Visualization of Longitudinal Data by Fitness Orbits: Application to Data on Rural Households at Agincourt

C.L. Mberi Kimpolo, D. Sherwell and M. Collinson.

1 Introduction

In this paper we present and discuss first demographic results using the techniques developed in [1, 2] for analysis of longitudinal data of a population of social units. From the longitudinal data of each social unit, using a mathematical strategy, we construct a geometric orbit in a fitness space that uniquely visualizes the welfare of the social unit. Mathematical modelling of time-dependent phenomena [3–6] conventionally proposes difference or differential equations and by choice of coefficients in these equations seeks to reproduce observed population behaviour. In this paper, we will begin with the observed data and induce the dynamical system. This is done in such a way that every observed orbit is a property of the dynamical system. We know of no similar study.

Mathematical models in demography usually seek population counts as a function of time [7–9]. Perhaps the most sophisticated models are those considered in [6]. There, the equations of statistical physics such as the Fokker-Plank partial differential equations are invoked that determine the probability for a social unit to be in some state of interest. The basic assumption in physics is that for large number of atoms (say $\sim 10^{23}$), the orbits of those particles can never be known and that statistical methods are then justified. However, it is precisely the importance of longitudinal surveys that they *are* the data of many social units that undergo change. It is a direct challenge to find a concise set of equations that describe change for any social unit that is surveyed.

The difficulty in designing longitudinal surveys of choosing the order in which questions are asked is well known. Following the ideas of [1, 2] we will use question order as a mathematical variable. We know of no similar usage. We will assume that questionnaires have been carefully designed, that responses have been honestly given and that data is clean.

2 Agincourt Hypothesized Data coding

We visualize orbits using longitudinal data from the Agincourt Health and Demographic Surveillance Site (HDSS). Here the social unit of interest is a household. We start by defining our purpose.

Purpose: To investigate the effect of change in three household characteristics on child educational progress. (1)

We define educational progress below. The data is intermittent. Under the purpose (1), we consider a questionnaire Q consisting of $n = 3$ questions with data for each household observation period, defined as follows.

$$\begin{aligned} q_0 &: \text{ Was there a child without a biological mother in the household?} \\ q_1 &: \text{ Was the head of the household a minor?} \\ q_2 &: \text{ Was there an adult death in the household?} \end{aligned} \tag{2}$$

We cannot include educational default in the question set (2) because it is observed approximatively every

5 years. We code each question answer a_i and determine its associated fitness value x_i as follows.

$$\begin{array}{lllll} a_0: & \text{Yes} & = & \text{Unfavourable} & \Longrightarrow \quad x_0 = 0 \\ a_1: & \text{No} & = & \text{Favourable} & \Longrightarrow \quad x_1 = 1 \\ a_2: & \text{Yes} & = & \text{Unfavourable} & \Longrightarrow \quad x_2 = 0 \end{array} \tag{3}$$

We have a sample of $K = 2669$ households with children of school-going age, observed over the period 1998 to 2007. We determine the population frequencies at Agincourt of changing answer values and find $\overline{f_1} < \overline{f_2} < \overline{f_0}$ where f_i denotes the frequency of change of answer value a_i.

3 Agincourt Population orbits

Demography is concerned with typical properties of populations or sub-populations. In Figure 1, we show in S_3 state orbits for sample population of $K = 2669$ households with children of school-going age.

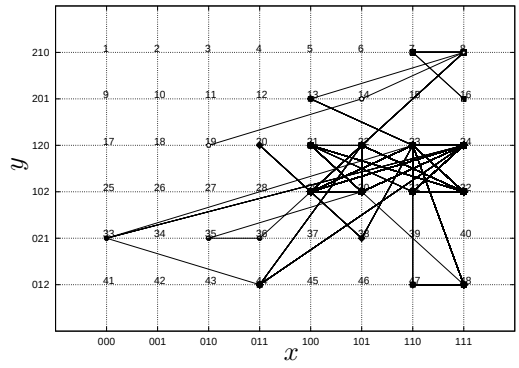

Figure 1: orbits in S_3 for the sample Agincourt population of $K = 2669$ households.

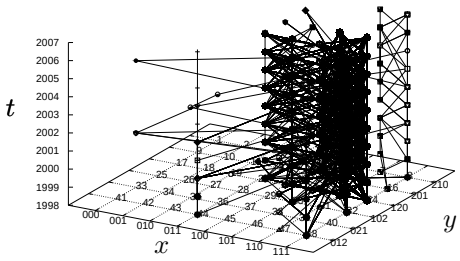

Figure 2: orbits of $K = 2669$ Agincourt households with time dependent in the vertical axis.

In Figure 2, we give all time series for the sample population. Clustering is clearest in Figure 2 and the most populous sub-space is identified by households where there is an adult head ($x_1 = 1$). Of particular importance is the time dependence of clustering. In Figure 2, it appears that the orbit *of the cluster* is a vertical straight column in this sub-space. We may refer to any orbit that stays within a cluster as a typical orbit of that cluster. Since there is only one cluster in the Agincourt data under our purpose, a typical orbit of Agincourt is any one in the dense column of Figure 2. The cluster does not change in time and for this questionnaire, we could sample less frequently, perhas every 5 years. Rigorous mathematical techniques of image processing may be used to identify typical orbits.

According to purpose (1), we now define educational default by those scholars who have failed more than three years in their school life. The sample population then splits into favourable and unfavourable sub-populations. We find that 54.17% of households in the sampled population are defaulting so that the two populations are roughly balanced in number (this proportion is a severe criticism of the quality of education offered in the Agincourt district).

In Figure 3 and Figure 4 we give collective phase space visualization for the two sub-populations. We note immediately that they are similarly clustered, show many similar transitions, and claims of cause of default must be carefully judged. In Figure 3, we note some severe jumps to unfit states, however the number of these transitions is insignificant.

Figure 5 and Figure 6 show the number of each transition in the most populous cases for the defaulting and non-defaulting sub-populations respectively. Here we consider the most active transitions on a sub-space of S_3. By doing this we ignore only 2.33% of transitions. The dominant transitions are clearly $23 \leftrightarrow 24$, in

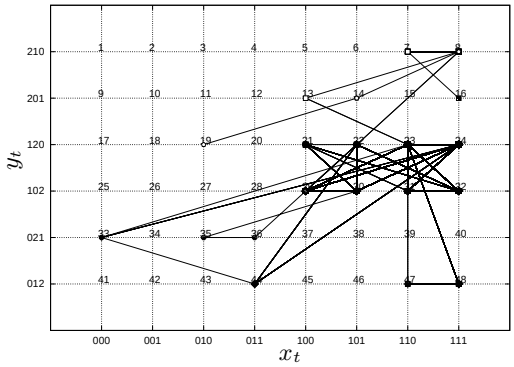

Figure 3: orbits in S_3 for defaulting Agincourt households.

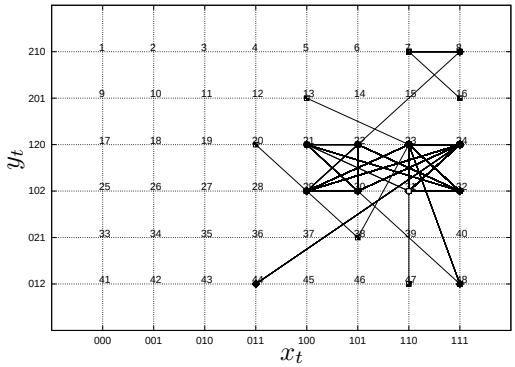

Figure 4: orbits in S_3 for non-defaulting Agincourt households.

both sub-populations. There are many idling states as indicated by the circles. In Table 1 we summarise these dominant effects, which finally comprise about 88% of all transitions.

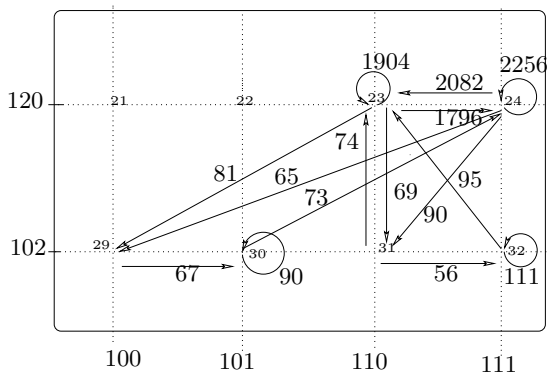

Figure 5: State space S_2 for defaulting Agincourt households. In this space, every household is headed by an adult.

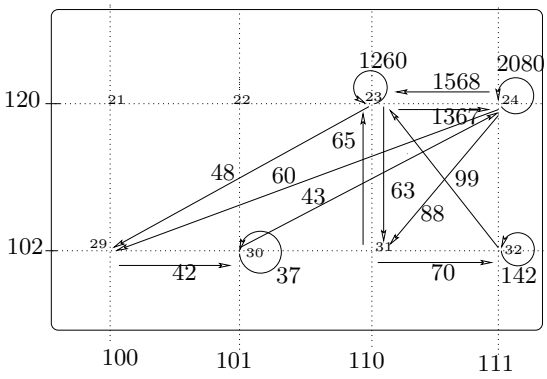

Figure 6: State space S_2 for non-defaulting Agincourt households. In this space, every household is headed by an adult.

Table 1: Typical sub-space transition counts.

Education measure	**Dominant transitions**	d_{ij}^{Ag}		$d_{ij}^{\overline{Ed}}$		d_{ij}^{Ed}	
	$i \to j$	#	%	#	%	#	%
Number of failure years = **4**	$24 \to 24$	4336	26.56	2256	52.03	2080	**47.97***
	$24 \to 23$	3650	22.36	2082	**57.04**	1568	42.96
	$23 \to 23$	3164	19.38	1904	**60.17***	1260	39.83
	$23 \to 24$	3163	19.37	1796	56.78	1367	**43.22**

Table 1 is used to make demographically useful conclusions. In Table 1, note that d_{ij}^{Ag} denotes the overall number of transitions $i \to j$ for the whole Agincourt population, for the whole period 1998 to 2007, $\overline{d_{ij}^{Ed}}$ (d_{ij}^{Ed}) for the educationally defaulting (non-defaulting) populations. In the defaulting population, note the two dominant transitions are into or at state 23, mother out-migration. In the non-defaulting population,

the dominant transitions are into or at state 24, mother at home. These results suggest that Agincourt educational default is related to out-migration of biological mothers.

Concerning cause and effect, it is tempting to suppose (as in physics) that a change to a state has a unique origin. In human affairs we cannot claim this. In the present case we cannot say that out-migration of mother causally precedes educational default, only that they tend to happen together. Given that educational default is observed roughly every 5 years, and the average observation time is 7 years, educational default is a property of the household and of the 10 years *period* of observation, not of household moments of observation, and we cannot expect to do better than this. It is clear that if we had the scenario annual education data, and, if there was a clear transition for households with mothers at home and non-defaulting children to mother out-migration and defaulting children, for a significant number of households, that we might reasonably claim a causal chain.

We come to these conclusions without statistical inference (but under the hypothesis that the questions (2) are appropriate and are sufficient for Purpose (1)). Our conclusions are for households with adult household head (because there are few contrary cases). Clearly out-migration of mothers puts children at risk of defaulting and has been coded correctly as unfavourable. We can make no clear conclusion about the effect of adult death and so the outcome is neutral to the coding (3). We speculate that independent sociologists would eventually agree and code and visualize the same orbits under Purpose (1) and questions (2). The strategy could now be extended by deleting q_1 and including different questions.

Acknowledgements

This work was supported by the Welcome Trust [085477/Z/08/Z].

References

[1] Mberi Kimpolo C.L. *Deterministic Dynamics in Questionnaires in the Social Sciences*. PhD thesis, Computational and Applied Mathematics, University of the Witwatersrand, Johannesburg, March 2010.

[2] Mberi Kimpolo C.L. and Sherwell D. A new mathematical framework for longitudinal demographic data analysis by fitness orbits of social units. *Submitted*, 2010.

[3] Bender Edward A. *An Introduction to Mathematical Modeling*. New York: Wiley, 1978.

[4] Murray Francis J. *Applied Mathematics: An Intellectual Approach*. New York: Plenum Press, 1978.

[5] Michael Olinick. Mathematical Models in the Social and Life Sciences: A Selected Bibliography. *Mathemntieal Modelling*, 2:237–258, 1981.

[6] Helbing D. *Quantitative Sociodyanmics-Stochastic and Models of Social Interaction Processes*. Dordrecht: Kluwer Academic Publishers, 1995.

[7] May Robert McCredie. *Stability and Complexity in Model Ecosystems*. Princeton University Press, Princeton, NJ, 1974.

[8] Rahim Moineddin et al. A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, 2007.

[9] Nigel Meade. A modified Logistic Model Applied to Human Populations. *Journal of the Royal Statistical Society.*, 151(3), 1988.