Supplemental Material for

"Nonparametric Regression Model with Tree-structured Response"

Yuan Wang, J.S. Marron, Burcu Aydin, Alim Ladha, Elizabeth Bullitt and Haonan Wang

A Cross-Validated Bandwidth Selection

A leave-one-out cross-validation method is implemented here. At the i-th step of this cross-validation procedure, the observation (x_i, t_i) is removed from the entire data set as the test set, and the rest of the $n-1$ trees form the training set. This training data set is smoothed with bandwidth h. The predicted tree object of x_i is denoted by $\widehat{t}_{h,-i}$, and performance is assessed through the distance to the test tree t_i. Thus, a good choice of bandwidth h_0 is

$$h_0 = \arg\min_h CV(h)$$

where $CV(h) = n^{-1} \sum_{i=1}^{n} d_I(t_i, \widehat{t}_{h,-i})$. By minimizing $CV(h)$ over all positive h, we are looking for a bandwidth that minimizes the approximated sample absolute deviation.

For the left middle cerebral artery data studied in Section 4, values of $CV(h)$ are computed for integer values $h \in \{2, 3, \ldots, 30\}$. As illustrated in Figure S.1, the minimum cross-validation value occurs when $h_0 = 18$. This large bandwidth is not surprising due to the high auto-correlation of the data. Note that, there are several local minima, which is well known for cross-validation (Chiu and Marron, 1990).

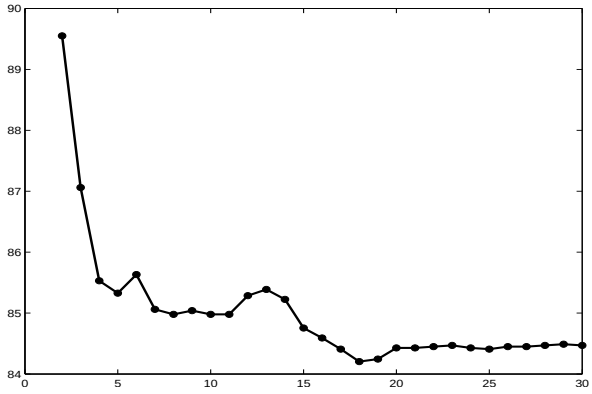

Figure S.1: A scatterplot of the cross-validation value $CV(h)$ (vertical axis) versus bandwidth h (horizontal axis) for the left middle cerebral artery system. The minimum is at $h_0 = 18$.

B Tree Smoothing Algorithm

Based on Theorem 1 and Theorem 2, the algorithm for solving the minimization of (7) from the main paper for a fixed bandwidth h consists of the following three steps:

1. Find the union tree t_U which consists of every node contained by at least one of the observed trees $t_1, \ldots, t_n$. Notice that $\mathrm{IND}(t_U) = \cup_{i=1}^{n} \mathrm{IND}(t_i)$;

2. For every node $k \in \mathrm{IND}(t_U)$, compute the score function

$$D_k(x) = \sum_{i=1}^{n} K_h(x - x_i) I\{v \in \mathrm{IND}(t_i)\} - \sum_{i=1}^{n} K_h(x - x_i)/2;$$

3. The sets

$$\{k : D_k(x) \geq 0\} \quad \text{and} \quad \{k : D_k(x) > 0\}$$

correspond to the maximal and minimal minimizer trees, respectively.

C Regression Analysis: Number of Nodes versus Age

Simple regression analysis on the number of nodes versus age can be carried out for each cerebral artery system. Results for the left middle cerebral artery system are shown in Section 4 of the paper. Here,

we will display the results for the other three cerebral artery systems. It shows that, in general, the number of nodes of the cerebral artery trees has a decreasing trend as age increases.

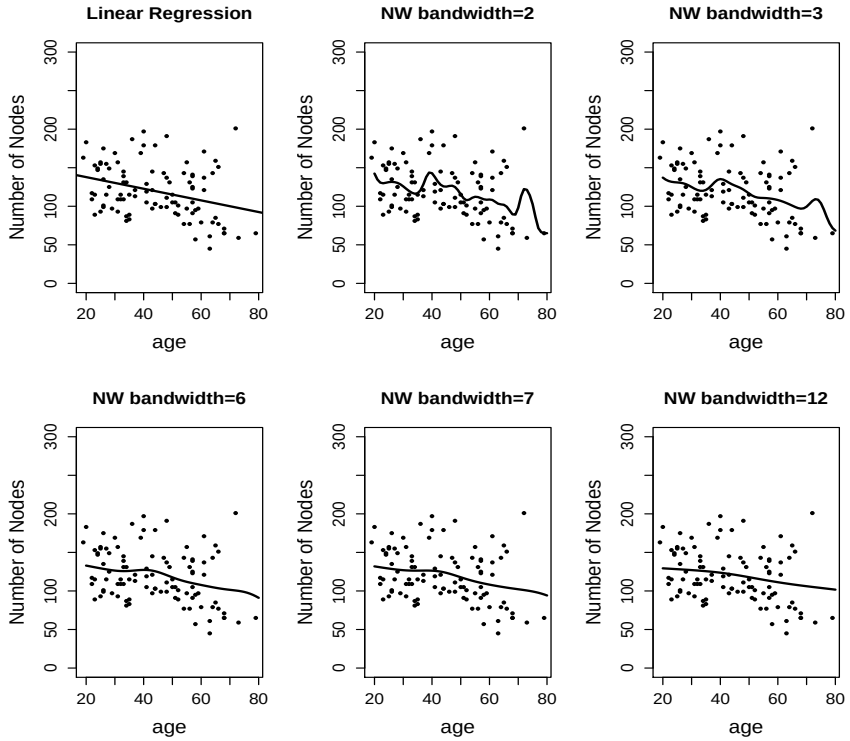

Figure S.2: Scatterplots of the number of nodes of the right middle cerebral artery trees. First panel: Simple linear regression with a significantly negative slope (p-value=0.001061). Panels 2 to 6: Nadaraya-Watson estimators for bandwidths $2, 4, 6, 7, 12$, respectively. Note that there is an overall decreasing trend in the number of nodes as age increases.

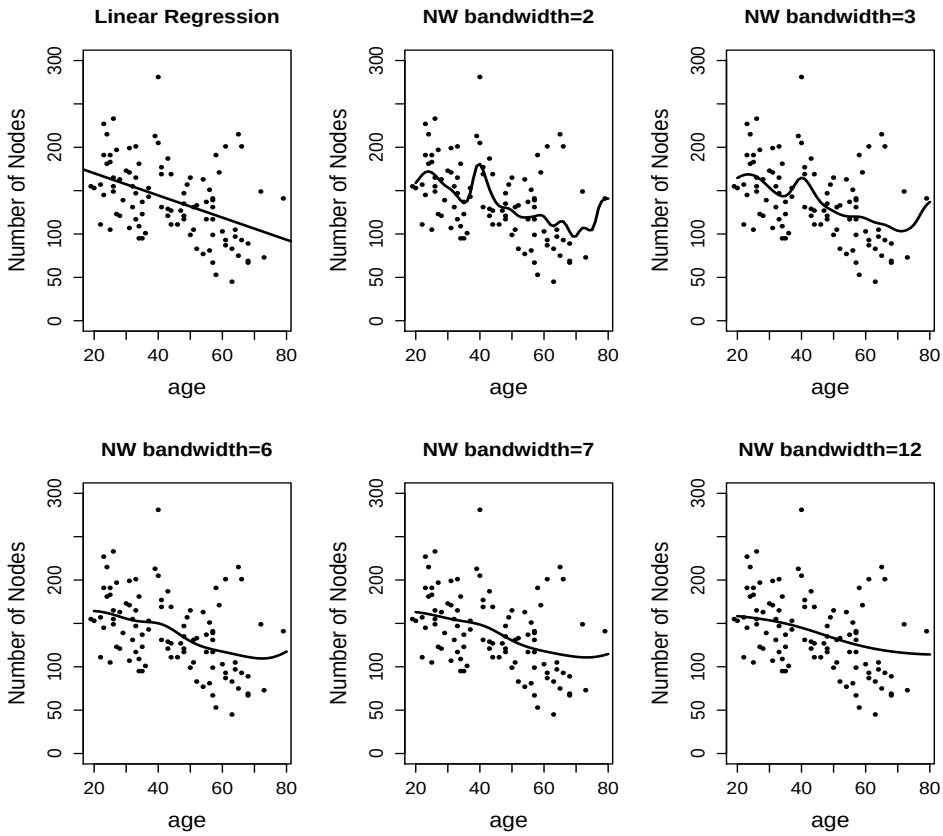

Figure S.3: Scatterplots of the number of nodes of the posterior cerebral artery trees. First panel: Simple linear regression with a significantly negative slope (p-value=7.636e-06). Panels 2 to 6: Nadaraya-Watson estimators for bandwidths $2, 4, 6, 7, 12$, respectively. During a long period at young ages, there is a decreasing trend in the number of nodes as age increases. However, for age greater than 75, the Nadaraya-Watson estimators show some increasing trend, but it is not clear the underlying population curve follows this trend due to the limited number of observations of old people.

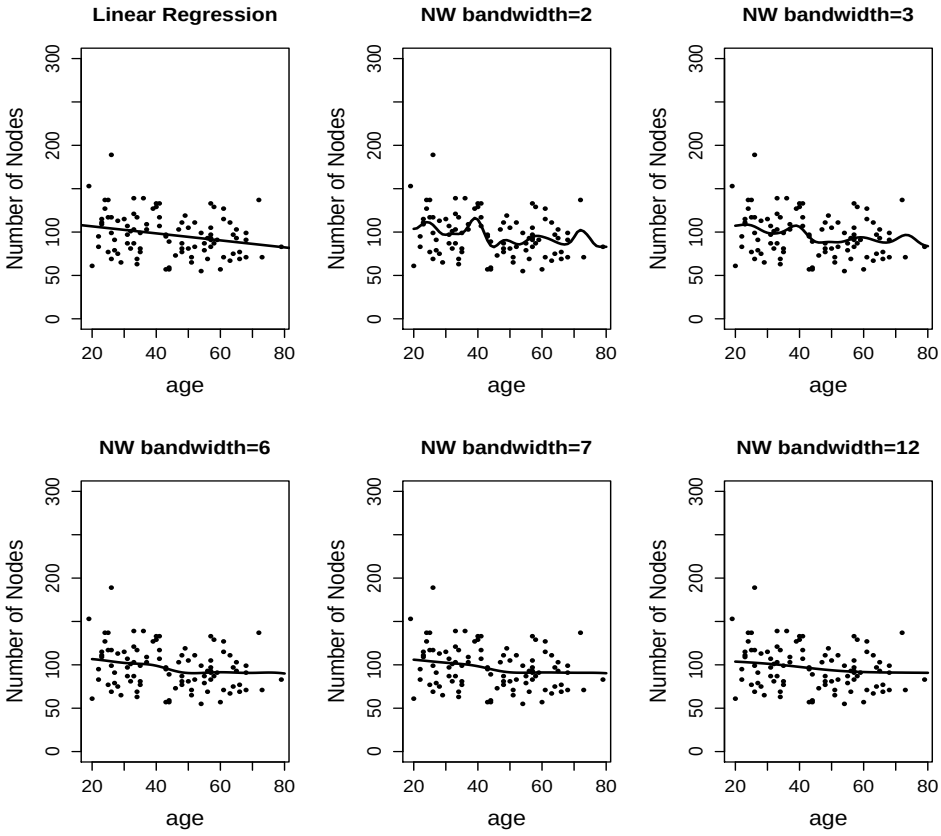

Figure S.4: Scatterplots of the number of nodes of the anterior cerebral artery trees. First panel: Simple linear regression with a negative slope (p-value=0.01475). At significance level 0.01, the decreasing trend is not significant. Panels 2 to 6: Nadaraya-Watson estimators for bandwidths $2, 4, 6, 7, 12$, respectively. Compared with the other three systems, the pattern for the front artery systems is relatively stable.

D More Results from Tree Smoothing for the Other Brain Artery Systems

We implement our proposed tree smoother on all four artery systems. The results for the left middle cerebral artery system are shown in Section 4 of the main paper. In this section, we will provide the results from analysis of the remaining three cerebral artery systems: the right middle cerebral artery, the posterior cerebral artery and the anterior cerebral artery.

For each system, we implemented our smoothing method for a set of bandwidths which are approximately equally spaced on a log scale and primarily intended to cover all the possible situations including undersmoothed, oversmoothed and moderately-smoothed. We first illustrate the smoothed results by summarizing the number of nodes of the fitted trees, at each selected bandwidth. By doing this, we try to describe how the cerebral artery systems grow over time and pick a reasonable bandwidth which balances the trade-off between variance and bias. Notice that small bandwidths usually yield more "wiggle" results while large bandwidths may oversmooth the data and obscure the underlying relationship between response and predictor variables. The next step, then, is to illustrate the topological structure of the fitted trees, at the selected bandwidth, at certain ages using the D-L view to find more interesting phenomena.

D.1 Results for Right Middle Cerebral Artery System

For the right cerebral artery system, we implemented our tree smoother methods for the set of bandwidths $\{2, 3, 4, 5, 6, 8, 12, 18\}$. In Figure S.5, each panel depicts a scatterplot of the number of nodes of the fitted trees versus age at the corresponding bandwidth. When the bandwidth is 2 or 3, the fitted tree seems "undersmoothed" since the number of nodes fluctuate strongly. When the bandwidth is 4 or 5, there are four distinct trends: one is decrease from 30 to 44, and one is increase from 44 to

57, and the third one is decrease from 57 to 70, and the fourth one is increase after 70. When the bandwidth is 6 or 8, the fluctuation at age less than 57 becomes stable. There is a decreasing trend from 59 to 68 and an increasing drift after 68. On the other hand, when the bandwidth is 12 or 18, the local smoother tends to oversmooth the tree data. For these large bandwidths, the dominant pattern is decreasing through the entire domain.

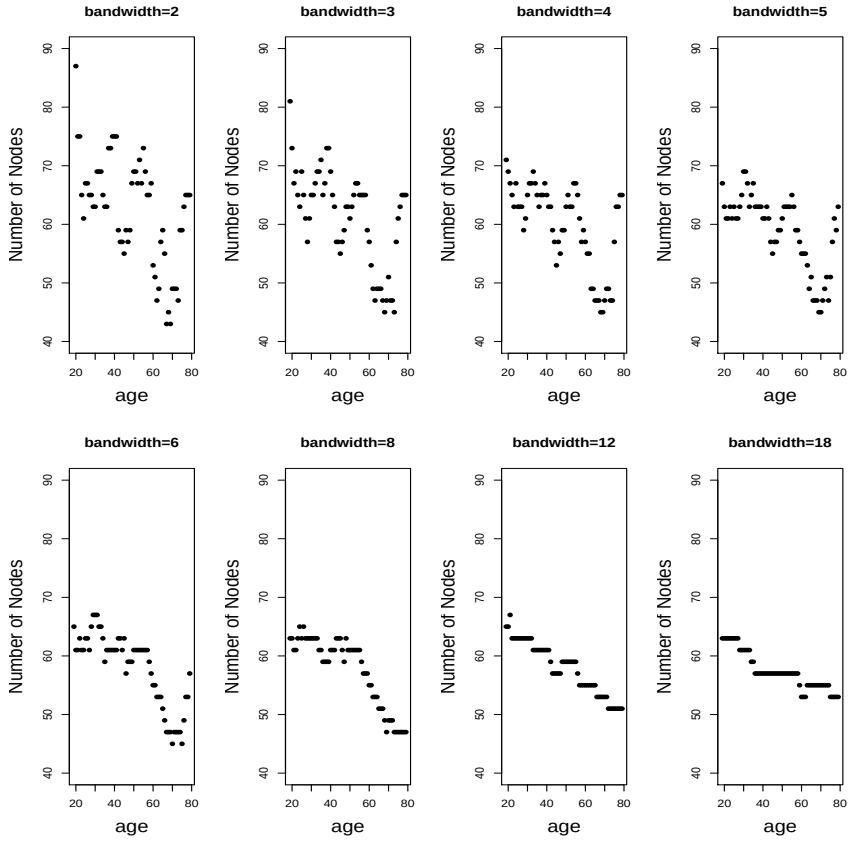

Figure S.5: Scatterplots of the number of nodes of the smoothed tree (right middle cerebral artery) versus age. There is a relatively stable pattern at young ages and a decreasing trend for older people.

Notice that trees with the same number of nodes could have very different structure; that is to say, the number of nodes only summarizes limited information contained in the tree. To reveal the change in characteristics of tree structure, such as branching pattern, as a function of age, we will present the

complete structure of the smoothed tree objects at bandwidth $h = 6$. We choose to display the results for bandwidth $h = 6$ because it seems to represent a reasonable trade-off between variance and bias among those bandwidths considered. Figure S.6 contains six subplots, in each of which, a D-L view of the topological tree structure of a fitted tree at a given age, obtained by our smoothing method with bandwidth $h = 6$ is given. Figure S.6 shows the tendency of change in the topological structure of the right middle cerebral artery trees. For our discussion, let $\widehat{t}_h(x)$ denote the fitted tree object at age x with bandwidth h. Compared with $\widehat{t}_6(20)$, $\widehat{t}_6(30)$ has more branching structure at low levels and less branching structure at high levels. This indicates a tendency that main artery segments grow and minute artery segments shrink or diminish over time. This tendency from 20 to 30 continues as age increases. Note that $\widehat{t}_6(40)$ has more branches at level 3 and less branches at levels $4, 6, 9$ than $\widehat{t}_6(30)$. When age increases from 40 to 50, the branches at levels $1, 2$, and 3 shrink while the branch at level 5 grows. Moreover, the number of nodes of the three estimated trees, $\widehat{t}_6(50)$, $\widehat{t}_6(60)$, $\widehat{t}_6(70)$, continues to decrease, and several branches at median levels shrink. This suggests for old people, the right middle artery segments have diminished over time.

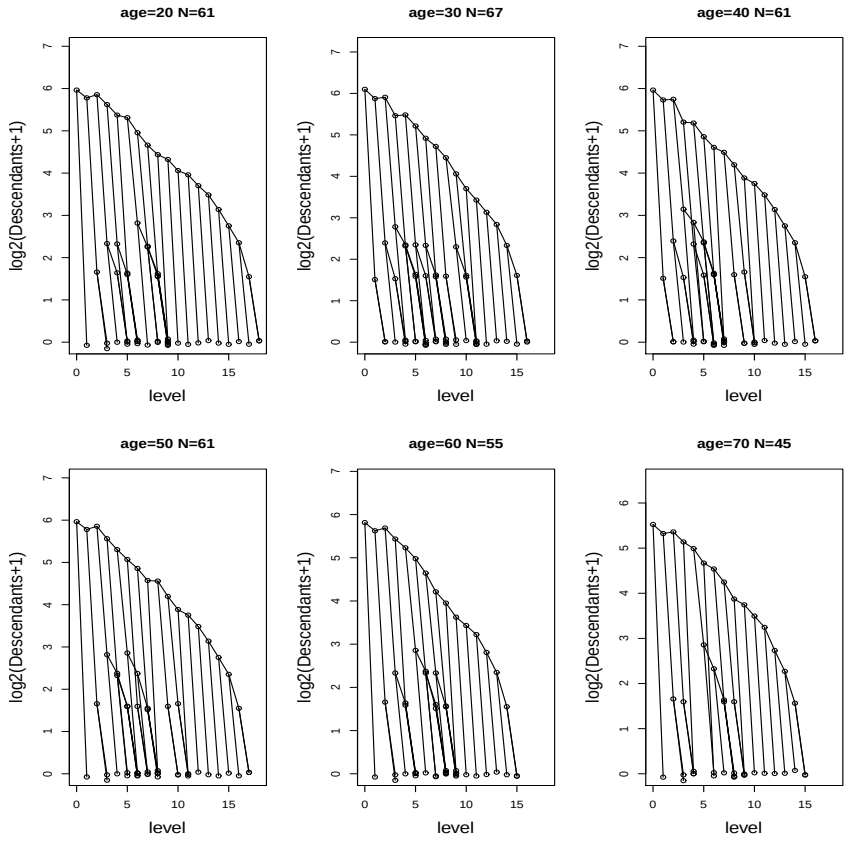

Figure S.6: A graphical illustration of the topological structures of the fitted tree-structured objects (right middle cerebral artery) at ages $20, 30, 40, 50, 60, 70$ when the bandwidth is 6. This shows an increase in structure for young ages, followed by a decrease in structure for older people.

D.2 Results for Posterior (Back) Cerebral Artery System

For the posterior cerebral artery system, we implement our tree smoother method for the same set of bandwidths. Figure S.7 shows the number of nodes of the fitted trees for each selected bandwidth. This result is rather different from the corresponding Figure S.5 for the right middle cerebral artery system. Now, for bandwidths 4, 5 and 6, the numbers of nodes decrease more monotonically. However, for larger bandwidths, the pattern is very similar as the right middle cerebral artery system. When the bandwidth is 8, there are three distinct trends: one is increase between 20 to 26, and one is decrease from 26 to 70, and the third one is increase after 70. Moreover, when the bandwidth is 12 or 18, the primary pattern is also decreasing through the whole domain.

Next, we will again focus our discussion on the smoothed trees with bandwidth 6. Figure S.8 is quite similar to Figure S.5, with D-L views of the same ages. This time, for $\widehat{t}_6(20)$, the two children of the root node both have rich branches while for the right artery system, the right child of the root node is a leaf node. As age increases, we can see that, in contrast with the growing pattern of the right system, $\widehat{t}_6(30)$ now has fewer branches at level 1 than $\widehat{t}_6(20)$, and $\widehat{t}_6(40)$ has further fewer branches at levels 1 and 2 than $\widehat{t}_6(30)$. This indicates a tendency for the posterior artery segments to diminish at young ages. The difference between the two systems is more obvious for older people with age greater than 50. First, the maximum level for the posterior system reduces more quickly. Second, for the branches at low levels, different from the shrinking pattern of the right system, the posterior system grows a lot from 50 to 60.

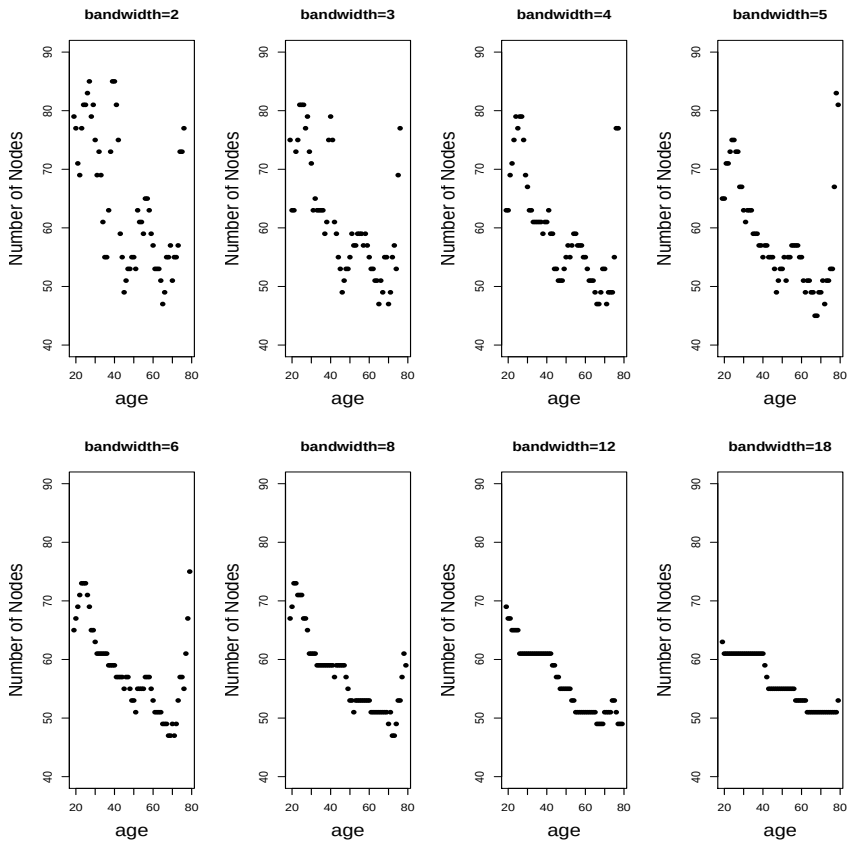

Figure S.7: Scatterplots of the number of nodes of the smoothed (posterior cerebral artery system) tree versus age. There is an overall decreasing trend, except for younger and older people at intermediate bandwidths.

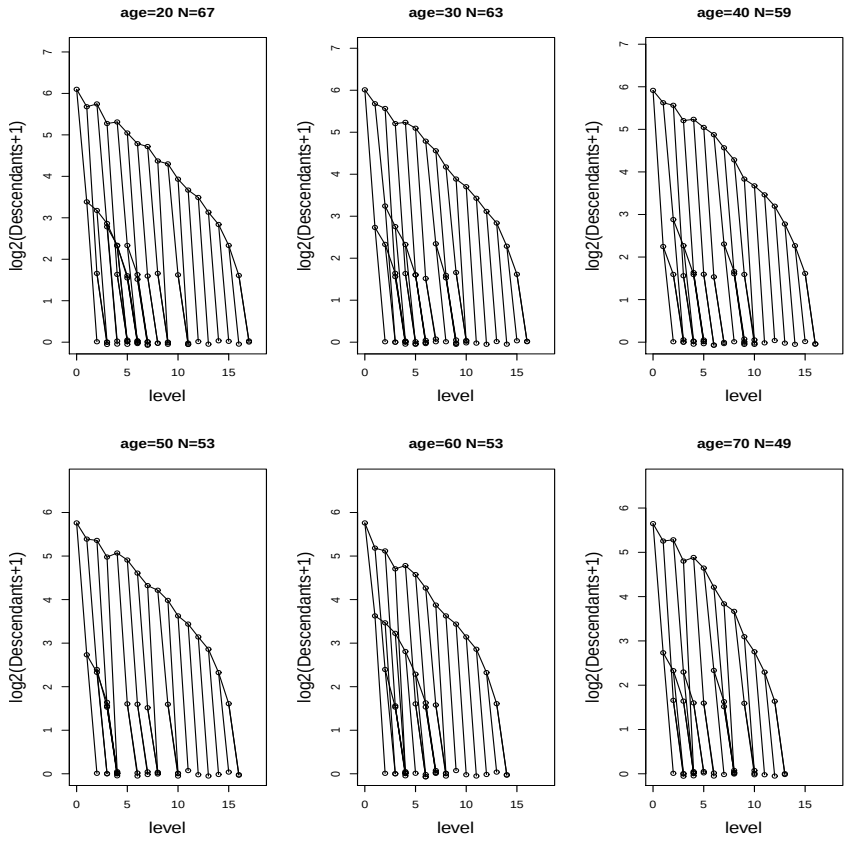

Figure S.8: A graphical illustration of the topological structures of the fitted tree-structured objects (posterior cerebral artery system) at ages $20, 30, 40, 50, 60, 70$ for bandwidth 6. This shows a decrease in structure for younger people.

D.3 Results for Anterior (Front) Cerebral Artery System

Application of our tree smoothing technique to the anterior cerebral artery data reveals a quite different pattern, as shown in Figure S.9. First, it can be seen that the smoothed estimates at small bandwidths 2, 3 or 4 are not so "wiggly" as the other three systems. Moreover, as the bandwidth increases (5, 6, 7, 8 and 12), in contrast with the mostly decreasing trend of the other systems, there is a V-shaped pattern: a decreasing trend from 20 to middle age and an increasing trend after that. When the bandwidth is 18, the V-shaped pattern is not as clear, which may be due to oversmoothing.

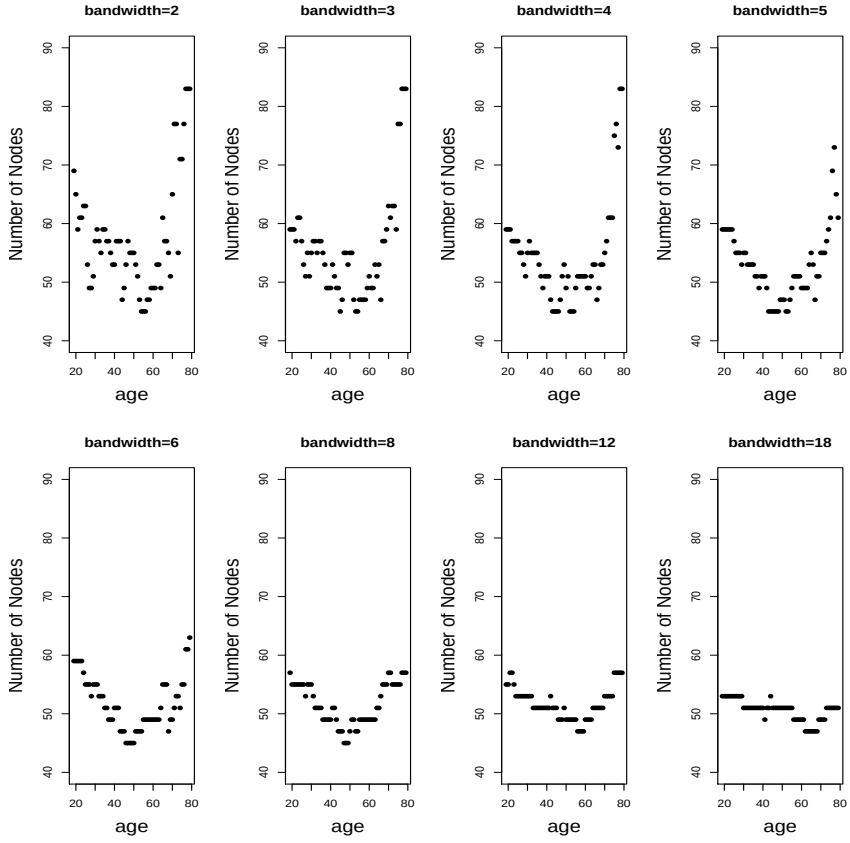

Figure S.9: Scatterplots of the number of nodes of the smoothed tree(anterior cerebral artery system) versus age. Note a V-shaped pattern except for the largest bandwidth.

Among all the selected bandwidths, $h = 5$ is chosen for further investigation. The corresponding fitted trees of the same ages are displayed in Figure S.10. When age increases from 20 to 50, the estimated central trees keep shrinking in a way that many branches at low levels disappear. This indicates a tendency for anterior artery segments to diminish at young ages. An increasing pattern appears after age 50. The right child of the root node of $\widehat{t}_5(60)$ has more branching activity among the offspring than for $\widehat{t}_5(50)$. Moreover, $\widehat{t}_5(70)$ has many branches at higher levels growing out. This suggests that, for older people, the anterior artery segments have split further.

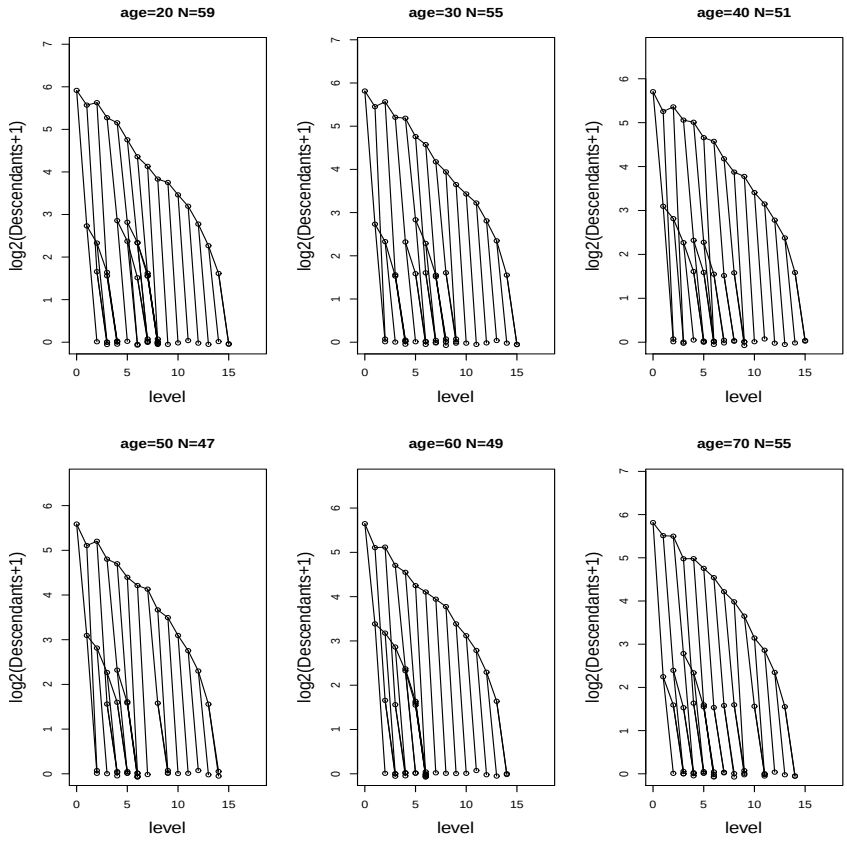

Figure S.10: A graphical illustration of the topological structures of the fitted tree-structured objects (anterior cerebral artery system) at ages $20, 30, 40, 50, 60, 70$ for bandwidth 5. This shows a decrease in structure for people younger than 50, and an increasing trend in structure after that.

In this section, we studied the three cerebral artery systems by summarizing the numbers of nodes of the fitted tree at different bandwidths and then characterizing the topological structures of the fitted trees at an interesting fixed bandwidth. In contrast to the simple regression on number of nodes (see Section C), in which we only observed an overall decreasing trend for all cerebral artery systems, our tree smoothing method reveals deeper relationships between cerebral artery systems and age. It can be seen that age systematically affects the structure of cerebral artery systems and such age effects vary across different artery systems. Further investigation is still ongoing for more biological explanation for such phenomena.

E Results for the Left Middle Cerebral Artery System

In Section 4 of the main paper, we depicted in Figure 6 a graph of the fitted tree-structured objects for the left middle cerebral artery system with bandwidth $h = 6$. In this section, we mainly display the results for the other bandwidths considered. Figures S.11 to S.17 display the fitted trees at bandwidth $h = 2, 3, 4, 5, 8, 12, 18$, respectively. In each figure, a D-L view of the fitted trees at ages $20, 30, 40, 50, 60, 70$ is depicted. The variable N indicates the number of nodes of each tree. Note that small bandwidths usually undersmooth the data while large bandwidths may oversmooth. By comparing the results for different bandwidths, we try to illustrate how bandwidth will affect the smoothed estimates. Brief description of our findings is stated in the caption of each figure. For our discussion, we will continue to let $\widehat{t}_h(x)$ denote the fitted tree object at age x with bandwidth h.

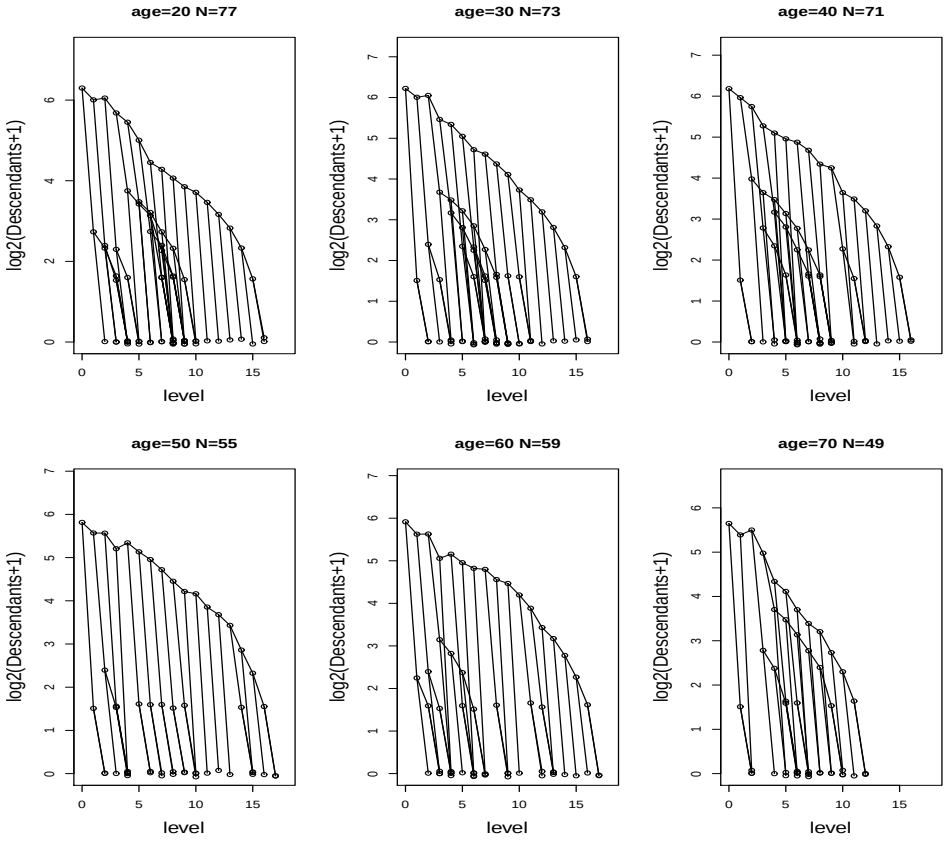

Figure S.11: A graphical illustration of the fitted tree-structured objects when the bandwidth is 2. For the top row, as age increase, many branches at low levels shrink. There is a big drop when age increases from 40 to 50 (60 to 70), as the number of nodes reduces from 71 to 55 (59 to 49, respectively). This strong fluctuation reveals the "undersmoothed" estimates.

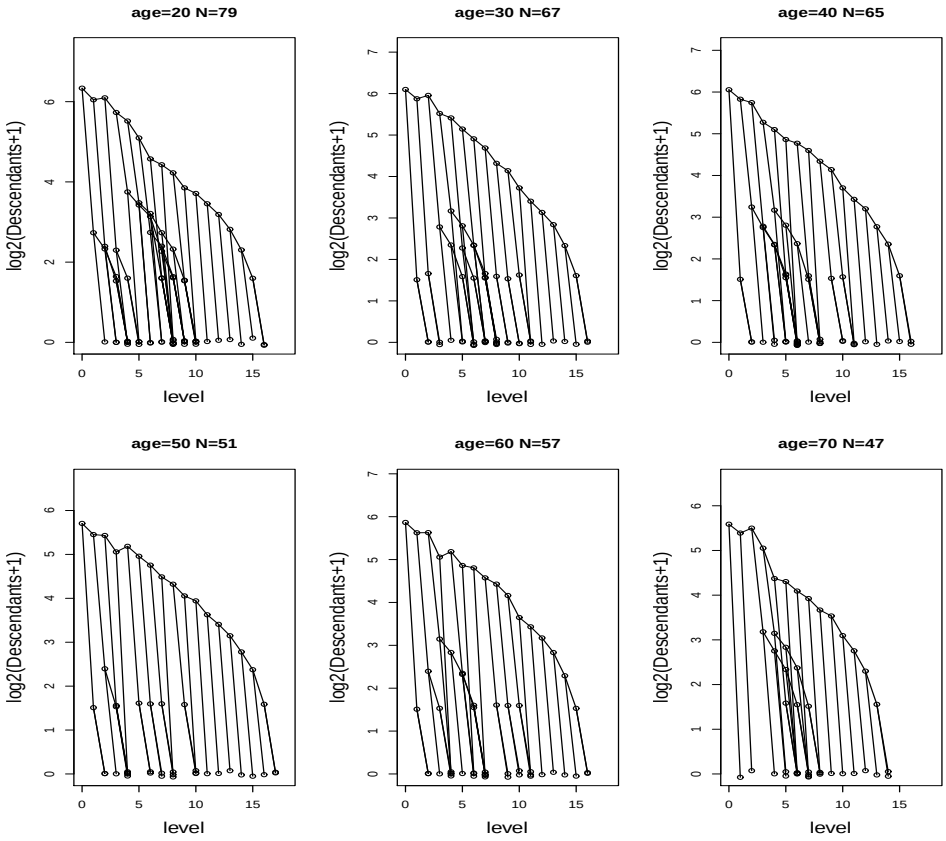

Figure S.12: A graphical illustration of the fitted tree-structured objects when the bandwidth is 3. The pattern is similar to for bandwidth 2.

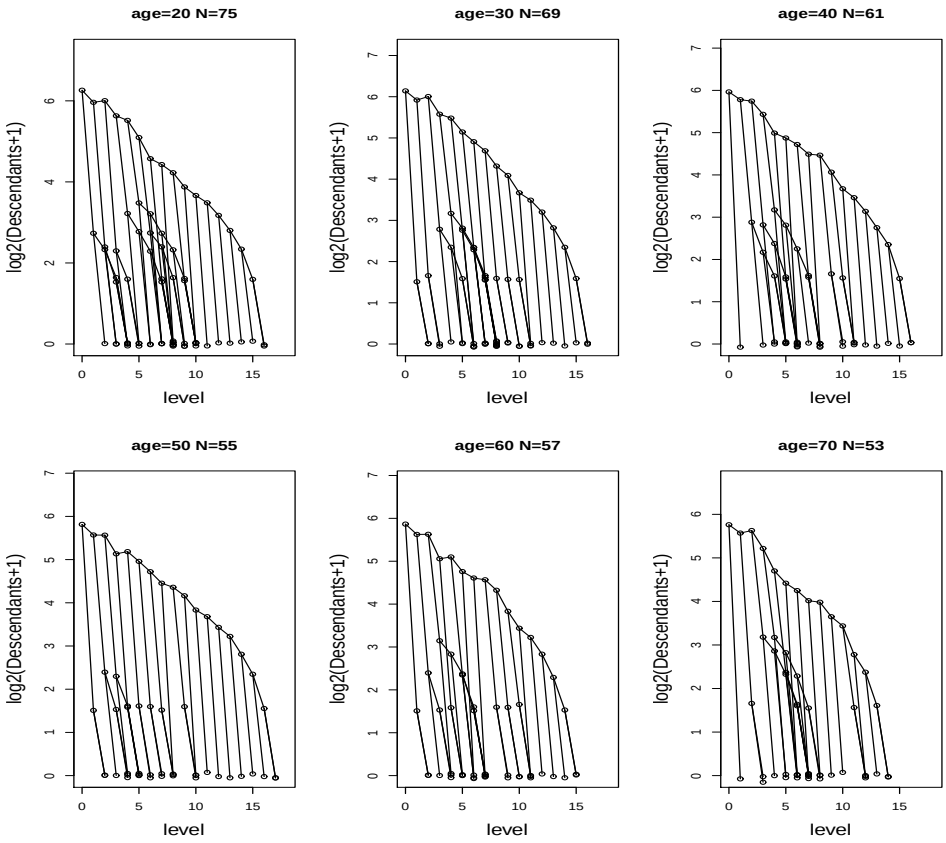

Figure S.13: A graphical illustration of the fitted tree-structured objects when the bandwidth is 4. As age x increases to 50, $\widehat{t}_4(x)$ continues to shrink at some low levels. When age increases from 50 to 60, there is more branching structure at low levels and fewer branches at higher levels, and this tendency continues as age increases from 60 to 70.

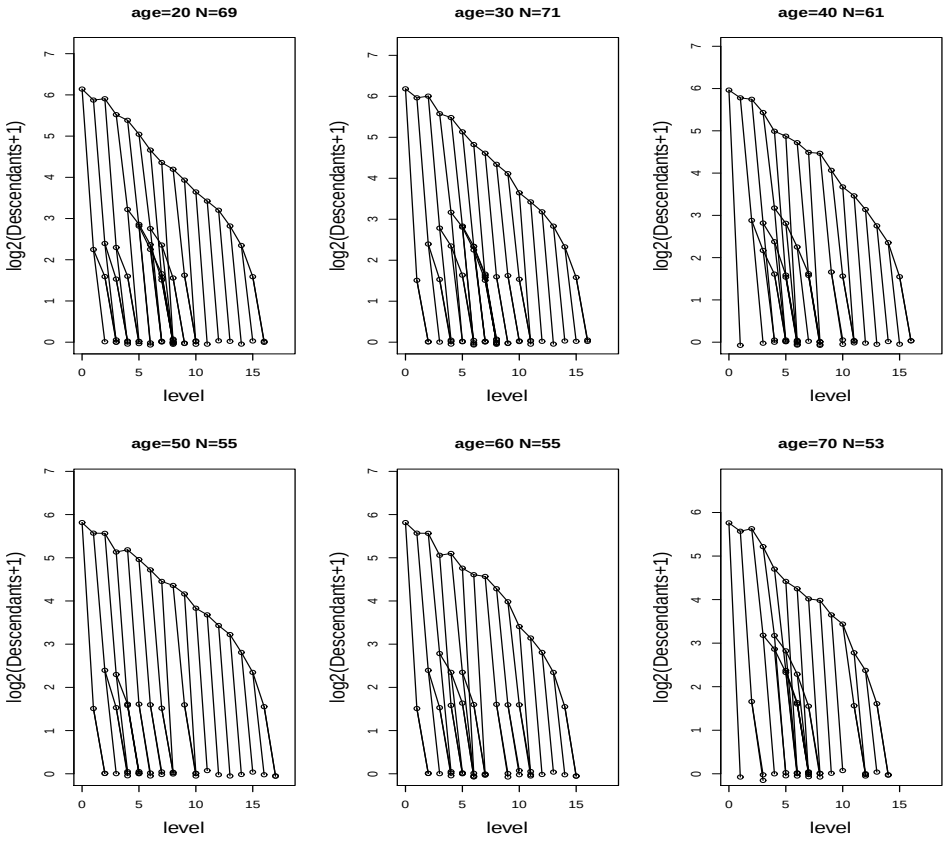

Figure S.14: A graphical illustration of the fitted tree-structured objects when the bandwidth is 5. Note that when age increase from 20 to 30, the fitted tree-structured object at $h = 5$ suggests an increase in structure. This phenomenon was not seen for bandwidths $h = 2, 3, 4$.

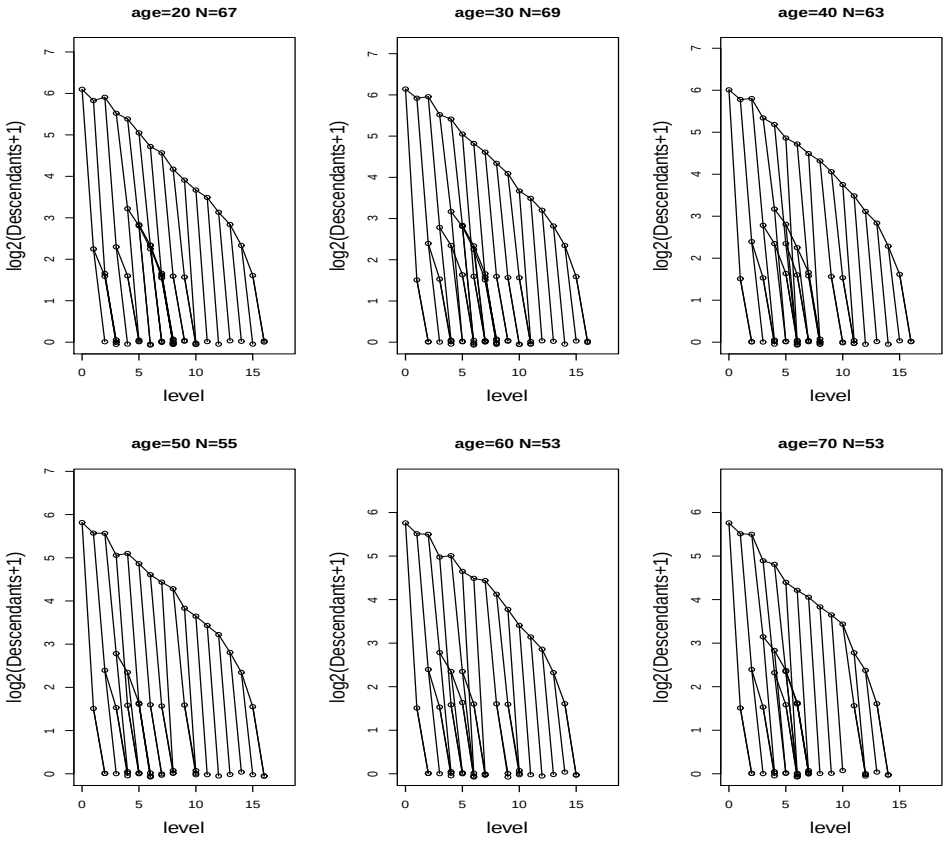

Figure S.15: A graphical illustration of the fitted tree-structured objects when the bandwidth is 8. When bandwidth increases from 5 to 8, the increasing pattern from age 20 to 30 continues. In general, the change of the fitted trees over each 10 year time period is smoother than for the smaller bandwidths.

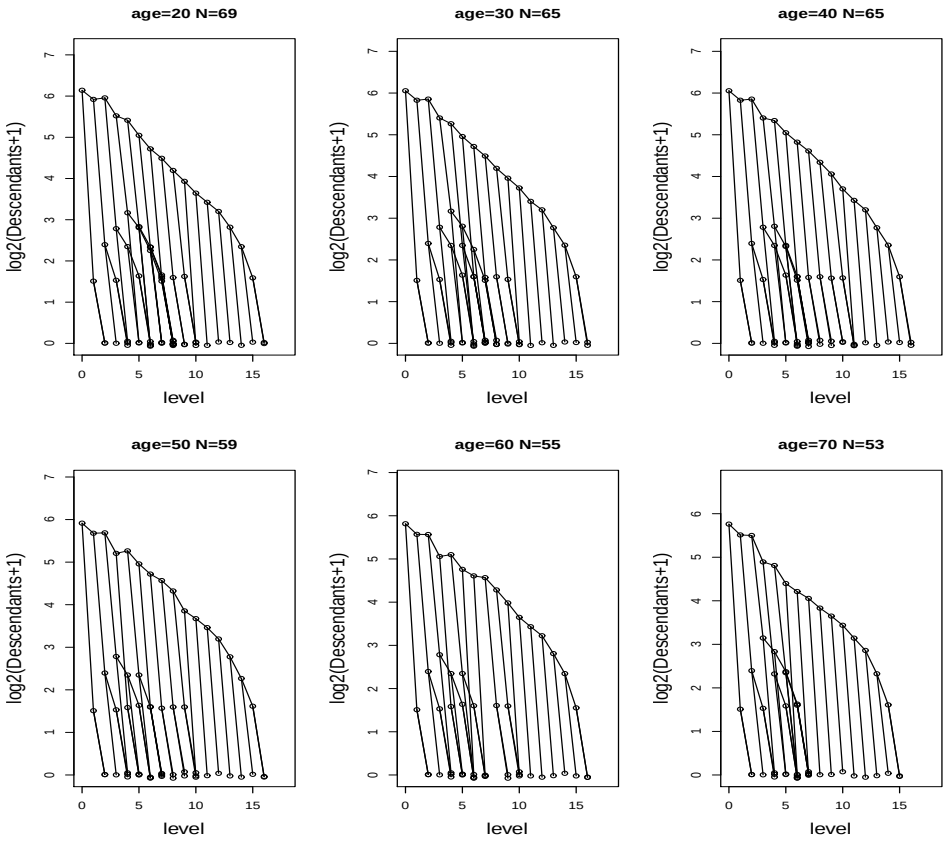

Figure S.16: A graphical illustration of the fitted tree-structured objects when the bandwidth is 12. First it can roughly be seen that the fluctuation is very small compared with the other bandwidths, especially note that $\widehat{t}_{12}(20)$ and $\widehat{t}_{12}(30)$ share almost the same structure. Second, the increasing pattern from $\widehat{t}_8(20)$ to $\widehat{t}_8(30)$ now disappears. We only see a slow decreasing trend over the entire domain.

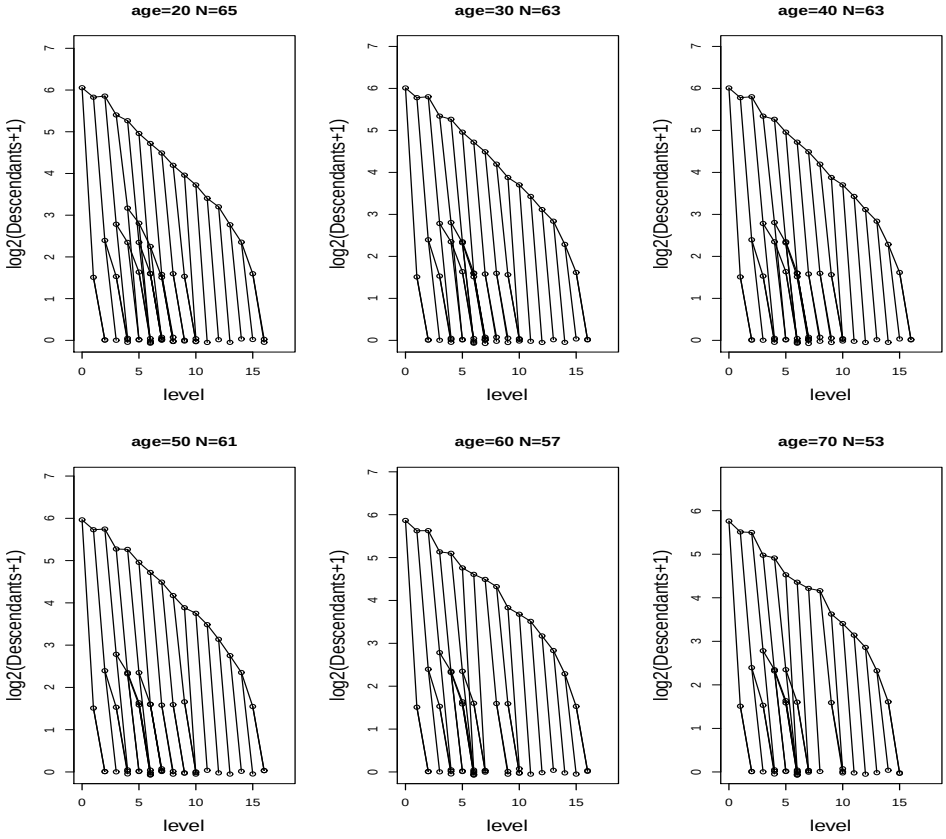

Figure S.17: A graphical illustration of the fitted tree-structured objects when the bandwidth is 18. The pattern is similar as that of bandwidth 12.

F More Simulation Results

In Section 5 of the main paper, we conducted some simulations to demonstrate the performance of our tree smoothing method. To save space, there we only present the smoothed tree estimates in one realization of scenario 1. In this section, we show the results for the other three scenarios discussed in Section 5. For each scenario, graphical illustration of the fitted tree overlaid with the randomly generated tree observations and the smoothed estimates is provided. A movie version is attached to this supplemental material. For each scenario $i = 1, 2, 3, 4$, the movie is named *scenario* i*.wmv*. Each movie includes the observation trees, the smoothed trees and the true parameter trees.

In Figures S.18-S.20, for the scenarios 2-4, the top row of each figure shows the D-L view of four tree realizations (long-dashed) at age around 36. Each panel also contains the population Fréchet central tree $\mu_F(x)$ (solid) and the resulting tree from our smoothing method (dotted). The second row shows the corresponding results for age around 66. It can be seen that, in general for all three scenarios, while the observations tend to be larger than the tree $\mu_F(x)$, the fitted tree captures the pattern of $\mu_F(x)$ quite well.

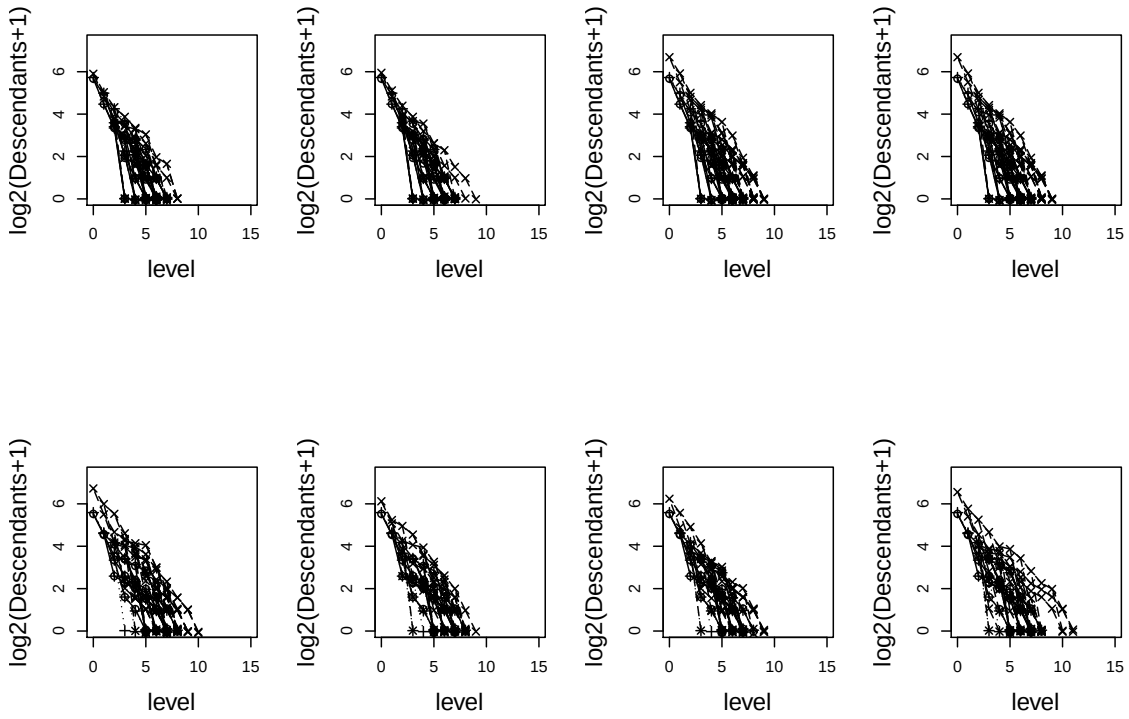

Figure S.18: Scenario 2: in the top row, the D-L view of four realizations (long-dashed) at age $x = 35$ are given in each subplot. The tree μ_F (solid) and our smooth predicted tree (dotted) at the same age are overlaid. The bottom row shows same results for age $x = 68$.

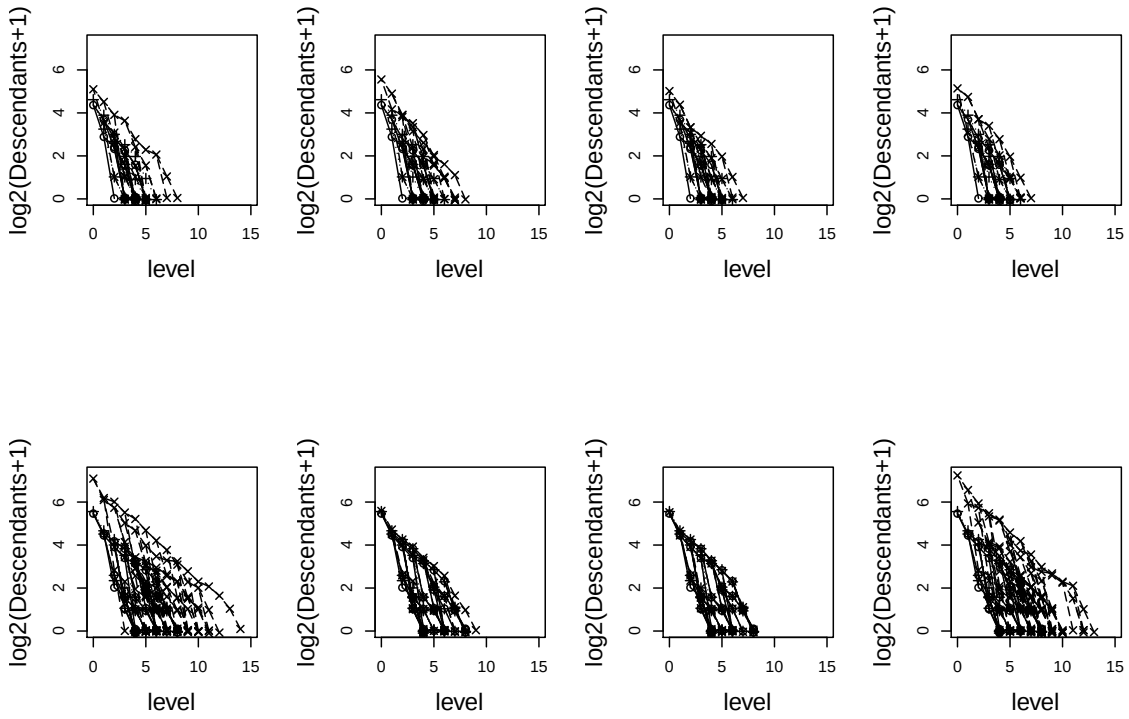

Figure S.19: Scenario 3, in the top row, the D-L view of four realizations (long-dashed) at age $x = 31$ are given in each subplot. The tree μ_F (solid) and our smooth predicted tree (dotted) at the same age are overlaid. The bottom row shows same results for age $x = 65$.

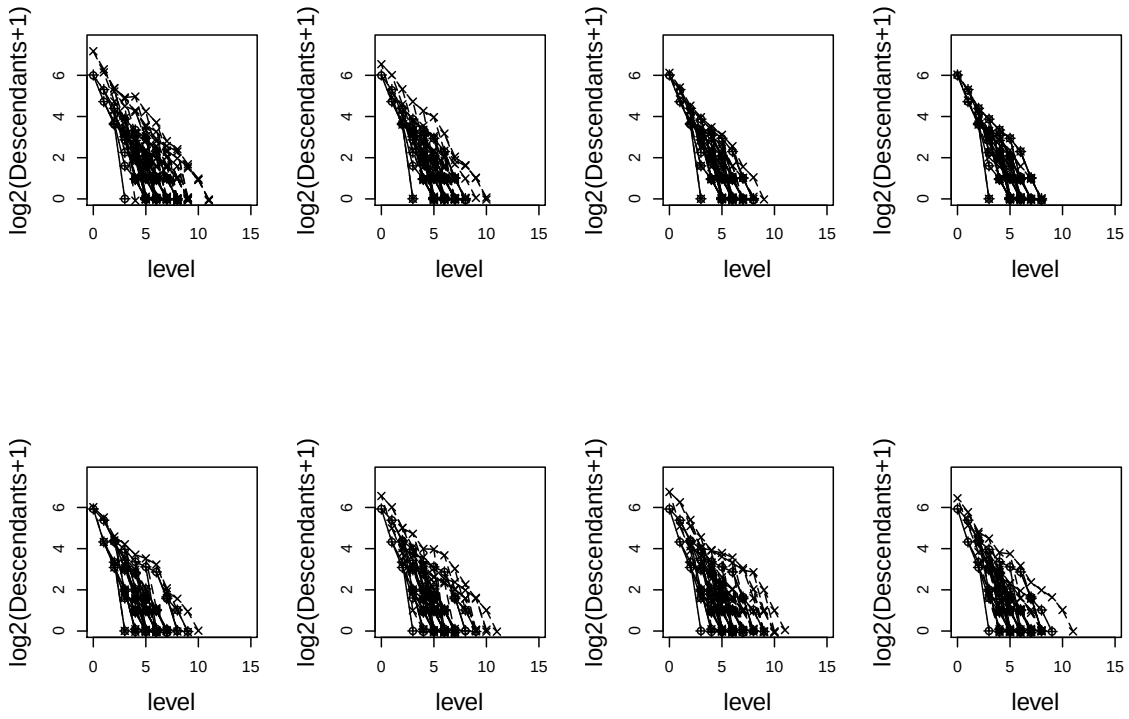

Figure S.20: Scenario 4: in the top row, the D-L view of four realizations (long-dashed) at age $x = 46$ are given in each subplot. The tree μ_F (solid) and our smooth predicted tree (dotted) at the same age are overlaid. The bottom row shows same results for age $x = 65$.

In the simulation study, we also conduct a Monte Carlo experiment of $R = 100$ replicates and measure the performance of our tree smoothing method by the absolute estimation error and the Fréchet variation. The Monte carlo estimates of these two measures are shown in Figure 12 of the main paper. Here, Figure S.21 shows boxplots, for each x, of the estimated Fréchet variations based on all 100 replicates. In all four scenarios, these boxplots indicate constant variation for the simulated data.

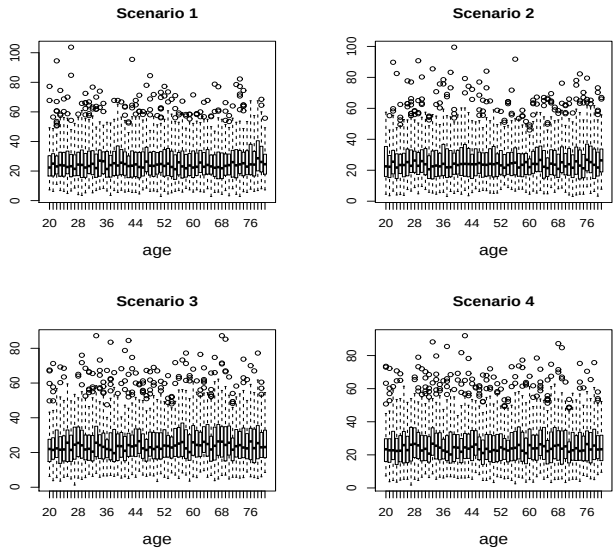

Figure S.21: Boxplots for estimated Fréchet variation over 100 realizations for all four simulation scenarios. These suggests no heteroscedasticity for the simulated data.