Using Matching Techniques with Pooled Cross-sectional Data

Paul Norris
Scottish Centre for Crime and Justice Research
University of Edinburgh

pnorris@staffmail.ed.ac.uk

What is Pooled Cross-sectional Survey Data?

“In the repeated cross-sectional design, the researcher typically draws independent probability samples at each measurement point” (Menard, 1991, p26)

- Samples will typically contain different individuals

- Each sample reflects population at the time it is drawn

- Asks comparable questions to each sample

For more details on this type of data, and possible approaches to analysis, see Firebaugh (1997), Menard (1991) Micklewright (1994) and Ruspini (2002)

Why Use Pooled Cross-sectional Data?

Repeated Cross-sectional surveys are much more common than panel based survey data

Data available covering a much wider range of topics

Cross-sectional data avoids issues such as sample attrition

Researchers often more used to analysing cross-sectional data

Can give increased sample size for cross-sectional models?

Limitations of Pooled Cross-sectional Data

Does not involve following the same individuals over time

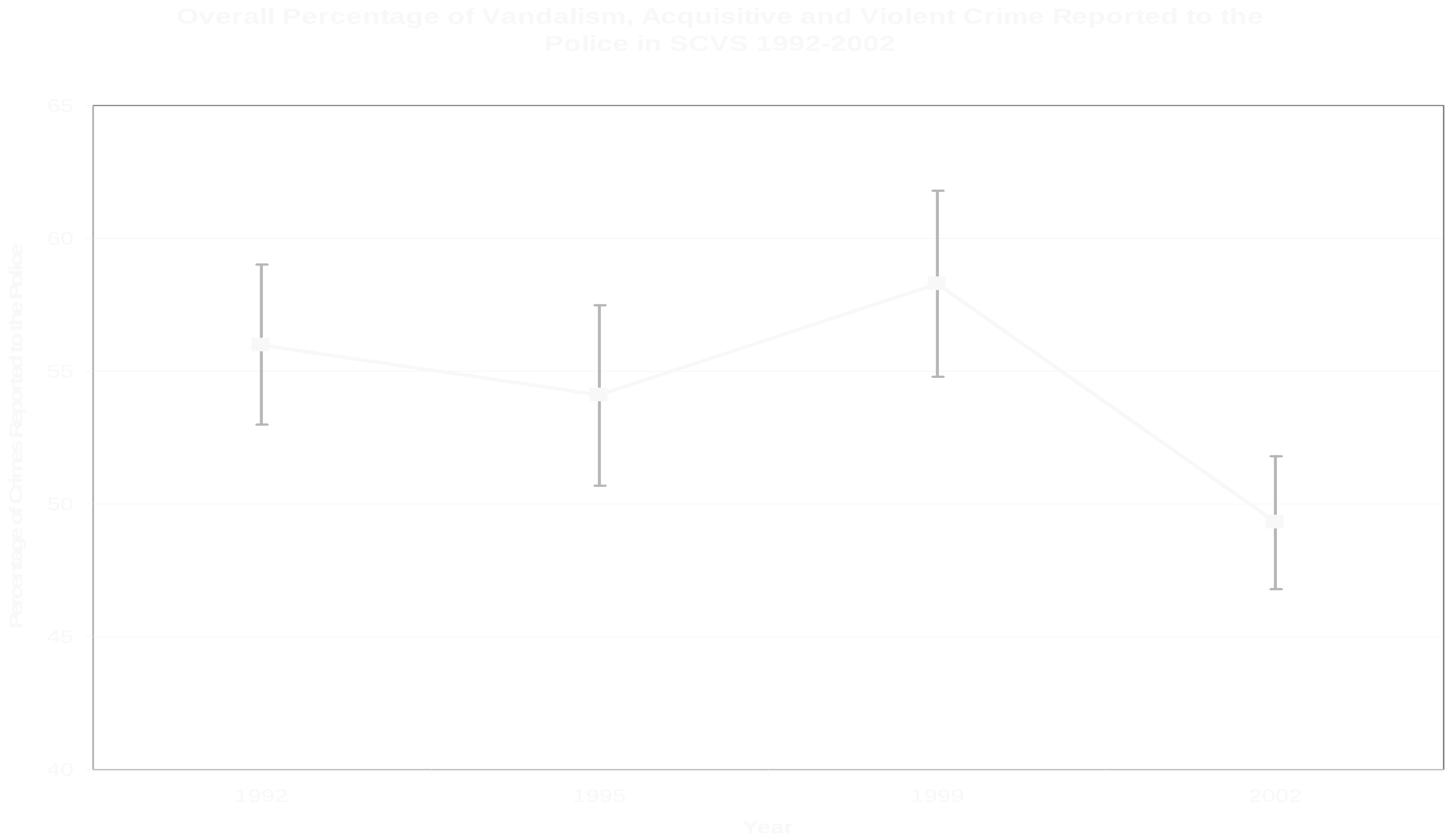
Most useful for exploring aggregate level change
– hard to establish intra-cohort changes

Difficult to establish causal order
- particularly at the individual level

Questions and definitions can change over time

For a discussion of the issues confronted when creating a pooled version of the General Household Survey see Uren (2006)

Studying Aggregate Trends



n:1992=1013, 1995=815, 1999=746, 2003=1251
Error bars show 95% confidence intervals

Which Shifts Underpin Aggregate Change?

Changes in an aggregate pattern can be attributed to two types of underlying shift:-

Model Change Effects – the behaviour of individuals (with identical characteristics) changes over time

Distributional Effects – the makeup of the "population" changes over time

For a more complete description of these terms see Gomulka, J and Stern, N (1990) and Micklewright (1994)

Separating Distribution and Model Change Effects

Estimates of distributional and model change effects can be created by considering what outcomes would occur if the behaviour from one time period was applied to the population from different time periods

Build up a matrix of predicted outcomes for different behaviours and populations

These figures allow us to see what would occur if population was constant and behaviour changed and vice versa

For an example of such a matrix see Gomulka, J and Stern, N (1990)

Comparing Reporting to the Police in 1992 with 2002

Imagine a simple case where the change in crime reported to the police is a function of two factors:

The mix of crime (Population distribution)

Willingness to report different crimes (Behaviour model)

		Reporting Behaviour	
		1992	2002
Mix of Crime	1992	55.7	
	2002		49.3

Estimating Alternative Reporting Rates

The missing figures on the previous slide can be calculated by applying the reporting rates for each crime from one year to the crime mix from the other year

1992	Proportion of Crime	Reporting Percentage
Vandalism	42.6	34.8
Acquisitive	40.2	79.3
Violence	17.2	51.9
Total	100	55.7

2002	Proportion of Crime	Reporting Percentage
Vandalism	54.5	42.6
Acquisitive	25.7	65.8
Violence	19.8	46.4
Total	100	49.3

Estimating Alternative Reporting Rates

The missing figures on the previous slide can be calculated by applying the reporting rates for each crime from one year to the crime mix from the other year

1992	**Proportion of Crime**	**Reporting Percentage**
Vandalism	42.6	42.6
Acquisitive	40.2	65.8
Violence	17.2	46.4
Total	100	52.6

2002	**Proportion of Crime**	**Reporting Percentage**
Vandalism	54.5	34.8
Acquisitive	25.7	79.3
Violence	19.8	51.9
Total	100	49.6

Updated Matrix With Estimated Reporting Rate

		Reporting Behaviour	
		1992	2002
Mix of Crime	1992	55.7	52.6
	2002	49.6	49.3

Both the change in the mix of crime and change in reporting behaviour appear to have lowered reporting between 1992 and 2002

Relative impact of distributional and model change effects depends on which year’s data is considered

What is Propensity Score Matching?

A method for identifying counterfactual cases across different samples

Employs a predicted probability of group membership—e.g., 1993 SCVS verses 2003 SCVS on observed predictors, usually obtained from logistic regression to create a counterfactual group

Matches together cases from the two samples which have similar predicted probabilities

Once counterfactual group is constructed – outcome is compared across groups

For a more complete description of propensity score matching see Sekhon (2007)

Using Propensity Score Matching to Estimate Distributional and Model Effects

		Reporting Behaviour	
		1992	2002
Mix of Crime	1992	55.7	52.6
	2002	49.6	49.3

The estimates provided by the propensity score matching are identical to those calculated earlier.

What a waste of a Thursday afternoon, or is it?

Generalising to More Factors

In reality changes in reporting are likely to be a function of more than just the two factors we have considered

Need to generalise the outcome matrix

		Reporting Behaviour	
		1992	2002
Population Distribution	1992	55.7	
	2002		49.3

Much harder to account for multiple factors in manual calculations

Factors Influencing Reporting to the Police

The decision to report crime to the police is likely to be a function of many factors

Type of Crime

Insurance

Attitude to the Police

Age

Quantity of Loss

Gender

Social Class

Income

Injury

Perceived Threat

Relationship to Offender

Culpability

Family Status

Social Context

Repeated Incident

Estimates Using “Full” Matching

		Reporting Behaviour	
		1992	2002
Population Distribution	1992	55.7	55.0
	2002	50.1	49.3

Matching on crime type, gender, age, social class, ethnicity, household income, weapon used, threat used, doctor visited, insurance claimed, value of damage/theft, Injury, took place at home, tenure and marital status

Change in population of crimes and victims seems to have lowered reporting rates

Reporting behaviour also slipped (but non-significant)

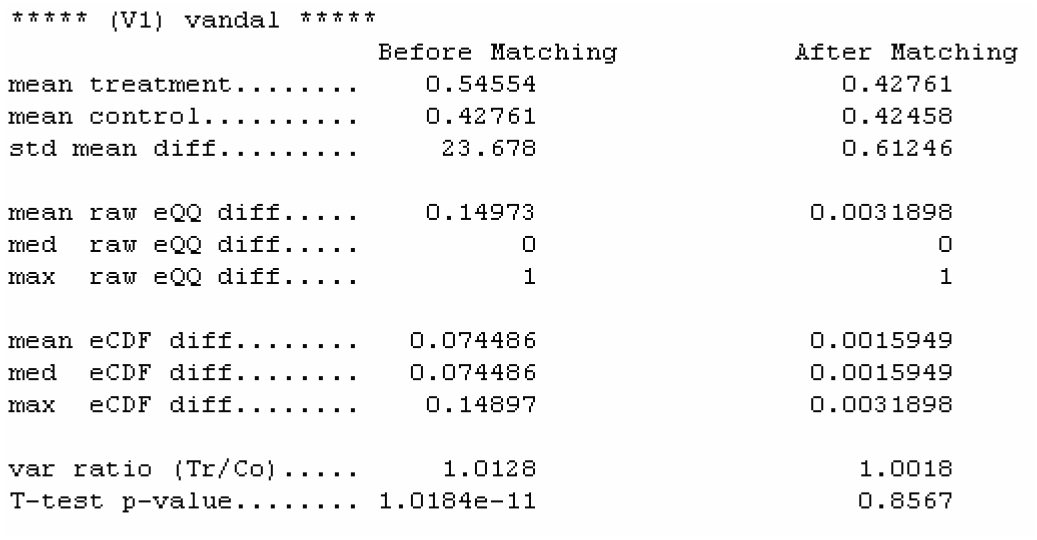
Change in reporting seems to be most related to distributional changes

Estimates appear more consistent across behaviour/distributional mixes

Balanced Samples

Propensity score refers to an “overall” indicator of differences between the two samples

Important to check characteristics of cases are evenly distributed across samples after matching



```
***** (V1) vandal *****
                       Before Matching          After Matching
mean treatment........    0.54554                  0.42761
mean control..........    0.42761                  0.42458
std mean diff.........     23.678                  0.61246

mean raw eQQ diff.....    0.14973                0.0031898
med  raw eQQ diff.....          0                        0
max  raw eQQ diff.....          1                        1

mean eCDF diff........   0.074486                0.0015949
med  eCDF diff........   0.074486                0.0015949
max  eCDF diff........    0.14897                0.0031898

var ratio (Tr/Co).....     1.0128                   1.0018
T-test p-value........ 1.0184e-11                   0.8567
```

Still issues of multivariate comparability

A more complete discussion of how to asses balance is given in Sekhon (2007)

Generic Matching

Achieving balance can prove difficult in propensity score matching

Generic matching is one possible approach to this problem

Aim is to maximise the p-value associated with the covariate which represents the greatest difference between the two samples

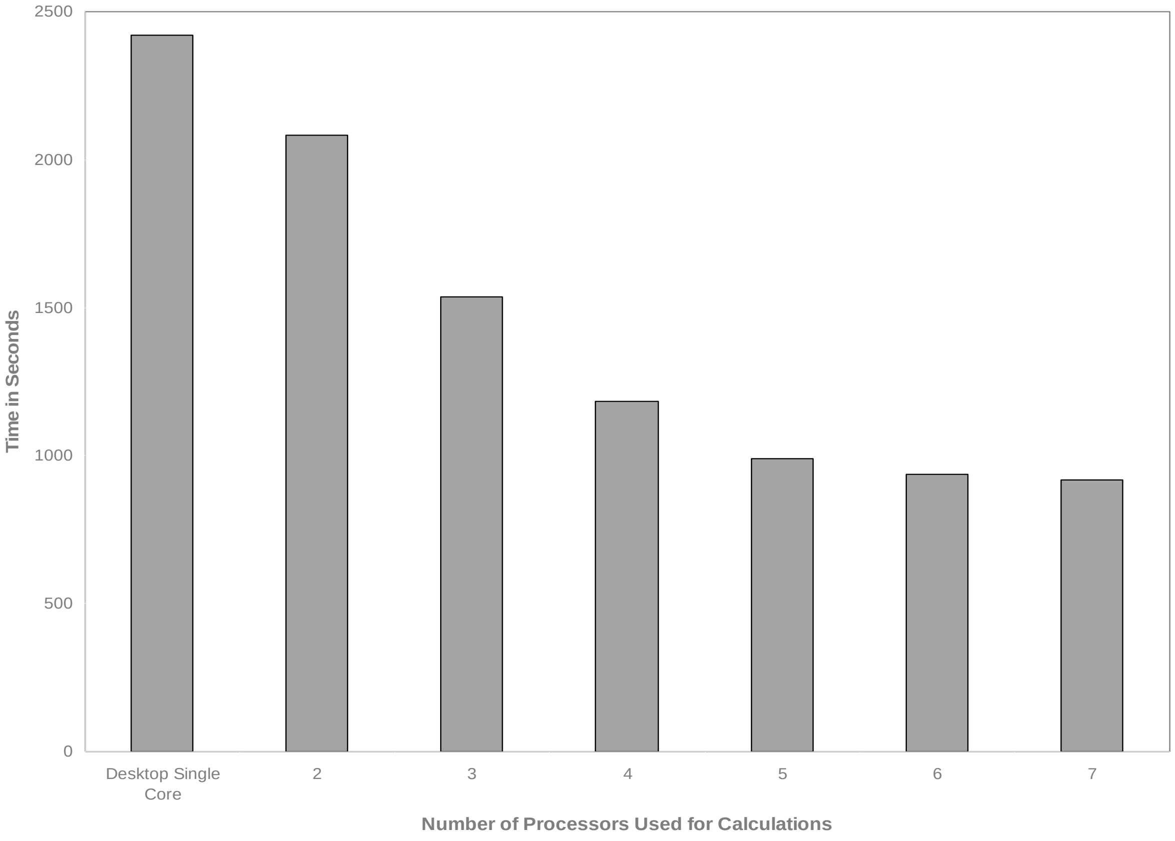
Uses an evolutionary algorithm to match cases

See Sekhon (2007) "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R." *Journal of Statistical Software.*

Generic Matching – Computational Issues

Generic matching is very computer intensive (both cpu and memory)

R routine can be used on a computer cluster



Analysis based on example dataset from Sekhon (2007) contains 185 treatment cases and matches on 10 variables

Strengths of Matching for Separating Distribution and Model Change Effects

Offers a perspective on social change over time

Intuitively simple – what is the change in outcome if we hold population constant?

Applicable to a wide range of data sources

Can be implemented in most standard software packages

Weaknesses of Matching for Separating Distribution and Model Change Effects

Only considers aggregate level change

Success relies on matching on all relevant factors

Comparability of data over time can be questioned

Issues around reliability of matching:-

- Can be difficult to achieve accurate matching using regression based methods
- Generic matching can be computer intensive

Bibliography

FireBaugh, G (1997) Analyzing Repeated Surveys. Sage Publications

Gomulka, J and Stern, N (1990) “The Employment of Married Women in the UK: 1970-1983" in Economica, 57(226): 171-200

Menard, S (1991) Longitudinal Research. Sage Publications

Micklewright, J (1994) “The Analysis of Pooled Cross-sectional Data" in Dale, A and Davies, R (1994) Analyzing Social Change. Sage Publishing

Ruspini, E (2002) Introduction to Longitudinal Research. Routledge

Sekhon, J (2007) "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R." *Journal of Statistical Software.*

Uren, Z (2006) The GHS Pseudo Cohort Dataset (GHSPCD): Introduction and Methodology http://www.statistics.gov.uk/articles/nojournal/Sept06SMB_Uren.pdf [cited 01/05/2008]