Bayesian Influence Diagnostics in Radiocarbon Dating

Fernández-Ponce, J.M.

Dpto. Estadística e Investigación Operativa

Universidad de Sevilla

41012 Sevilla, Spain

ferpon@us.es

Palacios-Rodríguez, F. and Rodríguez-Griñolo, M.R.

Dpto. de Economía, Métodos Cuantitativos e Historia Económica

Universidad Pablo de Olavide

41013 Sevilla, Spain

mrrodgri@upo.es, fpalrod@upo.es

May 29, 2012

Abstract

Linear models constitute the primary statistical technique for any experimental science. A major topic in this area is the detection of influential subsets of data, that is, of observations that are influential in terms of their effect on the estimation of parameters in linear regression or of the total population parameters. Numerous studies exist on radiocarbon dating which propose a value consensus and remove possible outliers after the corresponding testing. An influence analysis for the value consensus from a Bayesian perspective is developed in this paper.

AMS Subject Classification 2010: 62J20, 62P25

Key words and phrases: Conditional Bias, Influence Analysis, Outliers, Predictive Approach, Radiocarbon Dating.

1 Introduction

Radiocarbon dating is vital in the establishment of time lines for many archaeological studies. The calibration curves necessary to map radiocarbon to calendar ages were originally estimated using only measurements on known age tree-rings. The types of records available for calibration have since diversified and a large group of scientists (known as the IntCal Working Group) from a wide range of backgrounds has come together to create internationally-agreed estimates of the calibration curves (for more details see Blackwell and Buck, 2008, and the references therein). The radiocarbon community has participated in a number of interlaboratory checks over the last thirty years (see Hedeyat *et al.*, 2008, for details about statistical scoring procedures to laboratory performance evaluation). The most ambitious project to date was launched by the Glasgow group and supported by over 50 radiocarbon laboratories. This three-stage study was completed and the results published in 1990 (Aitchison *et al.*, 1990; Cook *et al.*, 1990; Scott *et al.*, 1990). The latter two studies have highlighted difficulties in the comparability of ^{14}C laboratories, and have quantified excess variability in the results. The Fifth International Radiocarbon Intercomparison (VIRI) continued the tradition of the TIRI (third) and FIRI (fourth) intercomparisons (Scott, 2003) as a ^{14}C community project, with samples provided by participants and a substantial participation rate. Scott *et al.* (2010) gave the final results of the VIRI where some outliers were detected and consequently omitted from the sample in order to compute the consensus value. In this paper, an influence analysis on a finite population from a Bayesian perspective is developed to contribute further information to the study of Scott *et al.* (2010). This influence analysis is based on the influence analysis in linear models from a frequentist viewpoint (see Muñoz-Pichardo *et al.*, 2000).

Linear models constitute the primary statistical technique for any experimental science. A major topic in this area is the detection of influential subsets of data, that is, of observations that are influential in terms of their effect on the estimation of parameters in linear regression or of the total population parameters. The influence on a model is considered through the examination of the variation that results from perturbing the model formulation. From among these variations, case-deletion is the most popular method

for the indentification of influential observations due to the intuition and simplicity involved. For correlated data, Preisser and Qaquish (1996) propose deletion diagnostics via generalized estimating equations. Banerjee and Frees (1997) developed partial influence diagnostics for longitudinal data in linear mixed models, based on the omission of a subject. Langford and Lewis (1998) explore techniques in terms of deviances, leverages and residuals to handle outliers for multilevel data.

Johnson and Geisser (1983) consider the problem of influential observations using a Bayesian approach and derive methods both for the estimation of parameters and the prediction of future observations. The approach involves the comparison of the posterior (predictive) distribution of the parameters (future observables) with and without the set of observations whose influence is to be determined. Geisser (1985) presents some results in predictive and estimative influence functions, data consistency, and model checking. The objective is the detection of those observations that are most influential with regard to decision making and inference, either in the estimative or predictive mode or both. Bayesian case-influence statistics have also been developed, for example, Johnson and Geisser (1982, 1983, 1985) discuss predictive influence; the influence of cases on predictive distributions. Kass *et al.* (1989) use an asymptotic approach to assess influence on point estimates.

Chambers (1986) considers the problem of robust estimation of a finite population total given sample data containing representative outliers, that is, sample elements with a value that has been correctly recorded and with a large random error generated by the model under consideration. He identified sample outliers as sample elements with values that have been correctly recorded and that cannot be assumed to be unique. Chaloner and Brant (1988) develop an approach for the detection of outliers in a Bayesian linear model where an outlier is defined as an observation with a large random error, generated by the linear model under consideration. If the parameters of the model are unknown, the posterior distribution can be used in the calculation of the posterior probability that any observation is an outlier. Weiss (1996) discusses marginal influence assessment procedures and Weiss and Cho (1998) give formulae for case deletion influence diagnostics in normal linear regression for joint and marginal posterior distributions using several divergence measures.

Bayesian procedures in finite population sampling are covered in Bolfarine and Zacks (1992), Ghosh and Meeden (1997) and Mukhopadhyay (2001). Fernández-Ponce and Infante-Macías (2005) study the influence diagnostics in superpopulation models in a simple example, whereby formulae for case deletion influence diagnostics in prediction theory are given.

The purpose of this paper is to apply the concept of Bayesian Conditional Bias for real data of radiocarbon dating. The paper is organized as follows. First, in Section 2, the concept of the Bayesian Conditional Bias (BCB) is shown, by using the Conditional Bias given in Muñoz-Pichardo *et al.* (2000), as seen in Fernández-Ponce and Infante-Macías (2005). An expression for the BCB of a statistic T that does not depend on the parameter is proved and a sufficient and necessary condition to obtain the Bayes estimator of the conditional bias of a statistic T is given. Some results of influence analysis from a Bayesian viewpoint are shown in Section 3. The BCB for predictors of population quantities, such as the total, and for the unknown variance of the random errors under the model are obtained. In Section 4, the results of the previous sections are applied in a radiocarbon dating problem to study the influence on the consensus value estimated by Scott *et al.* (2010). Finally, conclusions and new ideas for future work are discussed in Section 5.

2 The Bayesian Conditional Bias

The concept of Conditional Bias is used by Muñoz-Pichardo *et al.* (2000) as a tool in the study of the variation in an analysis that results from perturbing the problem formulation, and is applied to the general linear model from a frequentist point of view. These authors propose several influence measures for the linear general model based on the concept of conditional bias. For simplicity, the univariate case is assumed, that is, let $Y_1, \cdots, Y_n$ be a random sample from a random variable Y, $T = T(Y_1, \cdots, Y_n)$ be a statistic defined on the sample, $y_1, \cdots, y_n$ be a realization of the sample, and let $I = \{i_1, \cdots, i_m\}$ be a collection of subindices. The conditional bias of T, given the set of observations indexed

by I, is defined as

$$S(y_I;T) = E\left(T|Y_{i_1} = y_{i_1}, \cdots, Y_{i_m} = y_{i_m}\right) - E(T)$$

where $y_I = \{y_{i_1}, \cdots, y_{i_m}\}$. Since $S(y_I;T)$ is the average deviation from $E(T)$, which provokes the set of observations indexed by I, it can be considered as a measure of the joint influence of y_I on T. This approach does not presuppose any particular hypotheses on the distribution of the variables, and its theoretical foundation does not need any perturbation pattern. Using influence measures based on the conditional bias, Muñoz-Pichardo *et al.* (2000) carry out an application to the multivariate analysis of covariance. In this paper, a Bayesian approach to the conditional bias based on the corresponding frequentist concept is given together with an application in radiocarbon dating.

A family of distribution functions $\{G_\theta, \theta \in \Theta\}$ is now considered, where Θ is a subset of $\mathbb{R}$. Let $Y(\theta)$ denote a random variable with distribution function G_θ. For any random variable θ with support in Θ, and with distribution function π, let Y denote a random variable with distribution function H given by

$$H(y) = \int_\Theta G_\theta(y) d\pi(\theta), \quad y \in \mathbb{R}.$$

Henceforth, $F^\pi(\theta|y_I)$ is denoted as the distribution function of $\theta \in \Theta \subseteq \mathbb{R}$, given the observations y_I, when the prior distribution of θ is $\pi(\cdot)$, and $\mathbb{E}(Y)$ is denoted as the expectation of the random variable Y, that is, $\mathbb{E}(Y) = \int_{\mathbb{R}} y dH(y)$. This expectation can be considered as the overall expectation for $Y(\theta)$. Note that this model can easily be generalized when $Y(\theta)$ is a multivariate random variable with finite dimension and when Θ is a subset of $\mathbb{R}^k$. Our aim is now to define the concept of Conditional Bias when prior information is given by a parameter either explicitly or implicitly and this information is used as part of the model.

It is now assumed that $Y_1, \cdots, Y_n$ is a random sample of the random variable $Y(\theta)$ where $\theta \sim \pi(\cdot)$ and $\theta \in \Theta \subseteq \mathbb{R}$. Let $T = T(Y_1, \cdots, Y_n)$ be a real-valued statistic defined on the sample, let $y_1, \cdots, y_n$ be a realization on the sample, and let $I = \{i_1, \cdots, i_m\}$ be a collection of subindices. The BCB of T, given the set of observations indexed by I, is defined as

$$SB(y_I;T) = \mathbb{E}(T|y_I) - \mathbb{E}(T). \tag{2.1}$$

The BCB of T is not necessarily the Bayes estimator of the conditional bias using the squared error-loss function. As can be seen in Fernández-Ponce and Infante-Macías (2005), the BCB of T is the Bayes estimator of the conditional bias using the squared error-loss function if, and only if the observations y_I are invariant observations with respect to the overall expectation of the statistic T.

By using the expression for the BCB of q-valued statistic $\mathbf{T}$ which is given in Fernández-Ponce and Infante-Macías (2005), it is obtained for $q = 1$ that

$$SB(y_I; T) = \mathbb{E}(T - T_{(I)}|y_I). \tag{2.2}$$

Consequently, we propose the following measure to quantify the Bayesian influence:

$$Q_I = |SB(y_I; T)|.$$

3 Influence analysis in the model

Let $\mathcal{U}$ denote a finite population which consists of N units labelled $1, 2, \cdots, N$. It will be assumed that these labels are known and that they can often contain certain information about the units. Attached to unit i, let y_i be the unknown value of certain characteristics of interest. Let $y = (y_1, \cdots, y_N)^t$ be the unknown state of nature or parameter. It is assumed that y belongs to $\mathcal{Y}$, a subset of n-dimensional Euclidean space $\mathbb{R}^N$. A subset s of $\{1, \cdots, N\}$ is called a sample. Let $n(s)$ denote the number of elements belonging to s, and unless otherwise stated, it will be assumed that it is n, and let $\mathcal{S}$ denote the set of all possible samples. A (non-sequential) sampling design is a function p defined on $\mathcal{S}$ whereby $p(s) \in [0, 1]$ for every nonempty $s \in \mathcal{S}$, and $\sum_{s \in \mathcal{S}} p(s) = 1$. In many problems in finite population sampling there are additional known characteristics or variables associated with each unit. Let $\mathbf{X}$ denote the collection of these vectors for the entire population. The superpopulation or Bayesian framework for inference on the quantities of interest of the finite population can be provided by the knowledge of some random process that could generate the values of the characteristics associated with each unit of the population. A purely Bayesian model for a fixed finite population assumes a specific (prior) distribution of y and can be considered as a superpopulation model with a

single element of a specified parametric family F. Thus, in general, let $F = \{F_\psi; \psi \in \Psi\}$, where ψ is a parameter in a specified parametric space Ψ. The Bayes Superpolutation model imposes a prior distribution of $\psi, \zeta(\psi|\phi)$,on the family F where ϕ is a known parameter. By considering the model

$$\left.\begin{array}{c} y = \mathbf{X}\beta + \varepsilon \\ \varepsilon \sim N(0, \mathbf{V}) \end{array}\right\} \tag{3.1}$$

denoted as $\psi(\beta, \mathbf{V})$, where $\mathbf{X} = (x_{kj}, k = 1, \cdots, N, j = 1, \cdots, p), x_{kj}$ is the value of the auxiliary variable x_j on unit k, $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_N)^t$, $\beta = (\beta_1, \cdots, \beta_p)^t$ is a $p \times 1$ vector of unknown regression coefficients, and $\mathbf{V} = \sigma^2\mathbf{W}$ with unknown σ^2, and $\mathbf{W}$ is a known diagonal matrix of dimension N. By considering a non-informative prior distribution on (β, σ^2) :

$$\zeta(\beta, \sigma^2) \propto \frac{1}{\sigma^2} \quad \text{and } E(\beta) = \mathbf{b} \text{ is a parametric vector.} \tag{3.2}$$

The model $\psi(\beta, \mathbf{V})$ in (3.1) together with the prior (3.2) is henceforth denoted as ψ_R.

The BCB as an influence measure is now studied via the superpopulation model. Likewise, it is assumed that the sampling design is non-informative. That is, if $p(s)$ is independent of the model ψ_R, where $s \in \mathcal{S}$. Accordingly, if the sampling design is non-informative, the conditionality principle (Basu, 1975) implies that the influence on ψ should be based only on the observed part of y, y_s, and the model ψ which is the link between y_s and y_r. After the sample s is selected, the partitions of y, $\mathbf{X}$ and $\mathbf{V}$ are as follows

$$y = \begin{pmatrix} y_s \\ y_r \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} \mathbf{V}_s & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_r \end{pmatrix}.$$

Prediction of linear quantities $g_L(y) = \mathbf{l}^t y$ where $\mathbf{l} = (\mathbf{l}_s^t, \mathbf{l}_r^t)^t$ is a vector of constants is now considered. Moreover, it is assumed that some specific information about the unknown values y_r is obtained. This information will be noted as a random event A which belongs to the sigma-algebra generated by y_r, $A \in \sigma(y_r)$. It is known (see Theorem 3.3.3 in Mukhophadyay, 2001) that

$$\widehat{g}_{BL} = \mathbf{l}_s^t y_s + \mathbf{l}_r^t \mathbb{E}\left(y_r | y_s, A\right) \tag{3.3}$$

is the Bayes predictor under the squared error-loss, and under any ψ_R model for which $\mathbb{V}ar(y_r|y_s, A)$ exists. Given a collection of subindices $I = \{i_1, \cdots, i_m\} \subset s \subset \{1, \cdots, N\}$, the m-vector composed of the components of y_s subindicated by I is now denoted by y_I. Similarly, the matrix formed by the rows of $\mathbf{X}$ corresponding to the collection I is denoted by $\mathbf{X}_I$. Likewise, the omission of the observations indexed by I in the sample is indicated by the subindex $s - I$. Additionally, the predictive sample for the sample subset y_I, denoted by $y^*_{s,I}$, is defined as the vector of dimension $s \times 1$:

$$y^*_{s,I} = \begin{pmatrix} y_I \\ \mathbb{E}(y_{s-I}|y_I) \end{pmatrix}.$$

Following Bolfarine and Zacks (1992) (see eq(1.3.14), p. 16), it is obtained that

$$\widehat{\beta}_I = (\mathbf{X}_I^t \mathbf{V}_I^{-1} \mathbf{X}_I)^{-1} \mathbf{X}_I^t \mathbf{V}_I^{-1} y_I$$

is the least-squares estimator of β based on $\mathbf{X}_I$ and y_I, for the case where the matrix $(\mathbf{X}_I^t \mathbf{V}_I^{-1} \mathbf{X}_I)^{-1}$ exists. The following establishes the BCB of $\widehat{g}_{BL}$ given the observations y_I when the matrix $(\mathbf{X}_s^t \mathbf{V}_s^{-1} \mathbf{X}_s)^{-1}$ exists. Consequently, (3.3) can be expressed as

$$\widehat{g}_{BL} = \mathbf{l}_I^t y_I + \mathbf{l}_{s-I}^t y_{s-I} + \mathbf{l}_r^t \mathbb{E}(y_r|y_I, y_{s-I}, A).$$

Thus,

$$\mathbb{E}\left(\widehat{g}_{BL}|y_I, A\right) = \mathbf{l}_s^t y^*_{s,I} + \mathbf{l}_r^t \mathbb{E}\left[\mathbb{E}(y_r|y_s, A)|y_I\right].$$

Prediction of linear quantities $g_L(y) = \mathbf{l}^t \mathbf{y}$ is now considered, where $\mathbf{l} = (\mathbf{l}_s^t, \mathbf{l}_r^t)^t$ is a vector of constants. It is assumed that $A = [\mathbf{y}_r > \mathbf{c}]$, where $\mathbf{c}$ is a vector of known constants, and consequently

$$\widehat{g}_{BL} = \mathbf{l}_s^t \mathbf{y}_s + \mathbf{l}_r^t \mathbb{E}\left(\mathbf{y}_r|\mathbf{y}_s, \mathbf{y}_r > \mathbf{c}\right). \tag{3.4}$$

Hence,

$$\widehat{g}_{BL} = \mathbf{l}_s^t \mathbf{y}_s + \mathbf{l}_r^t \mathbf{z}_r \tag{3.5}$$

where $\mathbf{z}_r$ is a vector whose $i - th$ component is

$$\mathbf{z}_{r,i} = \frac{\int_{c_i}^{+\infty} x f(x) dx}{P(y_{r,i} > c_i)} \tag{3.6}$$

where $f(x)$ is the corresponding normal density function under the model and by taking into account that it is conditioned to $\mathbf{y}_s$. Consequently, an estimator of the Bayesian Conditional Bias of g by using (2.2) is

$$\widehat{SB}(\mathbf{y}_I; g) = \hat{g}_{BL} - \hat{g}_{BL(I)}$$

where $\widehat{g}_{BL(I)}$ is the linear predictor through omission of the subsample I of s, and the corresponding influence measure is

$$Q_I = |\widehat{SB}(\mathbf{y}_I; g)|. \tag{3.7}$$

In this case, the Bayesian Risk (BR) has the following expression:

$$BR = \mathbf{l}_r^t Var\left[\mathbf{y}_r | \mathbf{y}_s, \mathbf{y}_r > \mathbf{c}\right] \mathbf{l}_r. \tag{3.8}$$

Furthermore, $Var\left[\mathbf{y}_r | \mathbf{y}_s, \mathbf{y}_r > \mathbf{c}\right]$ is an $r \times r$ matrix, henceforth denoted by S, since the s_{ij} element is

$$s_{ij} = E\left[y_{r,i} y_{r,j} | \mathbf{y}_s, \mathbf{y}_r > \mathbf{c}\right] - E\left[y_{r,i} | \mathbf{y}_s, \mathbf{y}_r > \mathbf{c}\right] E\left[y_{r,j} | \mathbf{y}_s, \mathbf{y}_r > \mathbf{c}\right].$$

Specifically,

$$E\left[y_{r,i} y_{r,j} | \mathbf{y}_s, \mathbf{y}_r > \mathbf{c}\right] = \frac{\int_{c_i}^{\infty} \int_{c_j}^{\infty} z t f(z,t) dz dt}{P(y_{r,i} > c_i, y_{r,j} > c_j)} \tag{3.9}$$

where $f(z,t)$ is the density function of the $y_{r,i}, y_{r,j} | \mathbf{y}_s$ vector which can easily be obtained from the normal conditional densities by taking into account the value of the corresponding covariances. This BR depends on the unknown covariance-variance matrix in (3.8) and is estimated by using the corresponding approximation for the integral and the probability which appear in (3.9). These computations have been developed by using the **R** language for statistical computing, version 2.14.0. The `adapt` and `mtvnorm` packages have been used.

By using Theorem 3.1.1 from Bolfarine and Zacks (1992) and by incorporating the fact that the non-informative prior of β is obtained as the limit of $N(\mathbf{b}, \mathbf{B})$, when $\mathbf{B}^{-1} \to 0$ and $A = \mathbb{R}^r$, it is easy to show that

$$\mathbb{E}(y_{s-I} | y_I, A) = \mathbf{X}_{s-I} \widehat{\beta}_I; \quad \mathbb{E}(y_r | y_s, A) = \mathbf{X}_r \widehat{\beta}_s; \quad \mathbb{E}(y_r | y_{s-I}, A) = \mathbf{X}_r \widehat{\beta}_{s-I}$$

$$\text{and } \mathbb{E}\left(\hat{g}_{BL}|A\right) = \mathbf{l}^t\mathbf{X}\mathbf{b}.$$

Thus, by using (2.1),

$$SB(y_I;\widehat{g}_{BL}) = \mathbf{l}_I^t y_I + \mathbf{l}_{s-I}^t\mathbf{X}_{s-I}\widehat{\beta}_I + \mathbf{l}_r^t\mathbf{X}_r(\mathbf{X}_s^t\mathbf{V}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s^t\mathbf{V}_s^{-1}y_{s,I}^* - \mathbf{l}^t\mathbf{X}\mathbf{b}. \tag{3.10}$$

Consequently, this BCB depends on $\mathbf{b}$ and it should be estimated by using (3.7). Note that, under the ψ_R-model, sample subset sizes can exist whose Bayesian Influence cannot be computed. This fact depends on the existence of the inverse of the $(\mathbf{X}_I^t\mathbf{V}_I^{-1}\mathbf{X}_I)$ and $(\mathbf{X}_s^t\mathbf{V}_s^{-1}\mathbf{X}_s)$ matrices. Furthermore, the BR is given by

$$\mathbb{E}\left[\widehat{g}_{BL} - g\right]^2 = \sigma^2\left[\mathbf{l}_r^t\mathbf{W}_r\mathbf{l}_r + \mathbf{l}_r^t\mathbf{X}_r(\mathbf{X}_s^t\mathbf{W}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_r^t\mathbf{l}_r\right] \tag{3.11}$$

where $\mathbf{V}_s = \sigma^2\mathbf{W}_s$ and $\mathbf{V}_r = \sigma^2\mathbf{W}_r$. It is known (see Bolfarine and Zacks, 1992, pg. 62) that an unbiased estimator of the prediction variance is obtained by replacing $\mathbf{V}$ with $\widehat{\sigma}^2\mathbf{W}$ where

$$\widehat{\sigma}^2 = \frac{1}{n-p}\left(\mathbf{y}_s - \mathbf{X}_s\widehat{\beta}_s\right)^t\mathbf{W}_s^{-1}\left(\mathbf{y}_s - \mathbf{X}_s\widehat{\beta}_s\right).$$

Thus, the estimated BR is expressed as

$$\widehat{BR} = \widehat{\sigma}^2\left[\mathbf{l}_r^t\mathbf{W}_r\mathbf{l}_r + \mathbf{l}_r^t\mathbf{X}_r(\mathbf{X}_s^t\mathbf{W}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_r^t\mathbf{l}_r\right].$$

Influence on the variance σ^2 can also be established and analyzed (see Fernández-Ponce and Infante-Macías, 2005).

4 An example in radiocarbon dating

The radiocarbon community has participated in a number of interlaboratory checks over the last thirty years. The latter two studies have highlighted difficulties in the comparability of ^{14}C laboratories, and have quantified excess variability in the results. Not all laboratories that had previously participated in Phase 1 participated in Phase 2, since bone is not a routinely measured sample in all laboratories. A total of 42 laboratories reported results on various bone samples (see Table 1 in Scott *et al.*, 2010), which are used in this paper, are about different bone samples. The sample called E in Scott *et al.* (2010) is of mammoth bone from a site called Quartz Creek, Dawson City, Yukon

Territory. The bone is a portion of the pelvis of a *Mammuths* sp. specimen (for more details see Scott *et al.*, 2010). In our study, Sample E is analyzed in a different way since it has five *censored* values. These missing values are predicted by using the information of the data of the same laboratory.

Scott *et al.* (2010) gave a report for the consensus values for the samples following the procedure in Scott (2003). The exculsion of individual results from the final calculation was based on two criteria: 1) the absolute value; 2) the size of the quoted error. The consensus values were calculated as a weighted average of the remaining results. In this paper, an influence analysis from a Bayesian perspective is developed. This analysis permits a measure for the outliers to be established by quantifying the degree of effect of their omission on the final calculation of the consensus value. It is shown that the consensus value obtained is smaller than that proposed in Scott *et al.* (2010). This circumstance is not strange in this kind of analysis (for a similar case see Xu *et al.*, 2010).

In an initial test of the sample of mammoth bones, 0.58 g of collagen was recovered from 5 g of bone. The percentage of carbon of this collagen sample was 41%. For this sample, a small number of laboratories reported this figure as a minimum age. In this paper, our purpose is to predict the ages by using the information of corresponding laboratories and study the degree of influence on the consensus values when known data is omitted. Given that the quoted error for the *bounded* data is unknown, it is estimated by the median of quoted error for the corresponding dating technique. Bear in mind that the value of the quoted errors for the sample data might influence the final consensus value. All computations are made by using **R** software for statistical computing and graphics (v. 2.14.0). The consensus value is given by $g(y) = l^t y$ where

$$l_i = \frac{1/e_i^2}{\sum_{i=1}^{n} 1/e_i^2}$$

and e_i is the quoted error which can be seen in Table 2a (the fourth column) in Scott *et al.* (2010). The influence measure (3.7) of g is plotted in Figure 1. As it can be seen, there exists a data with the highest influence. This data corresponds to an outlier which was detected in Scott *et al.* (2010) (the 24^{th} item of data in Table 2a in Scott *et al.*, 2010) which corresponds to LSC with an estimation of 22810 yr BP. The effect of the

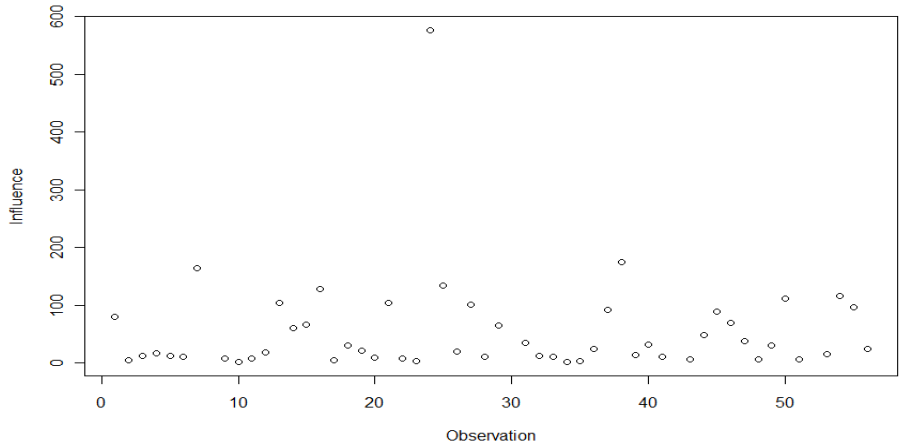

Figure 1: Influence on the consensus value in the Mammoth bone sample

omission of this data on the total quantity g is estimated at approximately 577 years. The value of the remaining influences can be seen in Table 1. The estimation of the BR is denoted as $\widehat{BR}$.

Table 1: Influence on the consensus value.

Observation number	Influence Measure (years)	$\widehat{g}_{(i)}$	$sqrt(\widehat{BR})$
24	577.36	36484.08	45.02
38	175.07	36081.79	51.55
7	163.52	36070.23	47.19
25	134.23	35772.49	48.03
16	128.75	35777.97	46.92
54	116.62	35790.10	47.42
50	112.20	36018.92	47.20
13	103.51	35803.21	46.50
21	103.24	35803.48	46.29
27	100.37	36007.09	45.43
55	95.78	36002.50	44.73
37	92.25	35998.97	51.69

Table 2: Prediction for the censored data.

Observation number	Censored data	Predicted value
8	> 41000	43992.55
30	> 42000	44375.6
52	> 35300	37378
57	> 25400	30470.5
58	> 37200	41263.28

The first column in Table 1 is the same as that in Table 2a of Scott *et al.* (2010). Note that the censored data has been removed in order to compute the influence of the data since this data is predicted by using the information of remaining the data and the corresponding model developed in this paper. Consequently, there are 55 items of data for the mammoth bones. The greatest twelve influences are represented in the second column of Table 1 in decreasing order. It is also interesting to note that there are three outliers among these twelve data; the 24^{th} item of data from LSC with a value equal to 22810; the 7^{th} observation from LSC with a value equal to 25530; and the 27^{th} item of data from LSC with a value equal to 24300. Two outliers which were detected by Scott *et al.* (2010), (the 29^{th} observation from LSC with a value equal to 26550 and the 55^{th} observation from LSC with a value equal to 21684), have an influence on the consensus value of less than 100 years apiece. Thus, these outliers are not considered as an influence on the consensus value. The estimation of the consensus value by omitting the ith corresponding observation $(\widehat{g}_{(i)})$ and the square root of the estimated BR is also added in Table 1. By using the equation (3.6), the prediction for the missing values in sample E are given in Table 2. The same analysis is now carried out but with the 24^{th} element removed from the data set since it is the data with the greatest influence. The influence measure (3.7) of g, by removing the element with greatest influence, is plotted in Figure 2. Note that the corresponding case for the data is represented along the OX axis. In this case, it can be said that there are no elements with a significantly high influence (see Table 3).

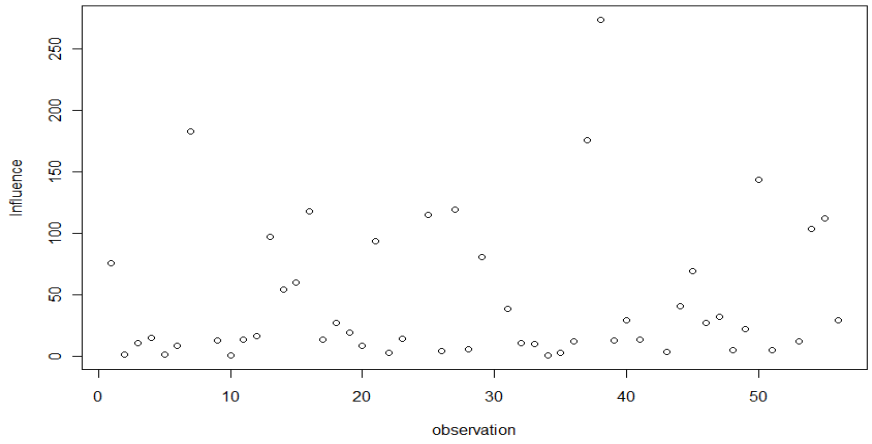

Figure 2: Influence on the consensus value in the Mammoth bone sample

Table 3: Influence on the consensus value on removing the 24^{th} observation.

Observation	Influence Measure (years)	$\widehat{g}_{(i)}$	$sqrt(\widehat{BR})$
38	273.58	36757.66	50.34
7	182.92	36667.00	45.07
37	175.47	36659.55	50.61
50	143.30	36627.38	45.93
27	119.09	36603.17	46.51
16	118.13	36365.95	45.72
25	114.71	36369.37	46.92
55	112.35	36596.43	44.71
54	103.36	36380.72	46.27
13	97.28	36386.80	45.28
21	93.40	36390.68	45.32

Table 4: Influence on the consensus value.

Observation	Influence Measure (years)	$\widehat{g}_{(I)}$	$sqrt(\widehat{BR})$
-	-	35906.72	46.06
24	577.36	36484.08	45.02
24, 38	850.94	36757.66	50.3
24, 38, 7	1068.13	36974.85	50.35
24, 38, 7, 25	958.17	36864.89	52.94
24, 38, 7, 25 ,16	823.82	36730.54	53.99
24, 38, 7, 25, 16, 54	698.96	36605.68	55.74

No rules exist which establish the best consensus value from solely observing Table 3, since influence analysis is not concerned with what data has to be removed. The consensus value and the BR (see Table 4) when data of great influence is sequentially removed are also shown, following the decreasing order in Table 1. Thus, by analyzing Table 4, the proposed consensus value in this paper is 36975 ± 50 yr BP for the mammoth bone sample. That is, the $24^{th}, 38^{th}$ and 7^{th} observation have been removed from the sample in order to propose a consensus value since these 3 observations have the greatest influence. Table 5 lists the summary statistics for the complete sample E, (including the size, mean, median, IQR and range) for ^{14}C and for various laboratories. Table 6 lists the same statistics but omitts the outlier cases of the 24^{th} and 7^{th} observations.

Table 5: Summary statistics for complete sample E.

	n	Mean	Median	ST dev	Q1	Q3	Min	Max
AMS	40	38950	39870	2654.26	36740	40730	33020	43990
GPC	6	38360	37820	4494.64	35580	41530	32570	44380
LSC	11	30190	30470	6206.02	24920	35420	21680	38350
Overall	57	37200	38350	5057.39	35500	40450	21680	44380

Table 6: Summary statistics for sample E with the omission of two outliers.

	n	Mean	Median	ST dev	Q1	Q3	Min	Max
AMS	40	38920	39870	2621.45	36740	40710	33020	43510
GPC	6	38300	37820	4392.39	35580	41530	32570	43990
LSC	9	31850	34150	6117.39	26550	35500	21680	38350
Overall	55	37700	38630	4376.05	35620	40470	21680	43990

Similar results are obtained if the bounded data is assumed with unknown values. In this case, the equations (3.5) and (3.7) are used to estimate the Bayesian Influence and (3.11) to estimate the BR. It is concluded that the three most influencial observations are the 24^{th} and 7^{th}, which are both outliers in Scott's analysis (see Scott *et al.*, 2010), and the 38^{th} item of data, in that order. Table 7 lists the summary statistics for the complete sample E (including the size, mean, median, IQR and range) for ^{14}C and for various laboratories. Table 8 lists the same statistics but omitts the outlier cases of the 24^{th} and 7^{th} observations.

Table 7: Summary statistics for complete sample E.

	n	Mean	Median	ST dev	Q1	Q3	Min	Max
AMS	40	38660	39470	2530.17	36720	40460	33020	42500
GPC	6	37330	36930	3420.69	35580	39510	32570	42060
LSC	11	29190	28420	5732.31	24920	34740	21680	38350
Overall	57	36690	37960	5024.20	35340	40320	21680	42500

Table 8: Summary statistics for sample E with the omission of two outliers.

	n	Mean	Median	ST dev	Q1	Q3	Min	Max
AMS	40	38660	39470	2530.17	36720	40460	33020	42500
GPC	6	37330	36930	3420.70	35580	39510	32570	42060
LSC	9	31160	32270	5673.01	26550	35340	21680	38350
Overall	55	37290	38180	4258.02	35520	40320	21680	42500

Table 9: Influence on the consensus value without *bounded* information.

Observation	Influence Measure (years)	$\widehat{g}_{(I)}$	$sqrt(\widehat{BR})$
-	-	35699.59	96.16
24	608.59	36308.18	89.05
24, 38	864.4	36564.09	99.03
24, 38, 7	1094.72	36794.31	96.74
24, 38, 7, 25	974.83	36674.42	101.26
24, 38, 7, 25, 16	833.74	36533.33	100.77
24, 38, 7, 25, 16, 54	702.21	36401.8	105.69

Obviously, descriptive measures with the bounded data (see Tables 5 and 6) are higher than the corresponding descriptive measures without this bounded information (see Tables 7 and 8). Table 9 lists influence measures when is the data items are omitted in a decreasing order following a similar process to that of Table 4. As can be seen, the same influence data is obtained. In this case, the proposed consensus value is 36794 ± 97 yr BP for the mammoth bone sample. Note that the difference between the above case and is not significant although the resulting consensus value is approximately 2500 yrs less than that proposed in Scott *et al.* (2010).

5 Conclusions

The preliminary analysis of results from Phase 2 of VIRI (see Scott *et al.*, 2010) highlighted the general and broad agreement amongst laboratories but also underlined the persistent problem presented by outlying data values from a relatively small number of laboratories. In this paper, an influence analysis on the consensus value is added to the analysis developed in Scott *et al.* (2010). As can be seen in this paper, some previous detected outliers are now viewed as influence data on the consensus value. Not only must this influence have to be interpreted as the effect of omitting each value or set of values on the consensus value but also on the predicted value for certain missing values given which should have been supplied by the laboratories. Likewise, this influence analysis

is based on a non-informative a priori distribution for the parameters in the model. It would be interesting to study the effect on the prediction values and the consensus value by taking into account a variety of a priori distributions in accordance with the a priori information. It is interesting to note that the Bayesian risk is not great since there are only five values to predict in our sample. Additionally, the forward search and our technique can be used jointly as a forward deletion formulae. However, these topics are beyond the scope of the current study and will be studied in future work.

References

[1] Aitchison, T.C., Scott, E.M., Harkness, D.D., Baxter, M.S. and Cook, G.T. (1990), Report on Stage 3 of the International Collaborative Program. *In* Scott, E.M., Long, A. and Kra, R.S., eds., Proceedings of the International Workshop on Intercomparison of ^{14}C Laboratories. *Radiocarbon,* **32**, 3: 271–278.

[2] Banerjee, M. and Frees, E. (1997), Influence diagnostics for longitudinal models, *Journal of the American Statistical Association,* **92**, 999–1005.

[3] Basu, D. (1975), Statistical Information and likelihood, *Sankhya, A,* **37**, 1–71.

[4] Blackwell, P.G. and Buck, C.E.(2008), Estimating radiocarbon calibration curves, *Bayesian Analysis* **3**,2 225–248.

[5] Bolfarine, H. and Zacks, S. (1992), *Prediction Theory for Finite Populations,* Springer-Verlag.

[6] Cook, G.T., Harkness, D.D., Miller, B.F., Scott, E.M., Baxter, M.S. and Aitchison, T.C. (1990), International Collaborative Study: Structuring and sample preparation. *In* Scott, E.M., Long, A. and Kra, R.S., eds., Proceedings of the International Workshop on Intercomparison of ^{14}C Laboratories. *Radiocarbon,* **32**, 3: 267–270.

[7] Chambers, R.L. (1986), Outlier Robust Finite Population Estimation, *Journal of the American Statistical Association,* **81**, No. 396, 1063–1069.

[8] Chaloner, K. and Brant, R. (1988), A Bayesian approach to outlier detection and residual analysis, *Biometrika,* **75**, 4 , 651–659.

[9] Fernández-Ponce, J.M. and Infante-Macías, R. (2005), A new approach to influence diagnostics in superpopulations, *Environmetrics,* **16**, 327–338.

[10] Geisser, S. (1985), On the prediction of observables: A selective update, *Bayesian Statistics,* **2**, 203–230.

[11] Ghosh, M. and Meeden, G. (1997), *Bayesian Methods for Finite Population Sampling,* Chapman and Hall, London.

[12] Hedayat, A.S., Guoqin, Su. and Streets, W.E. (2008), Statistical scoring procedures applicable to laboratory performance evaluation, *Journal of Statistical Planning and Inference,* **138**, 3336–3349.

[13] Johnson, W. and Geisser, S. (1982), *Assessing the predictive influence of observations. Statistics and Probability: Essays in Honor of C.R. Rao,* North-Holland, Amsterdam, 343–358.

[14] Johnson, W. and Geisser, S. (1983), A predictive view of the detection and characterization of influential observations in regression analysis, *Journal of the American Statistical Association,* **78**, 137–144.

[15] Johnson, W. and Geisser, S. (1985), Estimative influence measures for the multivariate general linear model, *Journal of Statistical Planning and Inference,* **11**, 33-56.

[16] Kass, R.E., Tierney, L. Kadame, J. (1989), Approximate methods for assessing influence and sensitivity in Bayesian analysis, *Biometrika,* **76**, 663–674.

[17] Langford, I.H. and Lewis, T. (1998), Outliers in multilevel data, *Journal of the Royal Statistical Society, Series A,* **161**, 121–160.

[18] Mukhopadhyay, P. (2001). *Topics in survey sampling.* Springer-Verlag, New York, Inc.

[19] Muñoz-Pichardo, J.M., Muñoz-García, J., Fernández-Ponce, J.M. and Jiménez-Gamero, M.D. (2000), Influence analysis in multivariate linear general models, *Communications in Statistics, Part B: Simulation and Computation* **29**(3), 529–547.

[20] Preisser, J.S. and Qaquish, B.F. (1996), Deletion diagnostics for generalised estimated equations, *Biometrika,* **83**, 551–562.

[21] Robert, C.P. (2001), *The Bayesian Choice,* Springer-Verlag, New York Inc.

[22] Scott, E.M., Aitchison, T.C., Harkness, D.D., Cook, G.T., and Baxter, M.S.(1990), An overview of all three stages of the international radiocarbon intercomparison. *Radiocarbon,* **32**, 3: 309–319.

[23] Scott, E.M.(2003), The Third International Radiocarbon Intercomparison (TIRI) and the Fourth International Radiocarbon Intercomparison (FIRI). *Radiocarbon,* **52**, 2-3: 846–858.

[24] Scott, E.M., Cook, G.T., and Naysmith, P.(2010), A report on Phase 2 of the Fifth International Radiocarbon Intercomparison (VIRI). *Radiocarbon,* **45**, 2: 135–408.

[25] Weiss, R.E. (1996), An approach to Bayesian sensitivity analysis, *Journal of the Royal Statistical Society, Series B,* **58**, 739–750.

[26] Weiss, R.E. and Cho, M. (1998), Bayesian marginal influence assessment, *Journal of Statistical Planning and Inference,* **71**, 163–177.

[27] Xu, X., Khosh, M.S., Druffel-Rodriguez, K.C., Trumbore, S.E. and Southon, J.R.(2010), Is the consensus value of Anu Sucrose (IAEA C-6) too high? *Radiocarbon,* **52**, 2-3: 866–874.