1352.0.55.062

Australian Bureau of Statistics

Research Paper

Issues Associated with Producing a Longitudinal Dataset of Businesses

www.abs.gov.au

New Issue

Research Paper

Issues Associated with Producing a Longitudinal Dataset of Businesses

Paul Sutcliffe, Martin Caruso and Helen Teasdale

Statistical Services Branch

Methodology Advisory Committee

18 June 2004, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) TUE 6 JUL 2004

ABS Catalogue no. 1352.0.55.062
ISBN 0 642 48166 0

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr Paul Schubert, Statistical Services Branch on Canberra (02) 6252 6591 or email <paul.schubert@abs.gov.au>.

ISSUES ASSOCIATED WITH PRODUCING A LONGITUDINAL DATASET OF BUSINESSES

Paul Sutcliffe, Martin Caruso and Helen Teasdale
Statistical Services Branch

EXECUTIVE SUMMARY

Recently, the ABS has been investigating the use of administrative data linked to ABS collected survey data to create a Business Longitudinal Database (BLD). The main aim of this project is to develop a longitudinal dataset containing information which will facilitate analysis of a range of policy issues based around business growth and performance. Based on discussions with external and ABS analysts, it appears that the preference is for a dataset which is smaller in terms of sample size yet denser in terms of data richness.

It is our belief that given a specific analytical outcome for a longitudinal dataset it is feasible to design the dataset to meet the analytical needs. The techniques for determining the design will vary depending on the analysis, but the process is analogous to the cross-sectional design and allocation problem. However, if the aim is to produce a dataset for general use by analysts who are using a wide range of techniques, such as what we are aiming for with the BLD, it is much harder to specify the size of the sample needed.

In the survey methodology field there is very little that links the design of a longitudinal survey to the analysis to be applied. Where decisions on sample size are mentioned at all, it is generally stated that these are dependent on available budget. In many of the papers we have reviewed, the problem of sample size becomes one of balancing costs associated with infrastructure, contacting and surveying each business and the number of items collected each period. Since the trade-offs cannot be considered in isolation of the costs, it is important to develop good cost models. Thus, like most authors, the problem is generally reduced to how best to use a fixed annualised budget to meet the analytical needs.

One of the primary aims of the BLD is to maximise data from existing sources leveraging off the available tax data. Following advice from experts, we have focused initially on getting the right data into the BLD. In the paper we describe the likely scenario in which we will aim to maximise the data available from the Innovation survey, by coupling this with a new collection of business characteristics, and augmenting with financial data from tax sources. The starting sample size would be

around 8,500 businesses. This is comparable to the sample size from a previous ABS longitudinal survey which analysts have found suitable for their analysis purposes.

In order to put some structure around the decision making process we favour an approach to developing a general longitudinal dataset which allows the designer to test likely scenarios against the objectives of the survey. In practice we believe developing good cost models is an important process for the BLD if long term costing implications are to be managed well.

DISCUSSION POINTS FOR MAC

The specific issues for discussion have been raised throughout the paper as they occur. For convenience they have been grouped here.

- Does MAC agree with our broad framework for determining the design of the BLD? Do you have any suggestions for improvements?
- From an analysts point of view, what are your thoughts on the choice for linking repeated panels for the BLD?
- Do you feel that the described population changes would have a great impact on the final longitudinal dataset produced if left unaddressed? If so, do you believe that our intended actions are appropriate/optimal?
- Do you feel that our approach of developing good cost models for assessing various sample size scenarios is the best way to tackle the sample size problem? Or is there an objective method available?
- Does our suggestion for addressing the issue of non-response seem reasonable? Do you have suggestions for the set of weights which should be included on the dataset?

CONTENTS

INTRODUCTION 1

DIFFERENCES BETWEEN LONGITUDINAL SURVEYS OF BUSINESSES AND HOUSEHOLDS 2

DATASET DESIGN 2

A PANEL SAMPLE 8

CHANGING POPULATION 11

SAMPLE DESIGN 14

TREATMENT OF MISSING DATA AND WEIGHTING 16

CONFIDENTIALITY 19

CONCLUSION 20

REFERENCES 21

ISSUES ASSOCIATED WITH PRODUCING A LONGITUDINAL DATASET OF BUSINESSES

Paul Sutcliffe, Martin Caruso and Helen Teasdale
Statistical Services Branch

Introduction

1 Recently the ABS has begun a project to create a Business Longitudinal Dataset (BLD) which will contain business information to facilitate analysis at the microeconomic level on a range of policy issues based around growth and performance. The aim was to create this dataset by combining annual administrative data, such as Australian Taxation Office and Australian Customs Service data, with ABS collected survey data. That is, without the need for data collection. However, during discussion with potential users of this dataset, the desire for business characteristic data has emerged. Such data is not collected by any current ABS business surveys and so the need for a specific survey has arisen. The current vision is a survey vehicle used to collect the characteristics data, which is then matched to available administrative data and other ABS survey data to form the BLD. Thus the BLD design issue is in fact a longitudinal survey design issue. It is the desire to produce a dataset which will be suitable for many different analytical purposes. This has presented a design challenge for us. The ABS has some experience at designing longitudinal surveys which will produce datasets for a particular analytical purpose, but much less experience in designing to produce a dataset which will be suitable for multiple analytical purposes.

2 At the Australian Statistics Advisory Council (ASAC) meeting in March 2000 the ABS presented an information paper which outlined the features and benefits of such surveys. The paper was in response to a review of the ABS Household Survey program in which users drew attention to longitudinal surveys. At the time the ABS had not undertaken any systematic work on business longitudinal surveys and the paper emphasised that the organisation had limited knowledge on the characteristics and uses made of longitudinal data.

3 On the business survey side the Business Growth and Performance Survey (GAPS) was conducted from 1996 to 1999 in collaboration with the Office of Small Business. This survey was designed to meet total RSE requirements at each wave (sample supplementation occurred at each wave to keep the sample representative of the population), rather than for analytical purposes, although it has been used extensively as an analytical dataset. The ABS has had involvement in longitudinal surveys of persons to a greater extent, designing for example the Survey of Employment and Unemployment Patterns (SEUP) from 1995 to 1997. The sample for this survey was determined assuming a particular analytical objective.

4 This paper discusses what we feel are the major issues associated with creating a business longitudinal dataset for multiple analytical purposes, describes how these impact on the development of the BLD, states the methodology we intend to use to overcome these issues, and then seeks the opinion of MAC on our proposed directions.

Differences between longitudinal surveys of businesses and households

5 There is a large amount of literature available which discusses longitudinal surveys of persons and households. The relative strengths and weaknesses of different techniques for data collection (i.e. repeated surveys, panel surveys, cohorts and so on) are well known and documented. See for example, Duncan and Kalton (1987). There are well-tested solutions available for common difficulties such as tracking respondents, minimising non-response and accounting for missing data. There appears to be much less information available for longitudinal surveys of businesses. While some of the issues that apply to household samples also apply to business samples – treatment of missing data being one, there are some that are not as relevant – such as tracking of sampled units, and other, different, issues that are unique to business samples – such as dealing with businesses which change structure during the life of the survey.

6 In this paper we concentrate on longitudinal surveys of businesses, specifically on issues that are either unique to such surveys, or which require a different solution to that applicable in surveys of persons.

Dataset design

7 Since the specific aim of conducting a longitudinal survey is to produce a dataset useful to economic analysts, it is important to consider what makes such a longitudinal dataset useful. That is, what would a dataset which is specifically designed for longitudinal purposes, and not to produce accurate cross-sectional information, look like. It is important to note that if this requirement for cross-sectional information were also placed on a longitudinal dataset it would restrict its ability to meet both needs. The literature suggests that different types of surveys are best suited to meet these two different needs. However, this does not preclude the longitudinal sample from being used as the core of a single point-in-time survey. Supplementing the longitudinal sample as needed and surveying the extra selected businesses will not impact on the success of the longitudinal sample, while still allowing cross-sectional estimates to be produced. For this reason, for the BLD we have chosen to specifically concentrate on how to produce a dataset which is optimal for longitudinal purposes, ignoring the need to produce cross-sectional estimates as well.

8 Based on discussions with external and ABS analysts, it appears that the preference is for a smaller-denser dataset rather than a larger-sparser dataset. That is, a dataset which is smaller in terms of sample size yet denser in terms of data items. This is because the richness of data across time from particular businesses is seen as more desirable than large samples with less data. In such a dataset all data items are available for all selected businesses for all time periods covered by the survey (the 'want it all principle'). This is a rather broad-brush description however, which masks the dilemma for dataset designers who need to resolve a number of difficult issues. Even if it is accepted that the ideal longitudinal dataset is one which is dense with information, there are many other characteristics that describe the dataset that also need to be considered. These include the sample size, sample distribution, number, spread and complexity of data items, number of waves, reference period covered and interval between waves.

9 Many of these will have no one solution, and in different situations a decision on say the number of waves needed to produce a "good" dataset will vary. However for most (perhaps with the exception of sample size) a decision can be made via discussion with various analysts on what would suit their particular needs. Availability of data, respondent reaction, standards and so on will all impact on decisions made about data items, reference period, number of waves etc. This decision-making process is the same as what would occur for cross-sectional surveys (with the obvious exclusion of discussion covering issues specific to longitudinal surveys).

10 Under a longitudinal framework the data is generally collected to measure changes over time for the individual businesses. Models are fitted to the data to describe the impact of policy changes to individuals whereas in the cross-sectional framework estimation at a single time point is the central aim.

11 Kalton and Citro (1993) highlight a number of general themes in the literature regarding longitudinal analysis:

- Measurement of gross change;
- Relationship between variables across time;
- Regression with change scores;
- Estimation of spell duration (survival analysis); and
- Structural equation models with measurement errors.

12 The opinion of the BLD technical reference group regarding the types of analyses which would be applied to the BLD was sought. The more academic researchers were interested in structural equation modeling and understanding the relationships between variables across time. On the other hand, there were other researchers who expect to use specific econometric packages which incorporate a

variety of techniques. None of the techniques listed by Kalton and Citro were discounted as being unsuitable for BLD purposes.

13 It is not the scope of this paper to describe all possible analyses, but rather to highlight our belief that given a specific analytical outcome for a longitudinal dataset it is feasible to design the dataset to meet the analytical needs. The techniques for determining the design will vary between different purposes, but the process is analogous to the cross-sectional design and allocation problem. The constraints to be solved can be expressed explicitly and can be minimised.

14 There are many examples whereby this approach has been taken. The drawback of this method of designing a dataset is that the possible ways in which it can be used are potentially limited. While the dataset will be optimal for a particular analysis, it may be quite unusable for other types of analyses. It may even be that the dataset isn't optimal for the specified analysis if various assumptions or information used in the design don't hold or are dated. An example of a longitudinal dataset which was designed for a specific analytical outcome is given below (note that this is for a survey of households rather than businesses).

Transition modelling measuring gross flows

An example of a longitudinal survey conducted by the ABS is the Survey of Employment and Unemployment Patterns (SEUP). This was a longitudinal survey run over 3 years. The survey was developed to assess the impact of particular government policy. In deciding the sample size for the SEUP longitudinal survey, a simple logit conditional model was used where the conditional probability of a person moving from unemployment to employment given whether their mother was Australian born was found to be the appropriate variable to fit. This logit model was fitted to historical data to determine the model parameters. The logit model was then used to decide whether the sample size was large enough to detect all the significant regressors.

Other models were looked at, such as transition models which examined the transition probabilities of moving from unemployment to employment from one year to the next; conditional models fitting other conditional probabilities and also unconditional cross-sectional models that looked at the probability of being in a particular employment state at a particular point in time.

15 If the aim is to produce a dataset for general use by analysts who are using a wide range of techniques, such as what we are aiming for with the BLD, it is much harder to specify the size of the sample needed. Theoretically, if the sample size can

be determined for a particular model then it can be determined for a set of possible models. We have considered two approaches while discussing the BLD design:

a. Treat the problem as a multivariate case of the specific dataset approach and try to optimise the sample with respect to multiple models; or,

b. Take the maximum sample size of the individual sample sizes optimised for each model.

16 On reflection both approaches have their limitations and we have come to the conclusion that any particular research effort on designing business longitudinal datasets would be better spent understanding the constraints for their design rather than developing a complex method for determining sample size.

17 We feel that the key to developing a good general purpose business longitudinal dataset will be to determine how the BLD might be analysed and to understand the key data themes that users agree are important. The process will yield general constraints that will give us a direction. For the BLD the most important dimensions are the data items which should be considered as broad themes (e.g. innovation and management practices) and the demographic classifications (e.g. industry, sector and size). The third dimension of sample size can only be resolved once the first two dimensions are well understood. Given a good understanding of the data items and their frequency the trade-offs in terms of the design of the dataset are endless.

18 The framework that we propose to follow when developing the BLD is:

a. Fully understand and document the analysts needs by:

- getting expert advice on how the dataset is likely to be analysed;
- understanding the various economic analysis techniques;
- understanding data themes and linkages between data items;
- developing clear specifications for the data items and their frequency;
- understanding the trade-offs for industry, size and sector;
- developing appropriate cost models for scenario testing (expanded on further in the Sample Design section below)

b. Scenario testing

c. Determine a final design

d. Test the final design against the original objectives

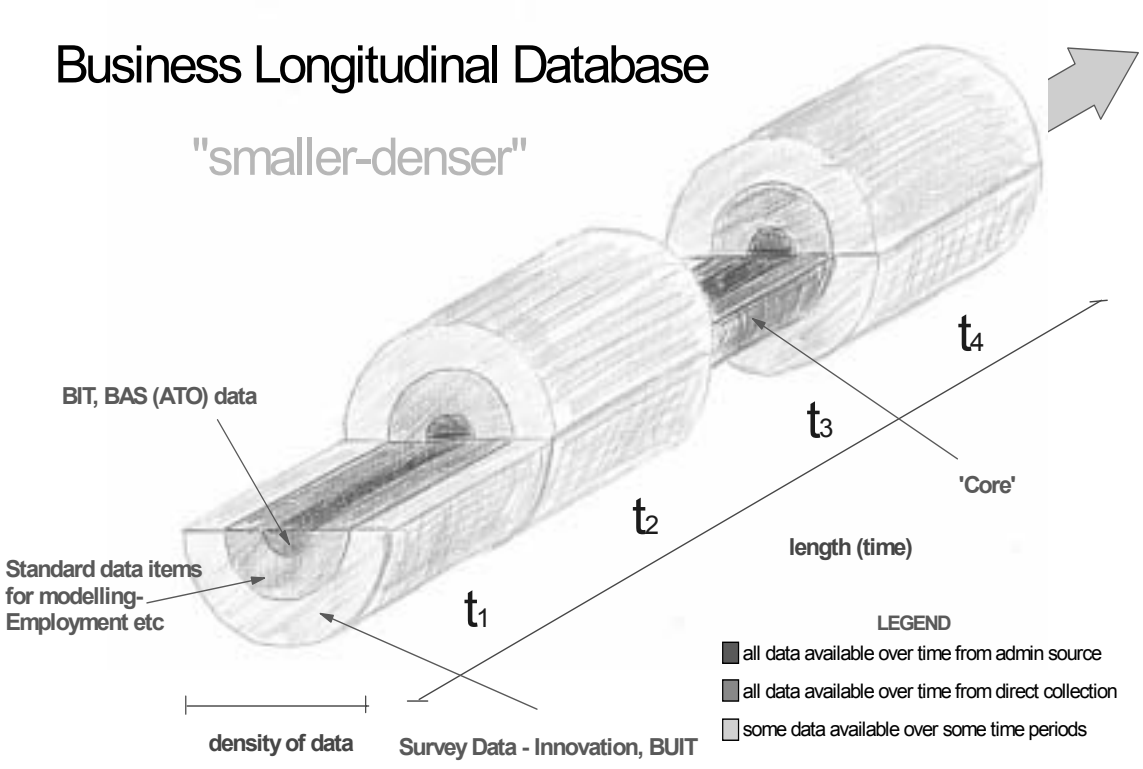
e. After a period of time review the design against the models being applied.

Question for MAC: Does MAC agree with our broad framework for determining the design of the BLD? Do you have any suggestions for improvements?

19 The ABS has conducted a number of sessions with experts which has given broad direction to the types of analyses that are likely to be performed on the BLD. They have provided advice on the trade-offs they would make under different budgeting scenarios. For example, they suggest it would be preferable to target particular industries rather than having sparse coverage in all industries. The main issue at the moment is that the ABS does not have an existing survey which collects information on business characteristics. Our current thinking is that if we want to maximise the amount of data available for each business included in the BLD then we have four main options:

- financial data from administrative sources such as BAS data (available annually);
- existing data collected via the Innovation Survey (to be conducted every two years);
- data from a business characteristics survey; and,
- data from any other ABS surveys.

Diagram 1: Proposed structure of the BLD



20 Diagram 1 explains the relationship of the four data sources over time. The 'inner core' (dark shade) would comprise data items available from administrative sources on an annual basis, generally for all businesses, such as BAS data. The next level, the 'outer core' (lighter shade) would be specific data collected in an annual business characteristics survey. The next level (lightest shade) would be data not available for all time periods, such as Innovation data which is not included in of an annual business characteristics survey.

21 From an analysis point of view Diggle (1994) provides a good summary of the quantities investigators must provide to enable determination of the required sample size:

1. Type I error rate – the probability that the study will reject the null hypothesis when it is correct.
2. Smallest meaningful difference to be detected – investigators typically want their study to reject the null hypothesis with high probability when the parameter of interest deviates from its value under the null hypothesis by an amount d or more.
3. Power – the power of a statistical test is the probability that the study rejects the null hypothesis when it is incorrect.
4. Measurement variation – for a continuous response variable this quantity measures the unexplained variability in the response.

In longitudinal studies the following additional quantities are also needed:

1. Number of repeated observations per unit – this number may be constrained by practical considerations, or may need to be balanced against the sample size.
2. Correlation among the repeated observations.

22 In practice these are very difficult to obtain during the exploratory phase of the dataset development. For this reason, we believe that it is better to incorporate this into the Scenario Testing stage ((b) in the framework given in paragraph 18) by applying various combinations of the above quantities to the preliminary sample size determined during the cost-modelling stage. This will enable fine tuning of the sample to be undertaken.

A panel sample

23 An important issue for longitudinal surveys that doesn't exist for single cross-sectional surveys is how to track the selected businesses from one period to the next. There is extensive literature covering this issue, detailing two main types of surveys that could be conducted. Repeated surveys are where a new sample is selected at each time point, for which there may or may not be controlled overlap with the previous sample. These are the standard used for cross-sectional surveys as they produce accurate point-in-time estimates, and with some level of overlap between time points are able to produce reasonable measures of change from one period to the next. They are not suited to longitudinal surveys as there is no effort made to retain particular units in sample for a defined number of time points. Panel surveys are ones in which the same sample is measured at different time points and are obviously more suitable for longitudinal output. They will not be able to produce data for future time points with the same level of accuracy as repeated surveys since the representativeness of the sample over time will diminish. Panel surveys are sometimes referred to as cohort studies.

24 Kalton and Citro (1993) discuss the abilities of these two survey designs to meet the various objectives of surveys which collect information across time. They argue that panel surveys are the optimal choice for longitudinal samples. The key advantage of a panel survey over a repeated survey is its abilities to measure gross change and to track data for individuals over time. They state that:

> "Repeated surveys are incapable of satisfying these objectives. The great analytic potential provided by the measurement of individual change is the major reason for using a panel design."

25 There is further discussion on the various types of panel surveys, the choice of which is dependent on the desired analytical outcome. A repeated panel survey is a series of panel surveys, each of a fixed duration, which may or may not overlap in the time dimension. Rotating panel surveys are equivalent to repeated panel surveys with overlap, but they tend to have longer duration and fewer panels. An overlapping survey is a series of cross-sectional surveys conducted at different time points with guaranteed sample overlap. Each have their own advantages, for example rotating panel surveys are able to estimate both current levels and net change, while repeated panel surveys focus more on longitudinal measures.

26 After considering such panel surveys, there is a section on design issues over and above those that exist for cross-sectional surveys. The major design decisions are presented as:

- Length of the panel – the longer the panel the greater the wealth of data available, but this needs to be balanced against the problems of

maintaining a representative cross-sectional sample at later waves, due to sample attrition and difficulties of including population births.

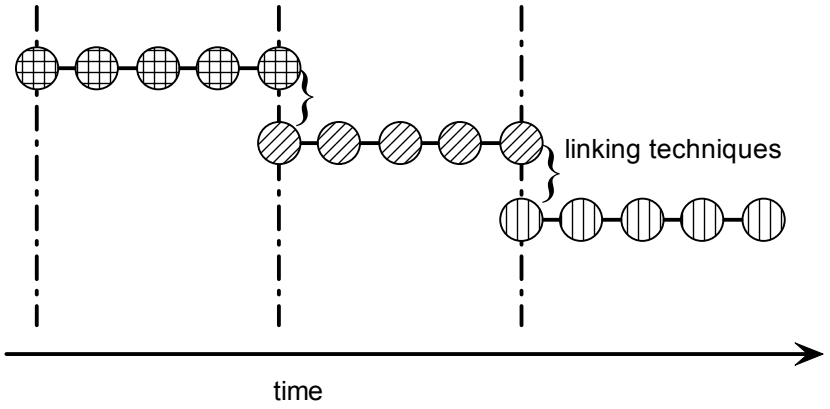
- Length of the reference period – this is discussed in terms of recall error which will only affect business collections if retrospective data is collected which isn't available by referral to historical records.
- Number of waves (time periods of collecting data within a panel) – this is usually determined by a combination of the length of the panel and length of reference periods. The greater the number of waves the greater the risk of panel attrition and the greater the degree of respondent burden.
- Overlapping or non-overlapping panels – the design of non-overlapping panels has the benefit of simplicity from a collection point of view. Overlapping panels permit the examination of biases through comparison of results.
- Panel sample size – this discussion is rather limited and states that the sample size for a given fixed annual budget is determined by the factors in the first four dot points.

27 As for many other factors in a longitudinal survey design, decisions on overlapping and rotating sample will depend on a number of other issues. The rate and degree of change in the population being studied, the likelihood of respondent fatigue and the length of the longitudinal survey will all drive the decisions. In most cases, we would expect that the optimal design would be one of overlapping and non-rotating samples. The overlapping component would address both the reality of respondent fatigue and the changing nature of the business population in Australia. The length of time between the overlapping samples would be dependent on how quickly this change occurred, as would the reaction from analysts to this concept.

28 When making decisions on the design of the BLD these various options have been considered and presented to an external reference group. The use of repeated surveys was dismissed, based on the fact that the ABS produces high quality point-in-time estimates from its current suite of economic surveys, and also that should estimates of data specific to the BLD be required for a specific time point, the sample could be supplemented to become representative of the current population (as mentioned in paragraph 7). Repeated panels were considered, where after a number of waves a new and independent panel sample would be selected, starting another cycle. This would require techniques for linking the two panels. This has the advantage of being simple to implement, but is limited by the ability to get matched samples at the change-over period and the possible need to collect retrospective data. The diagram below represents this concept, where the circles are data collection

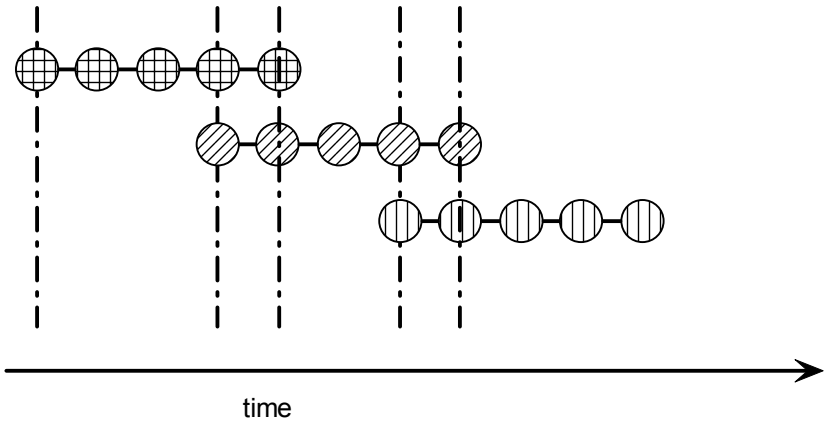
points. Different shadings indicate that each new panel is independent of the previous.

Diagram 2: Repeated panels



29 Overlapping panels were also considered, where rather than relying on a parallel sample at the change-over period only, a new panel is introduced to run in parallel with the current panel for a number of cycles. This allows for differences between the two panel samples to be measured. This is shown in Diagram 3. Here there are two cycles where the panels run in parallel, but this could be less or more.

Diagram 3: Repeated panels with overlap



30 A final decision on the choice of the linking of repeated panels, as well as the number of periods which the panels will be run in parallel, has not yet been made for the BLD.

Question for MAC: From an analysts point of view, what are your thoughts on the choice for linking repeated panels for the BLD?

Changing population

31 The business population in Australia is not static, changing constantly as a result of new businesses being created (often referred to as births), businesses which cease operation (deaths) and businesses which undergo structural change. These changes often occur faster than we are able to keep track of, and we have developed sophisticated methods of dealing with situations where the real world business is structurally different to that which we have recorded. Similar methods will need to be developed to manage structural change when surveying a business for longitudinal purposes.

Businesses that undergo structural change

32 As mentioned in paragraph 5, one of the major differences between a longitudinal survey of businesses and that of persons/households, is the changing structure of businesses over time which make following a given unit quite difficult. In some sense, a business can be thought of as analogous to a family in a social survey which changes due to marriage, divorce, children leaving home and so on. However, while it is easy to continue to follow the various persons of the original family in the social survey to avoid issues of how to follow an altered family, it is much more difficult to do so in an economic survey.

33 There are a number of reasons why a selected business will change in structure over the life of the longitudinal survey. A business may merge with another business, be wholly or partly taken over, split into multiple new businesses, take over part or all of another business or any combination of these. For any of these situations, it will be difficult to continue following the original business. Even where it may be possible to do so, this might not be the sensible approach as the business has undergone a change and should be treated as such. However, simply removing such businesses from the dataset removes information available to analysts who wish to study reasons for structural change.

34 For businesses included in the BLD sample which undergo structural change, we are proposing to link the old and new business entity wherever possible, which will allow the analyst to decide how to treat them in their analysis. Examples of structural change and our proposed treatment are:

- Splits – where a business has split into multiple new businesses each of the new businesses would be added as separate entities on the dataset, and flagged as being part of the original unit. Data for the original unit would not be recorded after the split data and data for the new units would not be recorded before the split.
- Merger – a new business created as a merger of two or more businesses will have different treatments depending on how many of the original

businesses were selected in the survey. If all businesses involved are part of the survey then the treatment will be the opposite of the split situation just described. If some of the businesses which merge are not included in the sample then a new business will be added to the dataset, with information on the original businesses appended.

35 For other structural changes it will not be as easy to track the old and new units, and each new case may require its own solution. A series of rules will need to be created to ensure that every type of structural change is treated in the same way. An initial set can be created based on structural change information that is already available within the bureau, with additions made as new changes are discovered. A method of detecting such structural change should be put in place, similar to that used in the cross-sectional business surveys, where respondents are asked if the business information provided on the front of the form is still applicable. The final dataset structure will need to be sophisticated enough to enable the proposed linking.

Population deaths

36 Another cause of population change is businesses who cease operating (i.e. 'die'). If this occurs part way through the life of the longitudinal survey the ceased business can in some ways be thought of as a form of attrition. However, we feel that this is not attrition in the true sense as the information after the time of death is not missing as it would be for a non-responding business, rather the fact that the business died is information in itself. It does however, affect the size of the live sample available at the end of the survey.

37 There may therefore be a need to increase the size of the original sample to ensure that the size of the live sample at the completion of the survey period is sufficient for analysis purposes. A longitudinal dataset with a large amount of deaths may not be very useful, except for perhaps prediction of business death. One method is to select an original sample larger than needed based on the expected death rate that will occur over the life of the longitudinal survey. This can only be estimated or modelled from previous information. It also suggests that the longitudinal survey will have a fixed number of waves.

38 The ABS has some experience in this area with the quarterly Labour Price Index survey (LPI). The original sample size is chosen using an expected death rate so as to ensure that at the end of each year the live and responding sample is of a given size. For the BLD, the optimal situation would be to use such a methodology, selecting a large enough original sample to ensure that sufficient live sample is available at the cessation of the current panel. However, as suggested in paragraphs 19 and 20, the most likely basis for the first panel of the BLD is the current Innovation survey. It is our intention to review the impact that deathing over time will have on this sample. If

this suggests that the available live sample after a given period of time will not be sufficient, we may need to augment the current sample. For the following BLD panels we will be in a better position to determine the optimal sample size needed.

Population births

39 Another major consideration when dealing with the business population is what to do with new businesses that begin operating after the commencement of the longitudinal survey. In a longitudinal social survey people new to a selected household are usually included in the sample from the point of entry onwards. New households which are formed by one or more people from a selected household leaving and joining or creating a new household are also followed. However, this behaviour is actually like that of businesses which split, merge and so on, rather than being analogous to true business births. A new business entering the population which is unrelated to any business currently selected (or not selected) is analogous to a new household being formed which has no links to any current households (say when a family migrates to Australia from overseas). If the number of such occurrences was similar in both populations (that is, the proportional number of totally new households in a given time period is equivalent to the proportional number of new businesses), then similar methods could be employed. We believe that it is a fair assumption however to state that the number of new businesses is greater. If this is the case, there remains the question of what should be done with such businesses.

40 There are a number of possibilities. Firstly, these businesses could be ignored and any analysis conducted on the resulting dataset relates to the original population only. Secondly, the weights of those selected businesses could be adjusted at each wave to show the differing population size. This however, doesn't allow for the characteristics of newly formed businesses to be analysed, except for those that were new at the time the sample was selected. The third possibility is therefore to add a number of new businesses to the sample in future waves. This will mean that longitudinal data is not available for such businesses for the whole of the period, however it could be argued that the missing data prior to the point of birth is valid information, just as missing data after a business dies is valid data. It is our intention to supplement the BLD sample at regular intervals to include a number of businesses that began operating after the previous supplementation occurred. This will provide analysts with information on the characteristics of new businesses, and possibly of more interest, their performance over the first few years of operation.

Question for MAC: Do you feel that the described population changes would have a great impact on the final longitudinal dataset produced if left unaddressed? If so, do you believe that our intended actions are appropriate/optimal?

Sample design

41 As we have discussed, it is our opinion that a good longitudinal dataset is one which is dense with information. We are also assuming that the size of the sample affects the usefulness of the dataset, but are not aware of any methods for objectively determining the size for general purpose datasets. In the survey methodology field there is very little that links the design of a panel survey to the analysis to be applied. Where decisions on sample size are mentioned at all, it is generally stated that these are dependent on available budget. There is however good discussion on the power of a given sample size for various analysis techniques, which can be used when considering sample size. However, it would be very useful to be able to construct a dataset based on how it should be sized and structured using an objective and robust methodology, rather than based on knowing all the various analyses that will be performed using it. This would mean that there is no (or little) restriction placed on the type of analysis that can be performed. This also potentially allows for the use of techniques that were not known or widely used at the commencement of the survey.

42 Associated with this question of sample size is the distribution of this sample. It may be that the optimal sample needed at various sub-levels should be determined and then summed to give the required total sample size, rather than allocating a required total across the sub-levels. This is similar to allocating sample for a cross-sectional survey to meet accuracy requirements at a state and industry level. It should be noted that the optimal sample for a longitudinal dataset created for modelling purposes is likely to be quite different to that of an optimal sample for producing point-in-time estimates. In usual business surveys the allocation tends to be disproportionate since the population is skewed (the small number of large businesses contribute the majority of the final estimates). Such a disproportionate allocation is unlikely to be optimal for estimating parameters of statistical models, rather a sample spread over the domain of study would be most useful. This was mentioned briefly in paragraph 7 where we stated that the BLD would not be designed to produce accurate cross-sectional estimates as this would compromise its ability to meet the longitudinal modelling requirements.

43 In many of the papers we have reviewed, the problem of sample size becomes one of balancing costs associated with infrastructure, contacting and surveying each business and the number of items collected each period. Since the trade-offs cannot be considered in isolation of the costs, it is important to develop good cost models. We have not undertaken much research in this area but expect that models which account for:

- fixed cost associated with producing a database each year;
- the cost of adding administrative data;

- the cost of adding data from other surveys;
- the cost of keeping businesses in the BLD panel after they rotate out from ABS surveys;
- the cost of conducting a specific survey to collect business characteristics; and
- processing costs (editing, validation);

would need to be developed in order to manage the long term funding arrangements of the BLD.

44 We have become convinced recently that this is the path that we need to take when determining the sample size for the BLD. With such a cost model various scenarios could be compared by cost and relative advantage. This relates to the framework for dataset design that was proposed in paragraph 18.

Provider load implications

45 The burden placed on respondents is of primary consideration when undertaking any survey, and even more so when undertaking a longitudinal survey. Respondent fatigue is a common issue in panel surveys, as respondents become bored or disinterested in taking any further part in the survey. There are various methods employed to attempt to reduce this level of fatigue, namely providing respondents with incentives (often monetary) and well developed procedures for reversing refusals. While the ABS has the advantage of the Census and Statistics Act, 1905 which will ensure higher response rates than would be expected if the survey was conducted elsewhere, the load placed on our providers is under constant scrutiny so methods to alleviate this burden will be needed.

46 The ABS policy on provider load states that small businesses will not remain in a given survey for more than three years, and that every effort is made to minimise the number of surveys that any given business is selected in. These two policies don't apply to large businesses as they contribute a significant proportion to the survey outputs. If the first of these constraints were adhered to it would be impossible to follow a given (small) business for more than three years. For any longitudinal survey where the desired length is greater three years this will be an issue. There is one ABS economic survey that retains small businesses in sample for five years via an exemption to the provider load policy, so it may be possible for a longitudinal survey to receive the same exemption. (It is for this reason that the diagram in the section on panel surveys showed each panel running for five years.) Given the lack of any precedence for more than five years it is unlikely that longer than this could be obtained. This effectively gives a maximum length for a longitudinal survey conducted by the ABS of five years.

47 A potential method for overcoming this is to use linked panels which were discussed in paragraph 28 above. This would only be a useful solution if a five yearly period between links was suitable to analysts, and further if the analysts were able to use techniques to link the data. It would also require some expertise within the ABS on such methods so that potential users of the data could be assisted.

Budget constraints

48 Another constraint placed on the design of a longitudinal survey is that of budget. In reality, there will be a limit on the amount of resources (monetary and others) that would be available to conduct such a survey. Budgeting for such a survey would also be significantly different to budgeting for a point-in-time survey as resources would need to be secured for years in advance. This may be quite difficult to do accurately given our limited knowledge of the processes needed to undertake such a survey.

49 Undertaking a longitudinal survey would require more resources in certain phases than would a point-in-time survey of similar size and complexity. Tracking of selected units would need to be undertaken. Literature suggests that tracking is a significant resource issue for household surveys, and although theoretically easier in the business sense, there would still be additional resources required over a point-in-time survey.

50 Designing for a fixed budget is not unique to longitudinal surveys, the majority of ABS economic surveys must work within a fixed annual budget which impacts on both the possible sample size, scope and amount of information collected. The difference here is that this budget will be longer term and possibly more difficult to manage. Any areas of overspend in a given year may affect the size or scope of the panel in future years if further funding cannot be secured.

Question for MAC: Do you feel that our approach of developing good cost models for assessing various sample size scenarios is the best way to tackle the sample size problem? Or is there an objective method available?

Treatment of missing data and weighting

51 Our proposed methods of managing businesses from which we can no longer collect information (i.e. deaths and some of those involved in structural change) were discussed in paragraphs 32 to 38. We will also receive no information from businesses that fail to respond to the survey in a given wave or waves. These require a different treatment to that which will be used for deaths and units involved in structural

change, as data is actually expected from these businesses. For this reason, non-response is referred to as a source of 'missing' information. There are generally two methods used in longitudinal (and point-in-time) surveys to account for missing data, weight adjustment and imputation. These are discussed in this section.

Businesses that fail to respond

52 There is potential for the level of non-response in a longitudinal survey to be higher than that in a single point-in-time survey. Rather than an individual business being flagged as either a respondent or non-respondent, each businesses included in a longitudinal survey will have a non-response pattern, representing whether they responded or not for each wave of the survey. In a longitudinal sense, a full respondent would be a business who responded to every wave of the survey. For all other possible non-response patterns, the question of how to treat the non-responding business arises.

53 One option is to leave the missing data on the dataset, however this makes micro level analysis more difficult. One simple approach is to exclude from any longitudinal analysis a particular business if data for it is not available for all time points under analysis. This effectively reduces the size of the dataset available for analysis, and doesn't make use of a large amount of data that has been collected (i.e. that from businesses in periods when they did respond).

Imputation and weight adjustment

54 Imputation, that is the substitution of data where it is missing, is used in the majority of ABS point-in-time business surveys. Using imputation to assign data to non-responding businesses has the advantage of preserving the size of the longitudinal sample. The method of deriving the impute varies depending on the relationship between the collected and available auxiliary variable, the nature of the data items, whether historical data is available, and so on. Note that in a longitudinal survey there is a greater potential for imputation using historical data than would generally occur in a cross-sectional survey (unless some form of rotation is employed whereby businesses remain in survey for a specified number of cycles).

55 The alternative to imputation is to adjust the weight of those businesses that are responding to account for those that aren’t. (See below for further discussion on weights.) This is a less common method in the ABS as it doesn't allow for different responses from different units, in effect all non-responding businesses receive the average value of those that did respond. Accounting for non-response via adjustment of weights reduces the size of the longitudinal sample, since a business need only non-respond in one wave to be excluded.

56 In the literature there is divided opinion on whether imputation is a preferable approach to weighting, effectively increasing the amount of longitudinal data available. See for example Lepkowski (1989) where he concludes that there is no clear favourite. During discussions on the BLD development there has been a suggestion that analysts would actually prefer no imputation to be undertaken and the non-respondents retained on the dataset with missing data. Analysts wanted to make their own decisions about imputation and suggested that modern software packages can handle imputation. Given the ABS's considerable experience at deriving optimal impute values, and the large amount of information about the non-responding business that we have available, we believe the most sensible option is to impute values for businesses which don't respond. These will be flagged on the dataset as having been imputed. Analysts are then free to include or remove these units from their analysis.

Weighting

57 For a point-in-time survey, every selected business is given a weight which is the inverse of its selection probability. This allows data collected from the sample to be inflated to represent the entire population. There is only this one weight that should be used when compiling estimates or undertaking analysis (this original weight may be adjusted to account for non-response as mentioned previously).

58 For a longitudinal survey there are multiple weights available. For each wave of the survey a different cross-sectional weight could be assigned to represent the current population size and distribution (even if the sample has not been updated). Also, at each wave a longitudinal weight could be assigned which reflects the sample design of the longitudinal dataset. This gives the potential for at least two weights for each wave of the survey. A further level of complexity can be added by including different weights for stock and flow data items.

59 From our reading and discussion with internal and external analysts, there doesn't appear to be a definitive position on the issue of weights, and it seems that the choice of weight is more dependent on the choice of analysis. For this reason, the BLD will have a number of commonly used weights attached. Decision on what should be included in this set of weights will need to be made in conjunction with major analytical users, and MAC's suggestions are sought.

Question for MAC: Does our suggestion for addressing the issue of non-response seem reasonable? Do you have suggestions for the set of weights which should be included on the dataset?

Confidentiality

60 It is ABS policy that no information will be released that compromises the undertaking of confidentiality we have made with providers. In practice this means that aggregated data will not be published or released at a fine level if the major proportion is from one business, or there are fewer than three businesses contributing. Data suppression occurs in these instances. When releasing unit record information, any identifying information is removed (i.e. name, address etc), units that are spontaneously identifiable are removed (such as very large businesses in certain industries who will be recognisable from other information on the dataset) and some data perturbation occurs to maintain both the confidentiality and structure of the dataset.

61 Although the initial purpose of the BLD was to provide information on the small and medium business sector, information from large businesses is still of interest. It is primarily the impact that the confidentiality policy has on the release of information from these large businesses that is of concern. If these cannot have their data released they will not be able to form part of the longitudinal sample, thereby effectively reducing the scope of the survey to small and medium businesses only. For some such small and medium businesses there may also be difficulties with releasing their data as spontaneous recognition will occur if they undertake an unusual activity for example.

62 There is little that can be done to avoid these outcomes without conflicting with ABS policy. Discussion with analysts on the impact of these scope restrictions should take place before agreement to undertake a longitudinal survey is made. During the development stage of the BLD discussion on this issue has occurred with expert analysts who will use the dataset. Their opinion is very strongly put,

> "that there is little point in developing the ultimate longitudinal dataset if we will not be able to gain access to it."

63 The one argument that can be made for retaining large businesses in the longitudinal survey is that this data will be available for use by internal analysts, and possibly in the future for external analysts via on-site access to the data or other yet to be developed methods. Therefore, results of analysis which don't require the releasing of unit record data could still be made available. The ABS has made significant developments recently in the area of allowing analysts access to data while still maintaining our confidentiality requirements. For this reason, we plan to include all sized businesses in the BLD. The sample may need to be somewhat skewed towards the small and medium businesses so that in the short term the size of the sample which will be available for analysis is sufficient for their purposes.

Conclusion

64 Designing longitudinal datasets is a difficult prospect because of the extra complexity added by the time dimension. We consider the best way to determine the optimal sample size is in a way analogous to the standard cross-sectional allocation problem. The logical conclusion is that this type of approach is only suitable when the longitudinal dataset has a specific purpose. Whenever a multipurpose dataset is required the decisions about what size the sample should be become more and more subjective. This is not necessarily a problem, but requires a more general approach to be taken. Thus, like most authors, the problem is generally reduced to how best to use a fixed annualised budget to meet the analytical needs.

65 One of the primary aims of the BLD is to maximise data from existing sources by leveraging off the available tax data. Users and other experts have emphasised the need to get the right data into the BLD. Unfortunately one of the main limitations presently is the lack of data on business characteristics. In the paper we discussed the likely scenario as trying to maximise the data which will become available from the Innovation survey, by coupling this with a collection of business characteristics, and augmenting with financial data from tax sources. The starting sample size would be around 8,500 businesses. This is comparable to the GAPS sample size which analysts have found suitable for their analysis purposes.

66 Since the data will come from a number of sources we believe that developing appropriate cost models will assist in determining what scenarios can be accommodated for an annualised budget. We will also have to overcome restrictions regarding provider load and access to the final dataset. At this stage we are expecting that panels will exist for five years before a repeated panel is required. The concept of having a core set of data enables the BLD to be flexible over time to changing needs. From a technical point of view the best we can do is to check various scenarios to determine if the final design and sample size meets analytical needs.

67 From a database point of view we expect to produce a smaller-denser dataset with business structural changes linked, missing data imputed and flagged, and with multiple weights available. In our view this would provide a comprehensive dataset for the next ten to fifteen years.

References

Australian Statistical Advisory Council Paper (2000) *Collecting Data Through Longitudinal Surveys*, Internal ABS Paper.

Duncan, G.J. and Kalton, G. (1987) "Issues of Design and Analysis of Surveys Across Time", *International Statistical Review*, 55, pp. 97–117.

Diggle, P., Lian, K. and Zeger, S. (1994) *Analysis of Longitudinal Data*, Oxford, Oxford Science Publications, pp. 27–32.

Kalton, G. and Citro, C.F. (1993) "Panel Surveys: Adding the Fourth Dimension", *Survey Methodology*, 19, pp. 205-215.

Lepkowski, J. (1989) "Treatment of Wave Nonresponse in Panel Surveys", in Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M.P. (eds), *Panel Surveys*, John Wiley, New York, pp. 348–374.

FOR MORE INFORMATION . . .

INTERNET	**www.abs.gov.au** the ABS web site is the best place for data from our publications and information about the ABS.
LIBRARY	A range of ABS publications are available from public and tertiary libraries Australia wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our web site for a list of libraries.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our web site, or purchase a hard copy publication. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE	1300 135 070
EMAIL	client.services@abs.gov.au
FAX	1300 135 211
POST	Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All ABS statistics can be downloaded free of charge from the ABS web site.

WEB ADDRESS	www.abs.gov.au

2000001524428
ISBN 0 642 48166 0

RRP $11.00