Annex

Jens Mehrhoff, Deutsche Bundesbank

In order to run the R code, it is necessary to download (and extract) all category-specific files, i.e. the UPC files and movement files, in the CSV format newly available at the Dominick's website. Note that there are no CSV files available for refrigerated juices ('RFJ'). Further note that the UPC file for paper towels ('PTW') includes an undocumented variable WSTART which has to be dropped. Furthermore, the two CSV files provided that prepare the information on the week variable and the stores included are needed.

The R code generates analysis-ready data and derives price indices by means of the weighted time-product dummy method. For the sake of exposition the weekly store-level UPC data is aggregated to chain-wide item codes at monthly frequency but this can be changed. The R code is restricted to one particular category, where the three-letter acronym for the category can be adapted (see instructions below).

The R code is equivalent to the orignal SAS codes (please refer to the documentation by Mehrhoff, 2018, *Promoting the use of a publicly available scanner data set in price index research and for capacity building*) with the difference that the SAS codes calculate results for each category. The upc part reads in the UPC file. The move part reads in the movement file and calculates total dollar sales; suspect data is dropped. The weeks & stores part reads in the week and store files, and merges them with the movement and UPC files. The wtpd example aggregates the data, calculates unit prices as well as expenditure shares, and derives price indices by means of the weighted time-product dummy method; finally, a CSV file is written with the analysis-ready data that allows basing calculations on the very same data, thus discounting the incomparability of different data sets in research and training.

The two main outputs of the R code are thus the analysis-ready data stored in the move_monthly data set and exported to a CSV file, and the weighted time-product dummy regression results stored in the wtpd1 data set. The analysis-ready data contains 5 variables (see table). The price index can be derived as the exponentiated time dummy variables, labelled 'factor(MONTH)'.

Variable	Description
COM_CODE	Commodity code (Dominick's version of categories)
MONTH	Month in which the week ends (weeks run from Thu. to Wed.)
NITEM	Item code (attempt at tracking products across multiple UPCs)
MOVE	Number of units sold
SALES	Total dollar sales

NB: An individual product should always be defined as the unique combination of COM_CODE × NITEM.

Instructions on how to use the R code:

- The code assumes that the CSV files are stored in C:\Data (Windows operating environment). Hence, this would need to be adapted should the location be different.
- The UPC file and movement file are read in the upc part and move part, respectively. Remember that the UPC files are named 'upcxxx' and the movement files 'wxxx', where 'xxx' is the three-letter acronym for the category.