

https://dspace.org/

Analysis of binary multivariate longitudinal data via 2-dimensional orbits: An application to the Agincourt Health and Socio-Demographic Surveillance System in South Africa.

Visaya, M.V.; Sherwell, D.; Sartorius, B.; Cromieres, F.

2015-04
Public Library of Science
http://hdl.handle.net/10539/20702
Visaya, M.V. et al. 2015. Analysis of binary multivariate longitudinal data via 2-dimensional orbits: An application to the Agincourt Health and Socio-Demographic Surveillance System in South Africa. 2015. PLoS ONE 10(4): e0123812.

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

PLOS | ONE

RESEARCH ARTICLE

Analysis of Binary Multivariate Longitudinal Data via 2-Dimensional Orbits: An Application to the Agincourt Health and Socio-Demographic Surveillance System in South Africa

Maria Vivien Visaya[1]*, David Sherwell[2], Benn Sartorius[3], Fabien Cromieres[4]

1 Department of Pure and Applied Mathematics, University of Johannesburg, Johannesburg, South Africa, **2** School of Computational and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa, **3** Discipline of Public Health Medicine, School of Nursing and Public Health, University of KwaZulu-Natal, Durban, South Africa, **4** Graduate School of Informatics, Kyoto University, Kyoto, Japan

* mvvisaya@uj.ac.za and v3isaya@gmail.com

CrossMark
click for updates

OPEN ACCESS

Citation: Visaya MV, Sherwell D, Sartorius B, Cromieres F (2015) Analysis of Binary Multivariate Longitudinal Data via 2-Dimensional Orbits: An Application to the Agincourt Health and Socio-Demographic Surveillance System in South Africa. PLoS ONE 10(4): e0123812. doi:10.1371/journal.pone.0123812

Academic Editor: Koustuv Dalal, Örebro University, SWEDEN

Received: November 9, 2014

Accepted: March 7, 2015

Published: April 28, 2015

Copyright: © 2015 Visaya et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

We analyse demographic longitudinal survey data of South African (SA) and Mozambican (MOZ) rural households from the Agincourt Health and Socio-Demographic Surveillance System in South Africa. In particular, we determine whether absolute poverty status (APS) is associated with selected household variables pertaining to socio-economic determination, namely household head age, household size, cumulative death, adults to minor ratio, and influx. For comparative purposes, households are classified according to household head nationality (SA or MOZ) and APS (rich or poor). The longitudinal data of each of the four subpopulations (SA rich, SA poor, MOZ rich, and MOZ poor) is a five-dimensional space defined by binary variables (questions), subjects, and time. We use the orbit method to represent binary multivariate longitudinal data (BMLD) of each household as a two-dimensional orbit and to visualise dynamics and behaviour of the population. At each time step, a point (x, y) from the orbit of a household corresponds to the observation of the household, where x is a binary sequence of responses and y is an ordering of variables. The ordering of variables is dynamically rearranged such that clusters and holes associated to least and frequently changing variables in the state space respectively, are exposed. Analysis of orbits reveals information of change at both individual- and population-level, change patterns in the data, capacity of states in the state space, and density of state transitions in the orbits. Analysis of household orbits of the four subpopulations show association between (i) households headed by older adults and rich households, (ii) large household size and poor households, and (iii) households with more minors than adults and poor households. Our results are compared to other methods of BMLD analysis.

Introduction

Binary multivariate longitudinal data (BMLD) is here exemplified by the binary responses in a Yes/No form to a set of $p \geq 1$ questions (variables) asked to each subject of a (sample) population over a period of time. As in the convenient convention of binary coding of 0 for a negative response and 1 a for positive response [1], the outcome of each of the binary variables here is coded as 0 if the outcome is unfavourable (by hypothesis) to a given purpose, and 1 if favourable.

Many BMLD studies use regression techniques [2] or Markov, transition and forecasting models ([3–5]). These methods involve parameter estimation for the explanatory variables. However, visual analysis of data is equally important as it presents initial insights about the data. Descriptive tools such as tables and charts give a visual summary and simpler interpretation. For visual analysis of multivariate longitudinal data, some analysis is given in ([6, 7]) but very few tools are available when the data is binary.

In [8], the focus is on visualizing the complex border between patterns of BMLD. The border in a multidimensional space is converted into visual 2-dimensional and 3-dimensional forms. However, it does not illustrate patterns and dynamics of the population over time. A technique that accounts for dynamics of BMLD and within subject information is the orbit method discussed in [9]. Orbit method here is distinct from the Kirillov orbit method used in representation theory [10]. By an *orbit* we mean a sequence of points related by the evolution function of the underlying system. The method of orbit is a technique based on deterministic outcomes with emphasis on geometric visualization of multivariate longitudinal data as 2-dimensional orbits. It considers the frequency of change of variables and uses the order of variables for constructing orbits that represent subjects from the population. Orbits give insight for data analysis and provide exact data visualization.

Here we use the orbit method to analyse binary demographic data of households from the Agincourt Health and Socio-Demographic Surveillance System (AHDSS) in South Africa. The longitudinal AHDSS data have been studied e.g. in [11] and [12] where a spatial-temporal model to analyse distribution of mortality and asset accumulation rate respectively were employed. Determinants of socio-economic status/poverty or the relationship between poverty and increased mortality were viewed from more of a static perspective i.e. not from the more dynamic approach offered by the orbit. The orbit approach presents a visualisation of the data in a truly longitudinal-temporal manner. The orbit method is briefly illustrated in [9] using variables regarding child educational progression in AHDSS. However, detailed interpretation of subsets in the subspace, nor analysis of orbits in the space, were not discussed. Aside from [9], we know of no other visual analysis employed for AHDSS, particularly involving household variables pertaining to socio-economic determination.

The AHDSS longitudinal data analysed here is of about 4,000 households from 2001–2007. With purpose

$$\textit{Purpose} : \textit{To determine the association of absolute poverty status with selected household variables}, \tag{1}$$

we consider the following questions:

$$\begin{aligned} Q_0(\mathrm{HH}): &\quad \text{Household head age} \geq 40 \text{ years old?} \\ Q_1(\mathrm{HS}): &\quad \text{Household size} \geq 3 \text{ individuals?} \\ Q_2(\mathrm{HD}): &\quad \text{Cumulative household deaths low (excluding household head death)?} \\ Q_3(\mathrm{AM}): &\quad \text{More adults (age} \geq 18\text{) than minors?} \\ Q_4(\mathrm{IF}): &\quad \text{Internal influx negative} \\ &\quad \text{(i.e. more people migrating into the household than leaving)} \\ Q_5(\mathrm{HN}): &\quad \text{Household head nationality} - \text{South African } \textit{or} \text{ former Mozambican refugee?} \\ Q_6(\mathrm{APS}): &\quad \text{Absolute poverty status} - \text{below poverty line } \textit{or} \text{ above poverty line?} \end{aligned} \tag{2}$$

Each question in (Eq 2) is associated to a variable, e.g. Q_0 to household head age (HH), Q_1 to household size (HS), and so on. We will sometimes refer to question Q_i ($i = 0, 1, \ldots, 6$) as variable i.

The advantage in representing the data of each subject as a two-dimensional orbit is that orbits capture the dynamics of change in response of each subject so it reveals information of change over time at both individual and population level while retaining the full information of the original data. Using orbits for data analysis give a way to visualize data, i.e. identify clusters associated to stable (less frequently changing) variables, and patterns in subpopulations associated to clusters. BMLD can involve hundreds of variables so visualising in $d \geq 4$ dimensions is difficult. Survey data (e.g. in the social sciences) is usually large both in dimension and in size but orbit representation is feasible for large numbers of variables and subjects. In this application of orbits to the analysis of AHDSS survey data, we hope to contribute in giving new insights in the analysis of binary multivariate longitudinal data.

Materials and Methods

Description of Data

The Agincourt Health and Socio-Demographic Surveillance System (HDSS) is located in Bushbuckridge in northeast South Africa and was established in 1992. Bushbuckridge is a poor rural sub-district that is made up of South African and former Mozambican refugees (approximately a third of the population) [13, 14]. There have been annual updates of births, deaths, in- and out-migrations of individuals identified as members of households, as well as regular special modules (e.g. household asset ownership) used to derive a socio-economic status index.

Recall our purpose and questions given in (Eq 1) and (Eq 2). Regarding absolute poverty status (APS), it is independent of the household variables associated to questions in (Eq 2). Here, households above the absolute poverty line was defined using the definition proposed in [15] for a sub-Saharan African setting, namely ownership of a radio and bicycle, a cement floor in the house, and access to public water and a pit latrine (toilet). Absolute poverty classification is thus independent of the 6 explanatory variables used in our orbit analysis. APS of households below the poverty line are coded 0, while APS above the poverty line is coded 1. Because APS is gathered only in 4 out of the 7 observation years (i.e. at $t = 2001, 2003, 2005$, and 2007), we use the mean APS of a household over 7 years, which we denote by $\overline{APS}$. If $\overline{APS} \in [0, 0.5]$, then $\overline{APS}$ is coded 0. Otherwise, $\overline{APS}$ is coded 1. Our sample population consists of 7715 household

Table 1. Favourable(= 1) and unfavourable(= 0) answer to questions Q_0 to Q_4.

Q_0 (HH: household head)	Q_1 (HS: household size)	Q_2 (HD: household death)	Q_3 (AM: adult(A) to minor(M) ratio	Q_4 (IF: internal influx)
HH<40 = 0	HS<3 = 0	HD high = 0	A<M = 0	$IF^+ = 0$
HH≥40 = 1	HS≥3 = 1	HD low = 1	A≥M = 1	$IF^- = 1$

doi:10.1371/journal.pone.0123812.t001

units, 4158 of which are either always above or below the absolute poverty line for all four years that APS was gathered. For these households, APS information not gathered for the three years 2002, 2004, and 2006 will not affect the coding of their $\overline{APS}$. Households with $\overline{APS} = 0$ are referred to as Poor households, and households with $\overline{APS} = 1$ as Rich households. Our analysis will only consider these 4158 households. Ethical clearance for the primary study was given by the University of the Witwatersrand Human Research Ethics Committee (Medical). The data used in this study does not contain clinical records (nor does the core Agincourt HDSS database). Individual and household identifiers are anonymized/de-identified by the data managers prior to handing it over to researchers for analysis to ensure confidentiality.

Aside from $\overline{APS}$, household head nationality is also *constant* throughout the survey period. In addition, former Mozambican (MOZ) refugees experience significantly higher levels of poverty compared to their South African (SA) counterparts and this gap has persisted over time [12, 16, 17]. It is then useful to extract Q_5 and Q_6 and analyse by these subpopulations of households where both poverty status and household head nationality are unchanging. We divide our population into four subpopulations, namely SA Rich, SA Poor, MOZ Rich, and MOZ Poor. Each subpopulation is analysed using $p = 5$ variables associated to Q_0 to Q_4. Binary data of the four subpopulations is given in S1 Dataset. From [12, 16], a 'yes' answer to questions Q_0 to Q_4 is assumed to be favourable to APS so we code all yes = 1, and all no = 0. Table 1 gives the favourable and unfavourable code for each of the five questions. Table 2 gives the number of households by constant variables (i.e. nationality and $\overline{APS}$).

The Method of Orbits

Given the number of variables $p \geq 1$, denote by

$$M_p = \{0, 1\}^p$$

the space of binary strings (responses) of length p. For a subject ℓ observed at times $t = 0, 1, \ldots, T$, we define

$$D^\ell = \{D_0^\ell, D_1^\ell, \ldots, D_T^\ell\},\ D_t^\ell \in M_p$$

Table 2. Distribution of households by nationality and mean APS.

Population	SA Rich	SA Poor	MOZ Rich	MOZ Poor
4158	2610	421	468	659

doi:10.1371/journal.pone.0123812.t002

Table 3. Concatenated coded answers of three subjects to questionnaire $Q = \{Q_0, Q_1, Q_2\}$.

t	ℓ	ℓ'	ℓ''
0	010	100	111
1	001	100	001
2	000	101	001
3	001	110	101
4	000	101	001

doi:10.1371/journal.pone.0123812.t003

the binary multivariate longitudinal data in $p \geq 1$ variables of subject ℓ. The binary longitudinal data in p variables from a population of $n \geq 1$ subjects observed over time T is the set

$$D_{[p,n,T]} = \{D^1, D^2, \ldots, D^n\}.$$

We will only consider subjects with complete data.

Analysis of BMLD *always* involves a fixed variable order where one can use the summary measure of the frequency of response pattern (elements of M_p) and perform factor analysis on the longitudinal data [1] or construct Markov models using information of change of time encoded in the matrix of transition probabilities [3]. The method of orbits [9] uses the fundamental information of *frequency of change* of variables and order of variables for analysis. The information of change is used to define a non-autonomous dynamical system from data of each subject, dynamically rearranging order of variables so that most stable least changing variable is eventually placed to the left, but keeping the full information in the original data. Mathematical properties of the map are discussed in [9].

To explain the orbit method, we illustrate for $p = 3$ variables. Let $Q = \{Q_0, Q_1, Q_2\}$ be a questionnaire and assign index i to Q_i, $i = 0, 1, 2$. Table 3 illustrates concatenated coded answers to Q of three subjects from a sample population. To Q_0, subject ℓ has constant answer 0 while ℓ' has constant answer 1. On the other hand, ℓ'' has constant answer 1 to Q_2. Observe that this property of subjects having constant answers to certain questions is not trivially illustrated in the time series of the three subjects given in Fig 1.

Suppose we order questions and give more weight to those that least frequently change. As in numbers, we let the digit in the left-most position of the question order be most significant, and digit in the right-most be least significant. Observe that for both ℓ and ℓ', answer to Q_0 is the most stable (i.e. it is constant), followed by Q_1 (changes once in ℓ and twice in ℓ'), and finally Q_2 as most frequently changing. Then question order for both ℓ and ℓ' is chosen as Q_0, Q_1, Q_2, which we will denote by 012. Now position lexicographically in increasing order as binary integers the states (responses)

$$000, 001, 010, 011, 100, 101, 110, 111 \tag{3}$$

along an axis, and denote this by X_3. For fixed question order 012, a one-dimensional dynamics on the states in X_3 arises, where answers of subjects ℓ and ℓ' are visualised jumping from one state to another, particularly staying in the distinct regions $0**$ and $1**$, the left and right half of X_3, respectively. However, different subjects may have different frequencies of change in answer values. Because ℓ'' has constant answer to Q_2, question order for ℓ'' is chosen such that Q_2 is given more weight. In particular, question order for ℓ'' is set to 210.

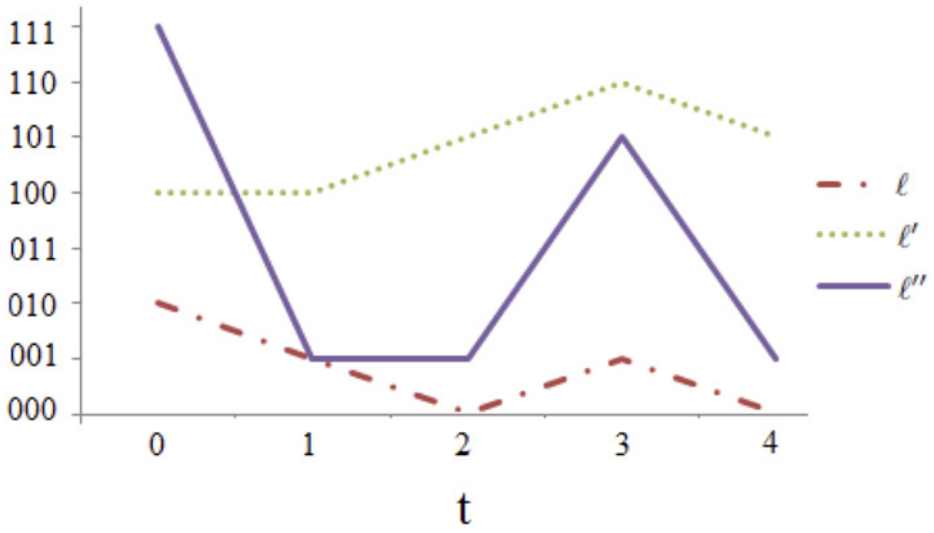

Fig 1. Time series of the three subjects in Table 3.

doi:10.1371/journal.pone.0123812.g001

We recall terms and notations as introduced in [9]. Let

p : number of variables

ℓ : subject from the population

n : number of subjects

Q : questionnaire

Q_i : question from Q, $i = 0, 1, \ldots, p-1$

f_i^l : frequency of change of Q_i in data of subject ℓ

Definition 1 *Given* $p \geq 1$, *the spaces of sequences*

$$X_p = \{(x_j)_{j=0}^{p-1} = x_0 x_1 \cdots x_{p-1} : x_j \in \{0,1\}\}$$

and

$$Y_p = \{(i_j)_{j=0}^{p-1} = i_0 i_1 \cdots i_{p-1} : i_j \in \{0,1,\ldots,p-1\},\ i_j's \text{ distinct}\},$$

both with the lexicographic ordering of sequences (i.e. as increasing integers) are the fitness axis and significance axis for p variables, respectively. An element $x \in X_p$ *is called a fitness state, and* $y \in Y_p$ *a significance state. The space*

$$\begin{aligned} S_p &= \{p = (x,y) : x \in X_p, y \in Y_p\} \\ &= X_p \times Y_p \end{aligned}$$

is the change space in p variables composed of $P = 2^p\, p!$ *states.*

For convenience, states in S_p are labeled from 1 to $P = 2^p\, p!$ starting from left to right, top to bottom. The labeled space S_p for $p = 3$ is illustrated in Fig 2. The space X_3 is the sequences in (Eq 3), $Y_3 = \{012, 021, 102, 120, 201, 210\}$, and the cardinality of $|S_3| = 2^3 3! = 48$.

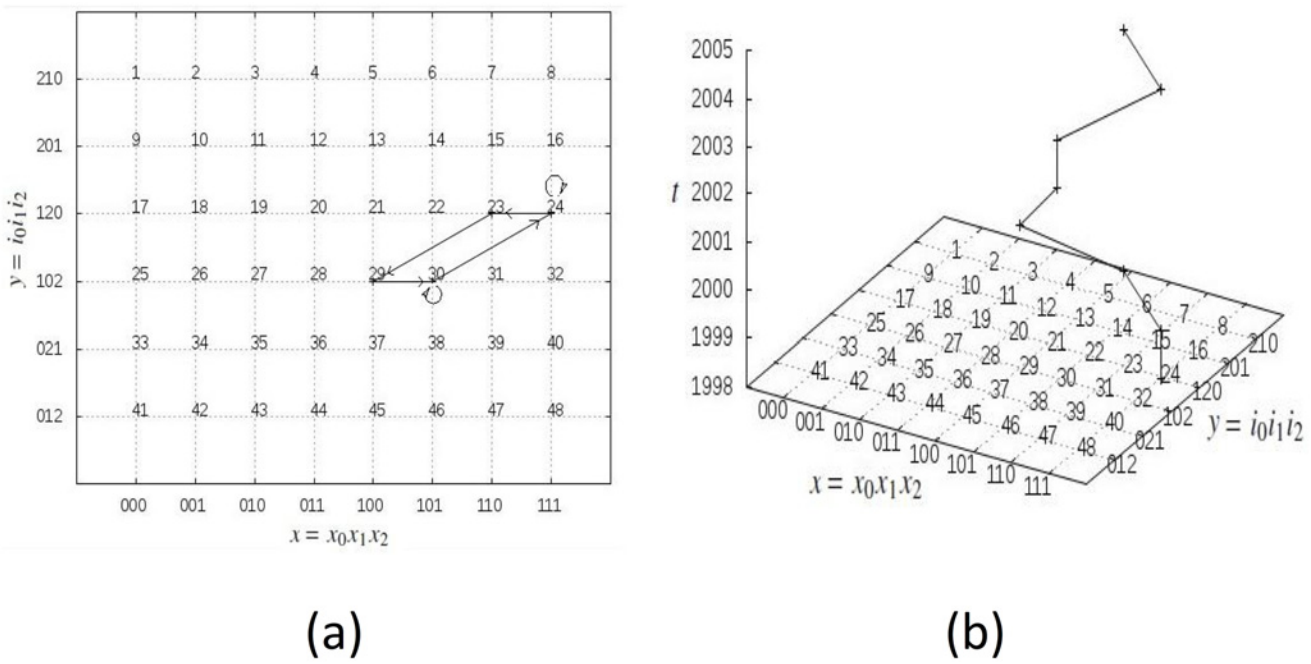

Fig 2. (a) Orbit of subject ℓ staying in subset of S_3 where variable 1 is favourable. (b) Time series of the orbit of ℓ.

doi:10.1371/journal.pone.0123812.g002

Definition 2 *1. Consider subject ℓ. Given a set Q* of $p \geq 1$ *questions,* $i_j \in \{0, 1, 2, \ldots, p-1\}$, *and* $Q_{i_j} \in Q$ let

$$f_{i_j}^{\ell} = \textit{number of times } Q_{i_j} \textit{ changes answer in data of } \ell \textit{ over the observation period}. \tag{4}$$

Suppose

$$f_{i_0}^{\ell} < f_{i_1}^{\ell} < \cdots < f_{i_j}^{\ell} < \cdots < f_{i_{p-1}}^{\ell}.$$

Then the initial question order of ℓ is

$$y_0^{\ell} = i_0 i_1 \cdots i_j \cdots i_{p-1}. \tag{5}$$

If $f_{i_j}^{\ell} = f_{i_{j+1}}^{\ell}$, *use population frequencies* $f_{i_j}^{n}, f_{i_{j+1}}^{n}$ *to determine order between* i_j *and* i_{j+1}. *If* $f_{i_j}^{n} = f_{i_{j+1}}^{n}$ *and* $i_j < i_{j+1}$, *choose question order* $i_j\, i_{j+1}$. *Otherwise, choose* $i_{j+1}\, i_j$.

2. Given initial question order $y_0^{\ell} = i_0 i_1 \cdots i_j \cdots i_{p-1}$, *the initial fitness state of ℓ is*

$$x_0^{\ell} = x_0 x_1 \cdots x_j \cdots x_{p-1}$$

where each x_j *is the answer to question* i_j *in* y_0^{ℓ}.

3. The initial state of ℓ is $s_0^{\ell} = (x_0^{\ell}, y_0^{\ell}) \in S_p$.

The algorithm for determining the next states s_t^{ℓ} $(t > 0)$ is as follows:

Step 1: [initial state s_0^{ℓ}] For $t = 0$ and subject ℓ, determine the initial significance state y_0^{ℓ}, followed by the initial fitness state x_0^{ℓ}.

Step 2: [state s_1^ℓ] Given initial state $s_0^\ell = (x_0^\ell, y_0^\ell)$ of ℓ, identify the questions that change answer values at t = 1. If there are none, then the next state $s_1^\ell = s_0^\ell$. Let

$$x_j^* = \begin{cases} 1 & \text{if } x_j = 0 \\ 0 & \text{if } x_j = 1. \end{cases}$$

If both Q_{i_j} and $Q_{i_{j'}}$ change answers at $t = 1$ and $j < j'$, then sequentially swap to the right i_j and $i_{j'}$ (resp. x_j and $x_{j'}$) of the question order (resp. answer order), starting with $i_{j'}$ (resp. $x_{j'}$). Change x_j to x_j^* and $x_{j'}$ to $x_{j'}^*$, i.e.

$$\begin{array}{lll} t = 0 & & t = 1 \\ x_0^\ell = x_0 x_1 \cdots x_j \cdot \cdot x_{j'} \cdot \cdot x_{n-1} & \longrightarrow & x_1^\ell = x_0 x_1 \cdot \cdot x_{j-1} x_{j+1} \cdot \cdot x_{j'-1} x_{j'+1} \cdot \cdot x_{n-1} x_{j'}^* x_j^* \\ y_0^\ell = i_0 i_1 \cdots i_j \cdots i_{j'} \cdots i_{n-1} & \longrightarrow & y_1^\ell = i_0 i_1 \cdots i_{j-1} i_{j+1} \cdots i_{j'-1} i_{j'+1} \cdots i_{n-1} i_{j'} i_j \end{array}$$

This new answer order and question order is the next state $s_1^\ell = (x_1^\ell, y_1^\ell)$.

Step 3: [edge color] Draw an edge from s_0^ℓ to s_1^ℓ. To show direction of transitions between states, color the edge red if transition is from right to left, green if transition is from left to right, and blue otherwise (i.e. same x-coordinate).

Step 4: [state $s_t^\ell, t \geq 2$] Update state s_0^ℓ as s_1^ℓ and time t as t = 2 in Step 2. Repeat Steps 2 and 3, and iterate until t = T−1.

Definition 3 *Let x_t^ℓ, y_t^ℓ, and $s_t^\ell = (x_t^\ell, y_t^\ell)$ be the fitness, significance, and state of subject ℓ at time t, respectively. The orbit of ℓ is the sequence of points*

$$\mathcal{O}(\ell) = \{s_t^\ell\}_{t \geq 0}.$$

Example 1 *Table 4 gives coded data of a subject ℓ to p = 3 questions. Recall that coding of answer is 0 = unfavourable and 1 = favourable according to purpose. The coded answer of ℓ to Q_i at time t is denoted by $a_{i,t}^\ell$. From (Eq 4), we have $f_0^\ell = 3, f_1^\ell = 0$, and $f_2^\ell = 2$, so initial significance of ℓ is $y_0^\ell = 120$, with corresponding initial fitness $x_0^\ell = 111$. This corresponds to state index 24 in Fig 2(a). No answer changes at t = 1 so $y_1^\ell = y_0^\ell$ and $x_1^\ell = x_0^\ell$ and state transition from t = 0 to t = 1 is denoted by 24 → 24. Now at t = 2, Q_0 changes answer so we swap 0 and 1 to the right of y_1^ℓ and x_1^ℓ respectively (note that both are already on the right), and then change answer 1 to 0. Hence, $y_2^\ell = 120$ and $x_2^\ell = 110$, which corresponds to state 23. Columns 3 and 4 give the rest of*

Table 4. Coded data and orbit of a subject ℓ. The number $a_{i,t}$ is answer to Q_i at time t.

t	$a_{0,t}^\ell$	$a_{1,t}^\ell$	$a_{2,t}^\ell$	$x_t^\ell = x_0x_1x_2$	$y_t^\ell = i_0i_1i_2$	State index
0	1	1	1	111	120	24
1	1	1	1	111	120	24
2	0	1	1	110	120	23
3	0	1	0	100	102	29
4	0	1	1	101	102	30
5	0	1	1	101	102	30
6	1	1	1	111	120	24
7	0	1	1	110	120	23

doi:10.1371/journal.pone.0123812.t004

the fitness and significance states respectively of the orbit. Observe that ℓ has favourable answer to Q_1 for all times so its orbit $\mathcal{O}(\ell)$ stays in the subset

$$\begin{aligned} L(\subset S_3) &= \{(x,y) : x_0 = 1 \text{ and } i_0 = 1\} \\ &= \{21, 22, 23, 24, 29, 30, 31, 32\} \end{aligned} \tag{6}$$

where question $i_0 = 1$ is favourable. The longitudinal data of ℓ in S_3 is visualised as the orbit in Fig 2(a) with its time series illustrated in Fig 2(b).

Example 2 *The orbits of the three subjects in Table 3 in* S_3 *and over time are illustrated in Fig 3(a) and 3(b) respectively. Observe that orbit of ℓ stays strictly on the left half of S_3, and the other two on the right half. Subject ℓ is unfavourable in stable variable 0, while ℓ′ and ℓ″ are favourable in stable variable 0 and variable 2, respectively.*

Remark 1 *By the 0/1 coding of data, it is reasonable to suppose that the (concatenated) answers composed of unfavourable values 00· · ·0 is 'less fit' than the answer composed of favourable values 11· · ·1. By the weighting of variables, 'relative fitness' is made precise so that ordering of elements from the space $M_3 = \{0, 1\}^3$ of multivariate binary responses has meaning. Because more weight (significance) is given to the left-most position, we can then, for a fixed question order, write $x = \mathbf{0}10 < x' = \mathbf{1}00$, where most significant variable is unfavourable in x, and favourable in x′. For a fixed question order, we say that 100 is fitter than 010 (or 000, 001). The same argument holds in stating that 110 is less fit that 111.*

Using orbits, the complete *p*-dimensional information of each subject at any moment is coded to a point in the 2-dimensional discrete space S_p. No information in data is lost nor approximated as each subject's orbit has a one-to-one correspondence with the subject's original data. Clearly, question order of each subject at each time step is monitored. The ordering is selected as frequently changing variables are swapped to the far right (less significant digit of *y*), thus pushing slow changing variables to the left (significant digit of *y*). This 'swapping-changing-variable-to-the-right' process exposes clusters associated to stable variables.

Remark 2 *The time complexity of computation of orbits for* n *subjects and time T scales like pnT and is feasible for large data. We also note that there are admissible and non-admissible state transitions in S_p[9], e.g. in Fig 2, a transition that starts at 23 can only end at 23, 24, 29, and 31.*

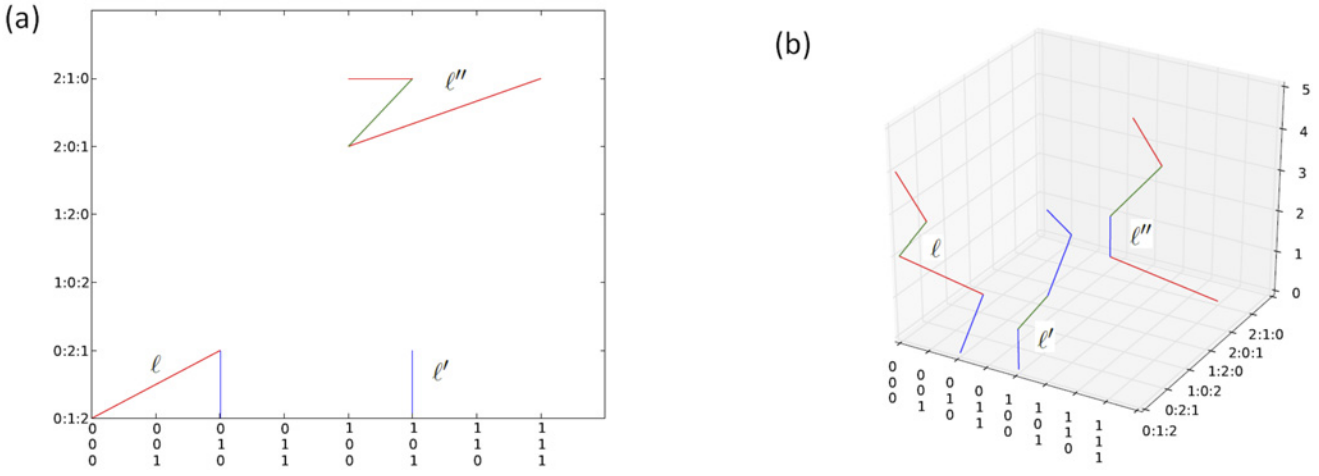

Fig 3. Orbit of the three subjects given in Table 3 (a) in S_3 and (b) over time.

doi:10.1371/journal.pone.0123812.g003

Remark 3 *The tendency of a subject to favour a particular state, or subset of states, is clustering in S_p. The strategy for choosing the initial question order in (Eq 5) places an orbit in its most likely position. This facilitates clustering and is useful for short data sets.*

Note that many households may share an edge (or orbit) in S_p. The following definitions are of interest regarding transitions in S_p.

Definition 4 *a. The accumulated number of transitions from state $s = (x, y)$ to $s' = (x' y')$ is called the **density** from s to s', denoted by $d(s, s')$.*

*b. The number of orbits at state s at time t is called the **capacity** of state s at time t and is denoted by $c_{s,t}$.*

Remark 4 *We can deduce correlation among variables from orbits in S_p. We explain for $p = 3$. Using Fig 4, if state transition of orbits most of the time stay in states 1 = (000, 210) and 48 = (111, 012) then we can test for positive correlation among the three variables 0, 1, and 2. If orbits spend most (if not all) of the time in a subset L of S_p such that $L \cong S_m$ for some $m < p$, then there is strong correlation between the first $p-m$ variables constant in S_m. In Fig 4 for example, if orbits stay strictly in the subset*

$$L = \{(x, y) : x = 11 * **, y = 01 * **\} \cong S_3,$$

then we can check for (positive) correlation between the first two variables 0 and 1. The asterisk $*$ in x (resp. in y) can take any binary value (resp. any question index except for i).

Remark 5 *The orbit method is not limited to binary data in p variables. For m-ary valued data, the space S_p is composed of same number p! of significance states but now with m^p fitness states. For the continuous case, data of each observation are binned, where bins are labelled from 0 to m−1 [18, 19]. For instance, binary coding can be done by assigning 0/1 if variable is above or below a given value, tertiary coding if the variable is in a good/neutral/bad range of values, and so on.*

Results

Orbit Results

Household orbits in S_5 for each of the four subpopulations are illustrated in Fig 4. The x-axis is composed of $2^5 = 32$ states but not all the 5! = 120 states in the y-axis are shown. There are no transitions between the four subpopulations as they are associated to constant variables. Recall that a red edge is used to denote a transition that goes from right to left on the next time step, a green edge for a transition that goes from left to right, and a blue edge a transition that goes to the same fitness state. The percentage of unfavourable answers for each question in each of the four subpopulations is given in Table 5 while the frequencies of answer change are given in Table 6. The frequencies of change for Q_0 (HH) and Q_1 (HS) are low, while Q_4 (IF) is the highest. This means that there is stability in the variables HH and HS in that most subjects will stay in the region where significance is either $y = 01i_2\, i_3\, i_4$ or $y = 10i_2\, i_3\, i_4$, $i_j \in \{2, 3, 4\}$. In addition, there is high activity of the IF variable, which means few transitions where $y = 4i_1\, i_2\, i_3\, i_4$, $i_j \in \{0, 1, 2, 3\}$. All of this is recognized in Fig 4.

There are immediately regions of interest in Fig 4. As an aide in interpreting regions in S_5, we present in Fig 5 the subsets of S_5 determined by the first significant variable i_0. A subject ℓ that spends most of its time in the region where $x = j****$, $y = i****$ means that answer of ℓ to Q_i is least frequently changing, with answer = j. The initial state of ℓ is chosen to lie in this

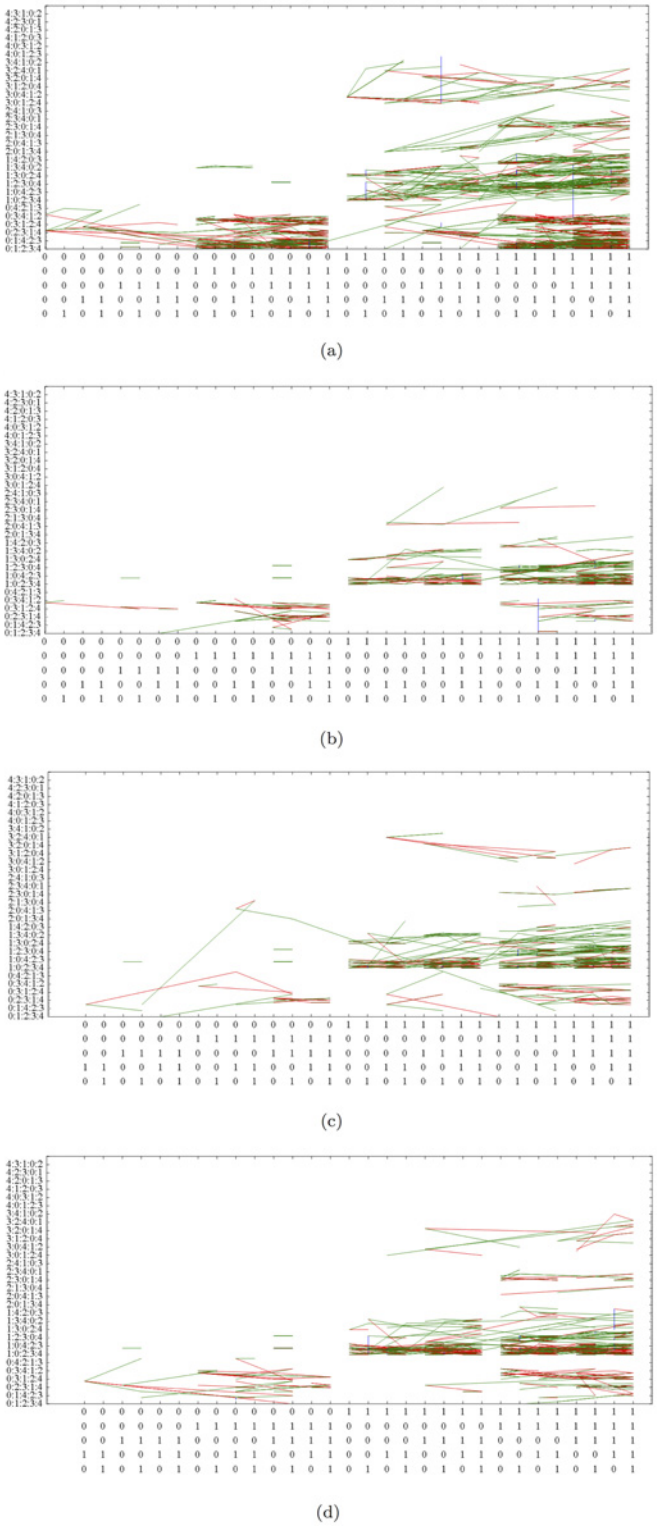

Fig 4. Orbits of (a) SA Rich, (b) SA Poor, (c) MOZ Rich, and (d) MOZ Poor households. Observe clusters formed in regions of each subpopulation. variable 4 (influx) is most frequently changing in all four subpopulations so orbits do not stay in the region where y = 4****. Not all 5! significance states are shown.

doi:10.1371/journal.pone.0123812.g004

Table 5. Percentage of unfavourable = 0 responses in Q_i for each of the four subpopulations.

SA Rich	Q_0 : 17.4%	Q_1 : 11.4%	Q_2 : 20.5%	Q_3 : 21.1%	Q_4 : 22.2%
SA Poor	Q_0 : 33.4%	Q_1 : 10.8%	Q_2 : 13.3%	Q_3 : 44.5%	Q_4 : 23.6%
MOZ Rich	Q_0 : 27.6%	Q_1 : 7.8%	Q_2 : 19.4%	Q_3 : 38.0%	Q_4 : 19.0%
MOZ Poor	Q_0 : 33.5%	Q_1 : 10.3%	Q_2 : 15.0%	Q_3 : 51.4%	Q_4 : 23.5%

doi:10.1371/journal.pone.0123812.t005

region (Definition 5). For example, a subject that stays in the region

$$R_{HH<40} = \{p = (x, y) \in S_5 : x = 0 * * * *, y = 0 * * * *\}$$

of Fig 5 frequently experiences younger (<40) household head age.

Regions in Fig 5 may be further analysed. A more detailed description of the regions HH≥40, HH<40, HS≥3, and HS<3 is illustrated in Fig 6. In each sub region, the two variables i_0 and i_1 are significant (i.e. less frequently changing answers), with i_0 being more significant. Of course these sub regions may be further subdivided.

In general, we say that a variable i is *stable* if orbits cluster in a subset of S_p determined by first significant variable i. Regions that are never visited (e.g. those associated to variables IF^+, IF^-, and A<M in Fig 4) are termed *holes*. Clusters are contained in regions where the leading significant variable is stable while holes are contained in regions with high activity of the leading variable. By visual inspection of orbits in S_p, we can immediately detect stable variables (via clusters) and unstable variables (via holes).

Clusters in the right half regions of S_p are fitter than clusters located on the left half of S_p as they are associated to stable leading variable with favourable condition. From the orbits of subpopulations in Fig 4, observe that there are no transitions between the left and right half of S_5 in both SA Poor and MOZ Poor subpopulations. This is verified by Fig 7, the orbits of the four subpopulations, in time. In addition, columnar structures over clusters correspond to variables that are stable over the survey period. Although there are few transitions between clusters, there is considerable activity within each. Household orbits in one cluster may then reasonably be analysed independently of households in other clusters. The time series representation of orbits reveals idle behaviour (sequence of vertical blue edges) that are not always visible in orbits in S_5.

As observed in Fig 4, some regions in a subpopulation appear denser than those of the other subpopulations (e.g. the region HH<40 appears to be more dense in MOZ poor than in MOZ rich, with the opposite phenomenon for the SA population). We use histograms to denote the

Table 6. Questions with corresponding frequency of answer change in each of the four subpopulations.

	SA Rich	SA Poor	MOZ Rich	MOZ Poor
Q_0 (HH)	374	78	110	102
Q_1 (HS)	466	53	50	85
Q_2 (HD)	1280	164	216	309
Q_3 (AM)	1334	308	299	405
Q_4 (IF)	3373	580	599	971

doi:10.1371/journal.pone.0123812.t006

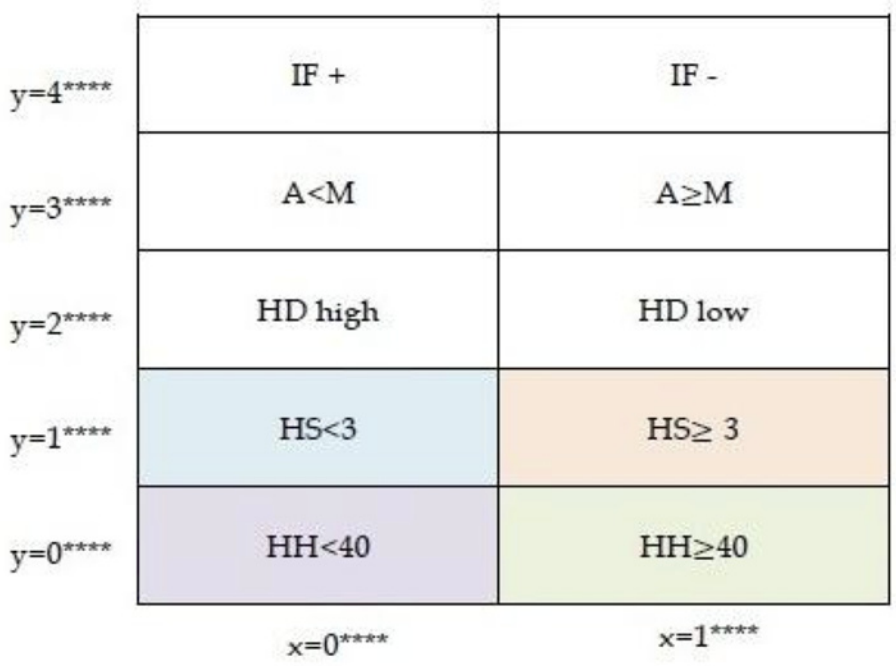

Fig 5. Regions in S_5 determined by the first significant variable. Observed units that often stay in a region determined by one significant variable often experience the property of that region.

doi:10.1371/journal.pone.0123812.g005

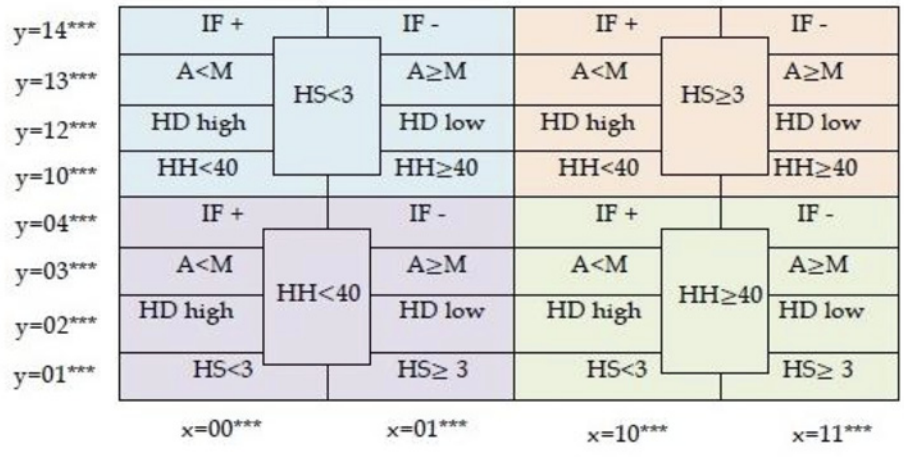

Fig 6. Regions in S_5 determined by the first and second significant variables.

doi:10.1371/journal.pone.0123812.g006

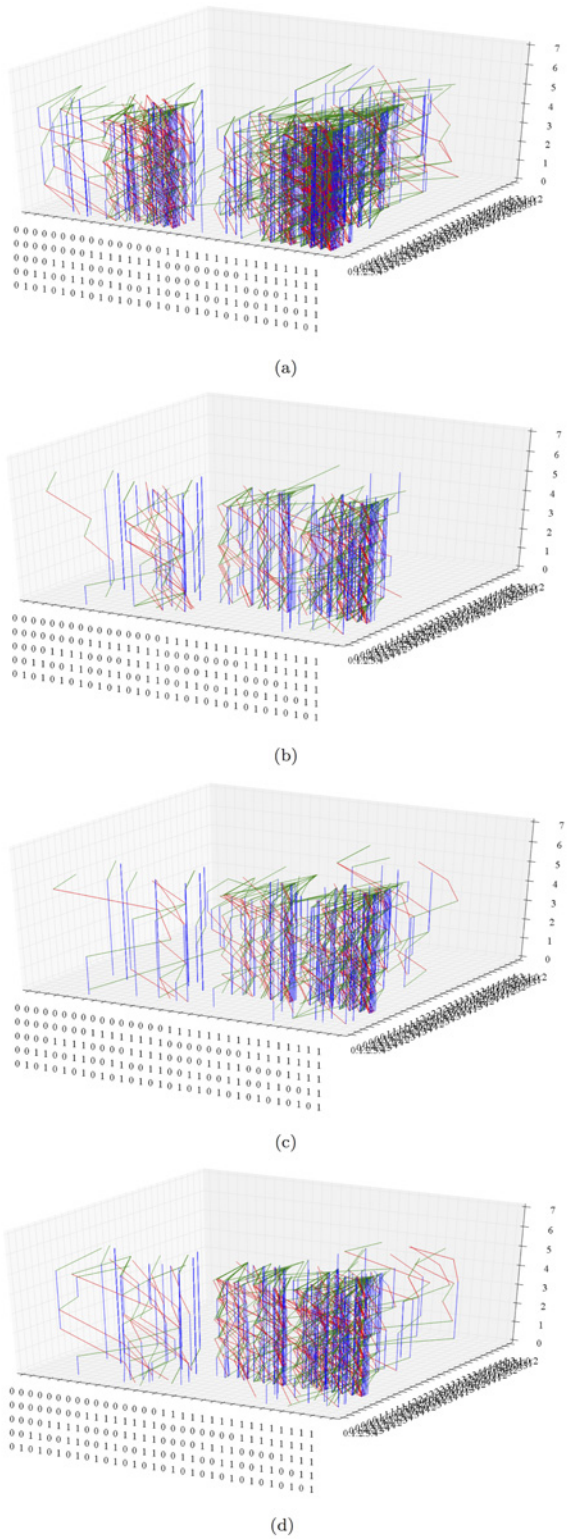

Fig 7. Column structures over clusters of (a) SA Rich, (b) SA Poor, (c) MOZ Rich, and (d) MOZ Poor household orbits in S_5.

doi:10.1371/journal.pone.0123812.g007

accumulated number of visits (i.e. capacity, Defn. 4) to each state in S_5. Fig 8 gives the accumulated capacity in states of S_5, represented by the height of bars. It is immediately noted that there are regions of high and low numbers. Density at each state at each time step can also be computed, and can be represented by bubbles. Fig 9 illustrates this case for the SA Rich subpopulation.

Fig 10 gives the percentages of visits in regions determined by one and two significant variables. The regions with no percentages are holes. The largest percentage in SA Rich is in the subregion

$$R_{HH\geq 40,HS\geq 3} = \{p = (x, y) : x = 11 * **, y = 01 * **\}$$

associated to older household head and larger household size (62%), with household head more stable. For the other three subpopulations, the largest percentage is in the subregion

$$R_{HS\geq 3,HH\geq 40} = \{p = (x, y) : x = 11 * **, y = 10 * **\},$$

associated again to older household head and larger household size, but with household size more stable.

Remark 6 *Population-level information is visible, but detailed individual information may be lost in the cluster. We may zoom into regions of interest (e.g. regions of high percentage of visit) to unclutter the display, as in the SA Rich region* $R_{HH\geq 40,\ HS\geq 3}$ *illustrated in Fig 11. As for individual orbits, of interest in Fig 4 are those that seem to be 'outliers'. They can further be analysed e.g. by using interactive techniques such as focusing and brushing, as in dynamic parallel component plots [20].*

Regarding Remark 6, we can further analyse orbits from the SA Rich subpopulation. Fig 12 shows dominant accumulated number of transitions ≥ 100 from state $s = (x, y)$ to $s' = (x', y')$ in SA Rich (i.e. density $d(s, s') \geq 100$). Most transitions 'idle' at state (11111, 01234) and correspond to household orbits that are constantly favourably in the five variables. As for non-idling transitions, it is dominant between states s = (11110, 01234) and s' = (11111, 01234) and involve change in variable 4 (IF) where $s \rightarrow s'$ indicate negative influx, and $s' \rightarrow s$ is non-negative influx.

Fig 13 shows the corresponding state indices for the subset $R_{HH\geq 40,\ HS\geq 3}$ of S_5 given in Fig 11. The capacity in SA Rich households at each time step for states with $c_{s,\,t} \geq 50$ is illustrated in Fig 14. We have the following observations:

1. The capacity at state 48 = (11111, 01234) is dominant. This is expected as most orbits idle in this state, as given by the numbers in Fig 12.
2. The capacity graphs for state pairs 48 and 47 = (11110, 01234), and 39 = (11110, 01243) and 40 = (11111, 01243), behave inversely and are almost symmetrical. Note that transition between state pair 47 and 48, and 39 and 40, are associated to change in variable 4 (IF) and 3 (AM) respectively. It is expected that capacity increase in 48 (more individuals migrating into households) result in decrease of capacity in 47. The same argument goes for exchange in capacity of states 39 and 40.
3. Transition between 23 = (11110, 01342) and 24 = (11111, 01342) are associated to change in variable 2 (HD). The capacity graph of 23 (HD = 0) is always above 24, except at t = 2007. The sharp increase in 24 (low household death) at this time corresponds to a drop in 23.

Results Regarding Purpose

We particularly use Fig 8 to draw conclusions with regard to our Purpose as stated in (Eq 1).

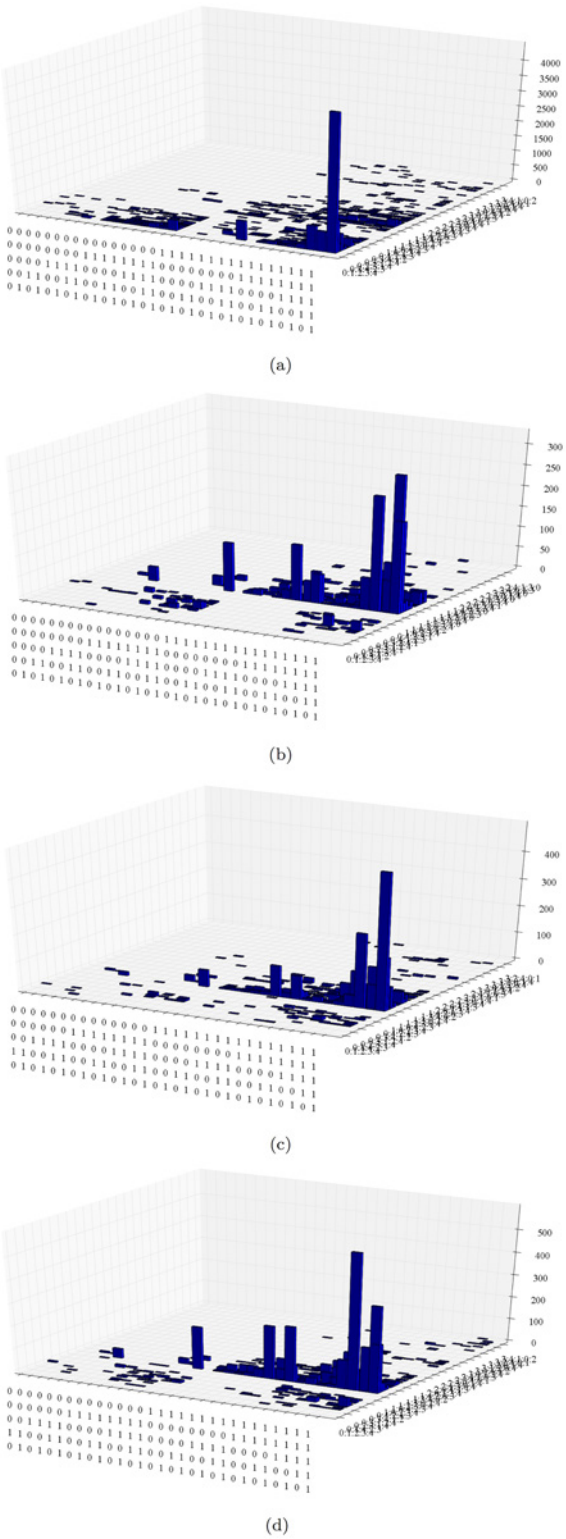

Fig 8. Accumulated number of visits (height of bars) in S_5 of (a) SA Rich, (b) SA Poor, (c) MOZ Rich, and (d) MOZ Poor household orbits.

doi:10.1371/journal.pone.0123812.g008

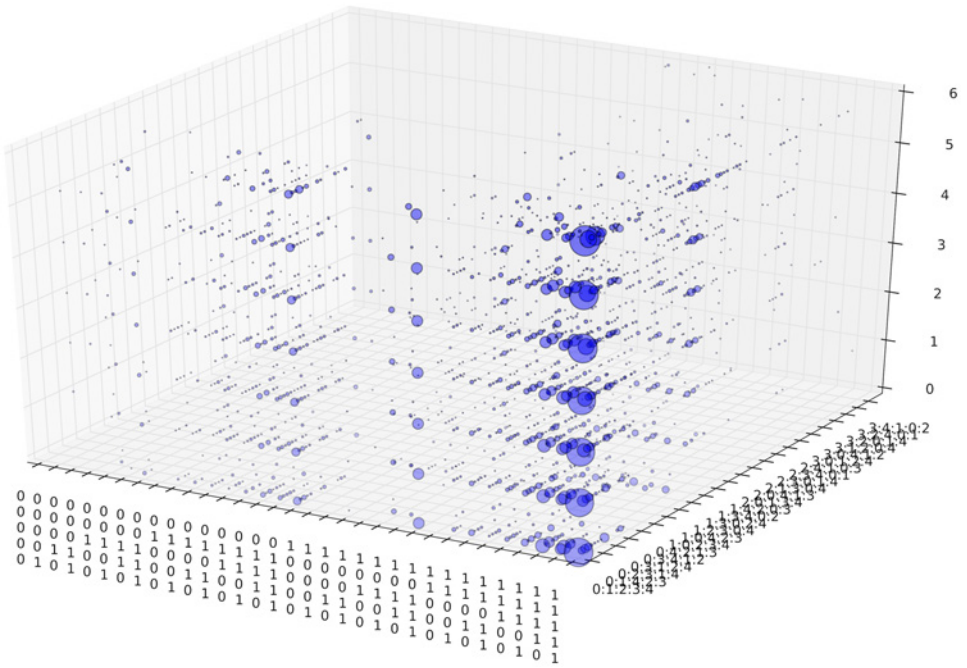

Fig 9. Capacity at each state and each time step in the SA Rich, represented by bubbles.

doi:10.1371/journal.pone.0123812.g009

1. There is one dominant peak in SA Rich. This occurs at the fully fit state (11111, 01234), where it is most stable in $Q_0 = 1$ (HH $\geq$ 40), followed by $Q_1 = 1$ (HS $\geq$ 3), and so on. For SA Poor and MOZ Rich/Poor we find fully fit states at (11111, 10234) characterized by stability of HS $\geq$ 3, followed by HH $\geq$ 40. We then associate (i) households headed by older adults to larger $\overline{APS}$, and (ii) larger households with lower $\overline{APS}$.
2. The peaks for SA Poor, MOZ Rich, and MOZ Poor are at states

$$\begin{array}{lll} \text{i.}(11111, 10234) & \text{iii.}(11111, 10243) & \text{v.}(10111, 10234) \\ \text{ii.}(11101, 10234) & \text{iv.}(11110, 10243) & \text{vi.}(10101, 10234) \\ & & \text{vii.}(01110, 10234) \end{array} \tag{7}$$

   For spikes at states ii., iv., and vi., unfavourable answer is either in Q_0 or Q_3 (i.e. HH<40 or A<M). We then associate young household heads and less adults to minors to poorer (i.e. not Rich SA) households. Now spike at state vii. which is unfavourable in Q_1 and Q_4 (HS<3 and IF$^+$) is also associated to poorer households. The condition of small households should be examined.

3. Spikes at states ii., iv., v., vi., and vii. are identified with relatively stable unfavourable states HH<40, A<M, and HS<3 with IF$^+$. We then associate absence of visits to these states with SA Rich, and their presence with the other three subpopulations.

4. For the two dominant peaks at states i. and ii. in MOZ Rich in Fig 8(c), A $\geq$ M has a higher peak than A<M. This is reversed in MOZ Poor in Fig 8(d). We associate MOZ Poor with a stable, dominant population of households with small adult component.

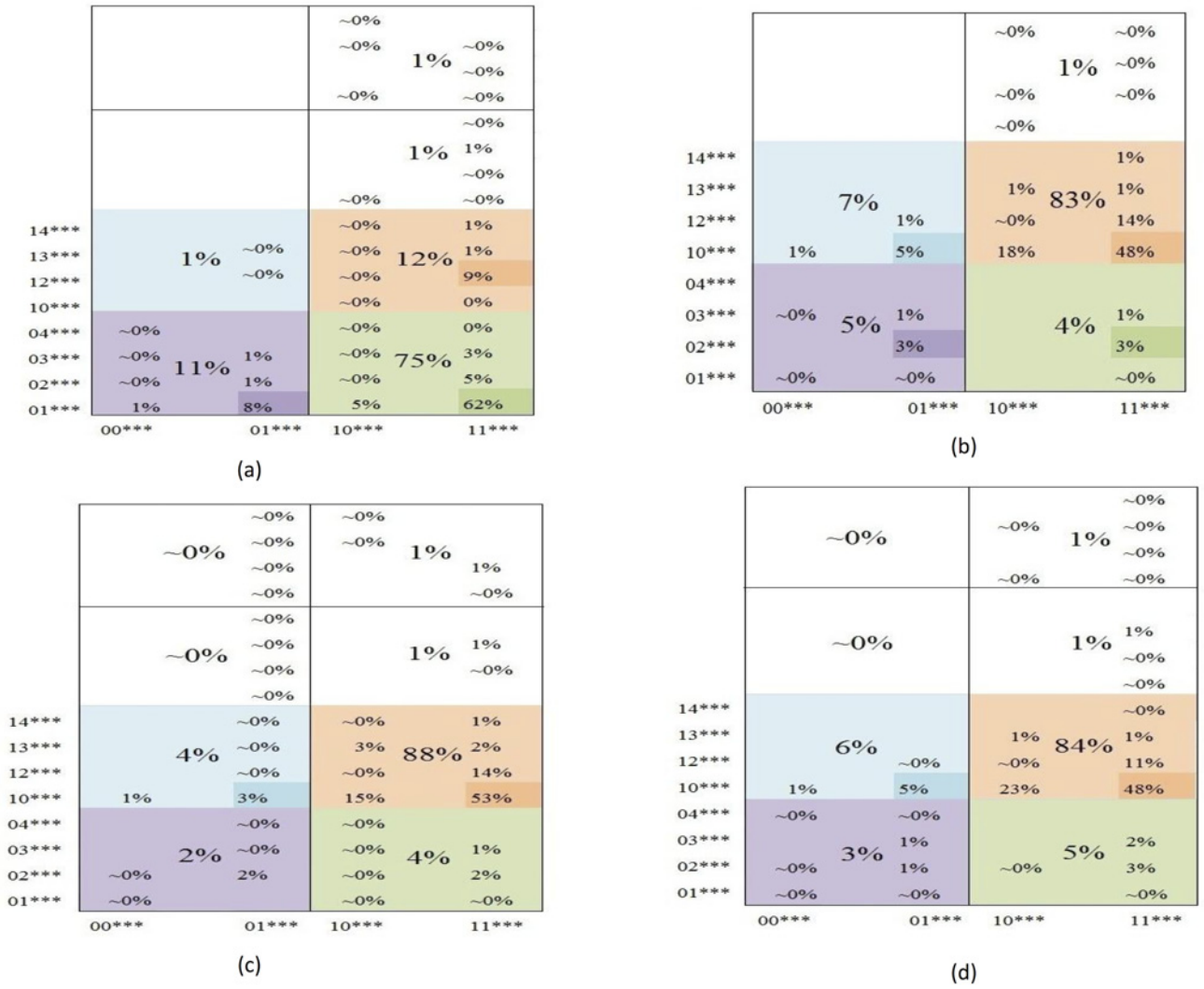

Fig 10. Percentage of visits of (a) SA Rich, (b) SA Poor, (c) MOZ Rich, and (d) MOZ Poor household orbits to regions in S_5 determined by the first and second significant variables.

doi:10.1371/journal.pone.0123812.g010

Other Methods of Binary Multivariate Longitudinal Data Analysis

We discuss the use of other conventional methods in analysis of BMLD and mention the advantage of using orbits.

Markov Chain Model. For p binary variables, Markov chain models considers the analysis of change over time measures in $M_p = \{0, 1\}^p$. Question order is arbitrarily fixed and a $2^p \times 2^p$ matrix of transition probabilities is constructed [3]. If a fixed question order alone is used for all times and for all subjects (say $012\cdots(p-1)$) in analysing binary multivariate longitudinal data, then some information (e.g. clusters and holes) may not be revealed as orbits overlap in a single row (question order) of S_p. For example, the six clusters visible in Fig 4(a) are not resolved in Markov analysis. This phenomenon of 'unfolding' states from a general case of a fixed question order is an advantage in analysing orbits in S_p. Given the fundamental weighting

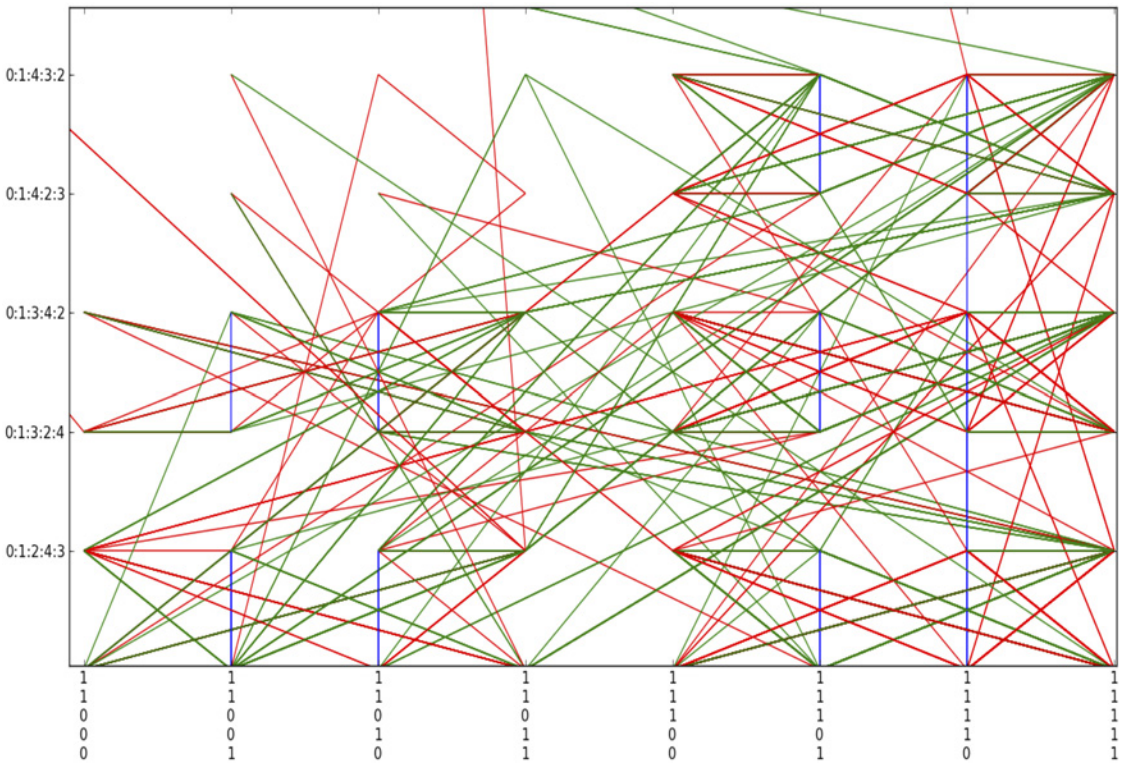

Fig 11. Orbits in SA rich population that cluster in the region $R_{HH \geq 40,\ HS \geq 3}$ where both household head and household size are favourably constant. This is the zoomed region in Fig 4(a) with high percentage of household visit.

doi:10.1371/journal.pone.0123812.g011

x	y	x'	y'	d(s,s')
11111	0:1:2:3:4	11111	0:1:2:3:4	3270
11111	0:1:2:4:3	11111	0:1:2:4:3	670
10110	0:1:2:3:4	10110	0:1:2:3:4	552
11101	0:1:2:3:4	11101	0:1:2:3:4	511
11110	0:1:2:3:4	11111	0:1:2:3:4	509
11111	0:1:2:3:4	11110	0:1:2:3:4	408
11110	0:1:3:4:2	11110	0:1:3:4:2	382
11110	0:1:2:4:3	11110	0:1:2:4:3	342
11101	0:1:3:2:4	11101	0:1:3:2:4	278
01111	0:1:2:3:4	01111	0:1:2:3:4	247
11111	0:1:3:4:2	11111	0:1:3:4:2	218

x	y	x'	y'	d(s,s')
11110	0:1:2:4:3	11111	0:1:2:4:3	208
11111	1:2:3:4:0	11111	1:2:3:4:0	157
11111	0:2:3:1:4	11111	0:2:3:1:4	151
11110	0:1:3:4:2	11111	0:1:3:4:2	140
11100	0:2:3:4:1	11100	0:2:3:4:1	140
11111	0:1:3:2:4	11111	0:1:3:2:4	128
11011	1:2:3:4:0	11011	1:2:3:4:0	117
11100	0:1:3:2:4	11101	0:1:3:2:4	113
10100	0:1:3:4:2	10100	0:1:3:4:2	109
11101	0:1:3:2:4	11111	0:1:3:4:2	108
01101	0:1:2:3:4	01101	0:1:2:3:4	102

Fig 12. Densities $d(s, s') \geq 100$ from state $s = (x, y)$ to $s' = (x', y')$ in SA Rich households. Highlighted lines are self-transition, i.e. $s = s'$.

doi:10.1371/journal.pone.0123812.g012

0:1:4:3:2	1	2	3	4	5	6	7	8
0:1:4:2:3	9	10	11	12	13	14	15	16
0:1:3:4:2	17	18	19	20	21	22	23	24
0:1:3:2:4	25	26	27	28	29	30	31	32
0:1:2:4:3	33	34	35	36	37	38	39	40
0:1:2:3:4	41	42	43	44	45	46	47	48
	11000	**11001**	**11010**	**11011**	**11100**	**11101**	**11110**	**11111**

Fig 13. State indices associated to states s_i in the subset $R_{HH \geq 40,\ HS \geq 3}$ of S_5 where both household head and household size are favourably constant.

doi:10.1371/journal.pone.0123812.g013

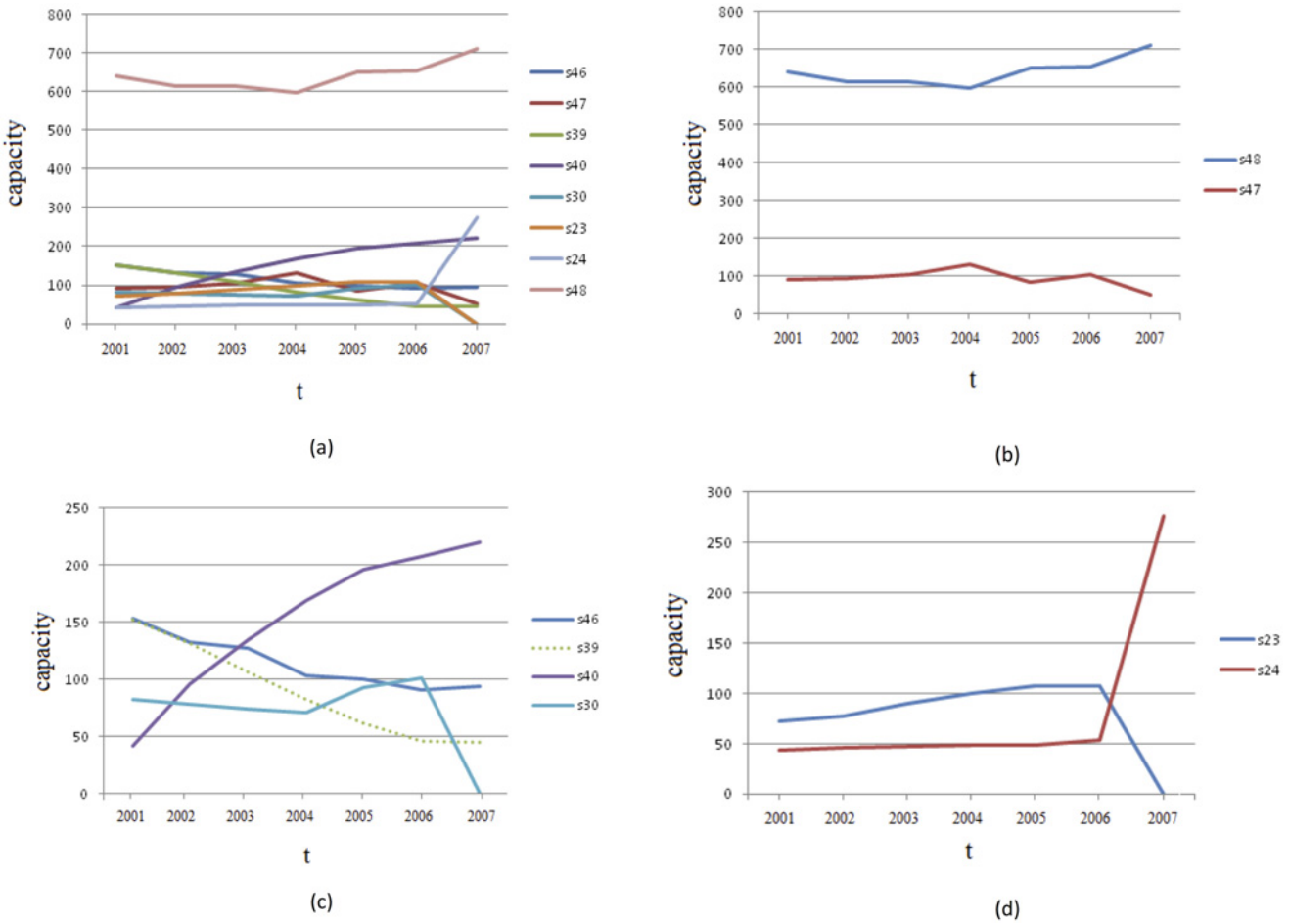

Fig 14. Number of SA Rich household orbits (capacity) at each time step in states (a) 23, 24, 30, 39, 40, 46, 47, and 48, (b) 47 and 48 associated to variable 4 (IF), (c) 39 and 40 associated to variable 3 (AM); 30, and 46, and (d) 23 and 24 associated to variable 2 (HD).

doi:10.1371/journal.pone.0123812.g014

Table 7. Variance Inflation Factor.

Variable	VIF	1/VIF
Q_1	10.18	0.0982
Q_4	7.14	0.1340
Q_2	4.73	0.2115
Q_0	3.24	0.3087
Q_3	2.69	0.3714
Mean VIF	5.60	

doi:10.1371/journal.pone.0123812.t007

by frequency of change of variables, it is of great interest that S_p is the space of all possible states to which subjects can change, and also captures change of significant variables. By prioritising slowly changing variables, orbits give a natural spatial ordering of states in S_p by fitness.

Generalized Estimating Equation Model. To compare the performance of the conventional statistical model to the deterministic orbit approach we have adopted a generalized estimating equation (GEE) population modelling approach. In [21], the estimation-equation approach is proposed for population average models. It is argued that in general, mixed models involve unverifiable assumptions on the data-generating distribution resulting to potentially misleading estimates and biased inference. We use the quasi-information criterion (QIC) to identify the best working correlation structure to be used for our data [22]. Maximum likelihood based model selection methods, such as the widely used Akaike Information Criterion (AIC), are not directly applicable to the GEE approach [23]. The exchangeable correlation structure proved to be the best when fitted to our data.

Before presenting the GEE model, we note that with regards to the correlated indicators, there is potential co-linearity between the household size and certain other covariates. This is suggested by the marginally high variance inflation factor (VIF) for this variable (Q_1) of $\sim$ 10 in Table 7. Further, the spearman rank correlation coefficient of 0.68 between Q_1 (household size) and Q_4 (influx) in Table 8 would be cause for further concern. Removing the co-linear effect of Q_1, the GEE model for a binary outcome (APS = 0/1) using a binomial family, logit link function and an exchangeable correlation structure is given in Table 9. The VIF without Q_1 is given in Table 10.

Remark 7 *The GEE model shows that HH $\geq$ 40, HS $\geq$ 3, HD low, A $\geq$ M, IF$^-$, and HN = SA are more likely in the rich households. This is consistent with our favourable/unfavourable orbit coding to APS. In addition, the model also informs us that variables associated to holes (not just clusters) in S_p should also be analysed. In particular, the Q_4 (IF) variable (associated to holes)*

Table 8. Spearman's Rank Correlation Coefficient.

	Q_0	Q_1	Q_2	Q_3	Q_4
Q_0	1.0000				
Q_1	0.0251	1.0000			
Q_2	-0.0226	-0.0231	1.0000		
Q_3	0.1068	-0.2919	-0.0302	1.0000	
Q_4	0.0381	0.6787	-0.0114	-0.1908	1.0000

doi:10.1371/journal.pone.0123812.t008

Table 9. GEE Model removing the co-linear effect of Q_1.

GEE population-averaged model				Number of obs	=	22270
Group variable:		hh_id		Number of groups	=	5567
Link:		logit		Obs per group: min	=	4
Family:		binomial		avg	=	4.0
Correlation:		exchangeable		max	=	5
				Wald chi2(5)	=	483.36
Scale parameter:		1		Prob > chi2	=	0.0000
Q_6 (APS = Rich)	Odds Ratio	Std. Err.	*z*	*P* > |*z*|	[95% Conf. Interval]	
Q_0 (HH $\geq$ 40)	1.2576	0.0315	9.14	0.000	1.1973	1.3209
Q_2 (HD low)	1.0118	0.0338	0.35	0.724	0.9477	1.0803
Q_3 (A $\geq$ M)	1.3439	0.0346	11.48	0.000	1.2777	1.4134
Q_4 (IF$^-$)	1.2846	0.0375	8.58	0.000	1.2132	1.3603
Q_5 (HN = MOZ)	0.7615	0.0180	-11.51	0.000	0.7270	0.7977
_cons	0.7559	0.0345	-6.13	0.000	0.6912	0.8266

doi:10.1371/journal.pone.0123812.t009

and Q_3 (AM) variable (associated to very few transitions) appear to be statistically significant and associated to households above the absolute poverty line.

Motion Charts and Heat Maps. A motion chart is a dynamic bubble chart that enables the display of large multivariate data with large number of data points [6]. The central object in motion charts is a blob, or in general a 2-dimensional shape, which represents one entity from the dataset. This allows for visualization of the data by using additional dimensions (e.g. time, size and color of the blobs) to show different facets of the data. The dynamic appearance of the data in a motion chart facilitates visual inspection of associations, patterns and trends in multivariate datasets. The problem with motion charts is that for many variables, there is not enough dimensions (e.g. size, shape, color, etc.) to represent different entities. The advantage of using orbits is that adding more variables is easily accommodated by the increase in the number of fitness and significance states in the change space S_p. Fig 15 show the proportion of households (by nationality and fitness sequence) above poverty line over the survey period while Fig 16 shows the proportion of households above poverty line by nationality and time, i.e. (HN, *t*), where HN = 0 = SA, HN = 1 = MOZ, and *t* = 2001, 2003, 2005, 2007. The labeling of the fitness states along the *x*-axis is given in Table 11. While this gives a sense of where more households fall with regards to relative poverty probability (stratified by household nationality and then by

Table 10. Variance Inflation Factor (removing Q_1).

Variable	VIF	1/VIF
Q_2	4.43	0.2257
Q_0	3.27	0.3059
Q_4	3.24	0.3089
Q_3	2.66	0.3762
Q_5	1.37	0.7301
Mean VIF	2.99	

doi:10.1371/journal.pone.0123812.t010

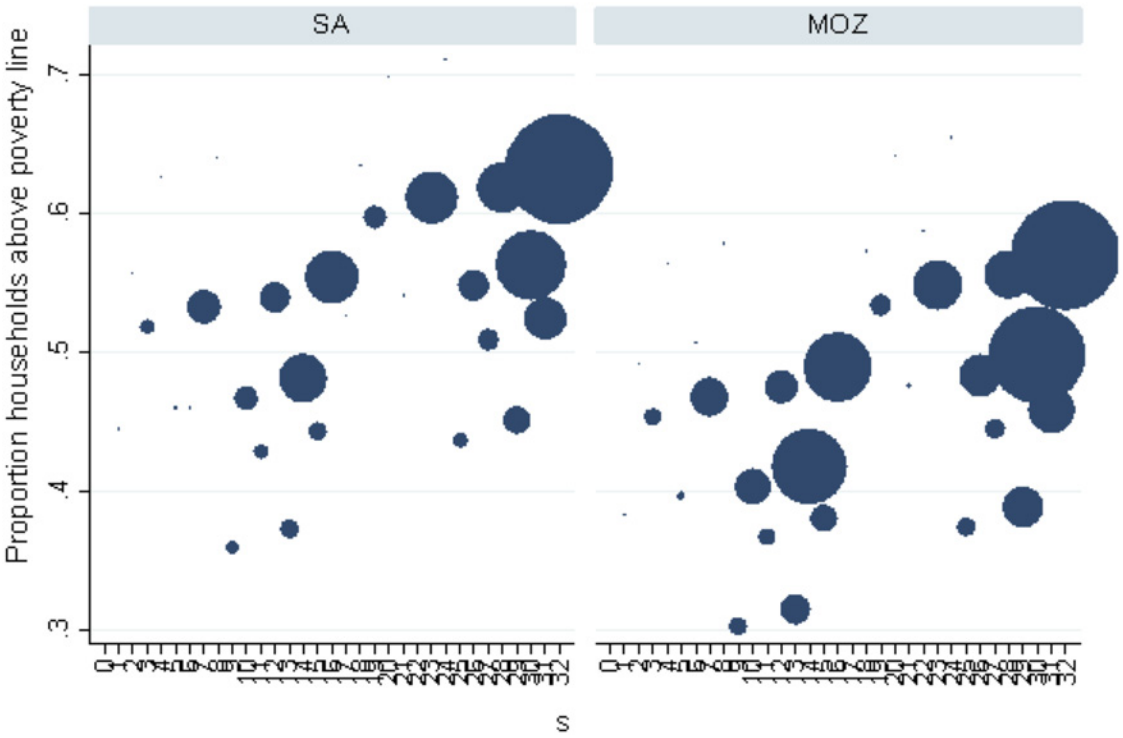

Fig 15. Proportion of households (by nationality and fitness sequence) above poverty line over the survey period 2001–2007.

doi:10.1371/journal.pone.0123812.g015

nationality and poverty line status classification), it does not convey the changing trajectory of households with time.

The heat map approach illustrated in Fig 17 reflects the observed proportion above the poverty line (by nationality) represented by the amplitude to graph at the point (x, y), where the fitness sequences are on the y axis and the 4 year time points (1 = 2001, 2 = 2003, 3 = 2005, 4 = 2007)are on the x-axis. While some differences can be observed by nationality, the clearer visualisation offered by the orbit approach is evident in our opinion. The heat map approach is not without merits (one being easy to implement) and would require more extensive and detailed application to longitudinal data such as ours to fully surmise its utility relative to the deterministic orbit approach.

Discussion

Using variables pertaining to socio-economic determination, we have illustrated via 2-dimensional orbits the dynamics and patterns of 4 subpopulations in the AHDSS. Stable and unstable variables (in terms of frequency of change) have been identified. The high frequency of change of IF variable (Q_4) in each of the four subpopulations intuitively, is an unfavorable phenomenon because it directly measures instability of household numbers i.e. the rapid flow of individuals into and out of households, within the community. Policies that might stabilize this phenomenon are of interest.

The value of using the method of orbits for analysis of binary multivariate longitudinal data is that it gives a picture of how subjects and the population behave. There are no known methods that show exact visualisation of such data. Orbits can be used as an additional tool for say demographers and social scientist in analysis of data. An additional value of the method is to

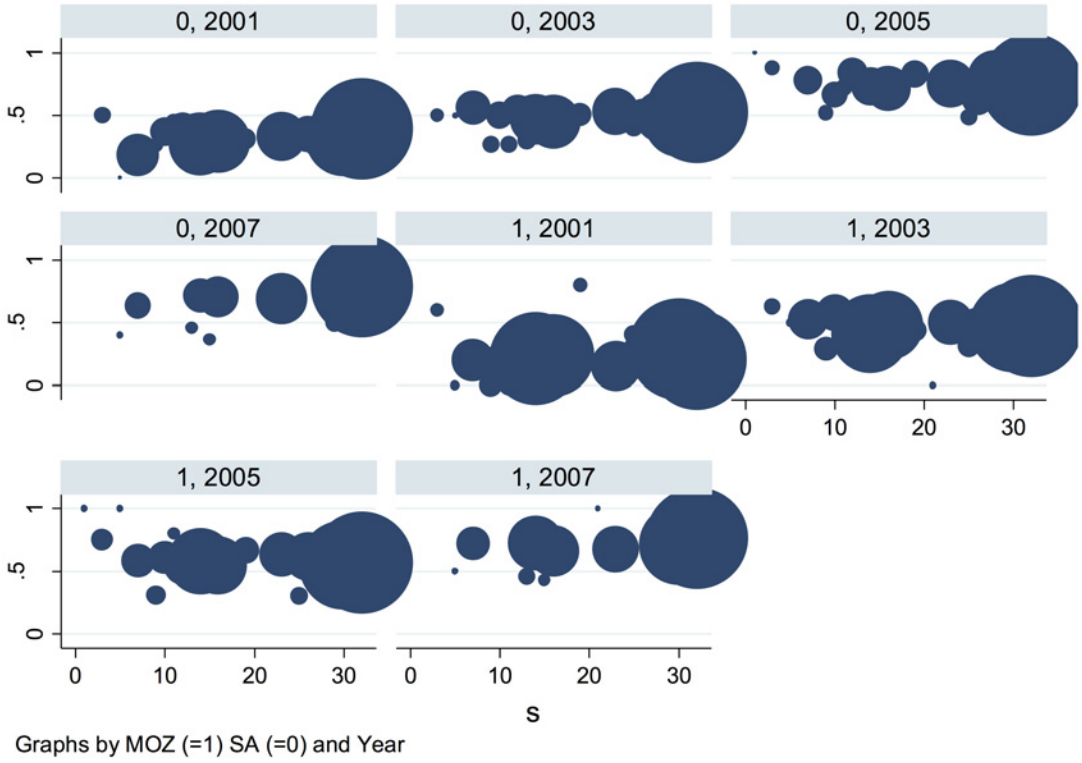

Fig 16. Proportion of households above poverty line by nationality at time (HN, *t*), where HN = 0 = SA, HN = 1 = MOZ, and *t* = 2001, 2003, 2005, 2007.

doi:10.1371/journal.pone.0123812.g016

give insight into possible cause and effect. Presentation of longitudinal data as a time-evolving geometric orbit naturally enables visual identification of possible cause and effect along the orbit (e.g. if only state *i* precedes *j*, then state *i* causes *j*). Using orbits for longitudinal data analysis is fundamentally different from conventional longitudinal statistical models in that it develops visible orbits for fitness states and therefore extracts more information from the data. For instance, the standard statistical model does not give a visual sense of the density of households in a given state, rather just the magnitude of association (odds ratio).

Table 11. Label for answer combination associated to Fig 15 and 16.

s	Answer	s	Answer	s	Answer	s	Answer
1	00000	9	01000	17	10000	25	11000
2	00001	10	01001	18	10001	26	11001
3	00010	11	01010	19	10010	27	11010
4	00011	12	01011	20	10011	28	11011
5	00100	13	01100	21	10100	29	11100
6	00101	14	01101	22	10101	30	11101
7	00110	15	01110	23	10110	31	11110
8	00111	16	01111	24	10111	32	11111

doi:10.1371/journal.pone.0123812.t011

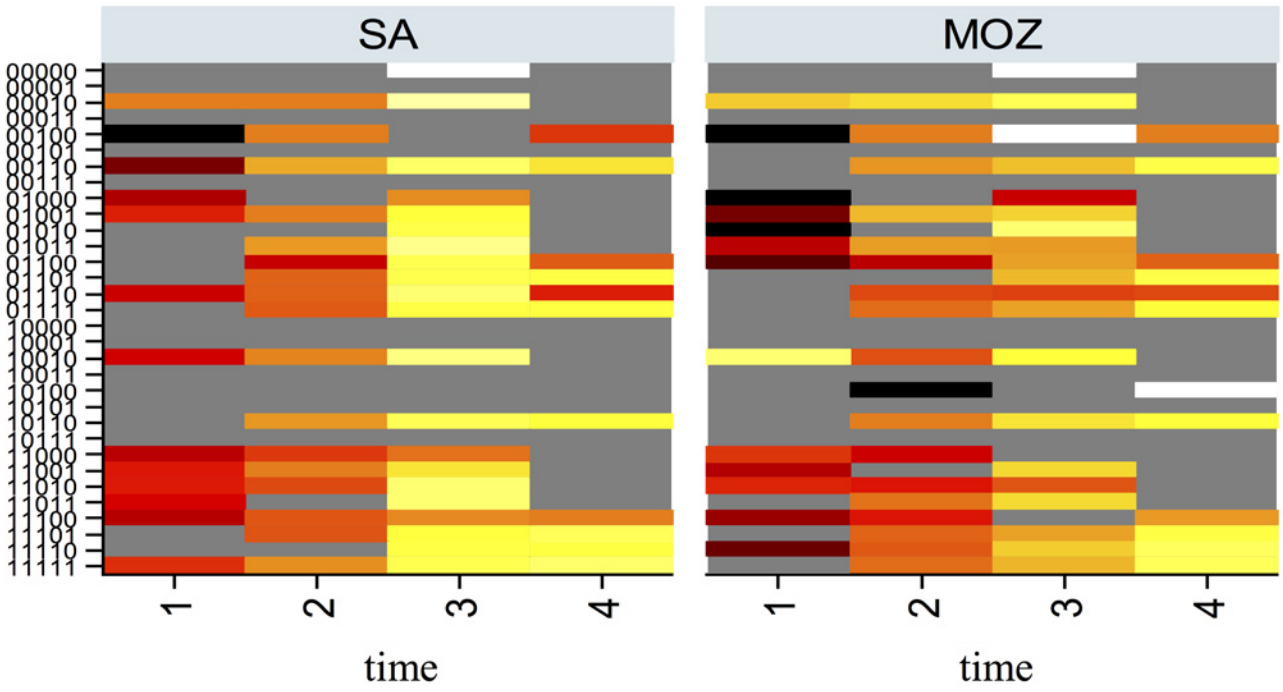

Fig 17. Heat map representing the observed proportion (density) above the poverty line (by nationality) and times 1 = 2001, 2 = 2003, 3 = 2005, 4 = 2007.

doi:10.1371/journal.pone.0123812.g017

One obvious limitation in using orbits is that it considers only complete data. Extending the method to accommodate missing data is necessary. Tools for (demographic) estimation from limited, deficient and defective data [24] may be used, where longitudinal data does not satisfy the assumption that there is no missing data, or that each variable and each subject is measured at the same times.

The primary confounder we included and stratified on in this analysis was household head nationality. Previous papers [12, 14, 16] on socio-economic status in Agincourt have identified the proximal importance of household nationality as a determinant/confounder for socio-economic status/poverty. Our GEE regression results confirm the importance of this confounder as a determinant of poverty status. As for potential confounders of socio data-economic determination such as occupation and income, they are rarely tracked in the Agincourt HDSS. In addition, given the large amount of missing data for these variables, we would not have be able to apply the orbit theory to the key indicators in the manner presented currently. Within our study period from 2001–2007, the education modules was run only in 2002 and 2006 i.e. not directly captured in the same time points. Mozambicans generally have a significantly lower number of education years compared to South Africans (e.g. [14]) so we believe the nationality would also capture any confounding effects of education status. However we cannot discount any residual confounding influence of occupation, income, and education on our results.

Supporting Information

S1 Dataset. Binary data of the four subpopulations SA Rich, SA Poor, MOZ Rich, and MOZ Poor.
(RAR)

Acknowledgments

The authors would like to thank the referees for their valuable comments. The data used in this study was supplied by the MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt). The Agincourt HDSS is funded by the Medical Research Council and University of the Witwatersrand, South Africa, Wellcome Trust, UK (grant no. 058893/Z/99/A, 069683/Z/02/Z, 085477/Z/08/Z), and National Institute on Aging of the NIH (grants 1R24AG032112-01 and 5R24AG032112-03).

Author Contributions

Wrote the paper: MVV DS. Wrote, prepared, and proofread the manuscript: MVV. Helped write and proofread the manuscript: DS. Performed orbit analysis of the data: MVV DS. Gathered data and performed detailed statistical analysis of the data: BS. Generated orbits and constructed the GUI/software for orbit visualization: FC.

References

1. Bartholomew D, Steele F, Moustaki I, Galbraith J. Analysis of multivariate social science data. 2nd ed. Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences; 2008.
2. Bandyopadhyay S, Ganguli B, Chatterjee A. A review of multivariate longitudinal data analysis. Statistical Methods in Medical Research. 2011; 20(4): 299–330. doi: 10.1177/0962280209340191 PMID: 20212072
3. Gottschau A. Markov chain models for multivariate binary panel data. Scandinavian Journal of Statistics. 1994; 21(1): 57–71.
4. Ilk O. Multivariate longitudinal data analysis: Models for binary response and exploratory tools for binary and continuous response. VDM Verlag; 2008.
5. Zeng L, Cook R. Transition models of multivariate longitudinal binary data. Journal of the American Statistical Association. 2007; 102(477): 211–223. doi: 10.1198/016214506000000889
6. Al-Aziz J, Christou N, Dinov I. SOCR motion charts: An efficient, open-source, interactive and dynamic applet for visualizing longitudinal multivariate data. Journal of Statistics Education. 2010; 18(3): 1–29.
7. Wang F, Ibarra J, Adnan M, Longley P, Maciejewski R. What's in a name? Data Linkage, Demography and Visual Analytics. In: Pohl M, Roberts J, editors. EUROGRAPHICS 2014: EuroVis Workshop on Visual Analytics; April 7–11 2014; Strasbourg, France.
8. Kovalerchuk B, Delizy F, Riggs L, Vityaev E. Visual discovery in multivariate binary data. In: Park J, Hao M, Wong P, Chen C, editors. Proc. SPIE 7530: Visualization and Data Analysis; 2010.
9. Visaya MV, Sherwell D. Dynamics from multivariable longitudinal data. Journal of Nonlinear Dynamics. 2014. doi: 10.1155/2014/901838
10. Kirillov A. Unitary representations of nilpotent Lie groups. Russian Mathematical Surveys. 1962; 17(4): 53–104. doi: 10.1070/RM1962v017n04ABEH004118
11. Sartorius B, Kahn K, Vounatsou P, Collinson MA, Tollman SM. Space and time clustering of mortality in rural South Africa (Agincourt HDSS), 1992–2007. Glob Health Action. 2010. doi: 10.3402/gha.v3i0.5225 PMID: 20838482
12. Sartorius K, Sartorius B, Tollman SM, Schatz R, Kirsten K, Collinson MA. Rural poverty dynamics and refugee communities in South Africa: A spatial-temporal model. Population, Space and Place. 2013; 19(1): 103–123. doi: 10.1002/psp.697 PMID: 24348199
13. Kahn K, Tollman SM, Collinson MA, Clark S, Twine R, Clark B, et al. 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. 2007; 35(69): 8–20. doi: 10.1080/14034950701505031
14. Tollman SM, Herbst K, Garenne M, Gear JS, Kahn K. The Agincourt Demographic and Health Study-site description, baseline findings and implications. South African Medical Journal. 1999; 89(8): 858–64. PMID: 10488362
15. Booysen F, Van Der Berg S, Burger R, von Maltitz M, du Rand G. Using an asset index to assess trends in poverty in seven Sub-Saharan African countries. World Development. 2008; 36(6): 1113–1130. doi: 10.1016/j.worlddev.2007.10.008

16. Collinson MA. Striving against adversity: The dynamics of migration, health and poverty in rural South Africa. Global Health Action. 2010. doi: 10.3402/gha.v3i0.5080

17. Rodgers G. Everyday life and political economy of displacement on the Mozambique-South Africa borderland. Journal of Contemporary African Studies. 2008; 26(4): 385–399. doi: 10.1080/02589000802481965

18. Daw CS, Finney CE, Tracy ER. A review of symbolic analysis of experimental data. Review of Scientific Instruments. 2003; 74(2): 915–930. doi: 10.1063/1.1531823

19. Dimitrova ES, Licona MP, McGee J, Laubenbacher R. Discretization of time series data. Journal of Computational Biology. 2010; 17(6): 853–868. doi: 10.1089/cmb.2008.0023 PMID: 20583929

20. Edsall R. The parallel coordinate plot in action: Design and use for geographic visualisation. Computational Statistics and Data Analysis. 2003; 43: 605–619. doi: 10.1016/S0167-9473(02)00295-5

21. Hubbard AE, Ahern J, Fleischer NL, Van der Laan M, Lippman SA, Jewell N. To GEE or not to GEE: Comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. Epidemiology. 2010; 21(4): 467–474. doi: 10.1097/EDE.0b013e3181caeb90 PMID: 20220526

22. Pan W. Akaike's information criterion in generalized estimating equations. Biometrics. 2001; 57: 120–125. doi: 10.1111/j.0006-341X.2001.00120.x PMID: 11252586

23. Cui J, Qian Q. Selection of working correlation structure and best model in GEE analyses of longitudinal data. Communications in Statistics—Simulation and Computation. 2007; 36: 987–996. doi: 10.1080/03610910701539617

24. Luo S, Lawson AB, He B, Elm JJ, Tilley BC. Bayesian multiple imputation for missing multivariate longitudinal data from a Parkinson's disease clinical trial. Statistical Methods in Medical Research. 2012. doi: 10.1177/0962280212469358