MAX-PLANCK-INSTITUT FÜR DEMOGRAFISCHE FORSCHUNG
MAX PLANCK INSTITUTE FOR DEMOGRAPHIC RESEARCH

Splitting abridged fertility data using different interpolation methods. Is there the optimal solution?

Pavel Grigoriev and Dmitry Jdanov

Max Plank Institute for Demographic Research, Rostock (Germany)

Human Fertility Database: Expanding research opportunities
Member-initiated meeting, PAA 2015 Annual Meeting,
Wednesday 29th April 2015, Hilton San Diego Bayfront

Background

Occasionally, there is a need to split aggregated fertility data into the fine grid of ages. This problem is not new and its solution is far from being trivial.

1) In 2009 Rodriguez and Philipov following McNeil et al (1977) proposed method based on spline interpolation of cumulative fertility rates. The modified version of this method is used in the HFD.

2) Using a sample of HFD countries Liu et al. (2011) tested 10 different methods which derive age-specific fertility rates from abridged data.

Conclusion

Beers method provides the best fit.

3) Using HFD and US Census International Database C. Schmertmann (2012) compared the performance of the Calibrated Spline (CS) estimator with the Beers and HFD methods.

Conclusion

Three methods perform very well but CS method provides the best fit and smoother schedules. The overall ranking places the CS method first, HFD second, Beers third.

Background (cont.)

For the purpose of the HFD/HFC, we need a method which meets the following criteria:

1) Fit (estimates should be close to the observed data)
2) Shape (estimated fertility curve should look plausible)
3) Balance (five-year age group totals should match the input data)
4) Non-negativity (estimates should be positive)

There are important differences between HFC and HFD that affect the choice of estimation strategy:

1) Heterogeneity of input data
2) Target measure to be estimated (births in the HFD vs. rates in the HFC)
3) The HFD provides high quality data for high quality research. Interpolation should not be transferred into smoothing which removes the real effects.
4) The original data included in the HFC may be erroneous (especially for countries with limited quality vital statistics). The smoothing is a solution.

Interpolation algorithm used in the HFD

1) Calculating cumulative fertility rate F(x) from age-specific fertility rates f(x)

2) Calculating logits of cumulative fertility rate

$Y(x)=\text{logit}[F(x)/F(x_{max})]=\log[F(x)/F(x_{max})-F(x)]$

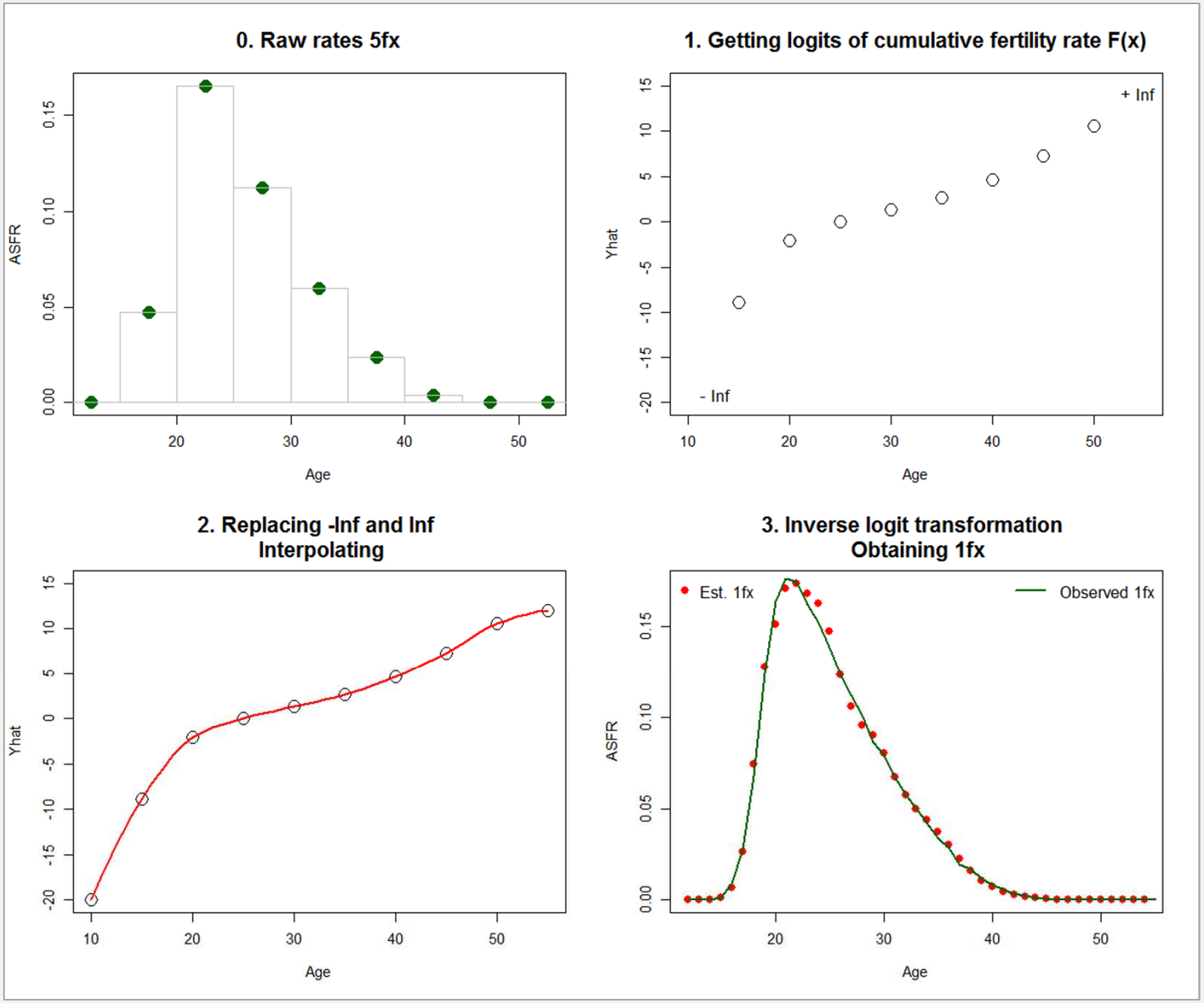
3) Setting $Y(x_{min})=-20$ and $Y(_{max})=12$ for the two data points (extremes) where the logarithm is not defined;

4) Estimating $Y_{hat}(x)$, continuous version of Y(x) using Hermite cubic spline interpolation (function 'interp1', method='pchip', R library 'signal');

5) Estimating F_{hat}, continuous version of F(x), using inverse logit transformation. $F_{hat}(x)=[\exp(Y_{hat}(x)/1+\exp(Y_{hat}(x)]*F(x_{max})$;

6) Obtaining single-year rates $1fx=F'(x)= F(x+1)-F(x)$.

Illustration of the HFD method: spline interpolation of single-year age-specific fertility rates on the basis of 5-year data; Russia, 1985



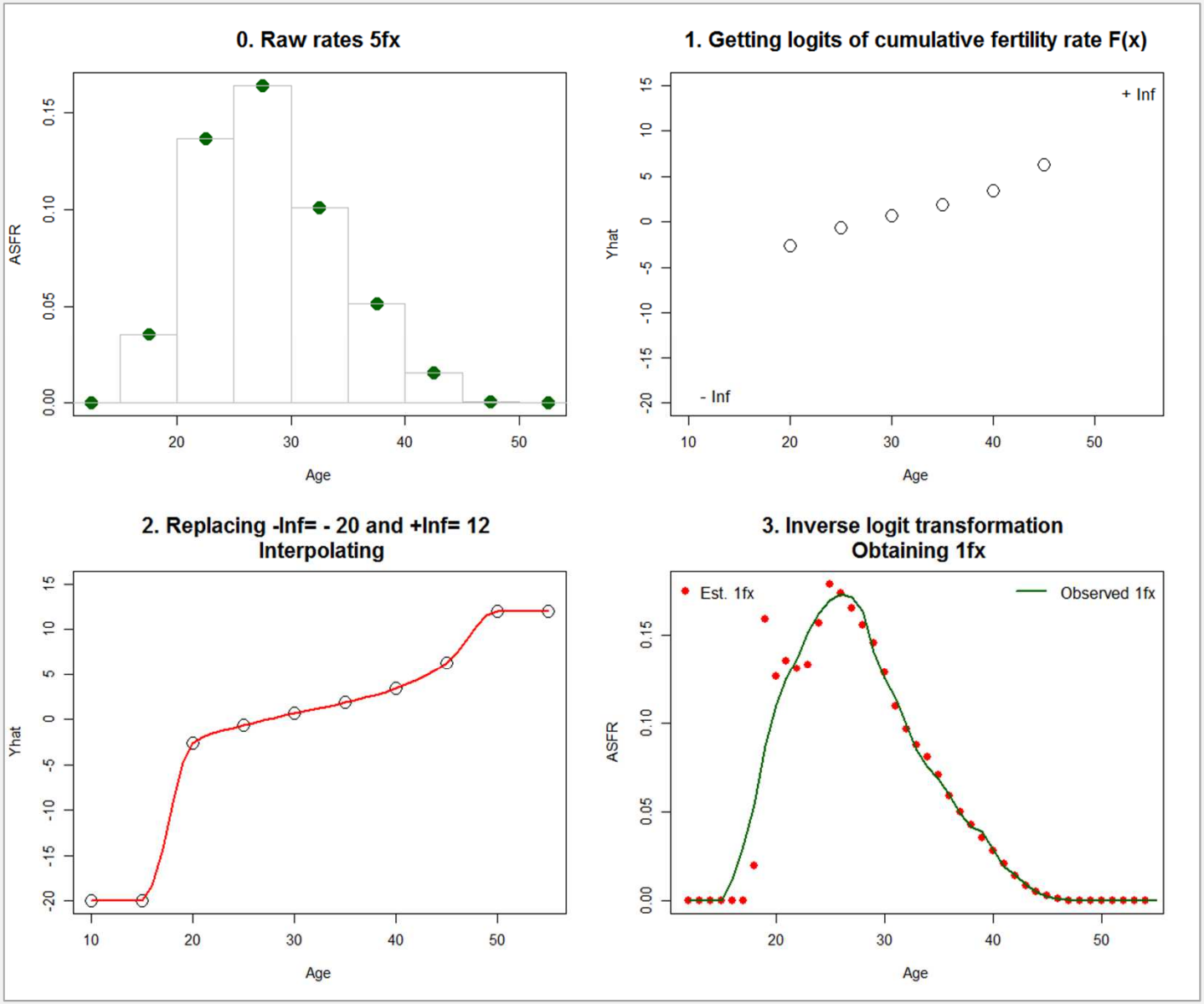
Limitations of the HFD method

Experiments based on the HFD data have shown that the current algorithm performs reasonably well. But sometimes (in very rare cases), it fails to produce plausible estimates because:

1) Sensitivity of the estimates to the values of undefined logarithms which occur during the LOG-LOG transformation.
2) Unlike many other spline function, Hermite spline does not guarantee the continuous second derivative, and thus the estimated fertility curve might have sudden twists.

Hermite spline ensures non-negative values of rates, i.e. to have non-decreasing cumulative function.

Spline interpolation of single-year age-specific fertility rates on the basis of 5-year data; Northern Ireland, 1975



Options to get things right

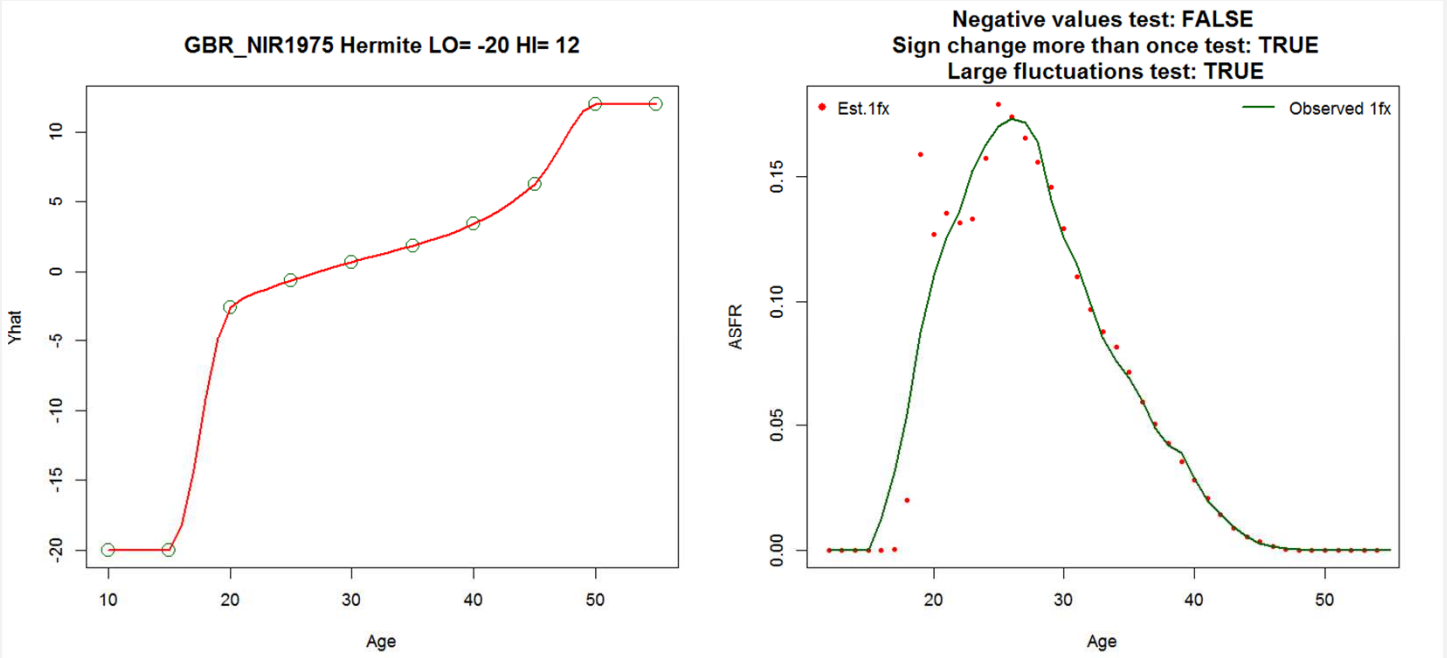
1) Calibrating the values of the lower (LO) and upper (HI) undefined logarithms
2) Adding a ‘phantom’ birth to the youngest age group (in our example, there are no births in the youngest age group).
3) Testing a different spline function
4) Trying something completely different

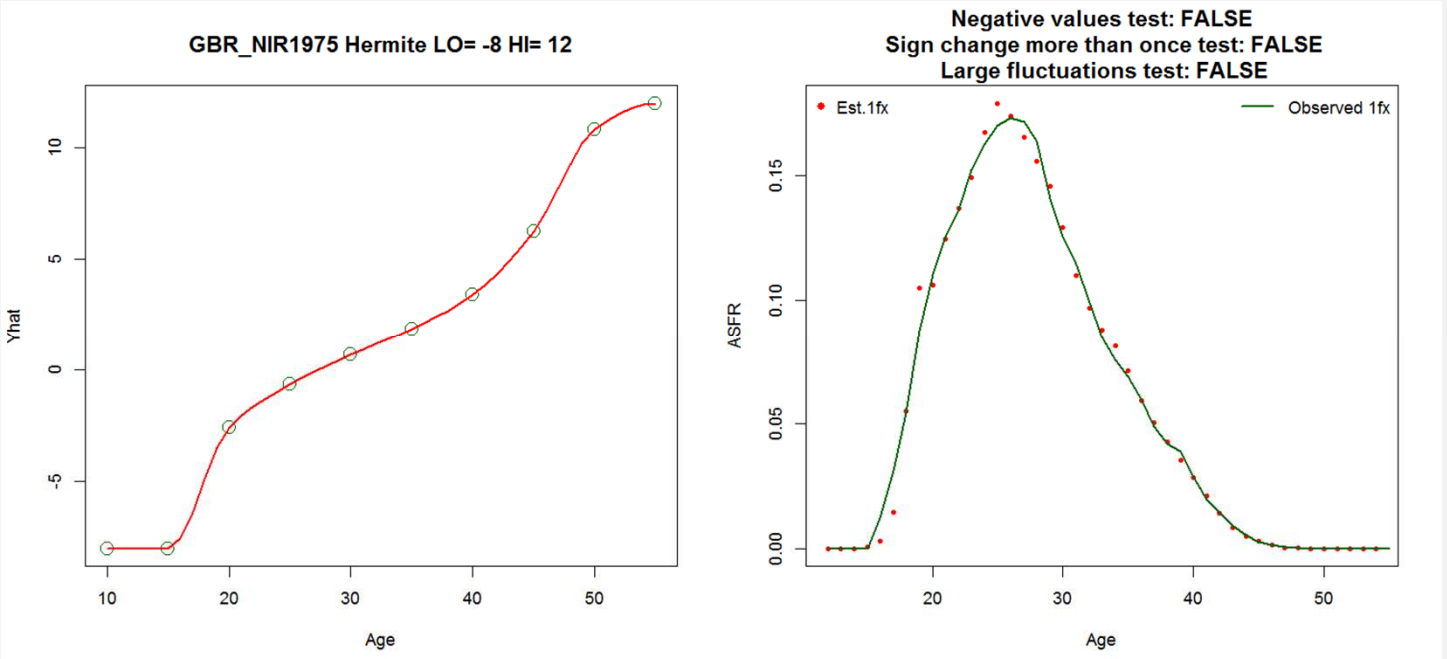
(1) and (2) to overcome the limitations of LOG-LOG transformation procedure

(3) is related to the main disadvantage of the Hermite spline, discontinuous second derivative

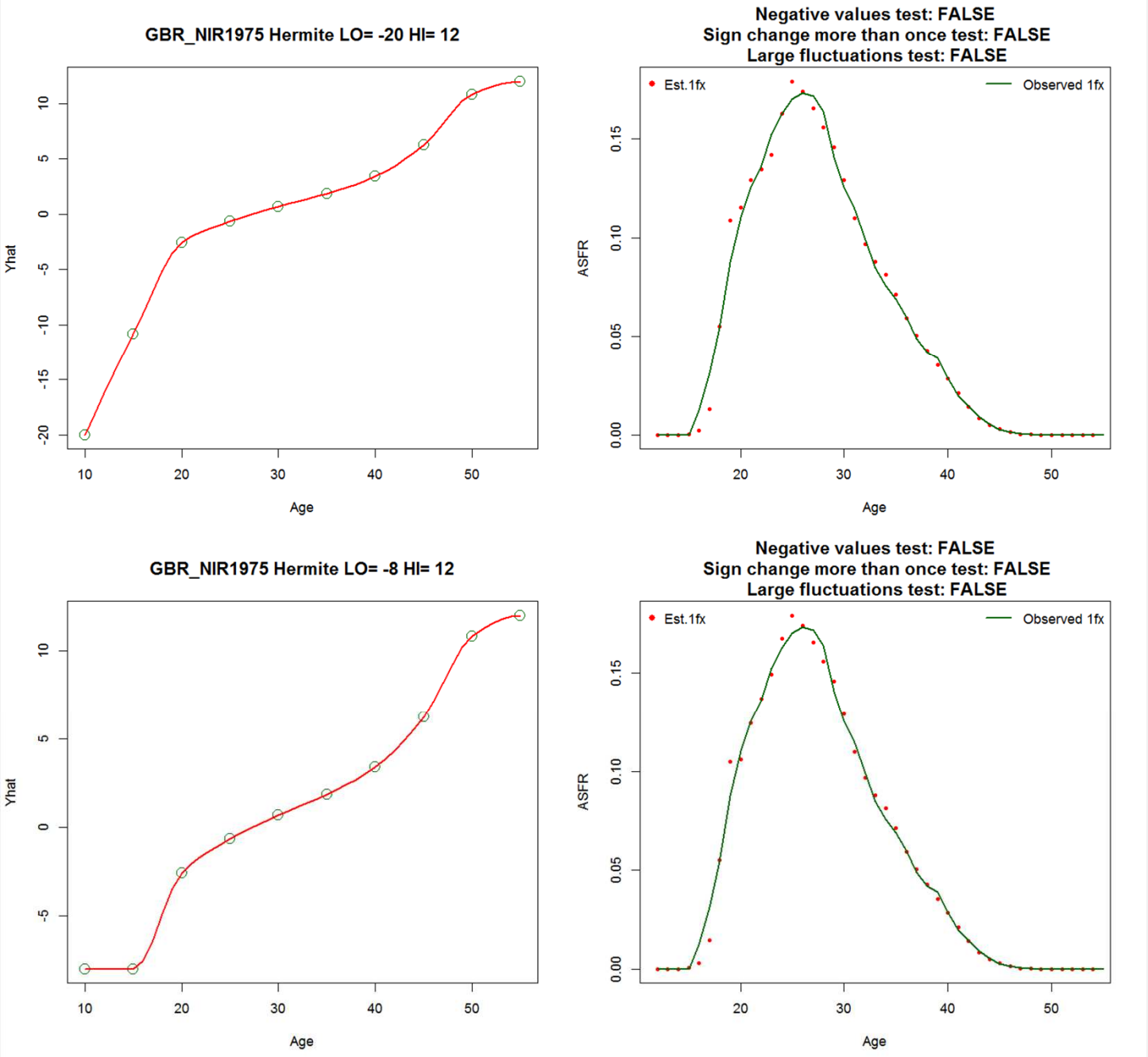
and (4) to fix the both.

Option 1: Calibrating LO and HI





Option 2: Adding 'phantom' birth



Option 3: Testing alternative spline functions but keeping LOG-LOG transformation procedure

Alternative 1

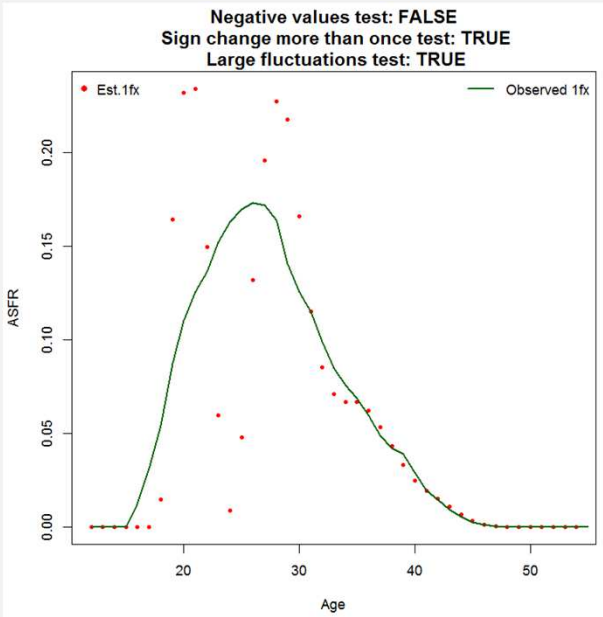
Cubic spline with Hyman filter. Ensures the monotonicity but the second derivative will be not continuous anymore at the knots where the Hyman filter changes first derivatives. [function 'spline' method='Hyman', R package 'stats']. See Smith, Hyndman, Wood (2004)

Alternative 2

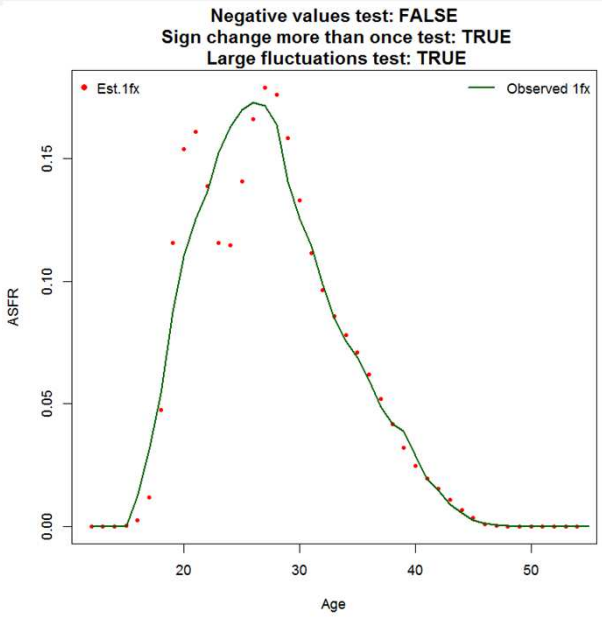
Polynomial Smoothing Spline
[function 'smooth.Pspline', R package 'pspline'). Possible to modify the order of the spline as well as to alter the smoothing parameter. If smoothing parameter is zero, function goes through the defined knots.

Behavior of the cubic spline with Hyman filter after adjusting different parameters

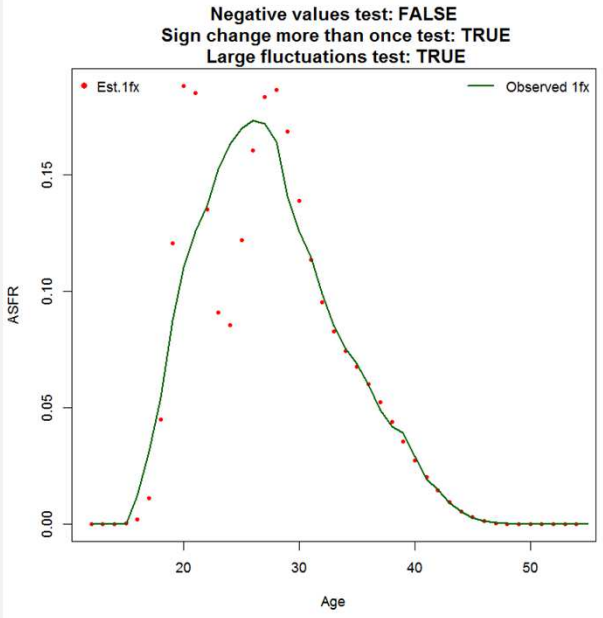
A. LO= - 20



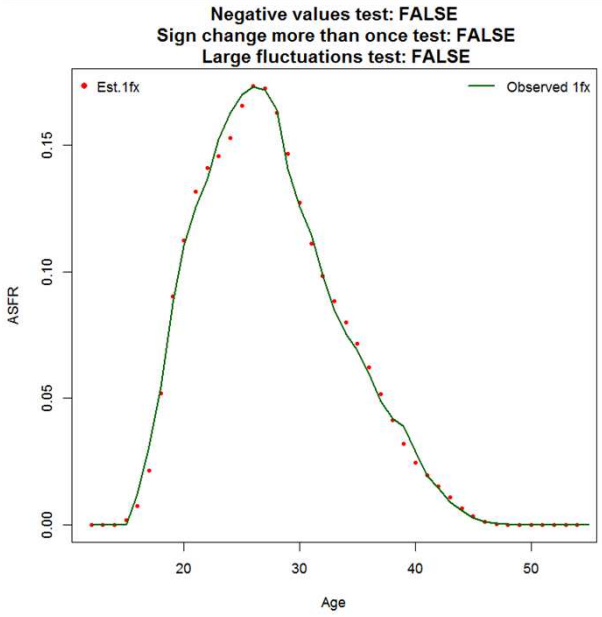
B. LO= - 8



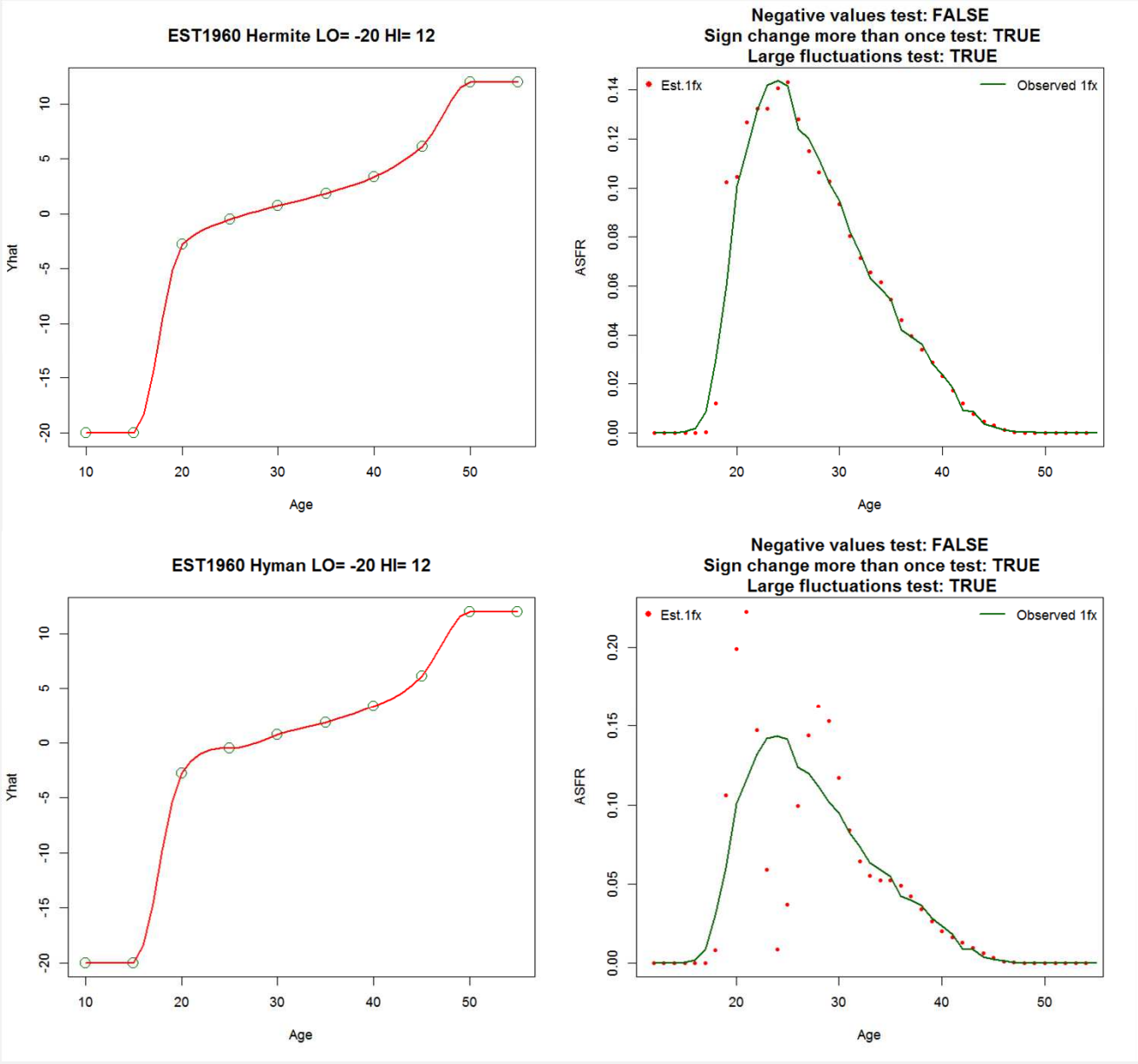
C. Phantom birth



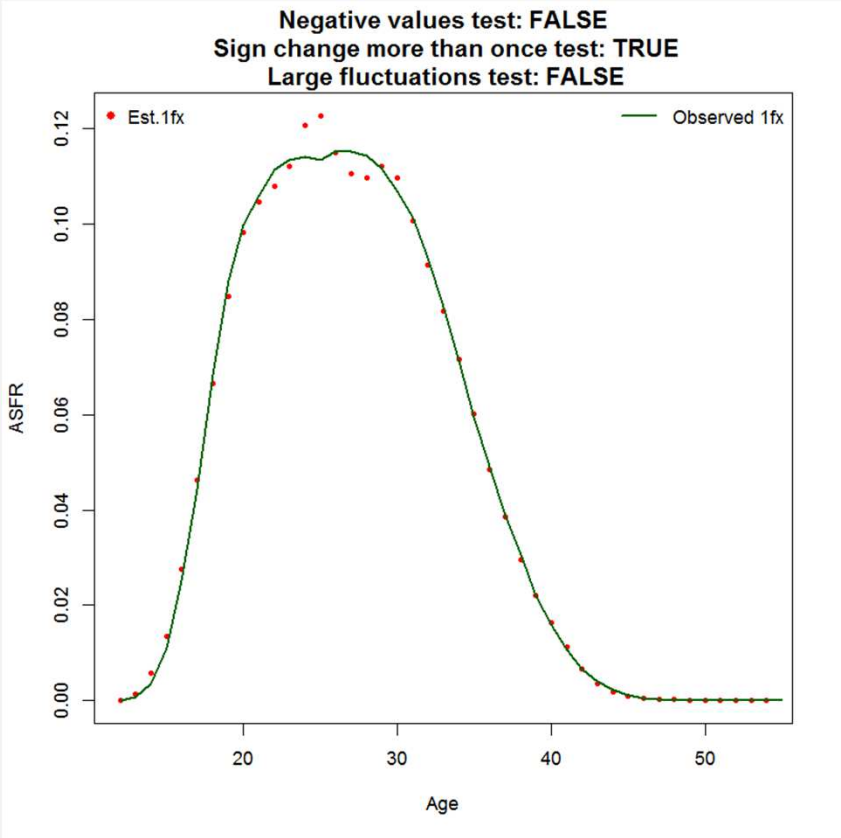
D. LO= - 6



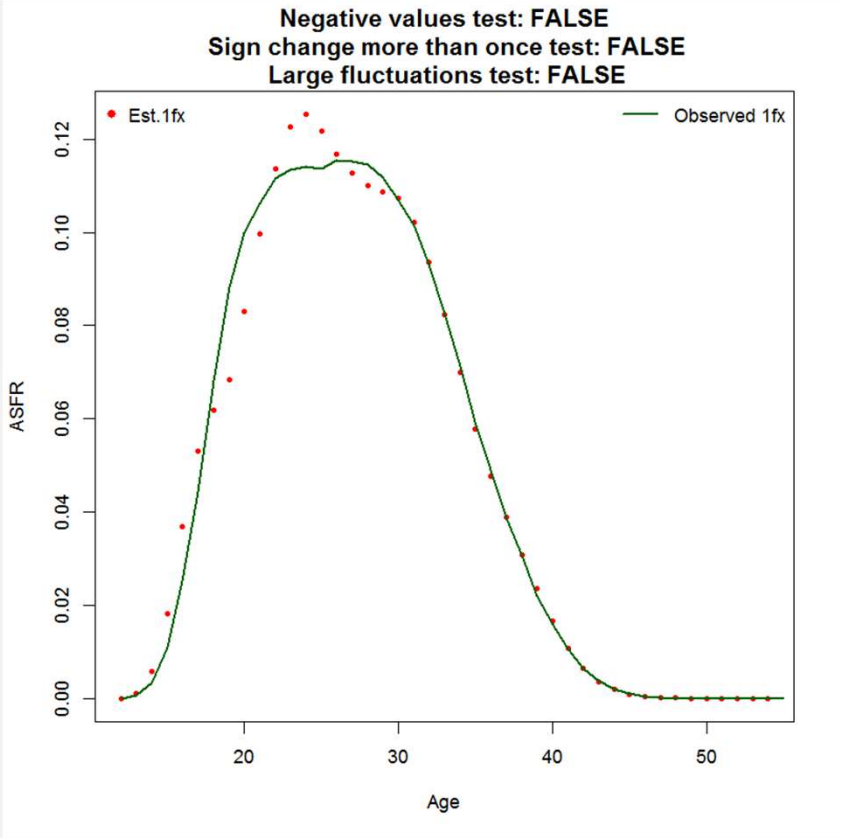
Hermite vs. Hyman; the case of Estonia, 1960



Spline interpolation of single-year age-specific fertility rates on the basis of 5-year data; USA, 2000



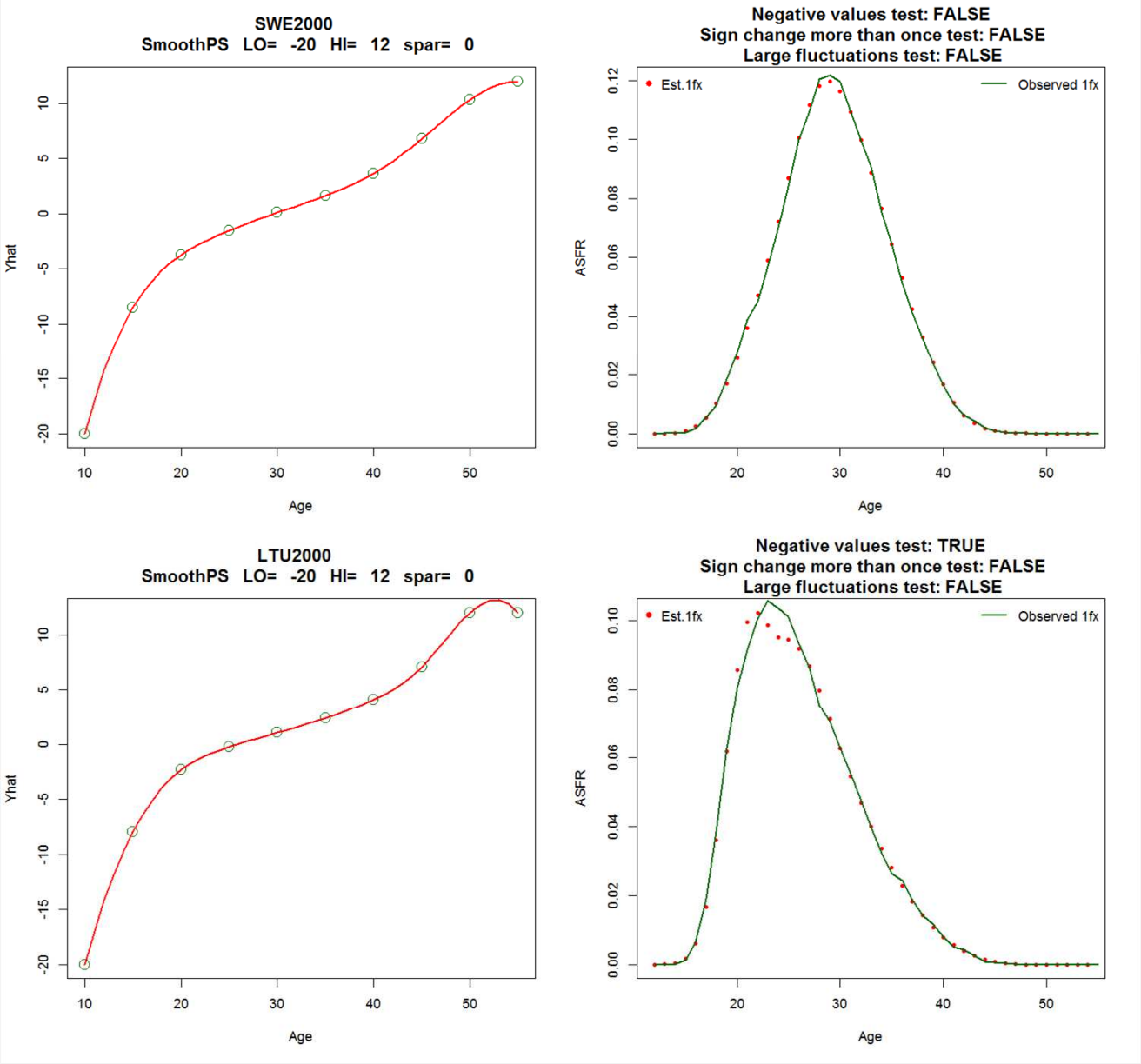
A. Hermite, LO=-20



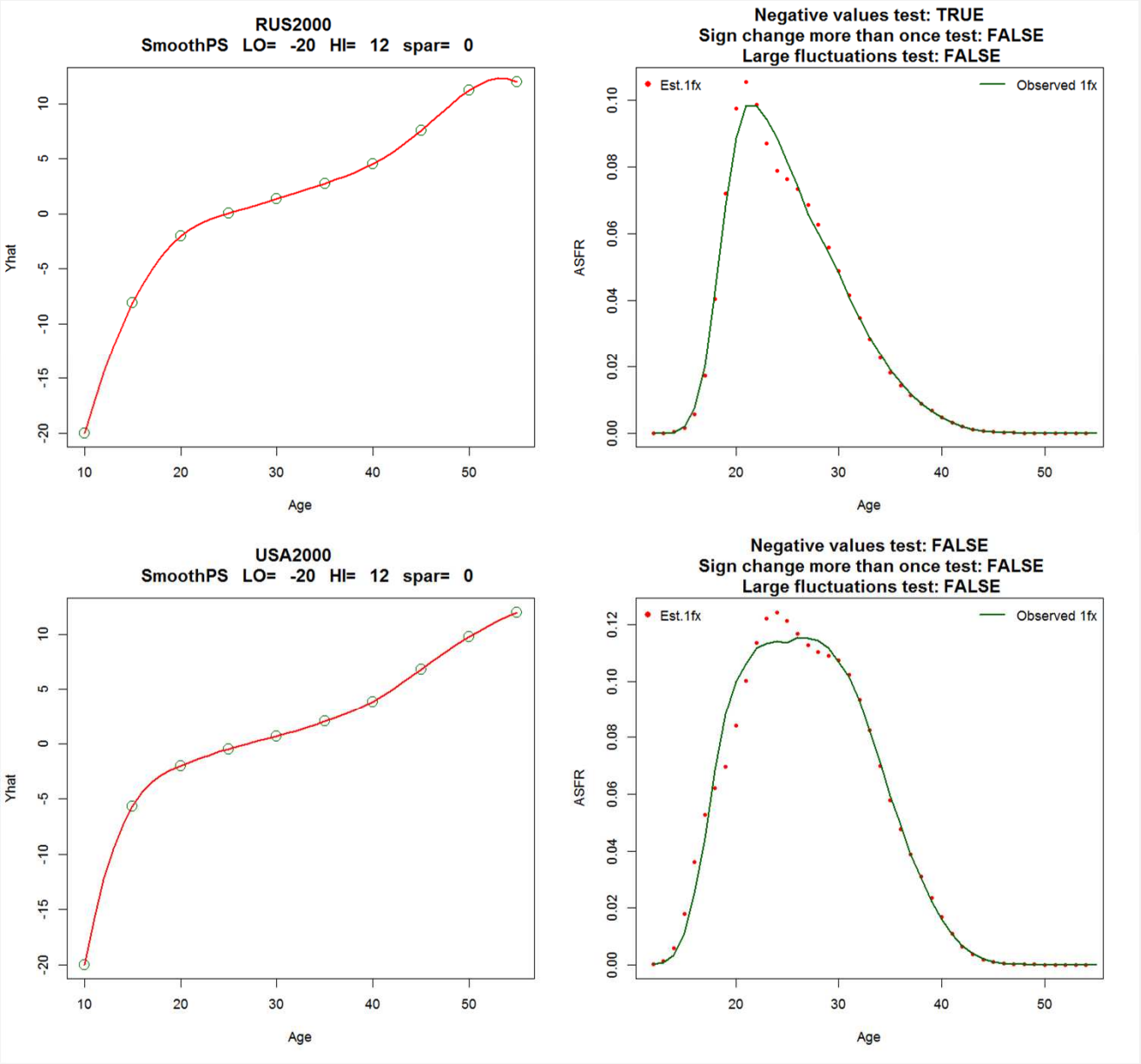
B. Hyman, LO=-20

In this case changes in the LO parameter did not result in any improvement

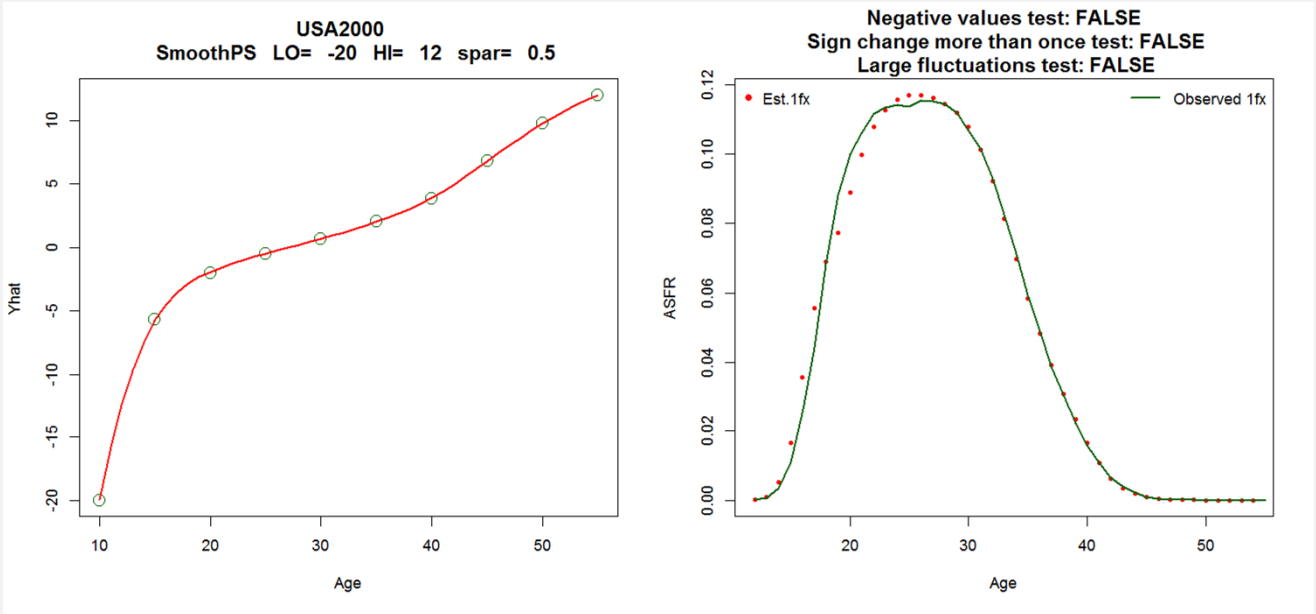
Spline interpolation of single-year age-specific fertility rates on the basis of 5-year data; Method SmoothPS

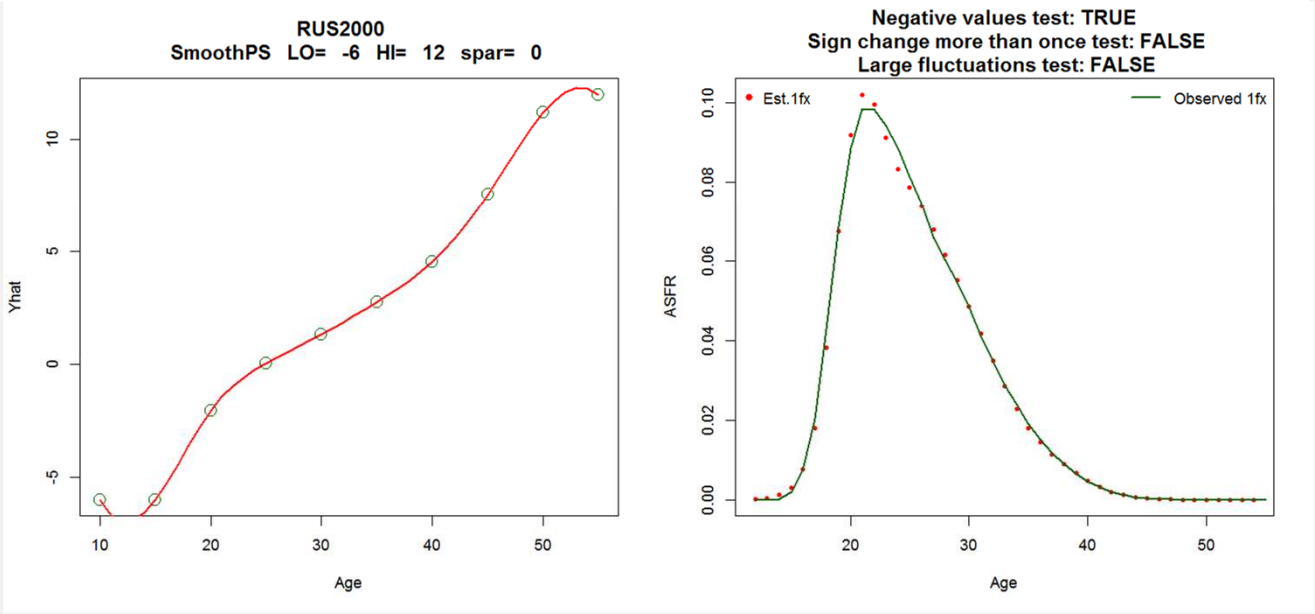


Spline interpolation of single-year age-specific fertility rates on the basis of 5-year data; Method SmoothPS (cont.)



Spline interpolation of single-year age-specific fertility rates on the basis of 5-year data; Method SmoothPS with adjusted parameters





Calibrated Spline (CS)* method

Two criteria of a good schedule:

1) 'Fit': close fit to the observed data

2) 'Shape': similarity of the known fertility patterns

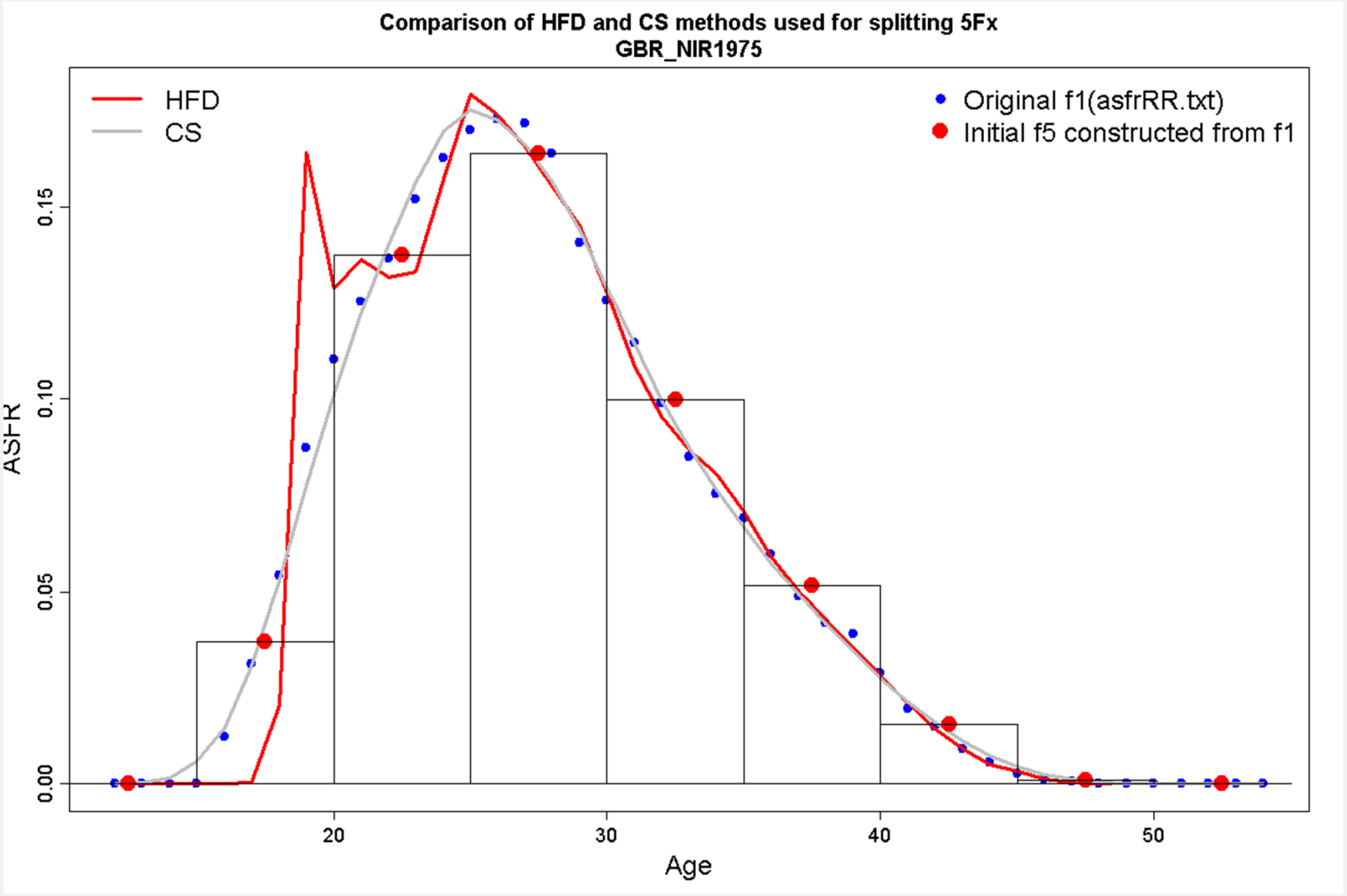
Finding the compromise between good fit and good shape by minimizing the squared error penalty.

The results of thorough testing has shown that CS replicates *known* 1fx schedules from 5fx data well. Also, its interpolated schedules look smoother compared to the current HFD method.

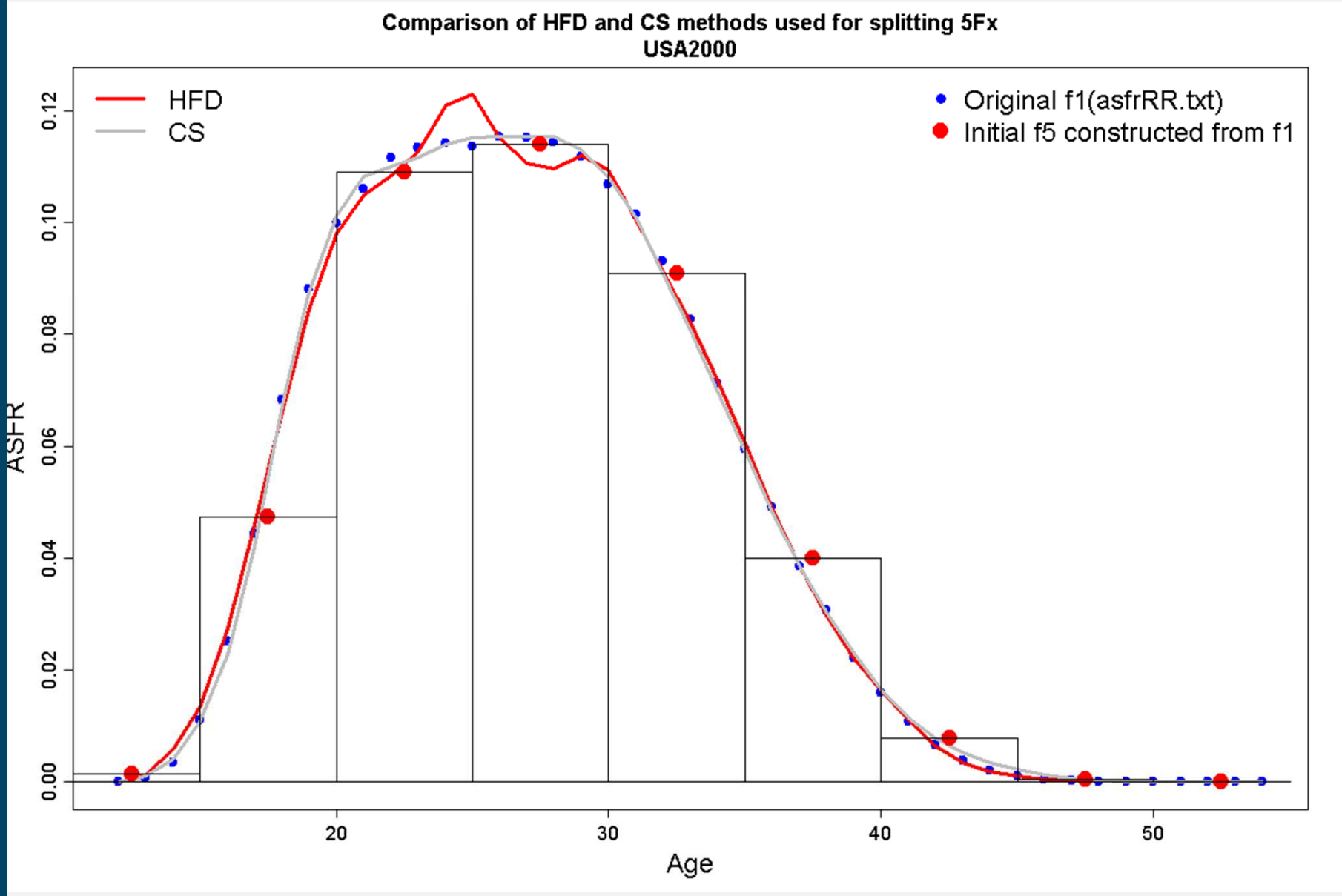
But the method uses the database of known fertility shapes formed by the original one-year age-specific fertility rates from the HFD

* Proposed by Carl P. Schmertmann
See http://calibrated-spline.schmert.net

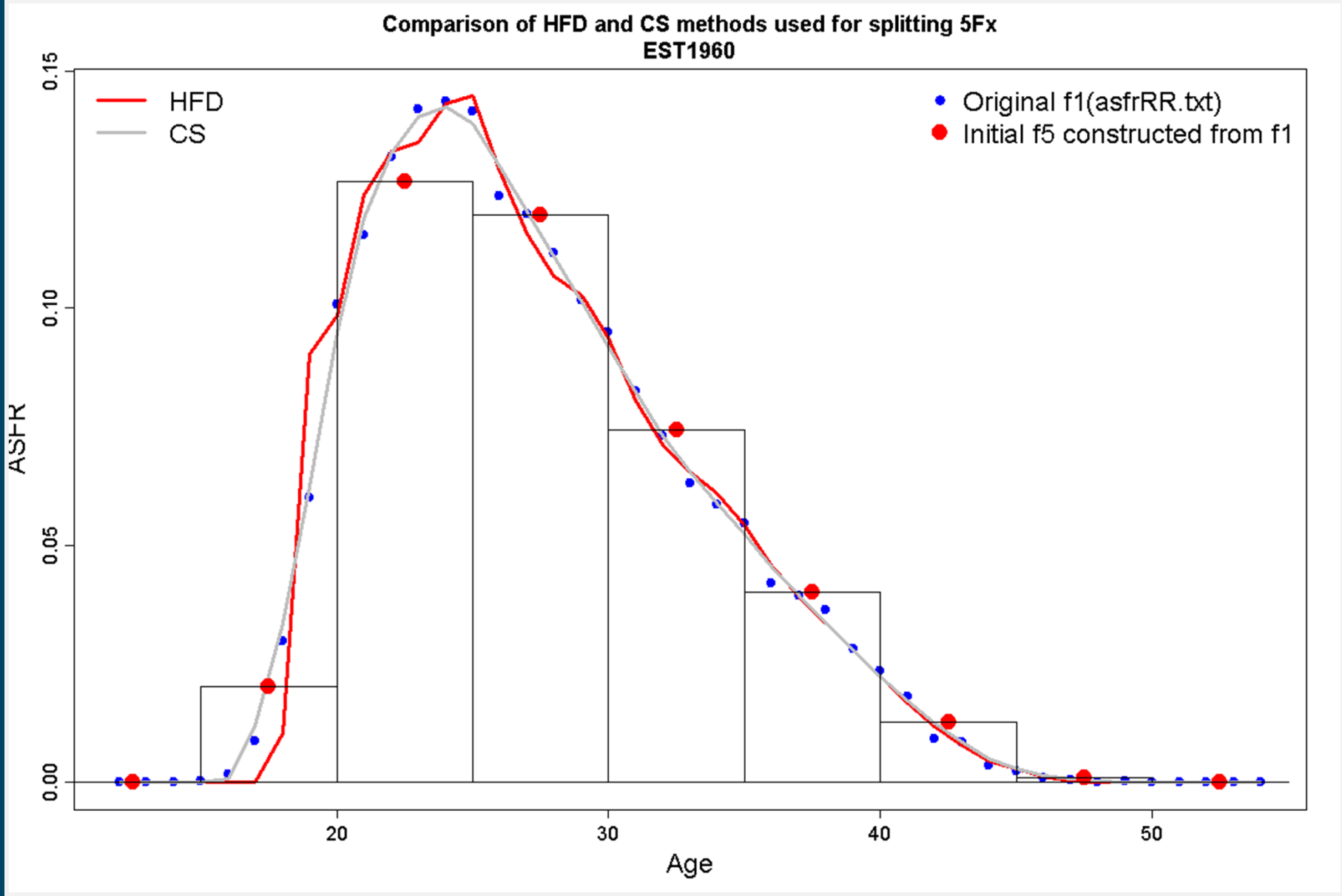
How does CS handle problematic cases?



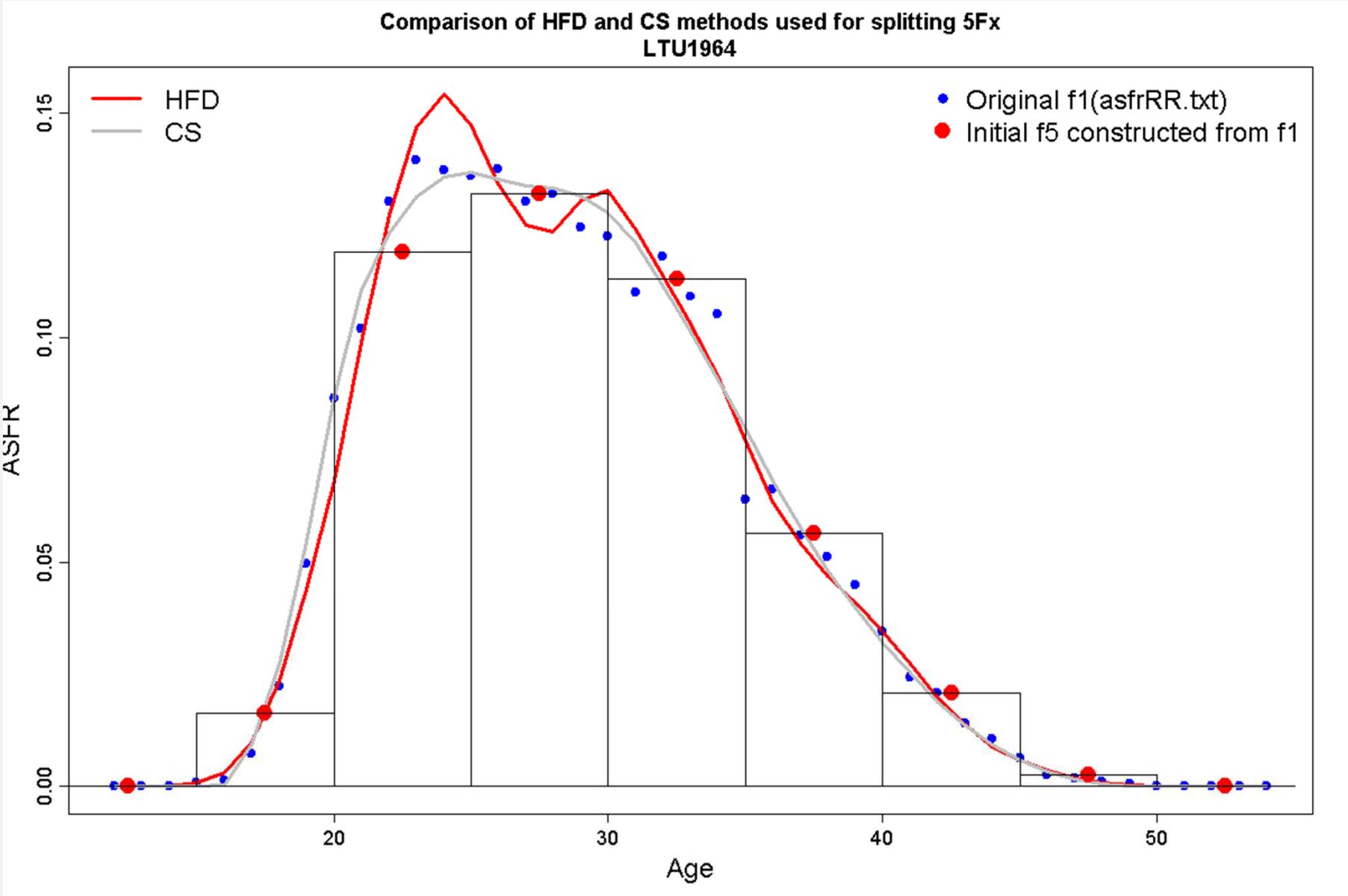
How does CS handle problematic cases? (cont.)



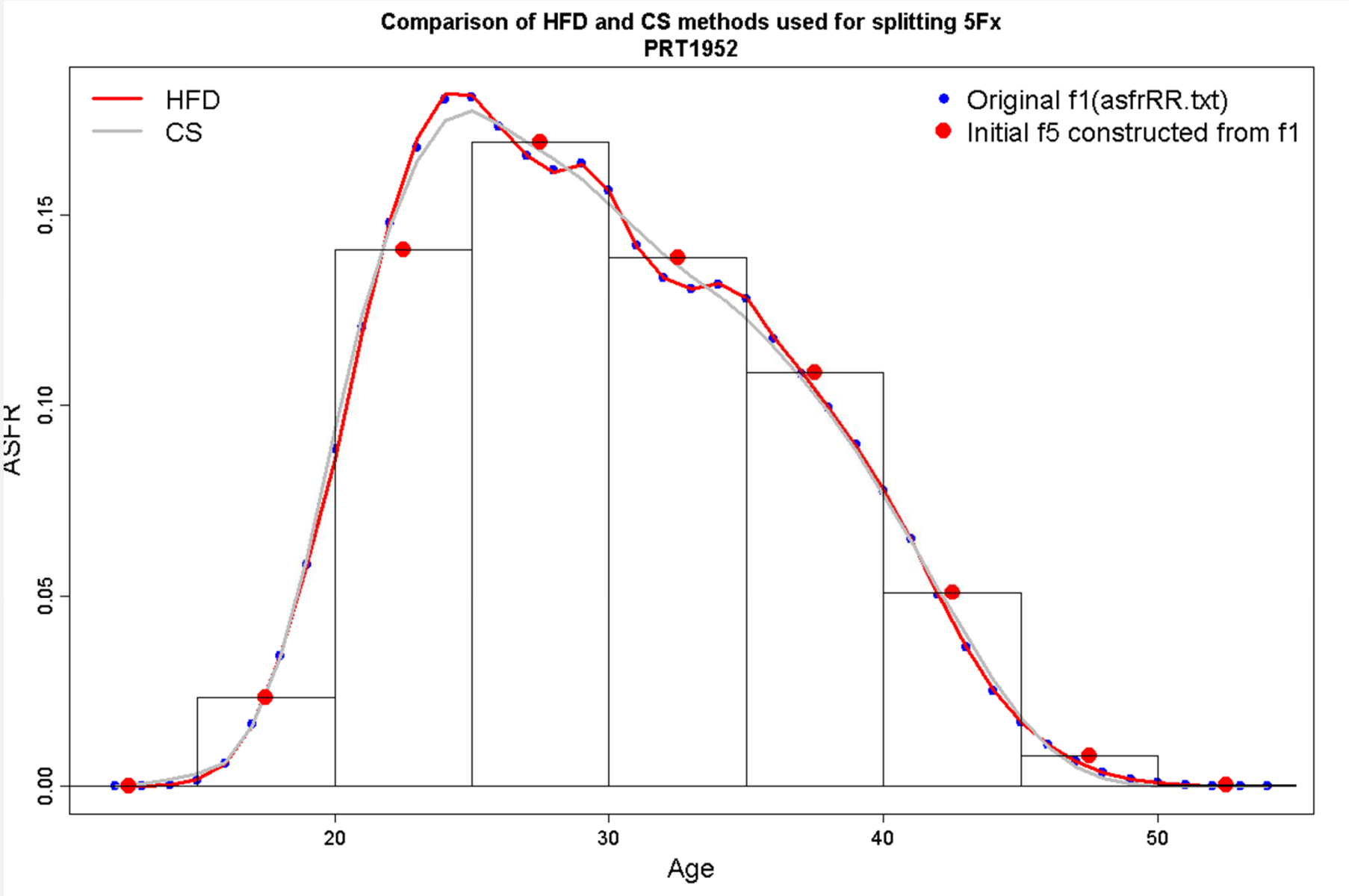
How does CS handle problematic cases? (cont.)



How does CS handle problematic cases? (cont.)



How does CS handle problematic cases? (cont.)



Main issues related to CS implementation in the HFD

1) The method relies on the database of known fertility shapes. Thus, input data which are not represented in this database might be smoothed out.
2) The method is not directly applicable to birth order data
3) There is a choice (or balance) between good shape and good fit. The nice looking shapes do not keep the sum of births count by age groups and in total.
4) Occasionally, the method produces negative values which then are being replaced by zeros with the disruption of the shape. The method might also produce positive values while in fact they should be zeros.

Last two points can be fixed but it makes the splitting procedure too complex and interpolated ASFR curves, in fact, less “smart”.

Example of the CS limitations: negative values

Switzerland, 1960

Age	CountryYear	Original	HFD	CSunadj	CSadj	Number of births (HFD)
12	CHE1960	0.00000	0.00000	0.00000	0	0
13	CHE1960	0.00002	0.00000	-0.00030	0	1
14	CHE1960	0.00002	0.00000	-0.00070	0	1
15	CHE1960	0.00064	0.00030	-0.00030	0	27
16	CHE1960	0.00194	0.00150	0.00230	0.00230	81
17	CHE1960	0.00789	0.00700	0.01060	0.01060	327
18	CHE1960	0.02361	0.02360	0.02580	0.02580	950
19	CHE1960	0.04805	0.04970	0.04850	0.04850	1805
20	CHE1960	0.07498	0.07240	0.07710	0.07710	2853

Key for the table heading

Original
ASFRs obtained from asfrRR.txt; used as the basis construction 5Fx

HFD
Hermite method of splitting the obtained 5Fx

CSunadj
Calibrated spline method before the adjustment for negative values

CSunadj
Calibrated spline method after the adjustment for negative values

Undesirable outcome

If estimated ASFRx=0 (while in fact it is not) ---> Bx=0 (missing births) or vice versa: if estimated ASFRx>0 (while in fact it is zero) ----> Bx>0 (additional (artificial births))

Another example

Age	CountryYear	Original	HFD	CSunadj	CSadj	Number of births (HFD)
12	DEUTE1980	0	0	0	0	0
13	DEUTE1980	0	0	0.0003	0.0003	0
41	DEUTE1980	0.00247	0.0024	0.0016	0.0016	333
42	DEUTE1980	0.00160	0.0015	-0.0001	0	201
43	DEUTE1980	0.00088	0.0009	-0.0014	0	107
44	DEUTE1980	0.00061	0.0005	-0.0020	0	75
45	DEUTE1980	0.00025	0.0003	-0.0019	0	29
46	DEUTE1980	0.00021	0.0002	-0.0012	0	22
47	DEUTE1980	0.00013	0.0001	-0.0006	0	13
48	DEUTE1980	0.00008	0.0001	-0.0001	0	8
49	DEUTE1980	0.00005	0	0.0001	0.0001	5
50	DEUTE1980	0.00003	0	0.0001	0.0001	3
51	DEUTE1980	0.00001	0	0	0	2
52	DEUTE1980	0.00001	0	0	0	1
53	DEUTE1980	0	0	0	0	0
54	DEUTE1980	0	0	0	0	0

And two more …

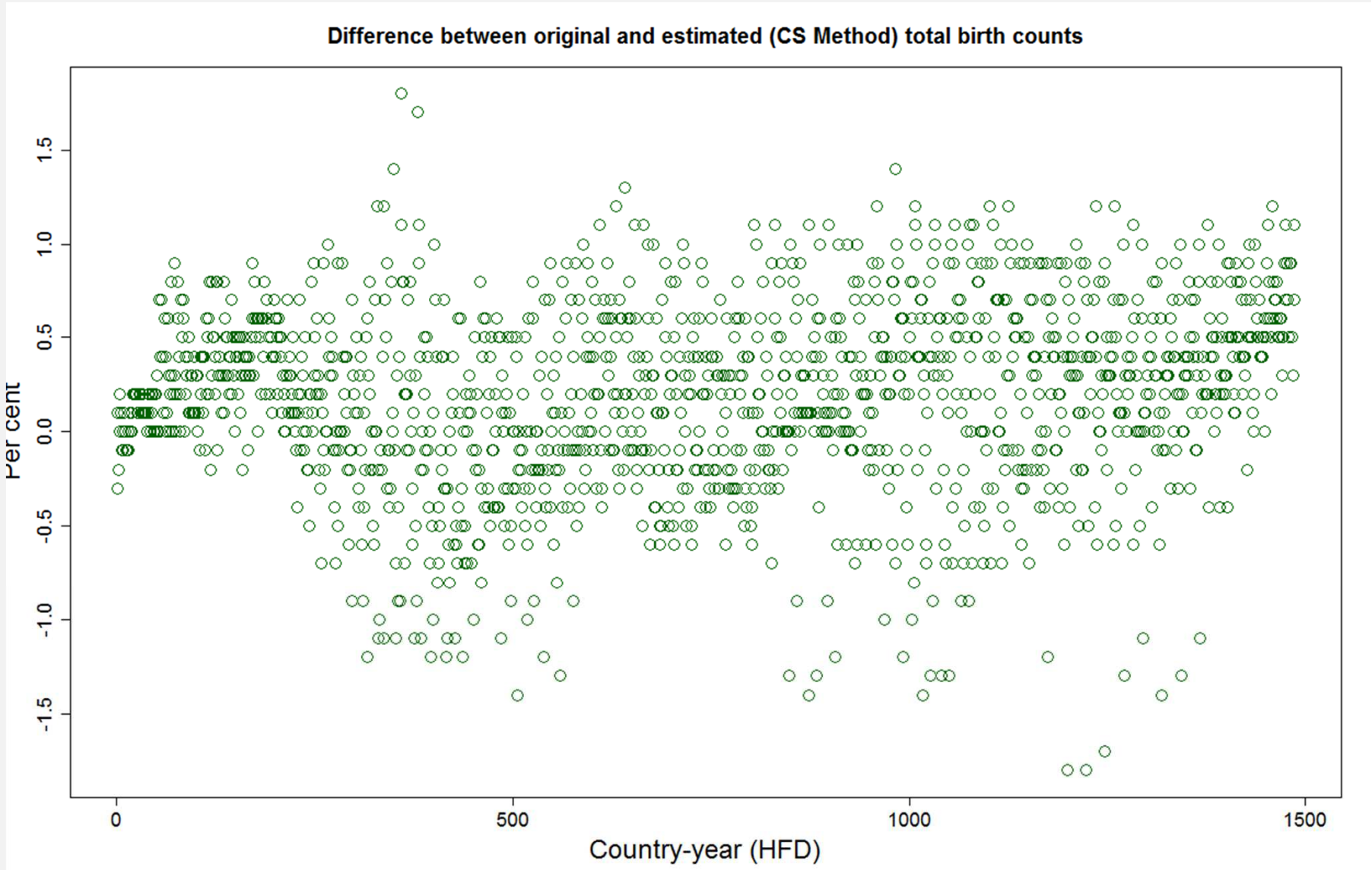
Age	CountryYear	Original	HFD	CSunadj	CSadj	Number of births (HFD)
12	GBR_NIR1975	0	0	0	0	0
13	GBR_NIR1975	0	0	0.0004	0.0004	0
14	GBR_NIR1975	0	0	0.0016	0.0016	0
15	GBR_NIR1975	0	0	0.0059	0.0059	0
16	GBR_NIR1975	0.01218	0	0.0145	0.0145	160
17	GBR_NIR1975	0.0311	0.0003	0.0305	0.0305	401
18	GBR_NIR1975	0.05435	0.0205	0.0533	0.0533	678
19	GBR_NIR1975	0.08733	0.1642	0.0775	0.0775	1053

Age	CountryYear	Original	HFD	CSunadj	CSadj	Number of births (HFD)
12	LTU1986	0	0	-0.0001	0	0
13	LTU1986	0	0	-0.0007	0	0
14	LTU1986	0.00005	0.0001	-0.0019	0	1
15	LTU1986	0.00049	0.0004	-0.0030	0	13
16	LTU1986	0.00264	0.0021	-0.0015	0	71
17	LTU1986	0.01115	0.0096	0.0123	0.0123	295
18	LTU1986	0.03063	0.0330	0.04	0.04	817
19	LTU1986	0.0704	0.0703	0.0775	0.0775	1901
20	LTU1986	0.11638	0.1036	0.1199	0.1199	3124
21	LTU1986	0.15065	0.1430	0.1502	0.1502	4134
22	LTU1986	0.17516	0.1713	0.1659	0.1659	4971

Difference between the original and estimated five-year birth counts before and after splitting

Code	Year	Age group	Before	After	Difference	Percent
DEUTE	1993	10-14	64	52	13	**19.7**
DEUTE	1993	15-19	4733	4481	252	5.3
DEUTE	1993	20-24	26488	26637	-149	-0.6
DEUTE	1993	25-29	26765	26364	401	1.5
DEUTE	1993	30-34	11082	11166	-84	-0.8
DEUTE	1993	35-39	3230	3372	-141	-4.4
DEUTE	1993	40-44	628	509	119	**19.0**
DEUTE	1993	45-49	20	6	14	**70.5**
DEUTE	1993	50-54	0	5	-5	Inf
DEUTE	1993	Total	73010	72592	418	0.6

The estimates are to be rescaled to match five-year input data.
It might affect the smoothness of the curve, particularly at the tails.


Difference between original and estimated (CS Method) total birth counts
Per cent
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0
500
1000
1500
Country-year (HFD)

Conclusion

Despite its limitations the HFD method is a reliable tool for splitting aggregated data.

Among other methods CS split is the best alternative to the current HFD protocol. It produces both better fit and more plausible shapes.

Because of its complexity and the need to perform post-estimation data adjustments CS method was not implemented in the HFD.

However, it turned out to be the optimal tool for the HFC containing ‘noisy’ and heterogeneous input data.

Acknowledgements

G.Rodriguez

D.Philipov

C.Schmertmann

Bibliography

Dougherty, R., Edelman A., Hyman, J. (1989). Nonnegativity-, monotonicity-, or convexity-preserving cubic and quantic Hermite interpolation. *Mathematics of Computation* (52), pp.471–494.

Hyman, J. (1983). Accurate monotonicity preserving cubic interpolation. *SIAMJournal on Scientific Computing* 4(4), pp. 645–654.

Jasilioniene, A., Jdanov, D.A., Sobotka, T., Andreev, E.M., Zeman, K., Nash, E.J. and Shkolnikov, V.M. (with contributions of Goldstein, J., Philipov, D. and Rodriguez, G.) (2012). Methods Protocol for the Human Fertility Database. Available at: http://www.humanfertility.org

Liu, Y., Gerland P., Spoorenberg T., Kantorova, V., Andreev K., (2011).Graduation methods to derive age-specific fertility rates from abridged data: a comparison of 10 methods using HFD data. Presentation at the First Human Fertility Database Symposium, Max Planck Institute for Demographic Research, Rostock, Nov 2011.

Schmertmann, C. (2012). Calibrated spline estimation of detailed fertility schedules from abridged data. MPIDR Working Paper WP-2012-022. Rostock.

Smith L., Hyndman R., Wood S. (2004). Spline interpolation for demographic variables: the monotonicity problem. *Journal of Population Research* 21 (1), pp. 95-97