse and Stanley Presser. *Survey Questions Handcrafting the Standardized Questionnaire*. Sage Publications, Inc, 1986.

[99] Peter Byass Edward Fottrell and Yemane Berhane. Demonstrating the robustness of population surveillance data: implications of error rates on demographic and mortality estimates. *BMC Medical Research Methodology*, 2008.

[100] Guy R. Clement and Levi T. Wilson. *Analytic and Applied Mechanics*, chapter 8, pages 279–282. The McGrraw-Hill Book Company, Inc., Department of Mathematics, United States Naval Academy, second edition, 1935,1943. Seventh Impression.

[101] Rizzo ML Espy KA Fang H, Brooks GP and Barcikowski RS. Power of models in longitudinal study: Findings from a full-crossed simulation design. *The Journal of Experimental Education*, 77(3):215–254, Apr 2009.

[102] Foster Morrison. *The Art of Modeling Dynamic Systems*, page 336. New York: J. Willey, 1991.

[103] Reuben Granich. *HIV, Health, and your Community : a guide for action*. Berkeley Calif. : Hesperian Foundation, 2001.

[104] Curtis Wilson. Newton's Orbit Problem: A Historian's Response. *The College Mathematics Journal*, 25(3):193–200, May 1994. Published by: Mathematical Association of America.

[105] D. K. Arrowsmith and C. M. Place. *Dynamical Systems: Differential Equations, Maps and Chaotic Behaviour*. Chapman and Hall, 1992.

[106] Morris W. Hirsch and Stephen Smale. *Differential Equations, Dynamical Systems, and Linear Algebra*. Academic Press, Harcourt Brace Jovanovich, Publishers, 1974.

[107] Ulrich K. Steiner Shripad Tuljapurkar and Stven Hecht. Dynamic heterogeneity in life histories. *Ecology Letters*, 12:93–106, 2009.

[108] Shripad Tuljapurkar Charlotte T. Lee and Peter M. Vitousek. Risky business: Temporal and spatial variation in preindustrial dryland agriculture. *Springer Science+Business Media*, 2006. Hum Ecol DOI 10.1007/s10745-006-9037-x.

[109] Cedric O. Puleston and Shripad Tuljapurkar. Population and prehistory ii: Space-limited human populations in constant environments. *Theoretical Population Biology*, 74:147–160, 2008.

[110] Bruce C. Berndt, Christian F. Krattenthaler, Gary L. Mullen, K. Ramachandra, A.K. Agarwal and Michel Waldschmidt, editors. *Number Theory and Discrete Mathematics*. Boston, MA : Birkhuser, 2002.

[111] Darlington P. J. *Evolution for Naturalists*, chapter 6. John Wiley and Sons, Inc., 1980.

[112] Friend J Diener-West M. Ybarra ML, Langhinrichsen-Rohling J. Impact of asking sensitive questions about violence to children and adolescents. *J Adolesc Health.*, 45(5):499–507, Nov 2009.

[113] Kalyan Chatterjee and William F. Samuelson, editors. *Game Theory and Business Applications*. Boston, Mass. : Kluwer Academic Publishers, 2001.

[114] Anatole Katok and Boris Hasselblatt. *Introduction to the Modern Theory of Dynamical Systems*, volume 54. Cambridge University Press, 1995.

[115] S. Wiggins. *Introduction to Applied Nonlinear Dynamic Systems and Chaos*, volume TAM 2. Springer-Verlag New York, Inc., Department of Applied Mathematics, California Institute of Technology, Pasadena, California 91125, USA, 1990.

[116] Ramon Roman-Roldan, Pedro Bernaola-Galvan, and Jose L. Olivier. Application of Information Theory to DNA Sequence Analysis: A Review. *Pergamon*, 29(7):1187–1194, 1996. Patern Recognition Society.

[117] Bethge Matthias, Rotermund David and Pawelzik Klaus. Second order phase transition in neural rate coding: Binary encoding is optimal for rapid signal transmission. *Phys. Rev. Lett.*, 90(8):088104, Feb 2003.

[118] Henning S. Mortveit and Christian M. Reidys. *An Introduction to Sequential Dynamical Systems*. Springer Science+Business Media LLC, New York, 2008.

[119] J. Llibre L. Alseda and M. Misiurewicz. *Combinatorial Dynamics and Entropy in Dimension One, Advanced Series in Nonlinear Dynamics*, volume 5. World Scientific, 2nd edition, 2000.

[120] C. Robinson. *Dynamical Systems: Stability, Symbolic Dynamics and Chaos*. CRC Press, 2nd edition, 1999.

[121] Cristopher Moore. Generalized one-sided shifts and maps of the interval. *Nonlinearity 4 (1991) 727-745.*

[122] Wael Bahsoun and Pawel Gora. Weakly convex and concave random maps with position dependent probabilities. *Stoch. Anal. Appl.*, 2003.

[123] W. H. Sulis and I. N. Trofimova, editors. *Nonlinear Dynamics in the Life and Social Sciences*, chapter Chaos Theory. IOS Press, 2001.

[124] Patrick Heuveline, Samuel H. Preston and Michel Guillot. *Demography : Measuring and Modeling Population Processes*. Malden, MA : Blackwell Publishers, 2001.

[125] John R. Weeks. *Population : An Introduction to Concepts and Issues*. Belmont, Calif. : Wadsworth Publishing, 7th edition, 1999.

[126] Susan Ziehl. *Population Studies*. Oxford University Press, 2002.

[127] Newell Colin. *Methods and Models in Demography*. New York : Guilford, 1988.

[128] Jurgen Moser. *Stable and Random Motions in Dynamical Systems*, volume Study 77 of *Annals of Mathematics Studies*. Princeton University Press and University of Tokyo Press, 1973.

[129] Pierre Gaspard. Brownian motion, dynamical randomness and irreversibility. *New Journal of Physics*, March 2005.

[130] Lin C.C. and Segel L.A. *Mathematical Applied to Deterministic Problems in the Natural Sciences*, volume 1. Society for Industrial and Applied Mathematics, 1988.

[131] Ravenstein EG. The laws of migration. *Journal of the Royal Statistical Association*, (48):167–227, 1885.

[132] Wolpert J. Behavioral aspects of the decision to migrate. *Papers of the Regional Science Association*, (19):159–169, 1965.

[133] Goodman JL. Information, uncertainty, and the microeconomic model of migration decision making. *New York: Pergamon*, pages 130–148, 1986.

[134] De Jong GF. Expectations, gender, and norms in migration decision-making. *Population Studies*, 54(3):307–319, 2000.

[135] Charles Hirsch. *Numerical Computation of Internal and External Flows*, volume 1, chapter The Concepts of Consistency, Stability and Convergence. John Willey and Sons Ltd., a willey-interscience publication edition, 1988.

[136] S. Yu. Kobyakov. Methods of Symbolic Analysis of Dynamic Systems. *Antomation and Remote Control*, 65(4):554–558, 2004.

[137] Cristopher Moore. Generalized shifts: unpredictability and undecidability in dynamical systems. *Nonlinearity 4*, 1991.

[138] Pingel Detlef, Schmelcher Peter and Diakonos Fotis K. Detecting unstable periodic orbits in chaotic continuous-time dynamical systems. *Phys. Rev. E*, 64(2):026214, Jul 2001.

[139] Pingel Detlef et al. Theory and applications of the systematic detection of unstable periodic orbits in dynamical systems. *Phys. Rev. E*, 62(2):2119–2134, Aug 2000.

[140] So Paul et al. Extracting unstable periodic orbits from chaotic time series data.

[141] Andrews J. G. and McLone R. R. *Mathematical Modelling*. London: Butterworths, 1976.

[142] Ronald L. Breiger, Scott A. Boorman and Phipps Arabie. An Algorithm for Clustering Relational Data with Applications to Social Network Analysis and Comparison with Multidimensional Scaling. *Journal of Mathematical Psychology*, 12:328–383, 1975.

[143] Chennaoui A. et al. Attractor reconstruction from filtered chaotic time series. *Phys. Rev. A*, 41(8):4151–4159, Apr 1990.