

Title: A Heap of Trouble? Accounting for Mismatch Bias in Retrospective Data

Authors: Dean R. Lillard, Hua Wang, Haim Yehuda Bar

Affiliation: Cornell University (all), DIW Berlin (Lillard)

Keywords: Heaping, rounding rules, Mismatch bias, retrospective data, cross-national comparability

Abstract

When researchers design and administer surveys they face the challenge of how to collect information about events that happened in the distant past. Such data enrich both cross-sectional and longitudinal survey but yield greater benefits when they can be merged to the rich history longitudinal data provide for a given individual.

Retrospective data are collected for a wide variety of topics that include marital events, births, deaths, purchase behavior, getting, changing, or losing jobs. In retrospective reports people tend to round up or down the year or time since the event occurred. Consequently, events tend to be “heaped” on multiples of chronological or calendar units (e.g. on units of five or ten for data that naturally occur over years). This pattern is evident in Figure 1 that shows the age that US and UK ex-smokers reported that they quit their habit. The GSOEP data in Figure 2 establish that respondents also use multiple heaping rules - some respondents heap on calendar time units (e.g. 2000, 1995, ...) while others heap on chronological time units (5 years ago=1997, 10 years ago=1992).

We use US (PSID), UK (BHPS), and German (GSOEP) data to document the heaping, show how it manifests itself and is shaped by the survey questions and response categories, and document the obvious and not so obvious rounding rules respondents appear to use. We describe the general problem heaping introduces. We investigate factors that predict who heaps. We develop and test algorithms that mitigate the mismatch bias associated with heaping. These range from a simple averaging rule to a more complicated adjustment to the likelihood function.

Our results confirm that heaping biases coefficient estimates downwards. Our algorithms reduce this bias. We conclude by discussing how our algorithms can be applied more generally to other types of retrospective data.

Figure 1 Distribution of “Age last smoked regularly” among US and UK ex-smokers

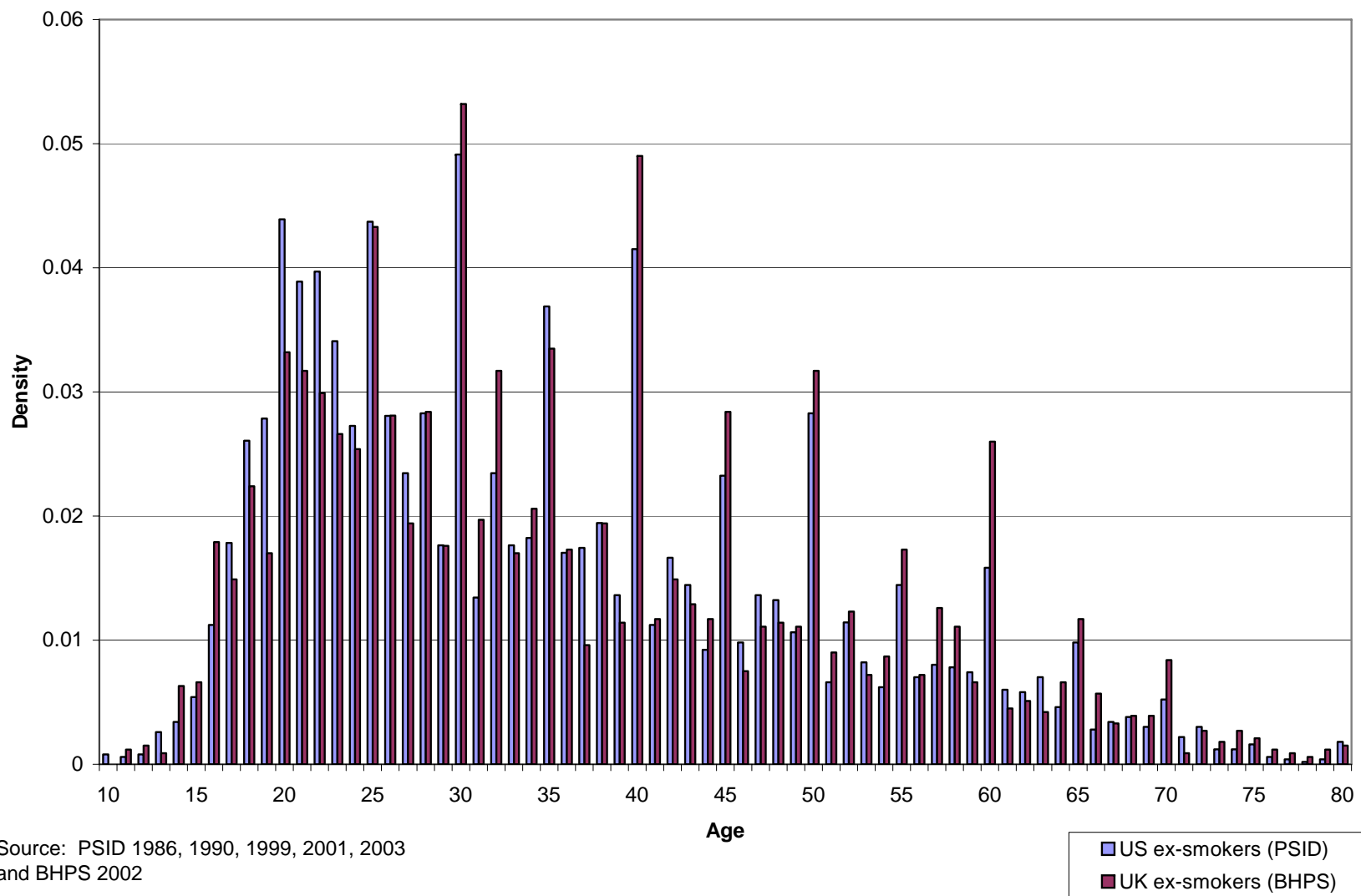
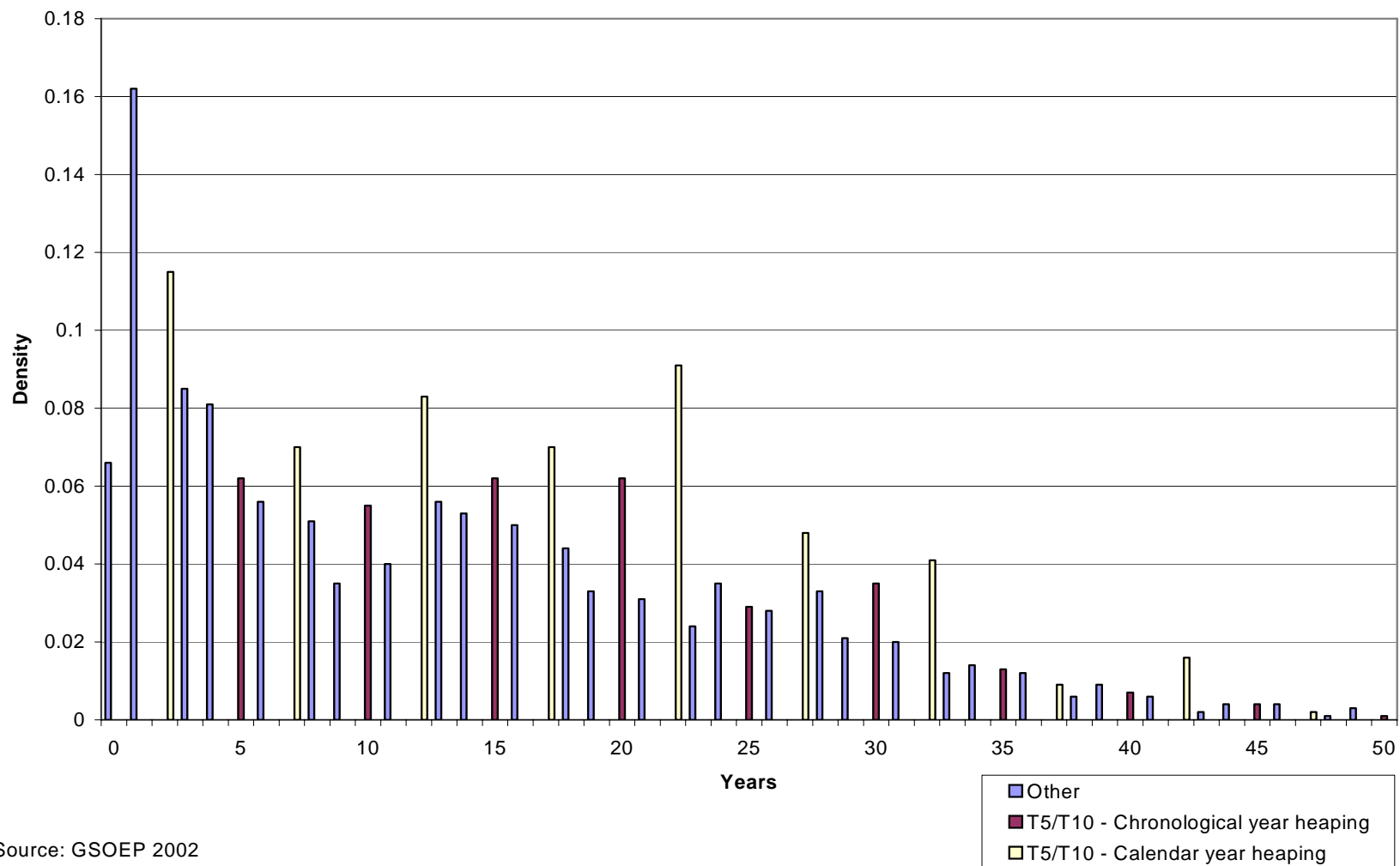


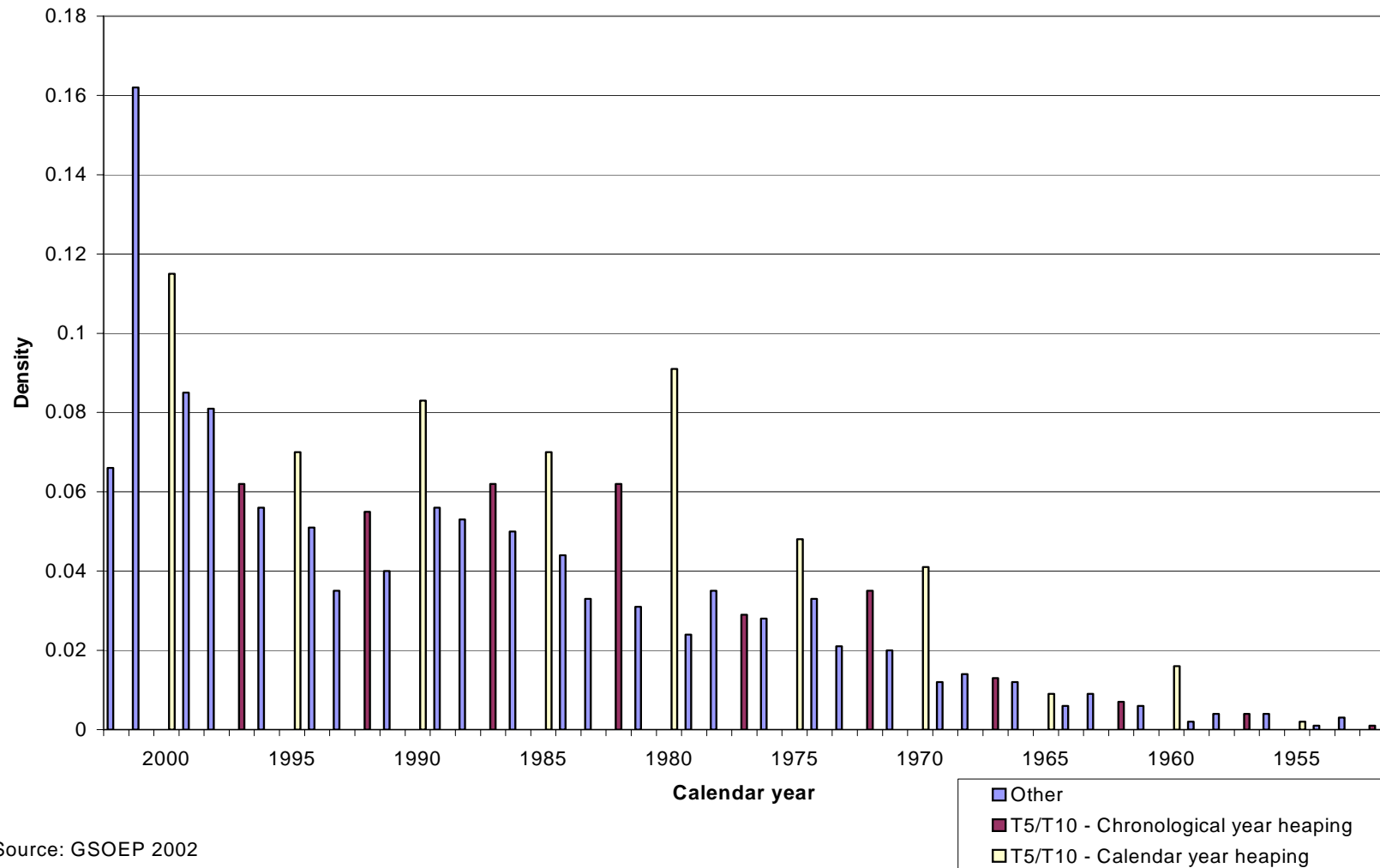
Figure 2 Evidence of mixed heaping rules in distribution of “Years since last smoked regularly” among German ex-smokers

A. Chronological time scale



Source: GSOEP 2002

B. Calendar time scale



Source: GSOEP 2002