Sequential methods for user choices: tests and properties applied to a panel database

G. Chilà
LAST – Laboratory for Transport Systems Analysis,
DIMET – Department of Computer Science, Mathematics,
Electronics and Transportation,
Mediterranea University of Reggio Calabria, Italy

Abstract

Systematic observation is an important method for measuring behaviour. Sequential techniques applied to systematic observation allow temporal analysis of the user choice process. In this work, we propose a sequential analysis of a sample family related to the number of vehicles owned. An analysis of recorded data is presented in order to ascertain whether current decisions are directly influenced by the most recent previous decisions. Results obtained by the test application are presented in the paper.

Keywords: sequential, sequence, significance, stationarity.

1 Introduction

Sequential techniques result from sequential analysis, which arose as a science applied to the observation of social behaviour in psychology. As reported in Gottman and Roy [4], in a seminal paper, in 1952, George Miller introduced Markov processes into psychology, noting that probabilistic methods had proved themselves in sensory psychology and test construction. The classic introductory piece in psychology was Fred Attneave's 1959 monograph, quoted in [4]. In this monograph, Attneave introduced the notions of temporal structure. However, the main mathematical principles behind sequential analysis were derived between 1957 and 1962. Particularly, maximum likelihood estimates for transitional probabilities in a Markov chain of any order were derived. Both likelihood ratio and chi-squared tests were also used for contingency tables to examine several properties of the chains [4].

WIT Transactions on The Built Environment, Vol 101, © 2008 WIT Press
www.witpress.com, ISSN 1743-3509 (on-line)
doi:10.2495/UT080121

The main question to be analyzed concerns whether or not specific transition frequencies from an antecedent to a consequent state differ from what would be expected if the two states were independent. Recently, new tests were developed and there emerged three tools that have begun to be applied in various fields: log-linear and logit analysis; time-series analysis; lag-sequential analysis [4].

Among the various fields, sequential analysis has started to be implemented also in the transport sector, particularly in demand models. In the literature, in the behavioural approach, demand models generally simulate user choices through discrete choice models. The consolidated approach is unable to explicitly simulate the variation in choice probability due to a variety of events that affect the system characteristics of users and of the transportation network. Demand models can be classified as:

- non-dynamic, if they give the choice probability without considering system evolution;
- dynamic, if they give the choice probability considering system evolution.

Dynamic models are termed sequential models if they give the choice probability according to the current and previous system condition, considering system evolution and earlier decisions. To this group belong models which result from sequential analysis.

In the transport sector, sequential models able to explicitly simulate transitions as choices taken in the present time in relation to choices taken previously are necessary in order to simulate, for example: path choice for high frequency service [5]; evacuation, when a population has to evacuate due to a forthcoming disaster [3]; vehicle ownership, when socio-economic properties of families and technical characteristics of vehicles change in time [6].

Here we report the main characteristics of sequential analysis, as regards vehicle ownership (section 2) and sequential tests (section 3). Then the panel database (section 4) used for significance and stationary test application (section 5) is described. The main conclusions and future objectives are presented in the last section.

2 Sequential analysis: general properties

Sequential analysis aims to demonstrate how to record observation data, related to user behaviour, in a way that preserves sequential information, and also how to analyze such data in a way that makes uses of its sequential nature [1].

There are at least two ways in which data are registered: continuous and discontinuous, in respect of time. Sequential analysis requires that data are collected in a systematic way, with a continuous approach in the time.

The analyst must fix the recording unit, which can be an interval or an event. In the first case, the analyst might choose to code time intervals, assigning codes to successive time intervals. In the second case, the analyst waits for an event of interest to occur. When one occurs, he/she codes it and perhaps records onset and offset times for the event as well. The recording unit choice depends on the number of factors, including the kind and complexity of the coding scheme, the desired accuracy for the data, and the kind of recording equipment available.

When the recording unit is defined, a coding scheme should be developed, defining:

- a correspondence between observed data kinds and codes, fixing the maximum number of codes (K);
- the sequence length (s), that is the number of codes that make a sequence.

Depending on how data were recorded, the investigator can extract different representations from the same recorded data for different purposes. Historically, there are at least four forms [1], termed the Sequential Data Interchange Standard:

- event sequences: a single stream of codes is presented without any information concerning time, whether onsets or offsets; this is the simplest way to represent sequential behaviour;
- state sequences: they are identical to event sequences, with the simple addition of timing information;
- timed-event sequences: this is a useful and general purpose format, used if codes can co-occur and if their onset and offset times were recorded;
- interval sequences: this is designed to accommodate interval recording in a simple and straightforward way; codes are simply listed as they occur and interval boundaries are represented by commas.

In this context, we suppose that the recording unit is the interval and the interest is in event sequence data: this means that data to be analyzed are obtained in each considered time interval and are represented as sequences or chains of coded events (or behavioural states but without time information). It is important to highlight that sometimes the same code in event sequences cannot be assigned.

Therefore, the total number of s-event sequences is:

- K^S if the same code in event sequences can be assigned;
- $K \cdot (K-1)^{S-1}$ if the same code in event sequences cannot be assigned.

Sequences are analyzed in respect of transitional probabilities. These probabilities are one kind of conditional probabilities. Conditional probabilities are defined as probabilities with which a particular target event occur, relative to another given event. A transitional probability is distinguished from other conditional probabilities in that the target and the given event occur at different times.

For example, we suppose that the recording unit is the interval, the data representation is like event sequences, with the same codes in succession allowed. We consider the following sequence: Q Q P Q P Q P P Q P. The same codes in succession (e.g. QQ or PP) represent a permanence; different codes in succession (e.g. QP or PQ) represent a transition.

Data collected sequentially are summarized using a frequency and probability transition matrix. The frequency transition matrix is a square matrix which has the same number of rows and columns as the number of events (or states); its generic element n_{ij} represents the transition frequency from state Q to state P (Table 1). These frequencies can also be converted by dividing each n_{ij} by the row total for row i. These probabilities can be arranged in a matrix called probability transition matrix (Table 2).

The rows of this matrix sum to one, and each entry gives the conditional probability that the system, which was in one state at time t, will be in some other specified state at time t+1 [4].

Tables 3 and 4 report the transition frequency matrix and the transition probability matrix for the sequence previously considered, respectively.

As regards transition probabilities, in order to indicate the temporal displacement between the target and a given event, the word lag is used. For example, if data are represented as event sequences, with events Q and P, the probability that given event Q, the target event P will occur immediately after (lag1), or after an intervening event, can be written, respectively, as $p(P_{+1}/Q_0)$, $p(P_{+2}/Q_0)$. If we consider state sequences, transitional probabilities can be written as $p(P_{t+1}/Q_t)$ or $p(P_{t+2}/Q_t)$ and so on.

Table 1: Transition frequency matrix.

	t:1	Q	P	
t:0	Q	n_{11}	n_{12}	n_{1+}
	P	n_{21}	n_{22}	n_{2+}
		n_{+1}	n_{+2}	

Table 2: Transition probability matrix.

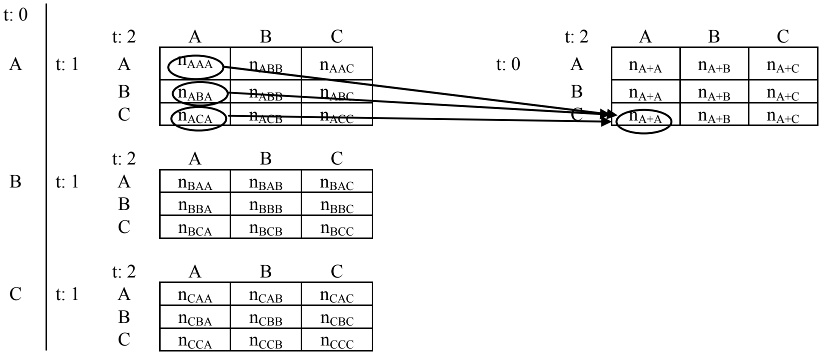
	t:1	Q	P	
t:0	Q	p_{11}	p_{12}	p_{1+}
	P	p_{21}	p_{22}	p_{2+}
		p_{+1}	p_{+2}	

Table 3: Transition frequency matrix.

	t:1	Q	P	
t:0	Q	1	4	5
	P	3	1	4
		4	5	

Table 4: Transition probability matrix.

	t:1	Q	P	
t:0	Q	0.20	0.80	1.00
	P	0.75	0.25	1.00
		0.95	1.05	



t: 0		t: 2	A	B	C
A	t: 1	A	n_{AAA}	n_{ABB}	n_{AAC}
		B	n_{ABA}	n_{ABB}	n_{ABC}
		C	n_{ACA}	n_{ACB}	n_{ACC}
		t: 2	A	B	C
B	t: 1	A	n_{BAA}	n_{BAB}	n_{BAC}
		B	n_{BBA}	n_{BBB}	n_{BBC}
		C	n_{BCA}	n_{BCB}	n_{BCC}
		t: 2	A	B	C
C	t: 1	A	n_{CAA}	n_{CAB}	n_{CAC}
		B	n_{CBA}	n_{CBB}	n_{CBC}
		C	n_{CCA}	n_{CCB}	n_{CCC}

	t: 2	A	B	C
t: 0	A	n_{A+A}	n_{A+B}	n_{A+C}
	B	n_{A+A}	n_{A+B}	n_{A+C}
	C	n_{A+A}	n_{A+B}	n_{A+C}

Figure 1: Transition frequency matrix for lag2.

If a lag>1 is considered, a three-dimensional matrix has to be constructed. Cells of this are denoted as x_{ijk}, where i, j and k represent the lag0, lag1 and lag2 dimensions, respectively. Traditional lag-sequential analysis would test for lag2 effects in the collapsed 02 table, that is the table whose elements are x_{i+k}, where

$x_{i+k}=\sum_{j=1,K}x_{ijk}$. For these cases a log-linear approach could be considered [1]. In figure 1 we report a three-dimensional transition frequency matrix, for a system with three events (A, B, C), and its corresponding two-dimensional collapsed matrix, in relation to final lag considered (lag0 and lag2). Similarly, a transition probability matrix can be obtained.

3 Sequential tests

The reduction in uncertainty through knowledge of past events is the basic concept in sequential analysis. In order to assess the significance of this reduction in uncertainty, several tests can be applied.

3.1 Significance of sequences with lag1

A particular model of event realization is supposed. The expected values that the model generates for a particular sequence are compared with those actually observed, by statistics:

- Chi-squared $X^2=\sum[(obs-eps)^2/eps]$ with (1)
  - obs observed frequencies;
  - eps expected frequencies;
  - degrees of freedom equal to $(K-1)^2$, with K number of codes.
- z score $z_{GT}=(x_{GT}-m_{GT})/(m_{GT}\cdot(1-p_{G+})\cdot(1-p_{+T}))^{1/2}$ (2)
  - if the same code in event sequences can be assigned, with
  - x_{GT} observed transition frequency between given events G and T;
  - m_{GT} expected transition frequency between generic events G and T, equal to $m_{GT}=(x_{G+}\cdot x_{+T})/x_{++}$;
  - x_{+T} sum of the observed frequencies in the Tth or target column; x_{G+} sum of the observed frequencies in the Gth or given row;
  - x_{++} total number of observed frequencies in the matrix;
  - p_{+T} sum of observed probabilities in the Tth or target column;
  - p_{G+} sum of observed probabilities in the Gth or given row.
  - degrees of freedom equal to $(K-1)^2$, with K number of codes.

If the test value is larger than that of reference, for a level α of significance (e.g. α=0.05), the supposed model of event realization is rejected and the dependence between target and given event is verified. Similarly, tests can be applied for sequences if the same code in event sequences cannot be assigned and with lag>1 [1].

3.2 Further statistical tests

3.2.1 Testing differences among individuals

Often more than a single individual, family or whatever is observed. In order to verify possible differences among these groups of individuals it is necessary to collapse a transition matrix into tables with two rows and two columns.

For example, suppose that we wish to know if event P is particularly likely to occur after event Q. In this case we would label rows Q and ~Q and columns P and ~P (where rows represent lag0, columns lag1 and ~ represents not). The collapsed table can then be represented as in table 5, where individual cells are labelled a, b, c, d as shown and represent cell frequencies.

Collapsed tables can be obtained for each group. The following tests may be applied:

- odds ratio=ad/bc (3)
  the odds ratio varies from 0 to infinity and is equal to 1 when it is the same for both rows (indicating no effect of the row classification). If it is greater than 1, the probability that event P is particularly likely to occur after event A is significant;
- Yule's Q=(ab-bc)/(ad+bc) (4)
  Yule's Q varies from -1 to +1, with zero indicating no effect.

Table 5: Collapsed table for testing differences.

	t:1	P	~P
t:0	Q	a	b
	~Q	c	d

3.2.2 Stationarity

In order to test whether the sequential structure of the data is the same regardless of where we begin in the sequence, we compare the actual data to the expected values under the null hypothesis that the data are stationary. Let us consider only the lag1 antecedent/consequent matrix. We apply the omnibus test, distributed as Chi-squared [1]:

- Omnibus $G^2=2\cdot\sum_t\sum_{i=1,K}\sum_{j=1,K} n_{ij}(T)\cdot\ln(p_{ij}(T)/p_{ij})$, with (5)
  - ○ n_{ij} (T) transition frequency for cell ij in temporal segment T;
  - ○ p_{ij}(T) transition probability for cell ij in temporal segment T;
  - ○ p_{ij} transition probability expected for stationarity hypothesis, that is transition probability pooled in T.

G^2 has degrees of freedom (T-1)(K)(K-1). If the test value is larger than that of reference, for a level α of significance (e.g. $\alpha=0.05$), the stationarity hypothesis is rejected.

4 Panel database

In this work, in order to test sequential methods, a panel database was used, obtained by surveying a sample of 50 Italian families, regarding their socio-economic features, from 1994 to 2007. The families were randomly selected among residents in the province of Reggio Calabria (Italy). It is worth noting that the sample size changes in time, since some families are formed after the first year considered in the survey (1994).

The survey design is characterized by [2]:

- definition of the sampling unit, that is one of the units into which an aggregate is divided for the purpose of sampling (people, family, …);
- definition of the survey questions.

Definition of the sampling unit is related to practical aspects, that is to the kind of survey and data availability. In this work, the sampling unit is the family. In relation to the second point, questions are grouped into several forms, shown in table 6. Each form was filled in for the period 1994-2007.

Form I is related to the socio-economic characteristics of the family. It can be subdivided into two parts:

- the first part includes data related to the number of components, workers, students, retired persons, those with driving licences, vehicles owned, motorbikes; it also includes data on family income, classified as in table 7;
- the second part includes data related to residence: address, distance to the bus stop and to the railway station, number of bus runs/day, in order to evaluate the availability of public transport in the area.

Form II is related to the socio-economic characteristics of each family member: sex, age, driving licence and work activity.

Form III concerns the mobility characteristics of each family member. It includes data related to the number of trips/day, the purpose of the most frequent trip and the main transport mode used. Form IV seeks to ascertain the number of trips per year with a distance between origin and destination of over 500 km, and the vehicle used for these.

Form V concerns the technical characteristics of vehicles owned. Each family fills in a form including brand, model, body, power, speed, acceleration, gasoline, load capacity, number of seats, consumption, production year, purchase year and price. Kilometres/month covered by the main vehicle user are also necessary.

Table 6: Survey questions of the panel database.

	Form	Filled in by	Description	Records
I	Family	The family as a whole	Socio-economic characteristics of the family	Components, workers, students, retired persons, those with driving licences, income, number of vehicles owned, number of motorbikes, residence and address, distance between residence and bus stop, distance between residence and railway station, bus runs/day
II	Personal	Each family member	Socio-economic characteristics of each family member	Sex, age, work
III	Mobility	Each family member	Mobility characteristics of each family member	Number of trips/day, most frequent trip purpose, transport mode of the most frequent trip
IV	Long-distance mobility	The family as a whole	Long-distance mobility characteristics of the family	Number of trips/year with a length>500 km, vehicle used for this trip
V	Vehicle	The family as a whole	Technical and performance characteristics of each vehicle owned by the family	Brand, model, body, power, speed, acceleration, gasoline, load capacity, number of seats, consumption, kilometres/month covered, production year, purchase year, price

Table 7: Income class.

Family income	Class	Income parameter
<20000 €/year	Low (B)	0
20000-40000 €/ year	Medium (M)	1
>40000 €/ year	High (A)	2

Below we propose, for the period 1994-2007, the analysis of families included in the considered sample, in relation to number of components (table 8) and number of vehicles owned (table 9).

5 Sequential tests applied to the panel database

In this paper an analysis of the number transition of owned vehicles is proposed.

We suppose that:

- the recording unit is a time interval, from 1994 to 2007;
- the data representation is like event sequences;
- there are three codes (K=3): adding a new vehicle (C); retaining the vehicle number currently owned (S); removing a vehicle (V);
- the sequence length (s) is equal to 2;
- the same code in event sequences can be assigned; hence the total number of obtainable sequences is $K^S= 3^2=9$;
- the model of event realization is equiprobable, that is we assume that codes occur with equal probability;

Table 8: Family analysis in relation to number of components.

Year	Family number with 1 component	Family number with 2 components	Family number with 3 components	Family number with 4 components	Family number with 5 components	Family number with 6 components	Total number
1994	2	4	6	19	9	2	42
1995	3	5	8	17	9	2	44
1996	3	3	10	17	9	2	44
1997	3	3	11	16	9	2	44
1998	3	8	8	16	9	2	46
1999	3	8	8	18	7	2	46
2000	3	9	8	19	6	2	47
2001	3	8	8	19	7	2	47
2002	3	7	8	20	7	2	47
2003	3	7	8	21	6	2	47
2004	3	11	7	20	6	2	49
2005	3	14	7	19	5	2	50
2006	3	12	13	17	3	2	50
2007	4	11	13	17	3	2	50

Table 9: Family analysis in relation to number of vehicles owned.

Year	Family number with 0 vehicles	Family number with 1 vehicle	Family number with 2 vehicles	Family number with 3 or more vehicles
1994	3	17	19	3
1995	4	14	22	4
1996	3	14	23	4
1997	3	12	23	6
1998	3	12	26	5
1999	3	11	27	5
2000	3	12	25	7
2001	2	12	26	7
2002	2	12	26	7
2003	2	12	24	9
2004	2	14	24	9
2005	2	18	22	8
2006	2	18	20	10
2007	2	18	19	11

- the level of significance is α=0.05.

The total number of observed two-sequences is 603. In table 10 the general transition frequency matrix is reported: each cell represents the family number which moves among C, S and V, from the previous year (t:0) to the current year (t:1). Below we propose significance and stationarity tests.

5.1 Significance of sequences with lag1

In order to test the significance of two-dimensional sequences, the chi-square test (sec. 2.1) is applied. To compare two-dimensional sequences collapsed transition frequency matrixes are computed. In table 11 a collapsed transition matrix is proposed for sequence CC. In table 12 observed and expected transition frequency, considering two-dimensional sequences, are compared. In table 13, for each two-sequence, the chi-square test is computed. In this case, degree of freedom is equal to 1, and the statistical value is 3.84, if α=0.05. All obtained values are higher than 3.84: each two-dimensional sequence is significant and the equiprobable model does not represent transition frequencies.

Table 10: Transition frequency matrix.

	t:1	C	S	V	Total
t:0	C	0	24	0	24
	S	28	528	13	569
	V	0	10	0	10
	Total	28	562	13	603

Table 11: Collapsed transition frequency matrix (CC sequence).

	t:1	C	~C	Total
t:0	C	0	24	24
	~C	28	551	579
	Total	28	575	603

Table 12: Transition frequency two-sequences.

Observed values				Expected values			
CC	0	~ CC	603	CC	67	~ CC	536
CS	24	~ CS	579	CS	67	~ CS	536
CV	0	~ CV	603	CV	67	~ CV	536
SC	28	~ SC	575	SC	67	~ SC	536
SS	528	~ SS	75	SS	67	~ SS	536
SV	13	~ SV	590	SV	67	~ SV	536
VC	0	~ VC	603	VC	67	~ VC	536
VS	10	~ VS	593	VS	67	~ VS	536
VV	0	~ VV	603	VV	67	~ VV	536

Table 13: Chi-square test for each sequence.

Sequence	CC	CS	CV	SC	SS
X^2	75.38	31.05	75.38	25.54	3568.45

5.2 Stationarity

In order to test whether the sequential structure of the data is the same regardless of where we begin in the sequence, we considered two temporal segments: T_1 [1994,2000] and T_2[2001,2007], comparing obtained values between these. We consider only the lag1 antecedent/consequent matrix. We apply the test described in section 2.3.2, with a chi-square distribution [1]. In this case T=2, K=3 and degrees of freedom are equal to 6. The corresponding statistic value is 12.59, if α=0.05. Frequency transition matrices for temporal segments T_1 and T_2 are reported in tables 14 and 15; probability transition matrices for the same periods are reported in tables 16 and 17. Computation gives G^2=2.02. As the obtained value (2.02) is lower than the statistic value (12.59) for α=0.05, the stationarity hypothesis is not rejected. Therefore the analyzed data can be judged stationary.

6 Conclusions and future objectives

In this paper general aspects of sequential analysis were described. Several tests applied to panel database were performed to evaluate whether or not current decisions are directly influenced by the most recent previous decisions. The results obtained by the test application confirm that event sequences, with lag1, are significant. Therefore there is a dependence between events in period t:0 (previous time) and events in period t:1 (actual time). The process is stationary, that is parameters of sequential connection over time are stable. These results allow stochastic patterns to be discovered in the data and justify, respectively: the sequential model definition, to simulate vehicle ownership choices in the time, considering two-event sequence data; the introduction, in the model, of parameters constant over time. Future objectives are related to the development of a behavioural model explaining sequential data, applied to simulate vehicle ownership choices, for each row of the proposed transition probability matrix.

Table 14: Transition frequency matrix for the period $T_{1.}$

t:0	t:1	C	S	V	Total
	C	0	13	0	13
	S	15	230	4	249
	V	0	4	0	4
	Total	15	247	4	266

Table 15: Transition frequency matrix for the period T_2.

t:0	t:1	C	S	V	Total
	C	0	11	0	11
	S	13	298	9	320
	V	0	6	0	6
	Total	13	315	9	337

Table 16: Transition probability matrix for the period T_1.

t:0	t:1	C	S	V	Total
	C	0.00	1.00	0.00	1.00
	S	0.06	0.92	0.02	1.00
	V	0.00	1.00	0.00	1.00
	Total	0.06	2.92	0.02	3.00

Table 17: Transition frequency matrix for the period T_2.

t:0	t:1	C	S	V	Total
	C	0.00	1.00	0.00	1.00
	S	0.04	0.93	0.03	1.00
	V	0.00	1.00	0.00	1.00
	Total	0.04	2.93	0.03	3.00

References

[1] Bakeman R., Gottman J.M., *Observing Interaction: An Introduction to Sequential Analysis*, Cambridge University Press, New York, 1997.

[2] Cascetta E., *Transportation systems engineering: theory and methods*, Kluwer Academic Press, 2001.

[3] Fu H. *et al.*, Sequential Logit Dynamic Travel Demand Model and Its Transferability. In *Transp. Res. Record*, vol. 1964, pp. 17–26, 2006.

[4] Gottman J.M., Roy K.A. *Sequential Analysis*, Cambridge University Press, New York, 1990.

[5] Russo F., Modelli di scelta del percorso sequenziali per l'assegnazione dinamica a reti di trasporto collettivo urbano. In *Metodi e Tecnologie dell'Ingegneria dei Trasporti*, Cantarella G.E. and Russo F. (Eds.), 151–165, FrancoAngeli, Milan, 1999.

[6] Russo F., Chilà G., Sequential models for the mobility decisions: experimentation for the vehicle holding choices. In *Proceedings of European Transport Conference 2007*, The Netherlands, 2007.