ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH

DOI: 10.5336/biostatic.2019-66285

Estimation of Right Censored Nonparametric Regression Solved by *k*NN Imputation: A Comparative Study

*k*NN Tamamlama Yöntemi ile Çözülen Sağdan Sansürlü Parametrik Olmayan Regresyon Modelinin Tahmini: Karşılaştırmalı Bir Çalışma

Ersin YILMAZ[a],
Dursun AYDIN[a]

[a] Department of Statistics, Mugla Sıtkı Koçman University Faculty of Science, Mugla/TURKEY

Received: 01.04.2019
Received in revised form: 10.05.2019
Accepted: 20.05.2019
Available Online: 22.05.2019

Correspondence:
Ersin YILMAZ
Mugla Sıtkı Koçman University Faculty of Science, Department of Statistics, Muğla, TURKEY/TÜRKİYE
ersinyilmaz@mu.edu.tr

ABSTRACT This paper introduces an estimation procedure for the right-censored nonparametric regression model using smoothing spline method. In this process, to overcome the censorship problem we used an imputation method based on *k*-nearest neighbors (*k*NN). Among some known censorship solutions, such as Kaplan-Meier weights (Kaplan and Meier, Miller) and Synthetic data transformation (Koul et al.), the most important advantage of the *k*NN imputation method is that it does not depend on a distribution. After solving the problem of censorship, the most important problem in obtaining the optimal estimation of non1parametric regression function by using smoothing spline will be the selection of the smoothing parameter. In order to achieve this aim, three commonly used criteria such as generalized cross-validation (GCV), Bayesian information criterion (BIC) and risk estimation using classical pilots (RECP) are considered in this study. A Monte-Carlo simulation study and a "kidney infection recurrence" data are carried out to realize the purposes of this study. Thus, it is determined that which selection criterion is more successful in estimating the non-parametric model with right censored data. Obtained results from both simulation and real-world studies show that BIC has remarkable performance among others. Also, it can be seen that GCV is better than BIC for large sample size. RECP has mediocre performance.

Keywords: Censored data; nonparametric regression; smoothing spline; *k*NN imputation; smoothing parameter

ÖZET Bu makalede, düzleştirici splayn yöntemi kullanılarak sağdan sansürlü parametrik olmayan regresyon modeli için bir tahmin prosedürü sunulmaktadır. Bu süreçte, sansür sorununun üstesinden gelmek için, en yakın komşulara (*k*NN) dayanan bir tamamlama (yerine koyma) yöntemi kullanıldı. Kaplan-Meier ağırlıkları (Kaplan ve Meier, Miller) ve Sentetik veri dönüşümü (Koul ve ark.) gibi bilinen bazı sansür çözümleri arasında, *k*NN değerleme yönteminin diğerlerine göre en önemli avantajı, bir dağılıma bağlı olmamasıdır. Sansür problemini çözdükten sonra, düzeltme parametresi kullanarak parametrik olmayan regresyon fonksiyonunun en uygun tahminini elde etmedeki en önemli problem, düzeltme parametresi seçimi olacaktır. Bu amaca ulaşmak için, genelleştirilmiş çapraz doğrulama (GCV), Bayes bilgi kriteri (BIC) ve klasik pilotlar kullanılarak risk tahmini (RECP) gibi yaygın olarak kullanılan üç kriter ele alınarak düzeltme parametresi seçilmiştir. Bu çalışmanın amaçlarını gerçekleştirmek için bir Monte-Carlo simülasyon çalışması ve "böbrek enfeksiyonun tekrar etmesi" verileri ile uygulama çalışması yapılmıştır. Böylelikle parametrik olmayan regresyon modelinin sağdan sansürlü verilerle tahmin edilmesinde hangi seçim kriterinin daha başarılı olduğu tespit edilmiştir. Hem simülasyon hem de gerçek veri çalışmalardan elde edilen sonuçlara göre, BIC yönteminin diğerleri arasında dikkate değer bir performansa sahip olduğu kolaylıkla görülmektedir. Ayrıca, GCV yönteminin büyük örneklem büyüklüğü için BIC'den daha iyi sonuçlar verdiği söylenebilir. RECP yöntemi ise diğer iki yönteme göre vasat bir performans sergilemiştir.

Anahtar Kelimeler: Sansürlü veri; parametrik olmayan regresyon; splayn düzeltme; *k*NN tamamlama; düzeltme parametresi

Censored data is an important problem in many different fields, especially in survival analysis including medical research studies. As is known, the completeness and quality of the data must be ensured in order to obtain a qualified model and inference, because the censored observations seriously affect the accuracy and reliability of the analysis. Therefore, in order to avoid biased and ineffective results, it is necessary to overcome the censored data in the modeling process. To achieve this goal, there are a number of ways in which censored observation can be removed from the data set, which is a primitive technique to overcome censorship, because it leads to loss of information and biased results. In the literature, there are more common methods to solve censorship such as Kaplan-Meier weights (Miller, Stute, ; Orbe et al.) and synthetic data transformation (Koul *et al.*, Leurgans, Aydın and Yılmaz) but, such methods have some theoretical limitations.[1-7] Although these common methods give efficient and consistent results, they cannot provide the accuracy of censored observations individually. The imputation methods are preferred in this context.

The imputation is expressed as a process that replaces censored observations with predicted values in a data set with the help of the specified method. In the literature, some examples of imputation include Schafer, Batista and Monard, Rubin and van der Laan, Yenduri and Iyengar and Andridge and Little.[8-12] There are also various imputation methods used for different types of data such as fuzzy K-means, singular value decomposition, multiple imputations by chained equations.[13,14] Most of these methods are developed to solve missing data problems but some of these methods are also suitable for solving the right-censorship problems. The *k*NN imputation can be counted as one of them. The *k*NN method computes the imputed value from average of measured *k* records in the dataset and it replaces it with the censored observation. Most important advantage of the *k*NN method is that it allows the imputation for all kinds of data such as high-dimensional data, right, left censored or interval censored data and it is independent from distribution assumptions. Some of the studies on the *k*NN imputation in the literature can be considered as Batista and Monard, Malarvizhi and Thanamani, and Chen and Shao.[9,15,16]

In this study, nonparametric regression model is estimated by smoothing splines under right-censored data problem solved by *k*NN imputation. However, as expressed earlier, in the smoothing method, the accuracy of the estimation is largely dependent on the smoothing parameter. The main purpose of this paper is to determine the optimum smoothing parameter by using *GCV*, *BIC* and *RECP* criteria. In order to observe how the criteria work, a comparative Monte-Carlo simulation study and then the kidney recurrence data set are considered as a real data example

The paper is organized as follows. Section 2 includes the smoothing spline method based on *k*NN imputation. Selection of the smoothing parameter is discussed in Section 2.3. Performance measurements for evaluate the estimated models are introduced in Section 2.4. Results of the simulation study and kidney recurrence dataset are presented in Section 3.1 and Section 3.2 respectively. Finally, the discussion and conclusions are given in Section 4 and Section 5.

MATERIAL AND METHODS

Consider the right-censored nonparametric regression model as follows

$$t_i = g(z_i) + \varepsilon_i, \quad 1 \le i \le n \tag{1}$$

where t_i'*s* are the data points of the right-censored response variable, z_i's are the values of nonparametric explanatory variable, *g(.)* is an unknown smooth function to be estimated and ε_i'*s* are the random error terms with mean zero and constant variance. Here, t_i's are censored by some random variable c_i. In this case, response variable is updated according to censor and new response variable can be written as

$$y_i = \min\ (t_i, c_i) \text{ and } \delta_i = I(t_i < c_i) \quad (2)$$

where y_i's are the values of the new response variable and δ_i's contain censor information about the data points. If y_i is censored then δ_i=0 and if not δ_i=1. In this context, modelling of the data can be possible with (y_i, z_i, δ_i) instead of (t_i, z_i). Thus, it can be said that in the real world the nonparametric regression model with right-censored data is given by

$$y_i = g(z_i) + \varepsilon_i,\ 1 \le i \le n \quad (3)$$

*k*NN IMPUTATION

As mentioned above, to solve censorship in model (3), *k*NN imputation is used. It should be noted that *k*NN works independently of the distributions of both y_i and c_i. Basically, *k*NN estimates each censored observation by using the average value of its *k*-nearest neighbours. Some important properties can be considered as follows: (i) the *k*NN does not manipulate data, it uses actual data points to estimate censored ones. (ii) the *k*NN works with both discrete and continuous variables. (iii) the *k*NN is fully nonparametric, which makes it preferable.

Working procedure of the *k*NN imputation is quite simple. It works with distances between data points and uses them as a measure of similarity. There are different distance measures such as Minkowski, Manhattan, Euclidean and Mahalanonbis. In general, the Euclidean distance is used in *k*NN problems and is used here, which can be defined as follows

$$\sqrt{\Sigma_{i=1}^{n}(a_i - b_i)^2} \quad (4)$$

In order to obtain the estimations of the right-censored response observations, an algorithm is developed for *k*NN imputation, given in the following Table 1.

TABLE 1: Algorithm for *k*NN imputation method.

Algorithm *1*: *k*NN imputation for right censored dara
Input: Right-censored dataset y_i
Censoring indicator δ_i
Number of nearest neigbours k
Values of explanatory variable z_i
Output:Imputed dataset $\mathbf{y}^{knn} = (y_1^{knn},..., y_n^{knn})^T$
1 **begin**
2 **for** (i = 1 to n) **do**
3 **if** (δ_i = 0) **do** (if data point is censored)
4 **for**(j = 1 to n) **do**
5 Find the Euclidean distances (4) between z_j and z_i for each censored data point
6 sort the distances from small to large
7 **for** (j = 1 to k) **do**
8 Take the first $uncensored$ k values of y_i associated to sorted distances
9 Calculate the ith imputed value (y_1^{knn}) with average of nearest k records of y_i
10Replace the imputed values (y_1^{knn}) with censored data points (y_i,δ_i=0) in censored data set $\mathbf{y}$=(y_1,...,y_n)
11 Return $\mathbf{y}^{knn}$=(y_1^{knn},..., y_n^{knn})^T
12 **end**

As stated in the Algorithm given above, it is important the choice of number of neighbours k. In this context, Cartwright et al.[17] recommended as k = 1 or 2. However, if k=1, the method will be highly sensitive to outliers. Since this study uses medical data, it is more appropriate to use a k value between the interval (see, Batista and Monard).[2,10,9] Now, imputed values can be used in smoothing spline method to estimate the nonparametric model.

SMOOTHING SPLINES

Smoothing spline method produces the estimations that fit perfectly to the data by using each data point as a knot. A roughness penalty term is added to the goodness of fit to avoid this perfect approach. For some constant $\lambda > 0$, smoothing spline estimates are obtained by minimizing the penalized least squares, given by

$$PRSS = \sum_{i=1}^{n} (y_i - g(z_i))^2 + \lambda \int (g''(z))^2 dz \quad (5)$$

where the first term in the right side of the equation (5) represents the goodness of fit and the second term denoted as $\int (g''(z))^2 dz$ shows the penalty term controlled by a smoothing parameter λ where function $g(z_i)$ is continuous on the interval $[a,b]$ and notation g'' means second derivative of the function. As mentioned before, the choice of the parameter λ is highly important and it is discussed in Section 2.3.

Let y_i^{knn} be the values of the response variable which is formed by kNN imputation method. In this case, the equation (5) can be rewritten as follows

$$PRSS = \sum_{i=1}^{n} (y_i^{knn} - g(z_i))^2 + \lambda \int (g''(z))^2 dz \quad (6)$$

In matrix and vector form, the equation (6) can be defined as

$$PRSS = (\mathbf{Y}^{knn} - \mathbf{Ng})^T (\mathbf{Y}^{knn} - \mathbf{Ng}) + \lambda \mathbf{g}^T \mathbf{Kg} \quad (7)$$

where $\mathbf{N}$ is a $n \times q$ incidence matrix with elements $N_{ij}=1$ if $z_i = m_j$ and $N_{ij}=0$ otherwise where $m_1,\ldots,m_q$ are the ordered distinct values of $z_i's$. $\mathbf{K}$ is a $q \times q$ positive definite symmetric penalty matrix, given by

$$\mathbf{K} = \mathbf{Q}^T \mathbf{R}^{-1} \mathbf{Q}$$

where $\mathbf{Q}$ and $\mathbf{R}$ are the $(q-2) \times q$ and $q-2 \times q-2$ dimensional symmetric and tri-diagonal matrices, respectively. Elements of these matrices can be calculated as

$Q_{jj} = 1/h_j$, $Q_{j,j+1} = -(1/h_j + 1/h_{j+1})$, $Q_{j,j+2} = 1/h_{j+1}$,

$R_{j-1,j} = R_{j,j-1} = h_j/6$, $R_{jj} = (h_j + h_{j+1})/3$,

After some mathematical calculations, the fitted values of function $\mathbf{g}$ that minimizes the equation (7) can be written as follows

$$\hat{\mathbf{g}}^{knn} = (\mathbf{N}'\mathbf{N} + \lambda \mathbf{K})^{-1} \mathbf{N}' \mathbf{Y}^{knn} \quad (8)$$

for a chosen parameter $\lambda > 0$. Also, if z_i's are ordered and distinct, then the incidence matrix can be considered as $\mathbf{N}=\mathbf{I}$, which is also stated as a special case of the smoothing spline method. Using equation (8), the mean vector for the model can be written as follows

$$\boldsymbol{\mu} = \hat{\mathbf{g}}^{knn} = (\mathbf{H}_\lambda^{knn} \mathbf{Y}^{knn}) = (\hat{\mathbf{Y}}^{knn} = E[Y|z]) \quad (9)$$

where $\mathbf{H}_\lambda^{knn} = \mathbf{N}(\mathbf{N}'\mathbf{N} + \lambda \mathbf{K})^{-1} \mathbf{N}'$ is a smoothing matrix. From here, the estimates of the variance of the error terms (σ_ε^2) can be obtained by residual sum of squares

$$RSS = \sum_{i=1}^{n} (y_i^{knn} - \hat{y}_i^{knn})^2 = (\mathbf{Y}^{knn} - \hat{\mathbf{Y}}^{knn})^T (\mathbf{Y}^{knn} - \hat{\mathbf{Y}}^{knn}) \quad (10)$$

Here, if we replace $\hat{\mathbf{Y}}^{knn}$ with $(\mathbf{H}_{\lambda}^{knn}\mathbf{Y}^{knn})$ then

$RSS = (\mathbf{Y}^{knn} - \mathbf{H}_{\lambda}^{knn}\mathbf{Y}^{knn})^T\ (\mathbf{Y}^{knn} - \mathbf{H}_{\lambda}^{knn}\mathbf{Y}^{knn}) = ||(\mathbf{I} - \mathbf{H}_{\lambda}^{knn})\mathbf{Y}^{knn}||^2$

Thus, the estimate of σ_{ε}^2 for kNN imputation method based on smoothing spline is given by

$$\hat{\sigma}_{\varepsilon}^2 = \frac{RSS}{tr(\mathbf{I} - \mathbf{H}_{\lambda}^{knn})^2} = \frac{||(\mathbf{I} - \mathbf{H}_{\lambda}^{knn})\mathbf{Y}^{knn}||^2}{tr[(\mathbf{I} - \mathbf{H}_{\lambda}^{knn})^T(\mathbf{I} - \mathbf{H}_{\lambda}^{knn})]} \tag{11}$$

where $tr(\mathbf{I} - \mathbf{H}_{\lambda}^{knn})^2$ denotes the degrees of freedom for a smoothing parameter λ and $tr(.)$ presents trace of a square matrix.

SELECTION OF THE SMOOTHING PARAMETER

In this study, choice of the smoothing parameter, which is of a crucial importance for the model estimation, is performed by three selection criteria such as GCV, BIC and RECP, as mentioned earlier. Calculations of the values obtained from the criteria are given in the Table 2.

In the calculation of $RECP(\lambda)$, $\hat{\mathbf{Y}}_p^{knn}$ and $\hat{\sigma}_p^2$ denote the pilot estimates of fitted values and the variance of the error terms for chosen a pilot smoothing parameter (λ_p), respectively. In practice, since the variance σ^2 is generally unknown, $\hat{\sigma}^2$ is used and it can be easily calculated by the equation (11). Note also that variance of error terms ($\hat{\sigma}_{\varepsilon}^2$) is called as a variance of the regression model. In order to compare the selection methods, two evaluation measurements are considered here and they are discussed in the following section.

EVALUATION MEASUREMENTS

To evaluate the performances of the estimated model based on different selection methods, mean squared error ($MSE(\hat{\mathbf{g}}^{knn})$) and relative efficiencies ($RE(\hat{\mathbf{g}}_{M1}^{knn}, \hat{\mathbf{g}}_{M2}^{knn})$) are used in this study. Definitions of these measurements are given, respectively, as.

$$MSE(\hat{\mathbf{g}}^{knn}) = E||\mathbf{Y}^{knn} - \hat{\mathbf{g}}^{knn}||^2 = ||(\mathbf{I} - \mathbf{H}_{\lambda}^{knn})\mathbf{Y}^{knn}||^2 = (\mathbf{g}^{knn})^T(\mathbf{I} - \mathbf{H}_{\lambda}^{knn})\mathbf{g}^{knn} + \sigma^2(\mathbf{I} - \mathbf{H}_{\lambda}^{knn})^2 \tag{12}$$

See the study of Aydın and Yılmaz (2018) for more detailed discussions. In this context, depending on equation (12), relative efficiency can be written as follows

$$RE(\hat{\mathbf{g}}_{M1}^{knn}, \hat{\mathbf{g}}_{M2}^{knn}) = \frac{MSE(\hat{\mathbf{g}}_{M1}^{knn})}{MSE(\hat{\mathbf{g}}_{M2}^{knn})} \tag{13}$$

where $\hat{\mathbf{g}}_{M1}^{knn}$ represents the estimated function based on first method and similarly $\hat{\mathbf{g}}_{M2}^{knn}$ for the second one. It can be said that if $RE(\hat{\mathbf{g}}_{M1}^{knn}, \hat{\mathbf{g}}_{M2}^{knn})<1$ then $\hat{\mathbf{g}}_{M2}^{knn}$ is more efficient than $\hat{\mathbf{g}}_{M1}^{knn}$. The measurements given in the equations (12-13) used in the simulation study and real data example to compare the selection methods.

This study is carried out in accordance with the principles of the Helsinki Declaration.

TABLE 2: Formulations for the selection criteria.

Criterion	Formula
$GCV(\lambda)$ (Craven and Wahba)[18]	$n^{-1}\|\|(\mathbf{I} - \mathbf{H}_{\lambda}^{knn})\mathbf{Y}^{knn}\|\|^2/[n^{-1}tr(\mathbf{I} - \mathbf{H}_{\lambda}^{knn})]^2$
$BIC(\lambda)$ (Schwarz)[19]	$n^{-1}\|\|(\mathbf{I} - \mathbf{H}_{\lambda}^{knn})\mathbf{Y}^{knn}\|\|^2 + (\log(n)/n)tr(\mathbf{H}_{\lambda}^{knn})$
$RECP(\lambda)$ (Lee)[20]	$n^{-1}\|\|(\mathbf{Y} - \hat{\mathbf{Y}}_p^{knn}\|\|^2 + \hat{\sigma}_p^2 tr(\mathbf{H}_{\lambda}^{knn})$

RESULTS

SIMULATION STUDY

In this section, a Monte-Carlo simulation study is carried out to assess the performances of the smoothing parameter selection criteria under the right-censored data. As mentioned above, the response observations which is formed by *k*NN imputation are estimated by smoothing spline method based on different criteria. In this respect, it can be said that different modified spline estimators are also tested for the right censored data.

We first generate the response variable (t_i) from the following nonparametric regression model

$$t_i=g(z_i)+\varepsilon_i\ ,\ i=1,2,\ldots,n \tag{14}$$

where the nonparametric function $g(.)$ is generated by $g(z_i)=e^{z_i/2}\cos(5^{z_i 2})$ where $z_i\sim U[0,1]$, and $\varepsilon_i\sim N(0,\sigma^2=0.5)$. To censor the observations of the response variable in the model (14), the censoring variable c_i is produced from normal distribution with some specific calculations, given by

$u_i\sim N(\mu_t,\sigma^2=0.5)$ and $c_i=u_i*\theta$

where θ determines the censoring level and μ_t is mean value of t_i's. It should be noted that θ=(1.2, 1.1, 1) corresponds to the censoring levels (C.L.) at (2%, 20%, 50%), respectively. Hence, the censored response variable y_i is formed by the equation (2). Also, for each C.L in simulation experiments, we generated 1000 random samples of size n=30, 100, and 250.

Before starting analysis, we first obtain the response variable (y_i^{knn}) which is formed by *k*NN method. Note that the obtained right-censored (y_i), uncensored (t_i) and imputed (y_i^{knn}) data points are provided visually in the following Figure 1 for some simulation configurations.

The numerical outcomes from the simulation study are summarized in the following Table 3 and Figure 2.

From Table 3, we see that large *MSE* values are obtained for high censoring levels. It is also noted that there is an obvious negative correlation between the sample sizes and the *MSE* values. As for selection criteria, it is clearly seen that *BIC* gives better results than the others for small and medium sample sizes and *GCV* has the best MSE scores for large samples. In this simulation study, *RECP* did not perform well but it should be noted that it still has closer scores to *BIC* and *GCV*. All criteria have almost same values, especially for large sample of size n=250.

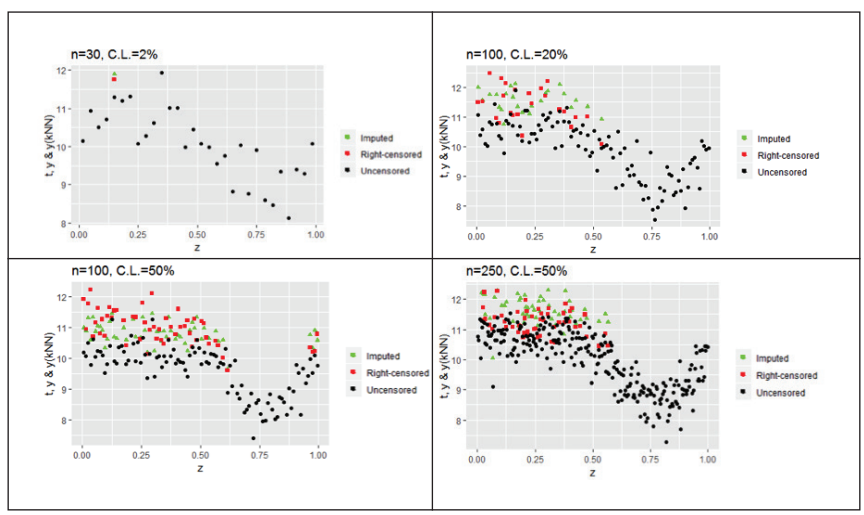

FIGURE 1: For a small, two medium and a large sample sizes with *C.L*=2%, 20%, 50%, and 50%, each panel shows the scatter plots of the simulated data points.

TABLE 3: The estimated MSE values and relative efficiencies from the smoothing spline estimators based on three different selection criteria for all simulation configurations.

n			30			100			250	
C.L.		2%	20%	50%	2%	20%	50%	2%	20%	50%
MSE	*GCV*	0.2902	0.3034	0.3893	0.2771	0.2811	**0.3260**	**0.2391**	**0.2671**	**0.2898**
	BIC	**0.2314**	**0.2233**	**0.3470**	**0.2749**	**0.2747**	**0.3260**	0.2394	0.2690	0.2917
	RECP	0.3614	0.3665	0.4803	0.3033	0.3110	0.3484	0.2541	0.2734	0.2962
Relative Eff.	*GCV*	1.3063	0.9624	0.8192	1.0157	0.9567	0.9197	**0.9959**	**0.9673**	**0.9776**
	BIC	**0.7667**	**0.7658**	**0.6659**	**0.9845**	**0.9531**	**0.9094**	1.004	0.9742	0.9842
	RECP	1.60	1.3146	1.2207	1.1177	1.0817	1.0875	1.0389	1.0391	1.0228

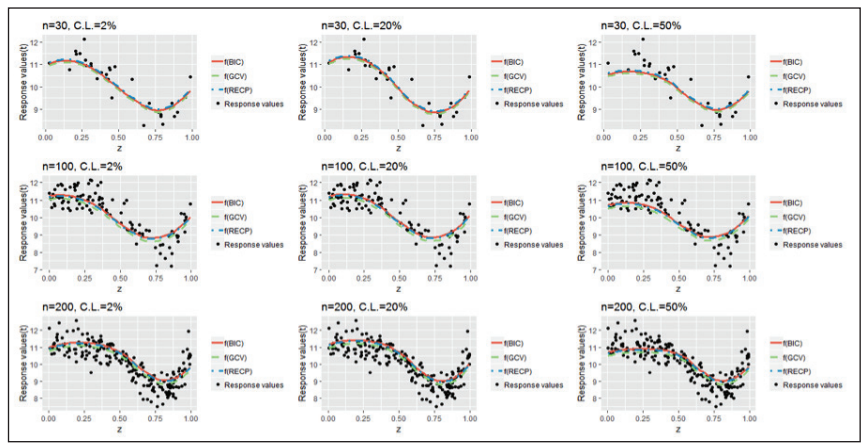

FIGURE 2: Each panel shows the response observations imputed by kNN, and the estimated curves from the smoothing spline estimators based on BIC, GCV, and RECP criteria using the data sets with different censoring levels for all simulation configurations.

The first row of the Figure 1 displays the results of the small sized samples (n=30) with three censoring levels ($C.L.$=2%, 20% and 50%). Similarly, the second row shows the results for n=100 and the third for n=250. As can be seen from Figure 3, all estimated curves give results that support the numerical outcomes given in Table 3. Note that as the sample sizes based on the same censoring levels get large, the spline estimates are get closer each other, as expected. When the estimated curves with high censoring rates are examined, it can be seen that these curves are forced to start from lower than others due to the effect of censorship. When the censoring level of 50% is really considered as heavy censored data, it is understood that *k*NN imputation can successfully overcomes this type of problematic data.

REAL DATA EXAMPLE

In this section, a kidney data set provided by McGilehrist and Aisbett is used for comparing the selection methods on real data.[21] The authors estimated the recurrence times of infection by Cox regression model with four predictor variables. Note that dataset is also supplied from *frailtyHL* package in *R* software. This dataset includes information about 76 kidney patients. In this study, only *frailty* (sensitivity to infection) is used as a nonparametric variable and recurrence time of infection (*rtime*) is used as a response variable

TABLE 4: MSE values and relative efficiencies from real data.

C.L.	GCV	BIC	RECP
MSE	1.2679	**1.2296**	1.3061
RE	1.0009	**0.9556**	1.0461

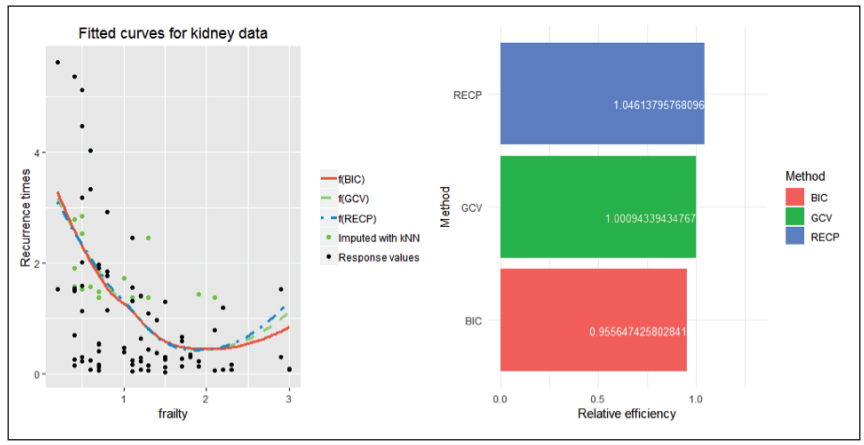

FIGURE 3: The imputed response values and their fitted curves obtained by smoothing spline using BIC, GCV and RECP. Also, the censored response values are marked by black points and imputed data observations are marked by green points in the left panel. The relative efficiencies of the selection criteria are displayed by bar plot in right panel.

for construction of the nonparametric regression model. This dataset also contains a censoring indicator (1= infection occurs; 0 = censored) associated with *rtime*. The number of censored observations is calculated as 18 from the censoring indicator variable, which means that the censoring rate in this data set is 23.68%. This rate indicates that kidney data is moderate censored.

The nonparametric model with kidney data is defined as

$$rtime_i = g(frailty_i) + \varepsilon_i, \quad i=1,\ldots,76 \tag{15}$$

Firstly, the imputed response values ($rtime_i^{knn}$) are provided by algorithm given in Table1. The estimation of nonparametric regression function in the model (15) can then be obtained by smoothing spline using different criteria. Accordingly, the outcomes from the real data set are given in Table 4 and Figure 3.

As can be seen from Table 4, it can be said that BIC criterion has good results in terms of MSE and relative efficiency. The *BIC* is followed by *GCV* and *RECP*, respectively. The same comments can be said for Figure 4. It should also be noted that it is important to say that the results obtained from simulation and real data studies support each other.

DISCUSSION

In this paper, right-censored data imputed by *k*NN method is modelled non-parametrically by smoothing spline method based on three different selection criteria. This study focuses on the selection of an optimum smoothing parameter for smoothing spline method under right censored data. Performances of the smoothing spline estimators based on three selection criteria are compared with the help of a simulation study and real-world data.

Finally, based on simulation and real data outcomes, the following conclusions can be noted:

• BIC method gives the best estimates for both simulation and real data studies, if right censored observations are completed with *k*NN imputation method.

• Although RECP does not show a good performance, GCV and RECP methods gives similar results. However, when they are inspected in terms of relative efficiencies, it can be seen that three criteria have satisfying results.

• When details of simulation study is examined, GCV has given better results than BIC for large sized samples. In this sense, GCV can be recommended as a good selection criterion for large sample sizes.

CONCLUSION

The results of this paper show that BIC is the most appropriate selection criterion for the choice of the smoothing parameter when the *k*NN imputation method is considered in order to cope with censored data.

Acknowledgment

This paper is supported by Scientific Research Project Office (BAP) in Mugla Sitki Kocman University with project number BAP-15/159.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Ersin Yılmaz, Dursun Aydın;* ***Design:*** *Ersin Yılmaz, Dursun Aydın;* ***Control/Supervision:*** *Dursun Aydın;* ***Data Collection and/or Processing:*** *Ersin Yılmaz, Dursun Aydın;* ***Analysis and/or Interpretation:*** *Ersin Yılmaz, Dursun Aydın;* ***Literature Review:*** *Ersin Yılmaz, Dursun Aydın;* ***Writing The Article:*** *Ersin Yılmaz, Dursun Aydın;* ***Critical Review:*** *Dursun Aydın;* ***Materials*** *Ersin Yılmaz, Dursun Aydın*

REFERENCES

1. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457-81. [Crossref]
2. Miller RG. Least squares regression with censored data. Biometrika. 1976;63(3):449-64. [Crossref]
3. Koul H, Susarla V, Van Ryzin J. Regression analysis with randomly right-censored data. Ann Stat. 1981;9(6)1276-88. [Crossref]
4. Stute W. Consistent estimation under random censorship when covariables are present. Journal of Multivariate Analysis (JMVA). 1993;45(1):89-103. [Crossref]
5. Orbe J, Ferreira E, Núñez-Antón V. Censored partial regression. Biostatistics. 2003;4(1):109-21. [Crossref] [PubMed]
6. Leurgans S. Linear models, random censoring and synthetic data. Biometrika. 1987;74(2):301-9. [Crossref]
7. Aydin D, Yilmaz E. Modified spline regression based on randomly right-censored data: a comparative study. Communications in Statistics-Simulation and Computation. 2018;47(9):2587-611. [Crossref]
8. Schafer JL. Analysis of Incomplete Multivariate Data. 1st ed. London: Chapman & Hall; 1997. p.448.
9. Batista G, Monard M. An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence. 2003;17(5-6):519-33. [Crossref]

10. Rubin DB, Van der Laan MJ. A general methodology for nonparametric regression with censored data, U.C. Berkeley Division of Biostatistics Working Paper Series. 2005. [Crossref]

11. Yenduri S, Iyengar SS. Performance evaluation of imputation methods for incomplete datasets. International Journal of Software Engineering and Knowledge Engineering (IJSEKE). 2007;17(1):127-52. [Crossref]

12. Andridge RR, Little RJ. A review of hot deck imputation for survey non-response. Int Stat Rev. 2010;78:40-64. [Crossref] [PubMed] [PMC]

13. Li D, Deogun J, Spaulding W, Shuart B. Towards missing data imputation: a study of Fuzzy K-means clustering method. International Conference on Rough Sets and Current Trends in Computing. 2004;573-9. [Crossref]

14. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520-5. [Crossref] [PubMed]

15. Malarvizhi R, Thanamani AS. Framework for missing value imputation. International Journal of Engineering Research and Development (IJERD). 2012;4(7):14-6.

16. Chen J, Shao J. Nearest Neighbor imputation for survey data. J Offic Stat. 2000;16(2):113-31.

17. Cartwright MH, Shepperd MJ, Song Q. Dealing with missing software project data. Proceedings of the 9th International Software Metrics Symposium. Sydney, Australia. 2003. p.154-65.

18. Craven P, Wahba G. Smoothing noisy data with spline functions. Num Math. 1979;31(4):377-403. [Crossref]

19. Schwarz G. Estimating the dimension of a model. Ann Statist. 1978;6(2):461-4. [Crossref]

20. Lee TCM. Smoothing parameter selection for smoothing splines: a simulation study. Computational Statistics & Data Analysis (CSDA). 2003;42(1-2):139-48. [Crossref]

21. McGilchrist CA, Aisbett CW. Regression with frailty in survival analysis. Biometrics. 1991;47(2):461-6. [Crossref] [PubMed]