Age Heaping in Population Data of Emerging Countries

Andres Barajas Paz, Andrew Cairns, Torsten Kleinow

Heriot-Watt University, Edinburgh

ab108@hw.ac.uk

16th International Longevity Risk and Capital Markets Solutions Conference, August, 2021

HERIOT WATT UNIVERSITY

Actuarial Research Centre

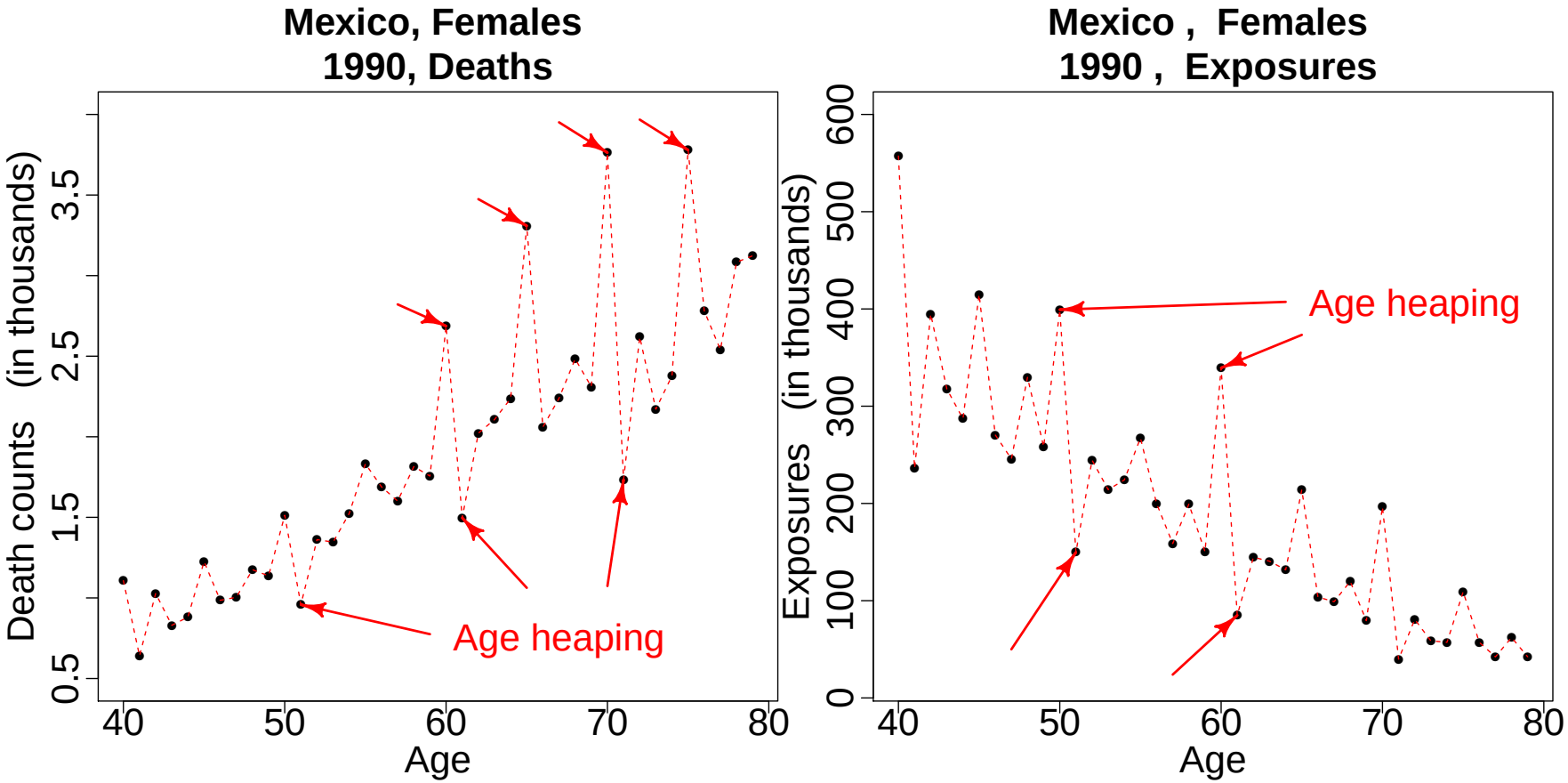
Institute and Faculty of Actuaries

Introduction.

- Motivation

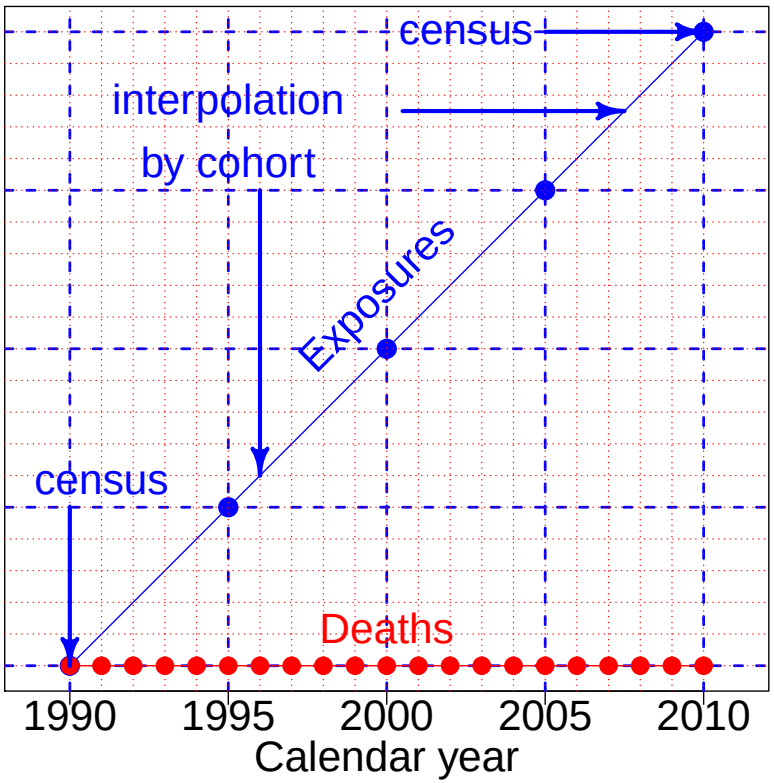
Mortality analyses have commonly focused on countries represented in the Human Mortality Database that have good quality mortality data. We address the challenge that in many countries population and deaths data can be somewhat unreliable. In many countries, for example, there is significant misreporting of age in both census and deaths data: referred to as "age heaping".

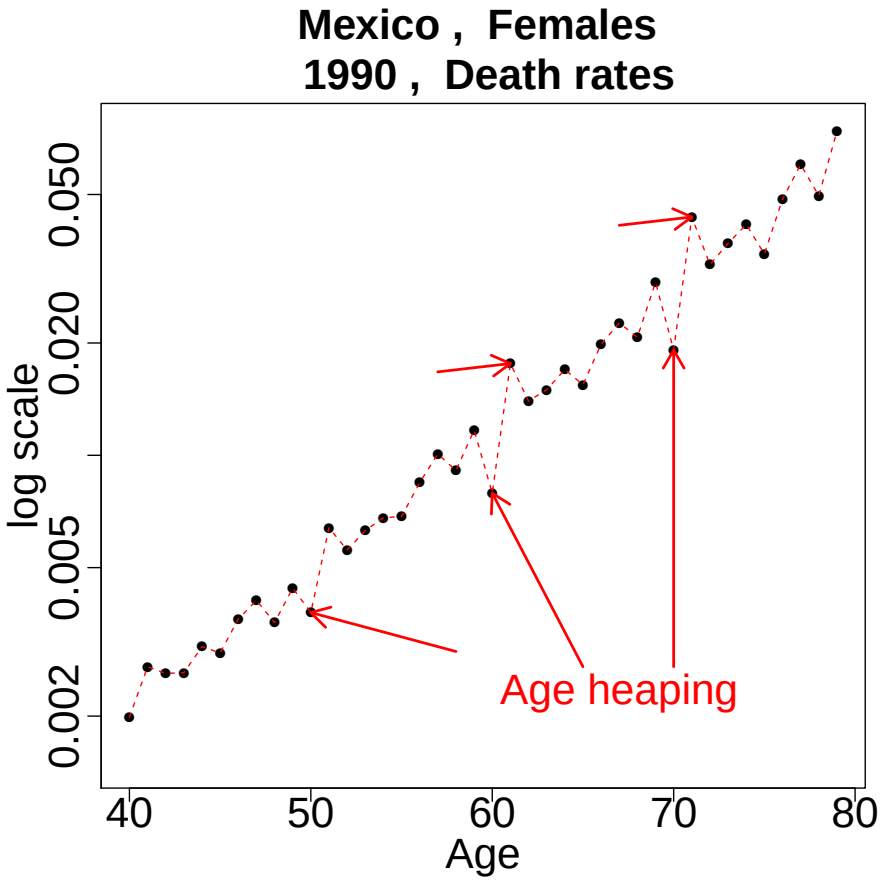
Introduction.



Age Heaping occurs when people misreport age.

Introduction.





Introduction.

- Main Objective

The purpose of our research is to develop Bayesian computational methods for fitting a new model for misreporting of age for emerging countries such as India and Mexico, where their population data is affected by age heaping.

Application: Reported data $\rightarrow$ Smoothed HMD $\rightarrow$ International Reinsurance.

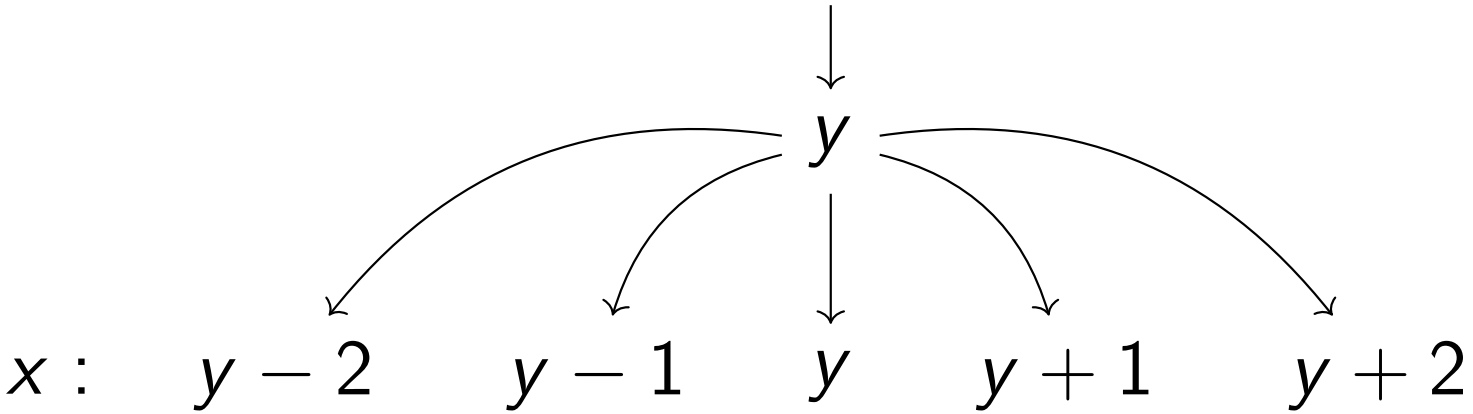
Data sources

- Death counts by single year and age from 1990 to 2015 (INEGI).
- Censuses with population counts by single age in 1990, 1995, 2000, 2005, 2010 and 2015 (INEGI).

Model and Notation

Our model for age heaping is based on the assumption that, reported exposures $\widehat{E}(t,x)$ have a Poisson distribution with

$$\mathbb{E}\left[\widehat{E}(t,x)\right] = E(t,y)g^{E}(t,y,x)$$



Model and Notation

Reported death counts:

$$\widehat{D}(t,x)\Big|\underline{m},\underline{E},\theta \sim Poisson\Big(\sum_{y} m(t,y)E(t,y)g^{D}(t,y,x)\Big),\ \forall t \in \mathcal{T}$$

$$\widehat{E}(t,x)\Big|\underline{E},\theta \sim Poisson\Big(\sum_{y} E(t,y)g^{E}(t,y,x)\Big),\ \forall t \in \mathcal{T}_c.$$

Mortality rate:

$$m(t,y) = \exp\Big[a(t) + b(t)(y-\overline{y}) + c(t)\Big((y-\overline{y})^2 - \sigma_y^2\Big) + d(t)\Big((y-\overline{y})^3\Big) + e(t)\Big((y-\overline{y})^4\Big)\Big].$$

Model and Notation

For deaths and exposures the model for g^* is:

$$g^*(t,y,x) = f^*(t,y,x)\exp\{H^*(x)\}k^*(t,y)$$
$$f^*(t,y,x) = \exp\{-\eta^*(t)(x-y)^2\}$$

$g^* \rightarrow \text{P[reported age } x \mid \text{true age } y], \sum_{x\in\mathcal{X}} g^D(t,y,x) = 1 \ \forall t,y$

$\eta^* \rightarrow$ Captures the increasing improvements across years,

$H^* \rightarrow$ Expresses individual preference, indifference or avoidance of certain ages x,

$f^* \rightarrow$ Propensity of misreporting.

Simulation for Canada.

- Given the true Canadian death counts $D(t, y)$ and exposures $E(t, y)$
- We assume $\eta^D(t)$ and $\eta^E(t)$ to be increasing

$$\left(\eta^D(1) = 2, \cdots, \eta^D(n_t) = 4\right) \ \forall t \in \mathcal{T},$$

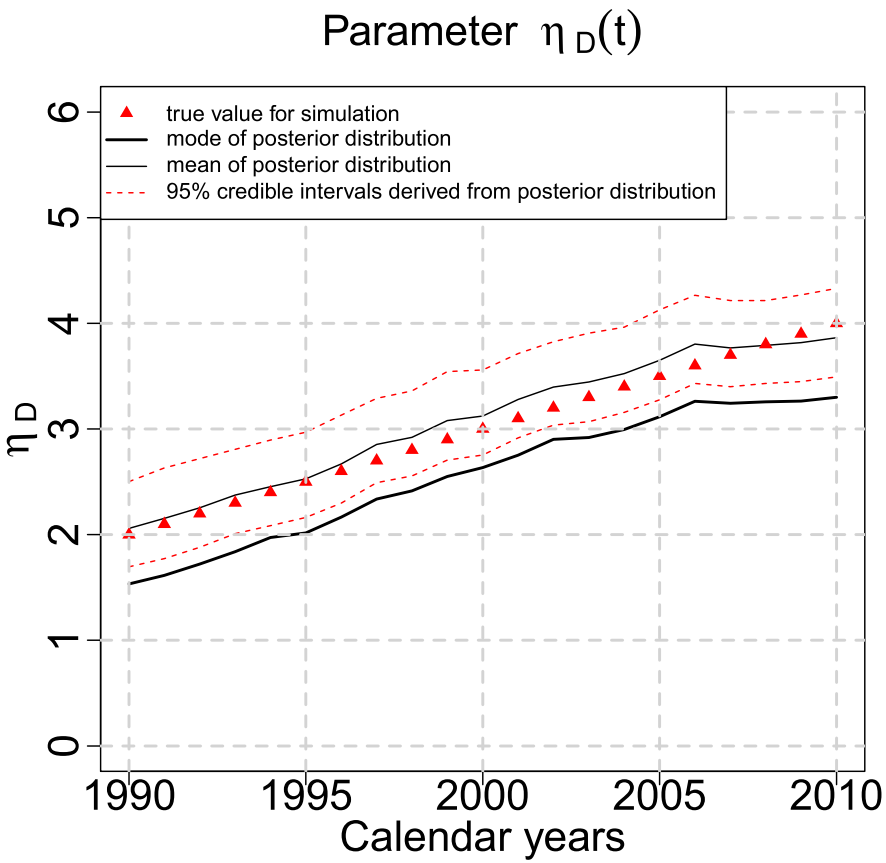
$$\left(\eta^E(1) = 2, \cdots, \eta^E(n_t) = 4\right) \ \forall t \in \mathcal{T}_c.$$

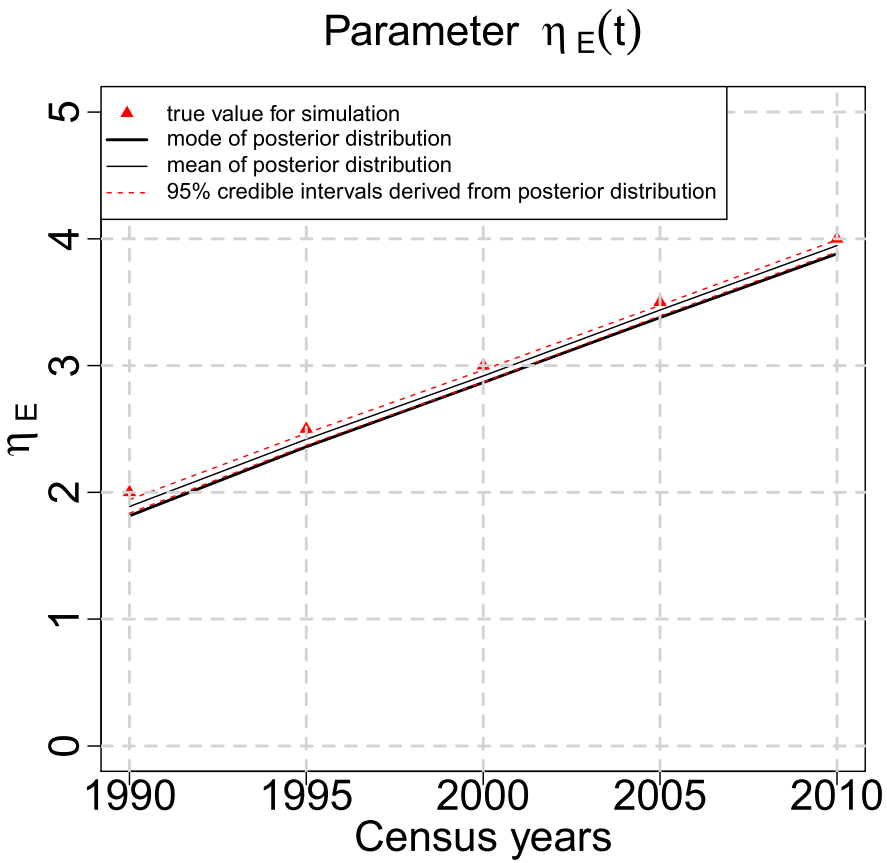
Simulation for $H^D(x)$ and $H^E(x)$, Canada.

$$H^D(x) = \begin{cases} 0.6 & \text{if last digit of x is 0} \\ -0.3 & \text{if last digit of x is 1 or 9} \\ 0 & \text{otherwise} \end{cases}$$

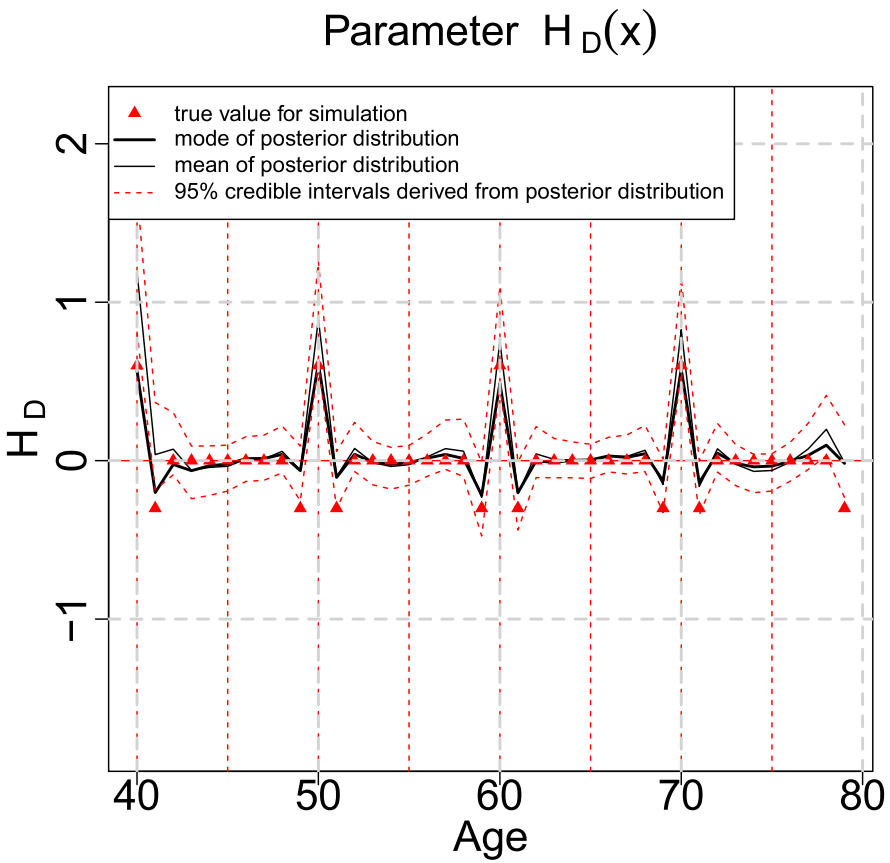
$$H^E(x) = \begin{cases} 0.8 & \text{if last digit of x is 0} \\ -0.4 & \text{if last digit of x is 1 or 9} \\ 0 & \text{otherwise} \end{cases}$$

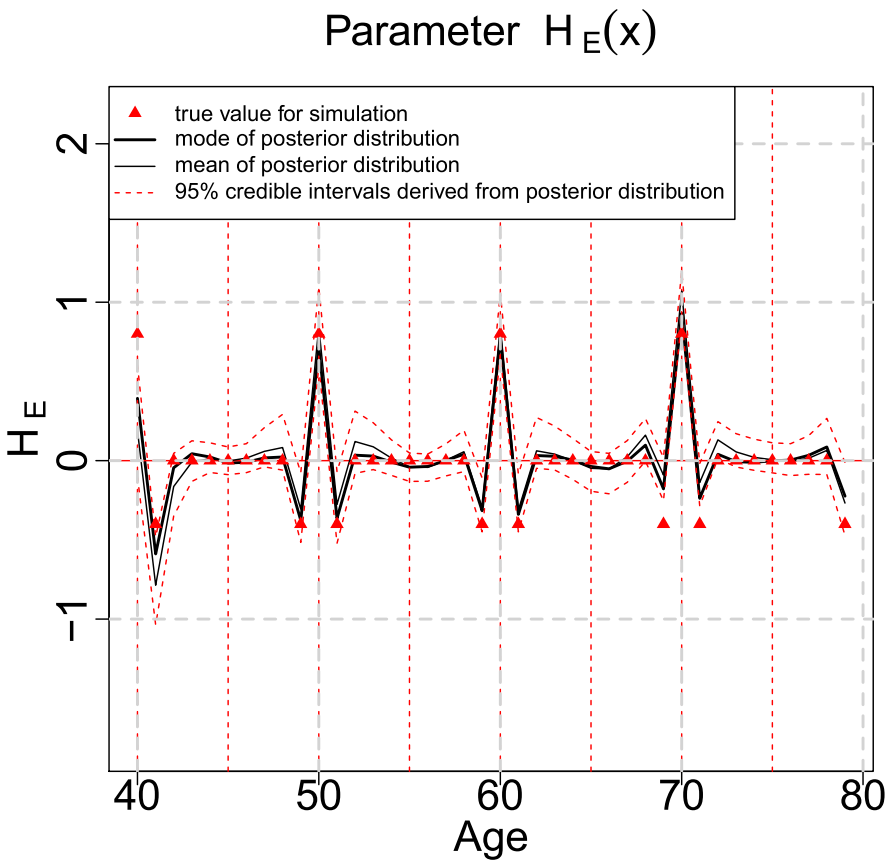
Parameter $\eta^D(t)$ and $\eta^E(t)$, Canada.



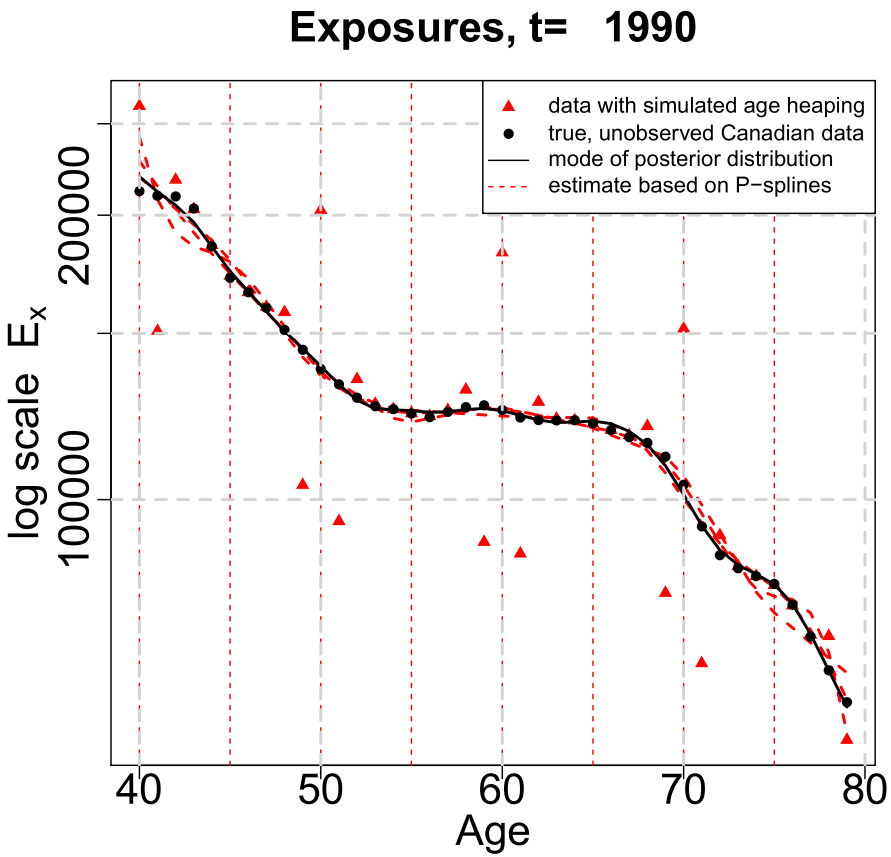


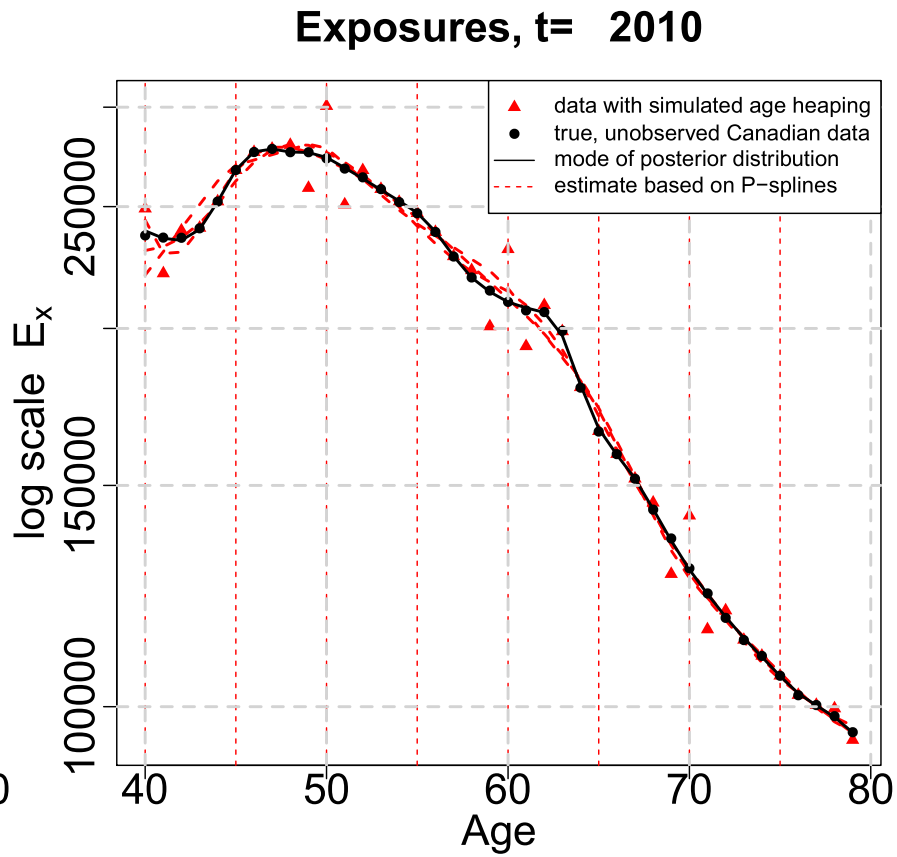
Parameters $H^D(x)$ and $H^E(x)$, Canada.



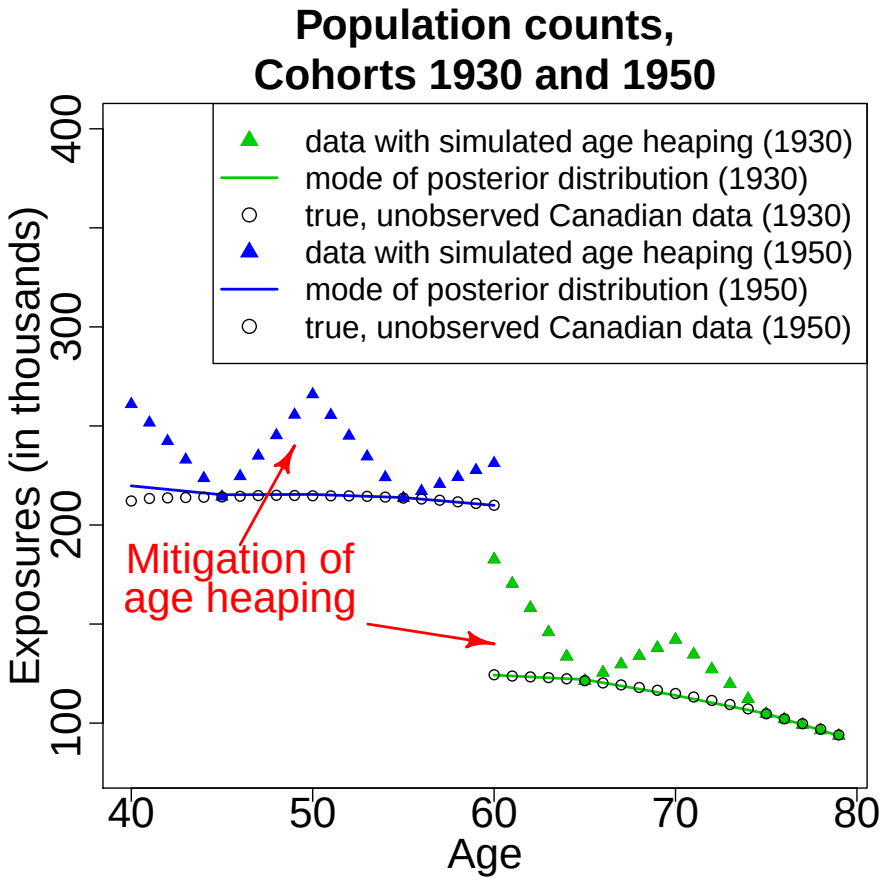


Exposures $E(t, y)$, Canada 1990 & 2010.

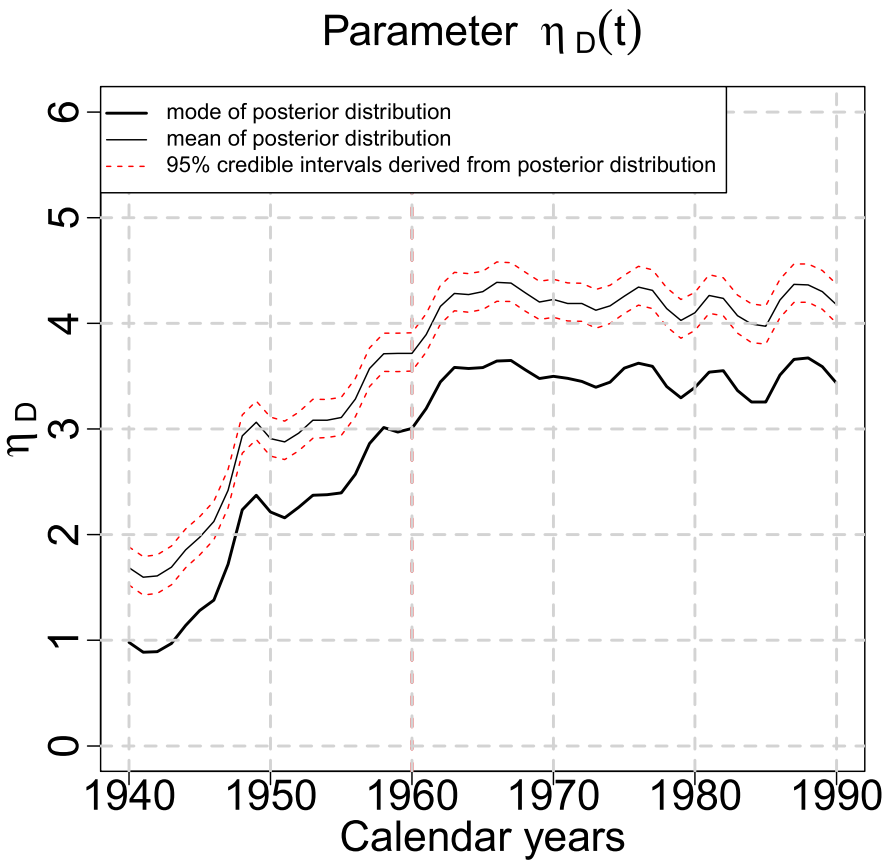


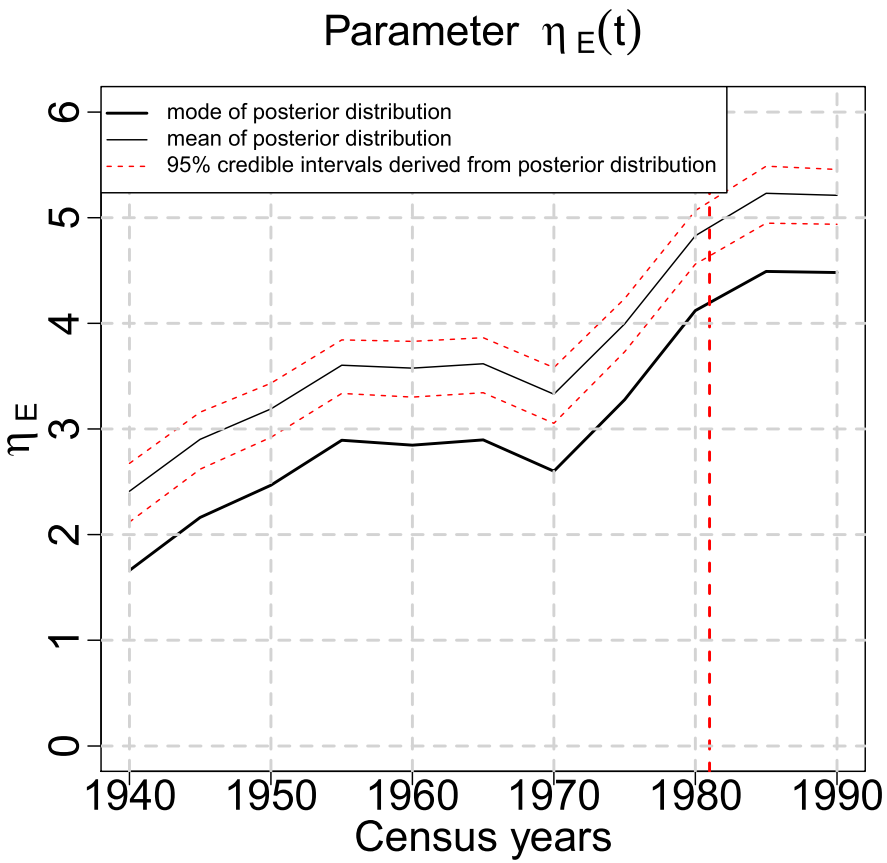


Exposures across cohorts, Canada.

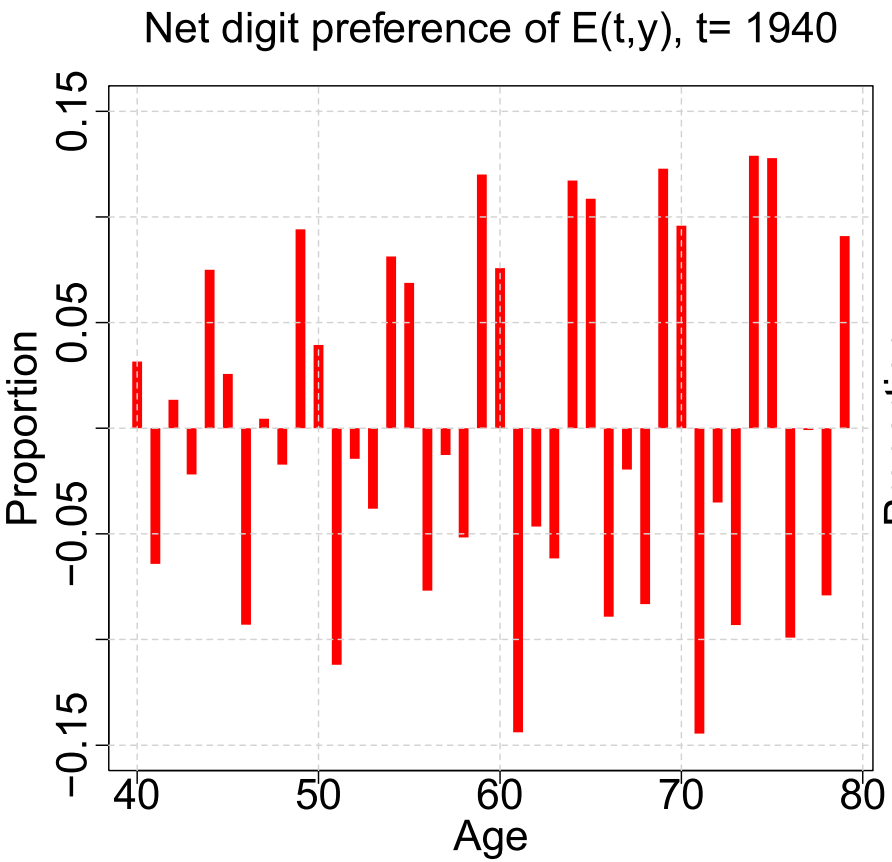


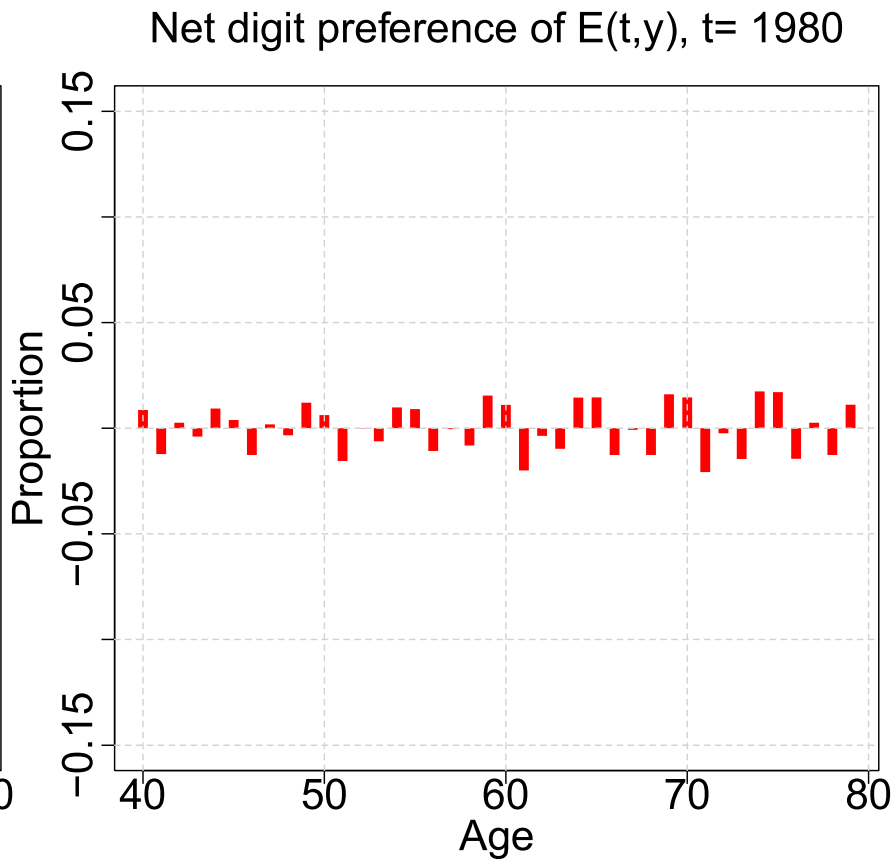
Parameter $\eta^D(t)$ and $\eta^E(t)$, Portugal.



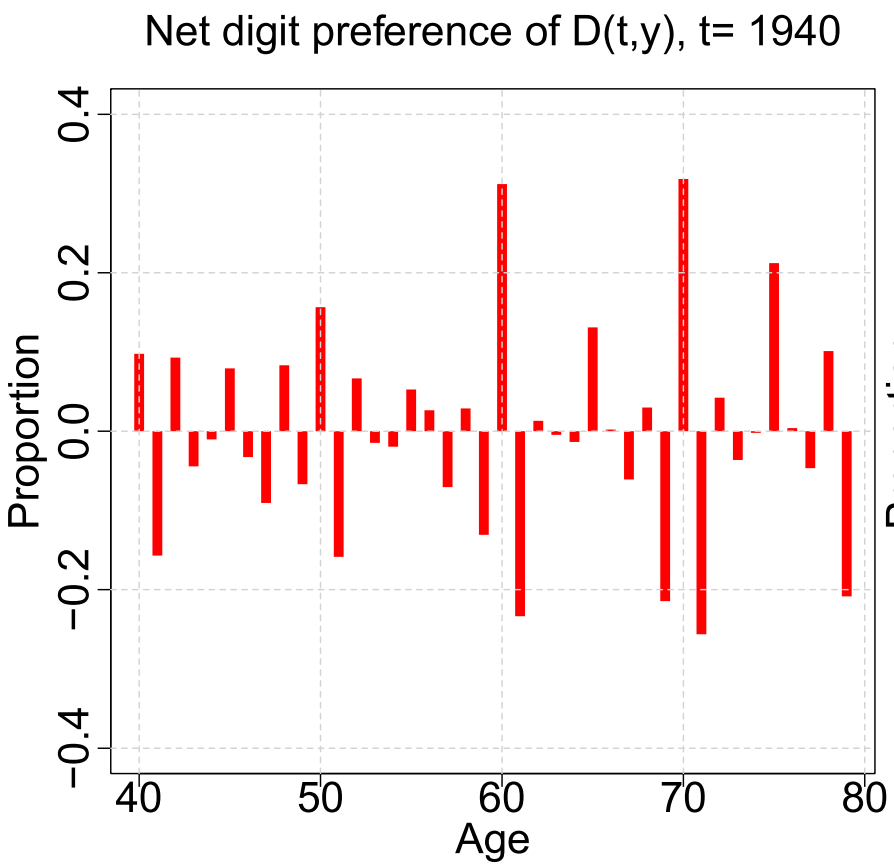


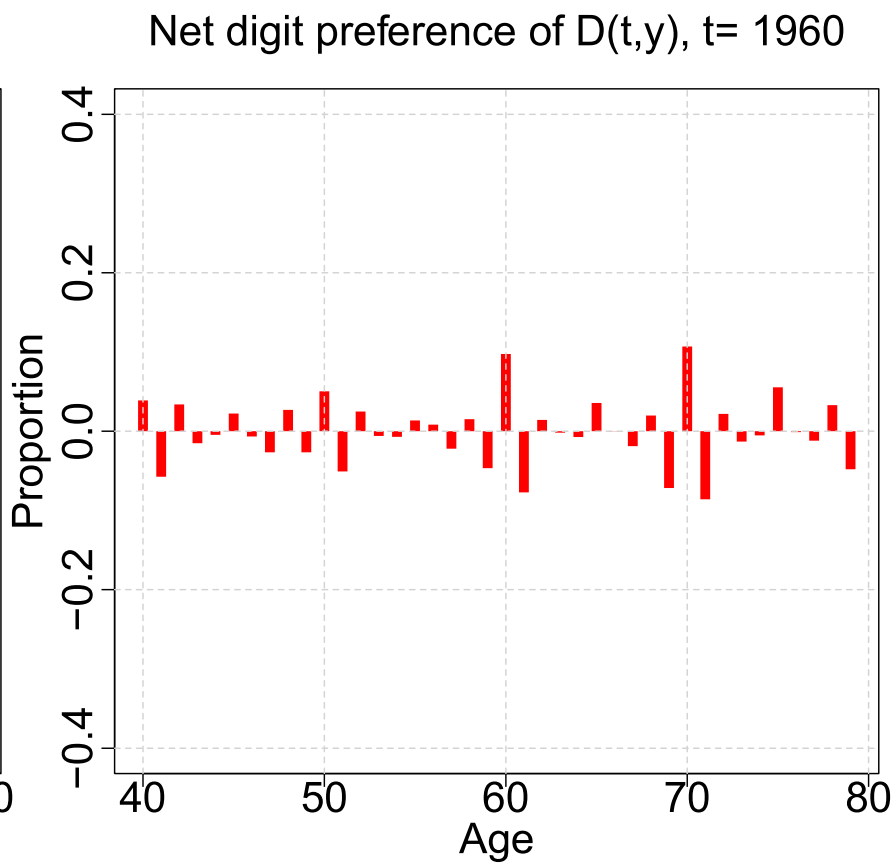
Net digit preference in exposures, Portugal 1940 & 1980.



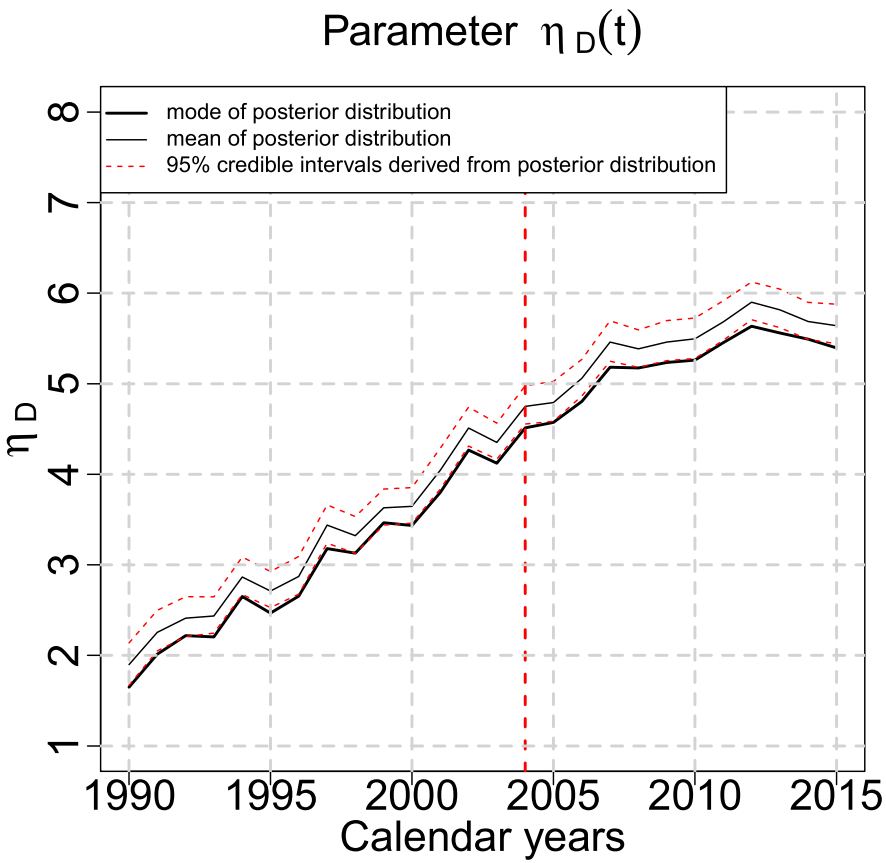


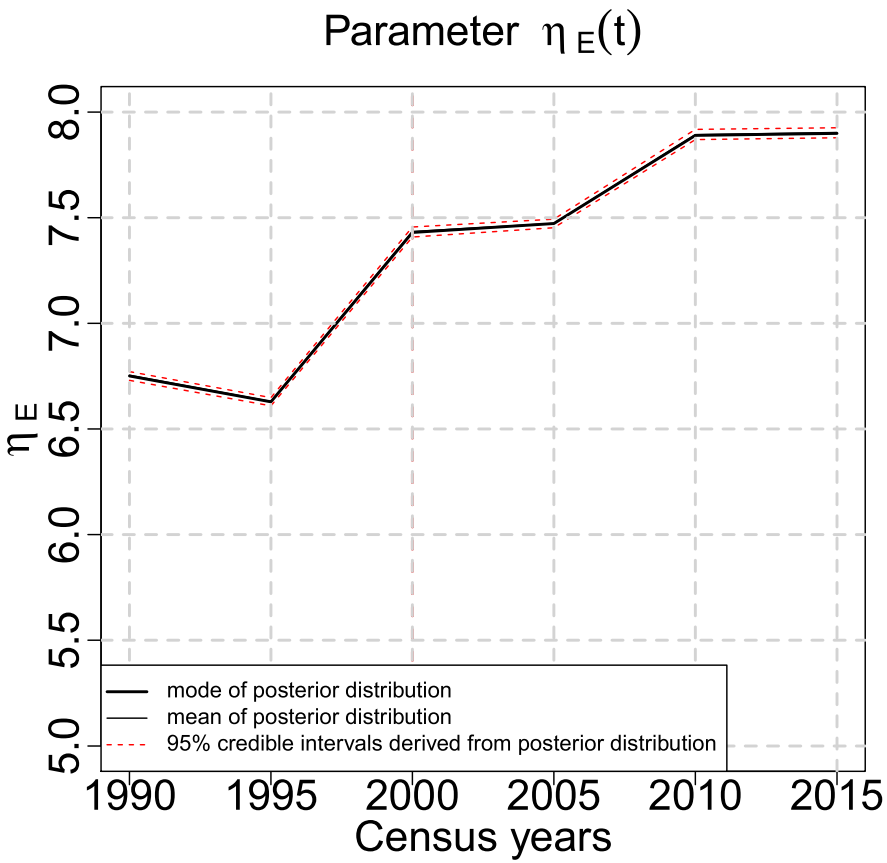
Net digit preference in death counts, Portugal 1940 & 1960.



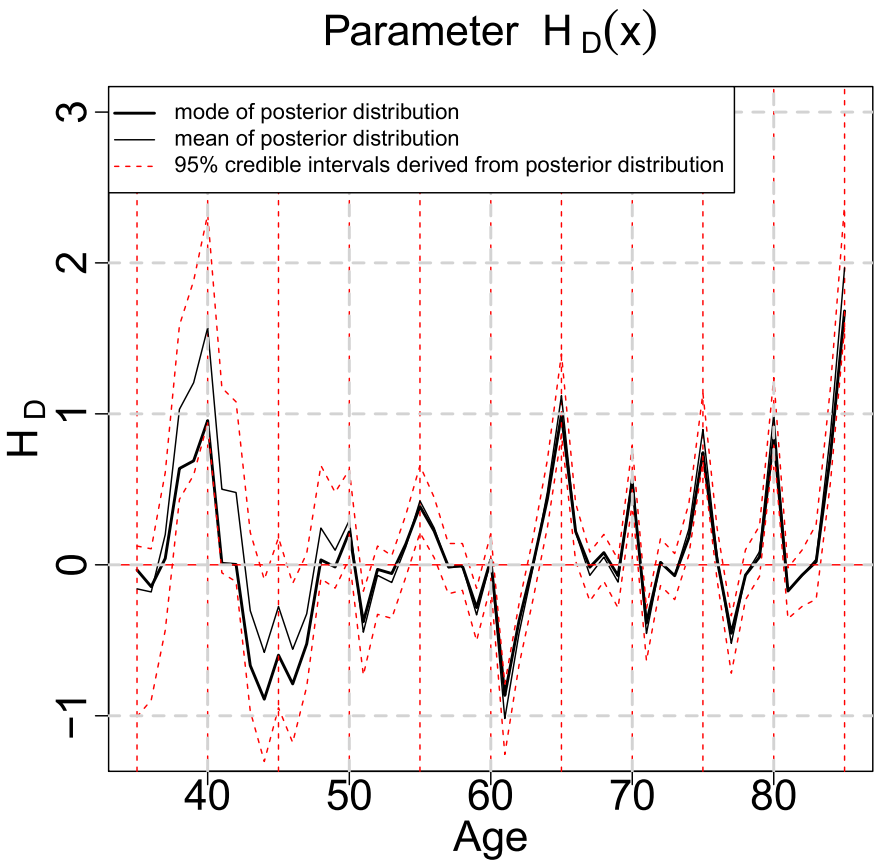


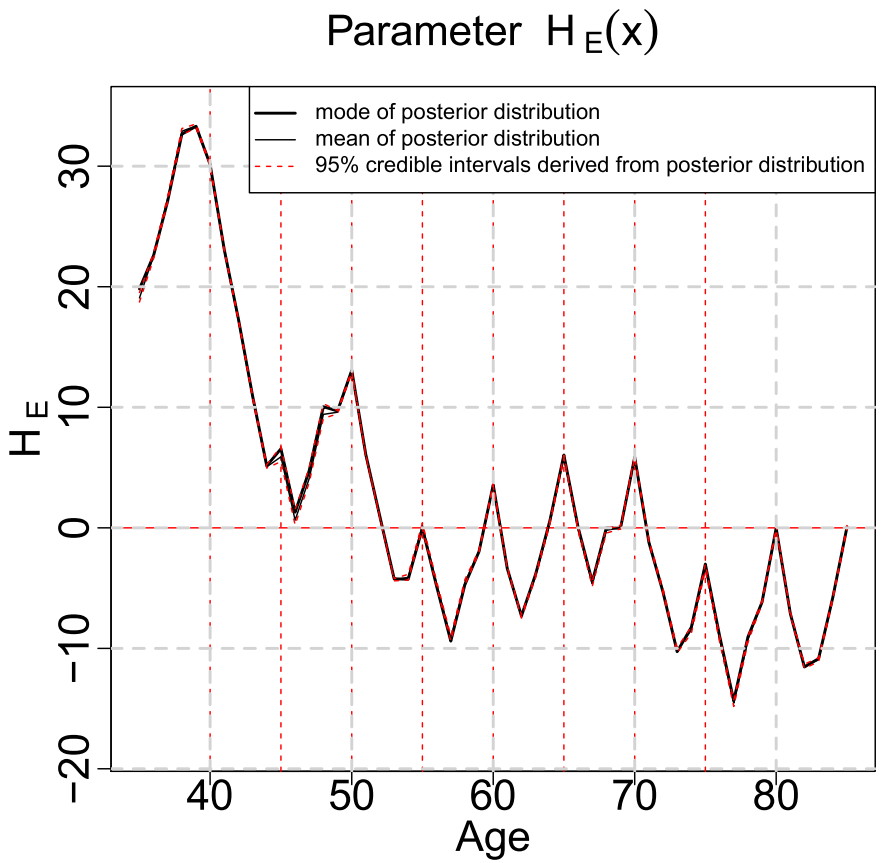
Parameter $\eta^D(t)$ and $\eta^E(t)$, Mexico.



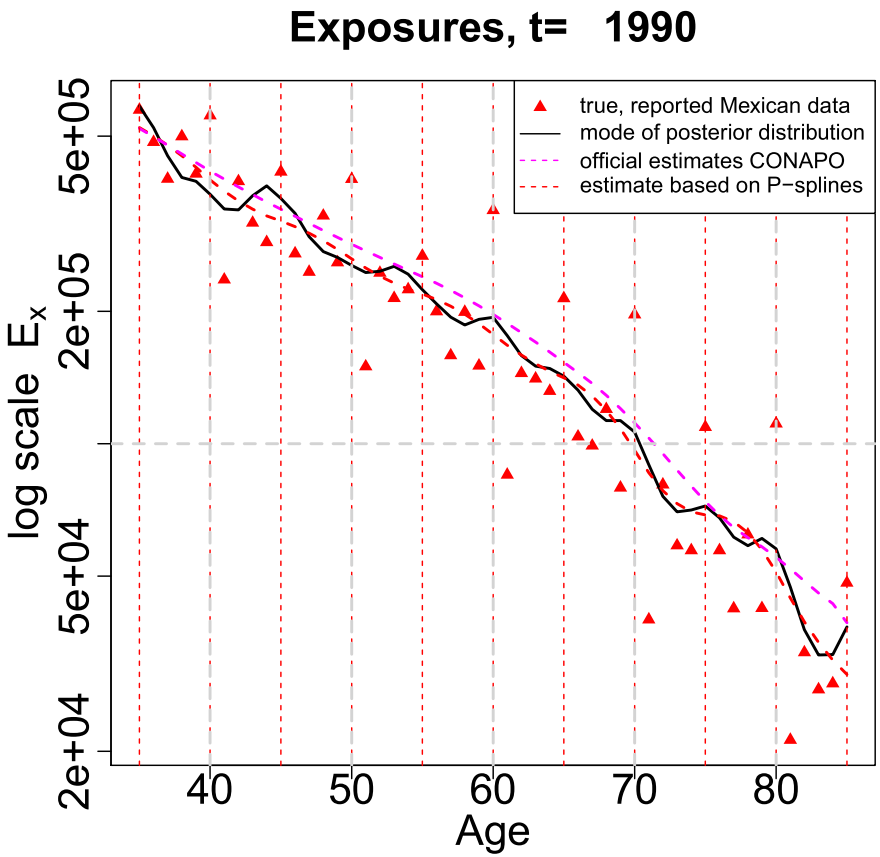


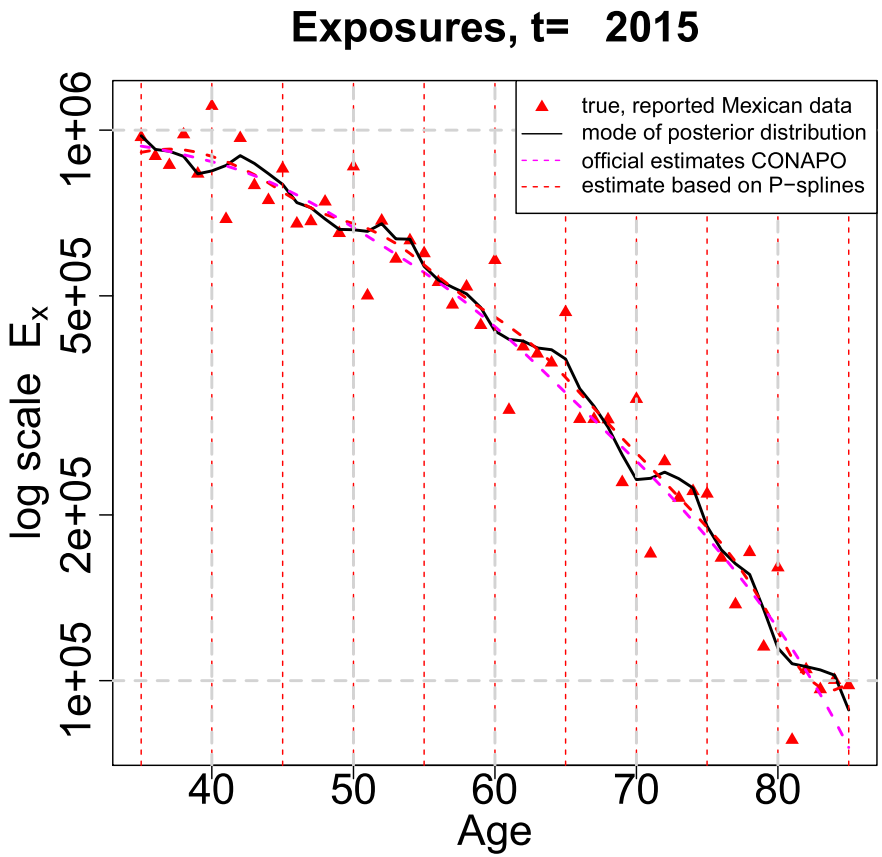
Parameter $H^D(x)$ and $H^E(x)$, Mexico.



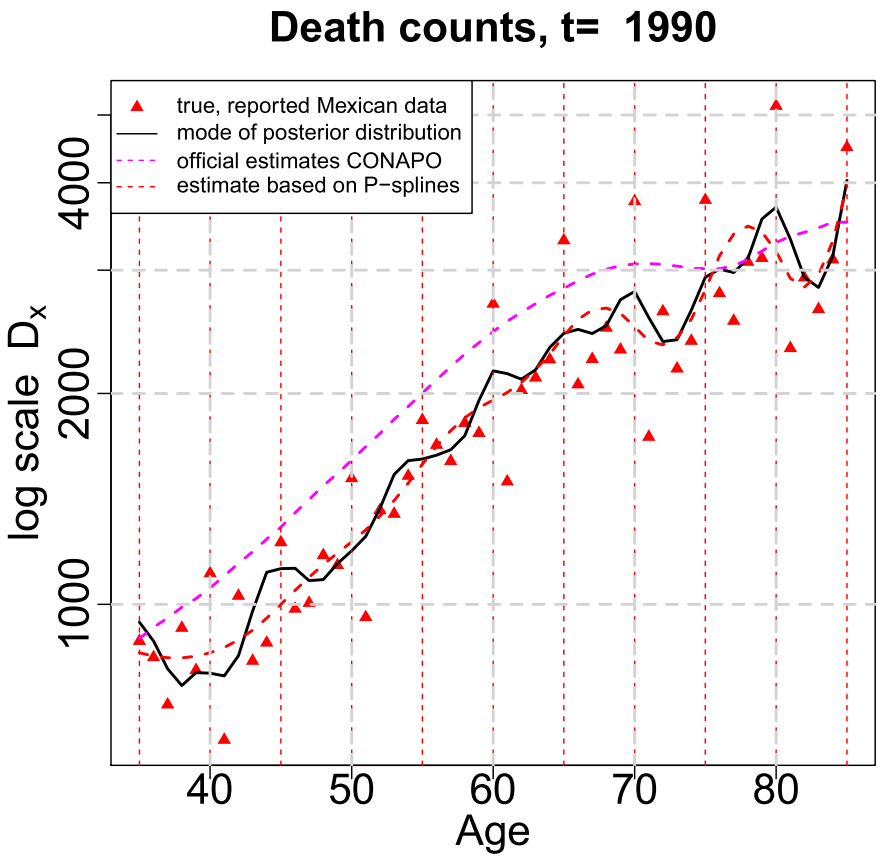


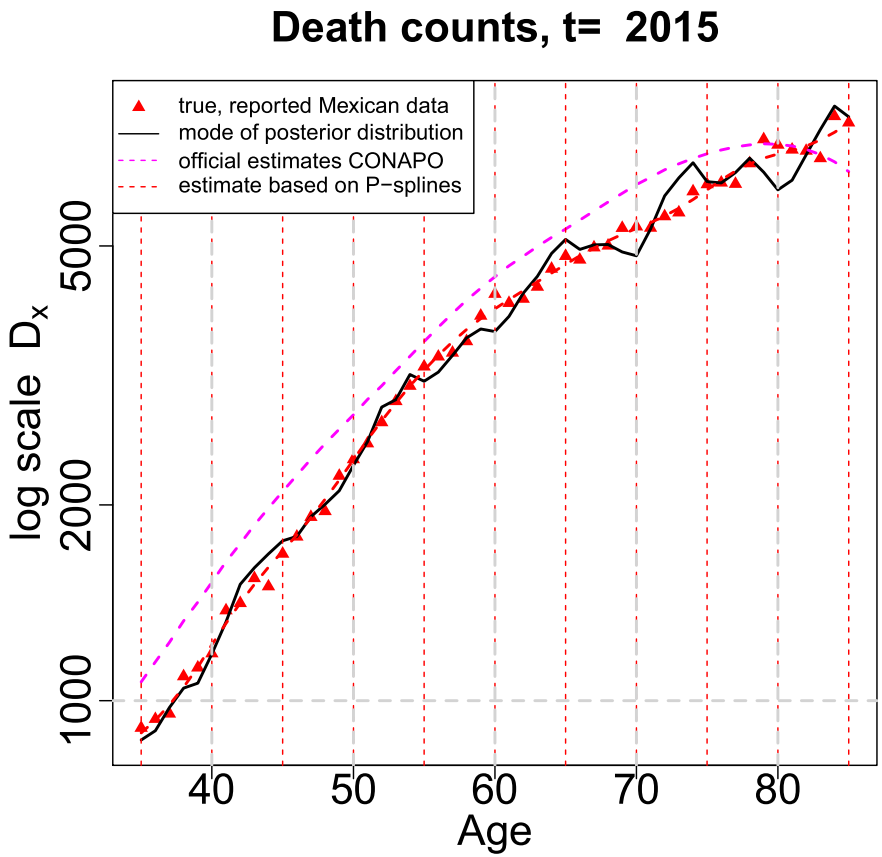
Exposures $E(t, y)$, Mexico 1990 & 2015.



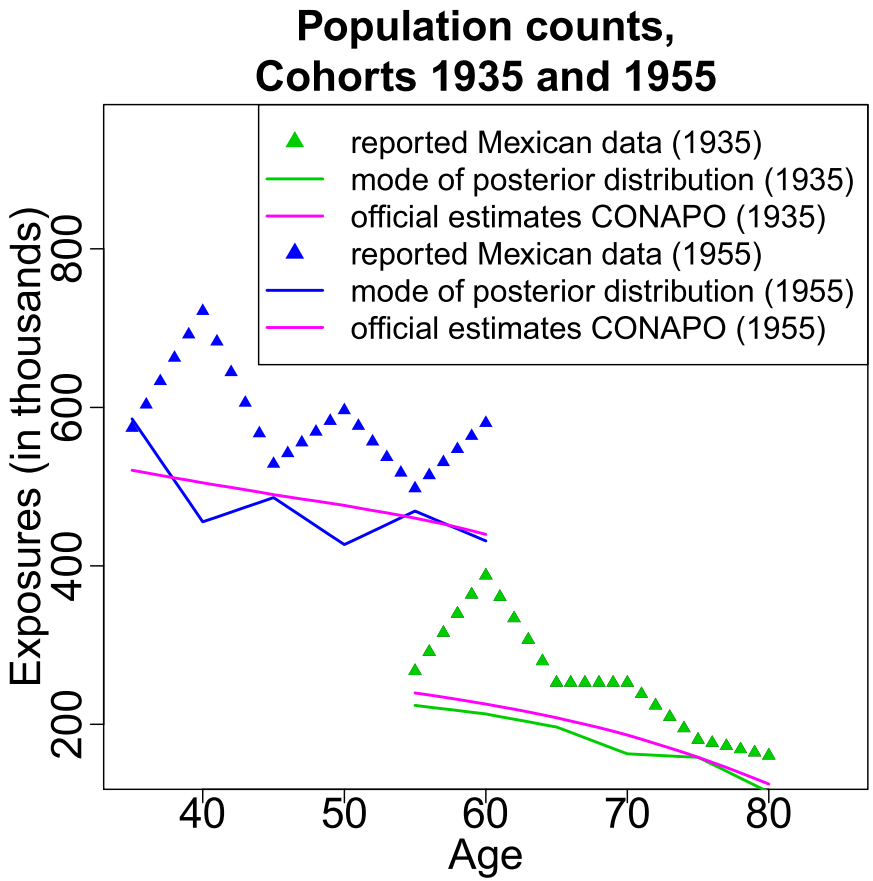


Death counts $D(t, y)$, Mexico 1990 & 2015.

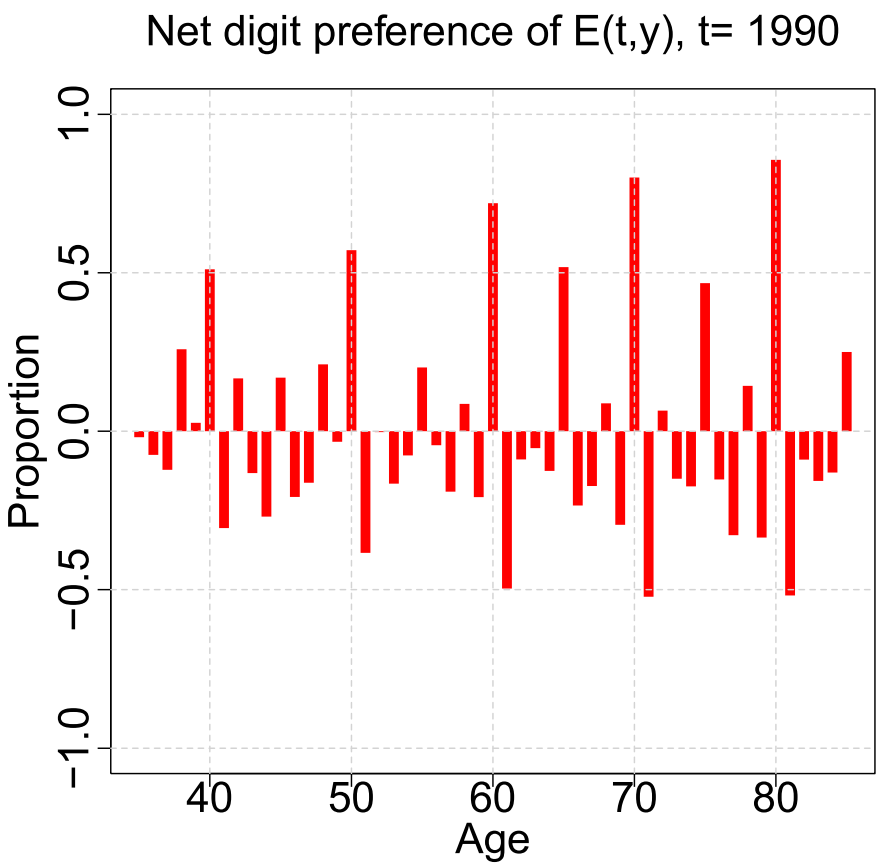


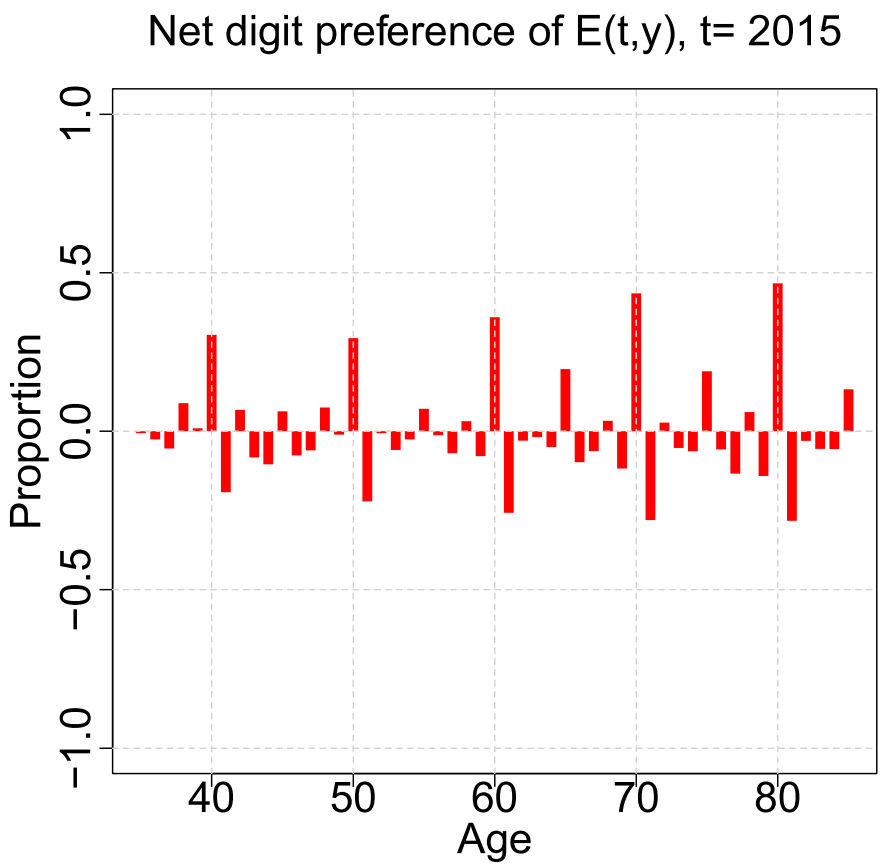


Exposures across cohorts, Mexico.

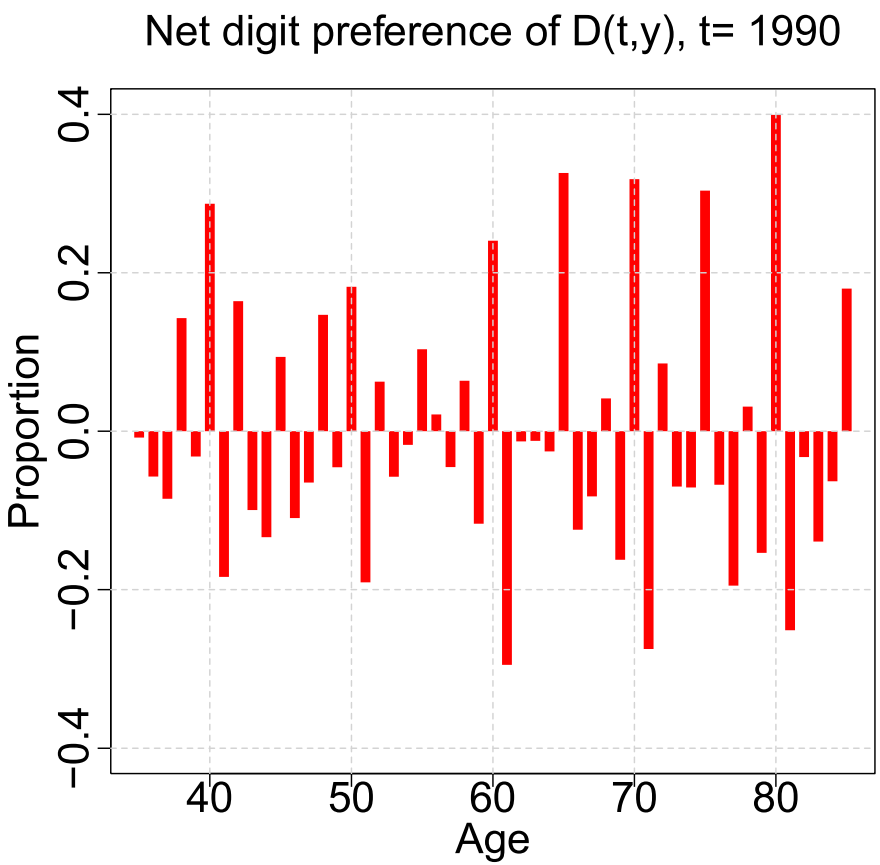


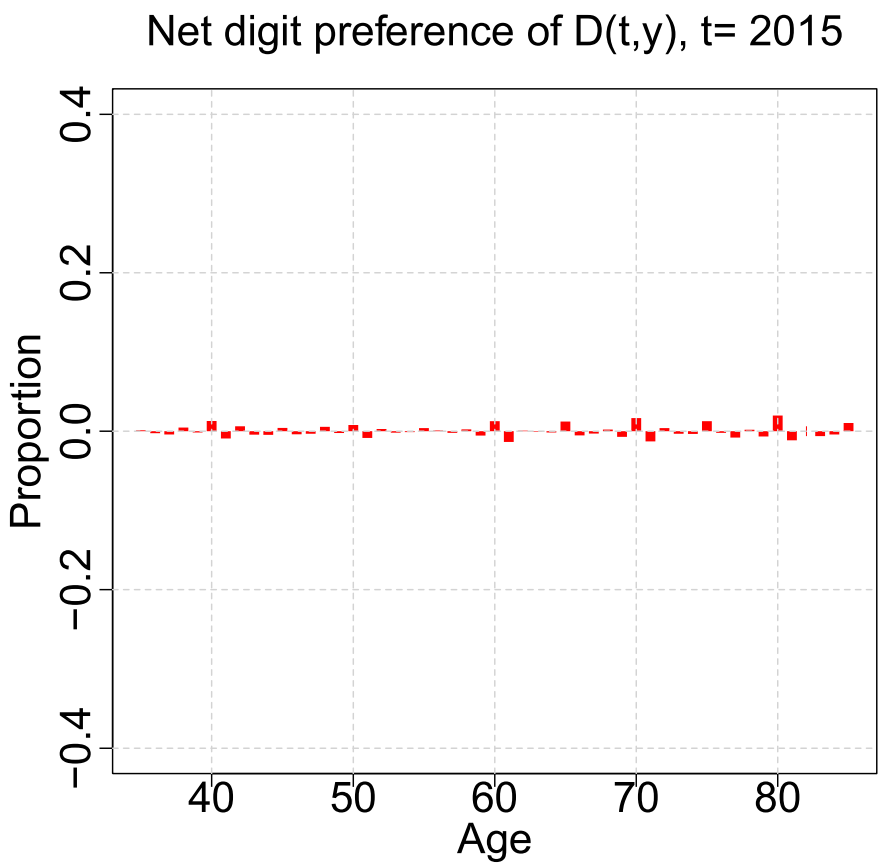
Net digit preference in exposures, Mexico 1990 & 2015.





Net digit preference in death counts, Mexico 1990 & 2015.





Conclusions.

- This model improves the quality of the Mexican data by reducing age heaping across all calendar years.
- we checked that our model gives us back the original data set that we started with for Canada and show that the results from our model are consistent with the HMD documentation regarding age heaping in Portugal.

Conclusions

- We observe there is a potential issue with possible local maximums or flat areas regarding the full log-posterior function $\log\left(\pi\left(\theta\right)\right)$ when we are trying to find the true mode.
- Parameter $\eta^*(t)$ reflects the improvement in the quality of the data over time. Hence, we expect $\eta^*(t)$ to increase over time. In other words, there is less age heaping by 2015 that there use to be, say in 1990.

Conclusions

- Parameter $H^*(x)$ expresses individual preference, indifference or avoidance
- We estimate the net digit preference across years for both deaths and exposures which is also consistent with the evolution of the improvement of the quality of the data.
- Our model could be used to improve historical data sets for gold standard countries such as Portugal which used to show age heaping in the past.

Future research

- We recommend to use our model as a quality control to monitor the quality of the data of gold standard countries that could experience deterioration of the quality of the data in the future.
- We will keep collaborating with HMD to extend our model to older ages where there is age exaggeration and improve the quality of the data for those ages.