文章编号：2096-1472(2021)-11-11-04
DOI:10.19644/j.cnki.issn2096-1472.2021.011.003

缺失数据插补方法性能比较分析

徐鸿艳[1]，孙云山[2]，秦琦琳[1]，朱明涛[2]

(1.天津商业大学理学院，天津 300134；
2.天津商业大学信息工程学院，天津 300134)
✉2552727224@qq.com; sunyunshan@tjcu.edu.cn; 3099141857@qq.com; 648191948@qq.com

摘　要：数据缺失问题在现实工作生活中不可避免，为保证信息完整度以便于后续统计分析，尽可能准确地预测填补缺失值则显得尤为重要。基于两组分别服从于高斯分布和伽马分布的模拟数据集和一组非洲地区部分国家预期寿命实际数据，分别预设5%、10%和20%三种缺失比例，利用计算机软件对四种插补方法统计结果进行比较分析。试验结果表明，模拟数据中自回归建模插补和均值插补整体效果略优于最近邻插补和线性回归插补，实际数据中当缺失数据比例较低时，最近邻插补和线性回归插补效果优于前两者，当缺失比例较高时与模拟数据效果无明显差异。

关键词：缺失数据；插补方法；自回归建模
中图分类号：TP399　**文献标识码**：A

Comparative Analysis of the Performance of Interpolation Methods for Missing Data

XU Hongyan[1], SUN Yunshan[2], QIN Qilin[1], ZHU Mingtao[2]

(1.*School of Science*, *Tianjin University of Commerce*, *Tianjin* 300134, *China*;
2.*School of Information Engineering*, *Tianjin University of Commerce*, *Tianjin* 300134, *China*)
✉2552727224@qq.com; sunyunshan@tjcu.edu.cn; 3099141857@qq.com; 648191948@qq.com

Abstract: Data missing is inevitable. In order to ensure information integrity and follow-up statistical analysis, it is particularly important to predict and fill in missing values as accurately as possible. Based on two sets of simulated data sets that are subject to Gaussian distribution and Gamma distribution respectively, and a set of actual life expectancy data of some countries in Africa, three missing ratios of 5%, 10% and 20% are preset respectively, and the statistical results of the four interpolation methods are compared and analyzed by computer software. The experimental results show that the overall effect of auto-regression modeling interpolation and mean interpolation in simulated data is slightly better than that of K-nearest neighbor interpolation and linear regression interpolation. In actual data, when the proportion of missing data is low, K-nearest neighbor interpolation and linear regression is better than the former two, and there is no significant difference in the effect of the simulated data when the missing ratio is high.

Keywords: missing data; interpolation method; autoregressive

1 引言(Introduction)

数据作为一种形式多变的信息载体，如今广泛存在并应用于各行各业中，尤其是进入大数据时代以来，各类数据信息的完整、准确、充足与否与各行业的发展息息相关。然而由于某些主观和客观原因(如数据统计或录入过程中被遗漏，数据获取渠道未公开等)，不可避免地会存在一些数据缺失的情况[1]。另外，现行的统计方法并不能直接对缺失数据进行统计分析，因而如何处理缺失数据并使其提供最大完整度的信息就成了重中之重。随着数据缺失这一现实问题逐渐受到重视，国内外相关学者也对其进行了一系列研究。相对而言，国外学者起步更早，早有学者于20 世纪便提出了缺失数据的相关问题。在经过无数次试验后，学者们发现缺失数据难以避免，因此一系列插补方法应运而生，如加权法、冷热平台插补、回归插补和EM算法等[2]。而我国相关问题的研究起步于21 世纪初，较有代表性的为金勇进教授在其文章《缺失数据的插补调整》中提出的一系列插补方法，而后随着其他学

者的不断深入学习，除传统插补方法以外，一些机器学习方法(支持向量机、神经网络和决策树等[3])在缺失数据插补上也得到了广泛应用。

本文旨在通过模拟数据和实际数据对现今应用较为广泛的几种数据插补方法进行比较分析，第二部分主要对缺失数据产生的原因和本文中应用到的几种插补方法进行简要概述，第三部分基于模拟数据和实际数据进行实证分析，最后针对试验结果得出结论。

2 缺失数据概述(An overview of missing data)

2.1 缺失数据产生原因及分类

从数据的收集、整理、处理到应用，任何一个环节出现问题都可能会造成数据信息的遗失，我们根据各类数据的不同阶段将其缺失原因主要分为以下几种：(1)调查人员调查不足造成资料中的数据丢失；(2)统计人员在数据录入过程中出现失误，或因数据采集设备故障等原因引起数据缺失；(3)被调查人员的主观失误或有意隐瞒造成的数据缺失[4]；(4)历史原因造成的数据缺失；(5)未公开数据难以获取和其他原因造成的数据缺失，等等。

而关于缺失数据的分类，随着近年来缺失数据问题研究的不断发展和相关学者研究的不断深入，我们既可根据缺失机制将其分为随机缺失、完全随机缺失和非随机缺失[5]三类，也可根据数据的缺失模式将其分为单变量缺失模式、多变量缺失模式[6]、单调缺失模式和一般缺失模式四类，各类别的具体含义于其他相关文献中均有较为详细的介绍，我们这里不作赘述。

2.2 缺失数据处理方法

对于缺失数据的处理方式，较为简单的主要有不处理和直接删除法。前者主要包括直接在含空值的数据上进行数据挖掘处理的贝叶斯网络和人工神经网络[7]等，但这种处理方式大多认为补齐后的数据不一定符合客观事实，错误地填充会导致错误的结果，因此仍希望在保持原始信息不发生变化的情况下对信息系统进行处理[8]。后者操作简单，但假设条件较高，容易产生估计偏差，且简单删除存在缺失的数据会损失大量信息进而影响信息的客观性与结果的准确性，因此人们在对缺失数据进行预处理时，除缺失比例极小的情况外通常也不会采用此方法。因此，插补法相对来说目前更符合常理且易被各界学者及研究人员接受，即基于数理统计等原理对数据集中的缺失部分作填充处理，使得数据集变得完整以便开展下一步工作，其主要用于处理项目无回答而造成的数据缺失，可保证数据分析的基本样本量。目前应用较为广泛的插补法主要有特殊值替换(均值替代、众数中位数替代等)、多重插补、冷热卡插补、KNN、EM算法和各类机器学习插补法等。此外，张量补全法、随机森林算法、朴素贝叶斯等新型插补方法也在某些领域的缺失数据处理上有着一定的应用，本文将对以下几种插补方法结果做比较分析。

2.2.1 最近邻插补方法

最近邻插补方法(K−Nearest Neighbor，KNN)，整体思路较为简单，即缺失数据插补过程中，缺失位置数值根据其特征空间内相邻最近的K个观测值决定，根据距离远近决定分类归属，其主要不足为计算量较大，且缺失比例较大或缺失数据点大量连续时计算机运行难以得到预测结果，因此其较适用于类域存在交叉和重叠的待估计样本点分类问题。该分类算法主要分为以下四个步骤：

步骤1：导入全部已知观测数据与待估计数据；

步骤2：计算每个待估计样本点(缺失数据点)到其他已知观测值的距离D；

步骤3：对每个计算得出的距离D进行排序，并选出距离最小的K个点；

步骤4：对上述选出的K个所属类别进行比较后，将待估计样本点归入在K个已知观测点中占比最高的那类。

2.2.2 均值插补与线性回归插补

均值插补作为一种操作简便且快速的缺失数据处理方式，与众数、中位数等插补方法同属于传统统计插补，主要分为单一插补和分层插补两大类，其缺点为容易造成变量方差和标准差变小，相对而言更适用于分布较为平均且已知样本量信息较多的数据插补问题。

而线性回归插补的主要思想则为，利用已知观测样本点建立线性回归模型，估计回归模型参数进而对缺失样本点进行预测填补，其主要局限在于当模型中的变量非线性相关或预测变量高度相关时，容易产生有偏估计。

2.2.3 自回归建模插补

自回归建模方法多用于传统统计学中处理时间序列预测的相关问题研究，后经过相关学者的不断深入研究，逐渐在信号处理中的缺失音频图文填充、缺失数据预测等方面也有了广泛应用。其主要特点在于不仅能处理因自身因素而受影响的预测问题，还能建立向量自回归模型处理因其他因素受影响的缺失数据预测插补问题。对于本文中非时间序列的预测问题，则可以利用已知观测数据进行正向和反向自回归拟合推断而来的估计值代替缺失数据，该方法主要运算步骤为：

步骤1：将已知观测数据$X(k)$和待估计数据$X(1)$的总数据$X(n)(n=1,2,3,\cdots,N,k+1=N)$变换为：

$$\tilde{X}(n)=X(n)-\bar{X} \tag{1}$$

其中，$\bar{X}=\frac{1}{N}\sum_{n=1}^{N}X(n)$。

步骤2：设定总样本数据自回归模型的阶数上限I，为避免排除有效模型，I应该设定得足够大。

步骤3：计算样本自协方差：

$$Cxx(i)=\frac{1}{N}\sum_{n=1}^{N-i}\tilde{X}(n+1)\tilde{X}(n) \tag{2}$$

其中，$i=0,1,2,\cdots,I$。

步骤4：通过最小二乘法拟合$M\left(M=1,2,\cdots,I\right)$阶自回归模型。

步骤5：通过比较FPE(Final Prediction Error，最终预报误差)大小得出最终预测结果，整个运算过程我们可通过计算机程序来实现。

3 基于模拟数据进行不同插补方法比较分析(Comparative analysis of different interpolation methods based on simulated data)

3.1 数据来源及统计指标说明

本部分我们首先基于服从高斯分布和伽马分布两种形式生成的2,000 个模拟数据，对其按照5%、10%、20%三种不同固定比例构造缺失数据后进行四种插补方法的比较，然后基于非洲地区47 个国家1993—2013 年(共21 年)的出生时预期寿命完整数据集，同样设置三种缺失比例对缺失部分进行插补预测，即对以上四种方法结果的适用性进行验证分析。本文主要以下面两种统计指标作为插补效果评判的依据：

指标1：均方误差(Mean Square Error，MSE)。

$$MSE\left(\hat{X}\right)=E\left(X-X\right)^{2} \tag{3}$$

其中，$\hat{X}$为预测值，X为实际值。

指标2：平均绝对百分比误差(Mean Absolute Percentage Error，MAPE)[9]。

$$MAPE=\frac{\left|X-\hat{X}\right|}{nX} \tag{4}$$

其中，n为样本容量。

本文用均方误差和平均绝对百分比误差这两种统计指标的大小来评判插补效果，其中MAPE值和MSE值的大小同样能反映插补值与真实值之间的差异[10]，数值越小则表示预测值与真实值之间的差异越小，即插补效果越好，反则反之。

3.2 基于多种方法不同缺失比例下的插补结果比较分析

3.2.1 高斯模拟数据集

本部分基于服从高斯分布的模拟数据集，分别运用最近邻(K=5)插补、均值插补、线性回归插补和自回归建模插补四种方法进行缺失数据的预测，其中缺失比例我们预设为5%、10%和20%三种情况，预测插补后两种评价指标均方误差和平均绝对百分比误差的具体结果如表1、图1和图2所示。

表1 高斯数据集缺失数据插补统计结果

Tab.1 Statistical results of missing data interpolation in Gaussian data set

方法	均方误差			平均绝对百分比误差		
	5%	10%	20%	5%	10%	20%
最近邻插补(K=5)	0.0722	0.1211	0.2409	8.8204	46.6813	118.2472
均值插补	0.0621	0.0984	0.1885	5.0949	11.5287	21.6918
线性回归插补	0.0855	0.1540	0.3175	11.4383	64.4391	107.8830
自回归建模插补	0.0663	0.1157	0.2001	6.7867	23.9425	46.5361

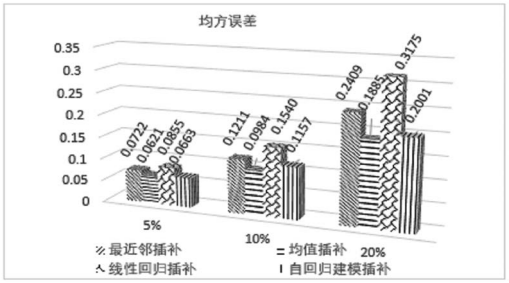

图1 高斯数据集插补结果MSE

Fig.1 Gaussian data set interpolation result MSE

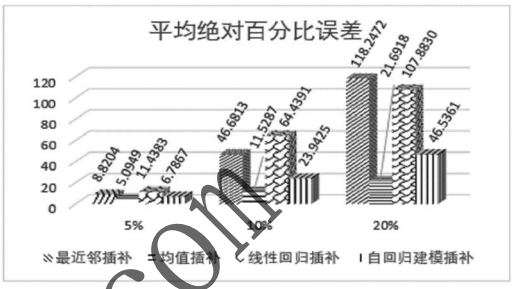

图2 高斯数据集插补结果MAPE

Fig.2 Gaussian data set interpolation result MAPE

由以上结果可知，三种缺失比例下的插补准确率效果整体趋势相同，相比较而言均值插补和自回归建模插补两种方法的效果较好。其中，评价指标MSE值二者更为接近，插补预测后两种插补方法在三种缺失比例下的均方误差分别为0.0621、0.0984、0.1885和0.0663、0.1157、0.2001。而最近邻插补和线性回归插补两者的结果较差，从统计结果来看与前两者尚存在较为明显的差距，尤其是当缺失数据比例为10%和20%时，劣势尤为明显。

3.2.2 伽马模拟数据集

与上一部分中试验过程类似，本部分试验基于服从伽马分布的模拟数据集，分别运用最近邻(K=5)插补、均值插补、线性回归插补和自回归建模插补四种方法进行缺失数据的预测插补，为控制变量以便作插补效果对比，缺失比例我们同样预设为5%、10%和20%三种情况，预测插补后两种评价指标均方误差和平均绝对百分比误差的具体结果如表2、图3和图4所示。

表2 伽马数据集缺失数据插补统计结果

Tab.2 Statistical results of missing data interpolation in gamma data set

方法	均方误差			平均绝对百分比误差		
	5%	10%	20%	5%	10%	20%
最近邻插补(K=5)	0.0923	0.2447	0.5267	5.5873	14.9253	27.6175
均值插补	0.0769	0.1981	0.4058	5.4763	12.8679	24.0488
线性回归插补	0.1068	0.2818	0.5772	5.5385	13.8301	27.7256
自回归建模插补	0.0730	0.2068	0.4525	5.2446	11.5582	24.8326

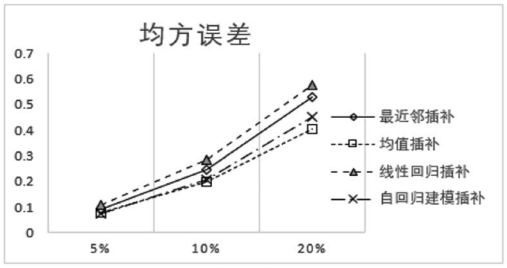

图3 伽马数据集插补结果MSE

Fig.3 Gamma data set interpolation result MSE

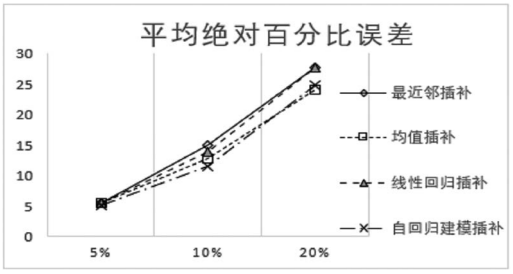

图4 伽马数据集插补结果MAPE

Fig.4 Gamma data set interpolation result MAPE

由以上结果可知，服从伽马分布的模拟数据集三种缺失比例下的四种插补方法预测效果与高斯模拟数据集无明显差别，同样为均值插补和自回归建模插补效果较好，最近邻插补和线性回归插补效果相对较差。另外，由图3和图4我们可观察到，当缺失比例较小时，四种插补方法的均方误差和平均绝对百分比误差结果均极为接近；而当缺失比例为10%时，四种方法的均方误差和平均绝对百分比误差数值虽未有明显差异，但已逐渐开始产生区别；当缺失比例为20%时，平均绝对百分比误差数值上均值插补和自回归建模插补，最近邻插补和线性回归插补分别两两接近，分别为24.0488和24.8366，27.6175和27.7256。

3.3 实例分析

进行了上文中四种插补方法对两种分布的模拟数据预测插补结果分析后，本部分选取了非洲地区47 个国家1993—2013 年(共21 年)的出生时预期寿命(Life Expectancy at Birth)实际数据进行实证对比分析，同样对其预设5%、10%和20%三种缺失比例，进行存在缺失部分数据的插补。在比较统计结果的同时检测以上几种方法在实际缺失数据插补当中的适用性，其中具体结果如表3、图5和图6所示。

表3 出生时预期寿命数据集插补统计结果

Tab.3 Statistical interpolation results of the life expectancy at birth data set

方法	均方误差			平均绝对百分比误差		
	5%	10%	20%	5%	10%	20%
最近邻插补	0.6545 (*K*=5)	0.4315 (*K*=4)	2.5687 (*K*=2)	0.2461 (*K*=5)	0.2805 (*K*=4)	0.7883 (*K*=2)
均值插补	4.3511	5.0685	13.2615	0.7712	1.0378	2.3595
线性回归插补	4.0008	11.9035	20.4462	0.7110	1.4215	2.9468
自回归建模插补	6.3643	8.4091	19.2148	0.8650	1.2979	2.7524

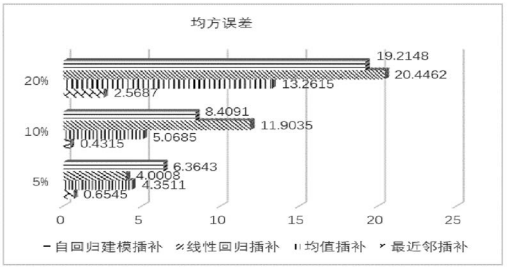

图5 出生时预期寿命插补结果MSE

Fig.5 Life expectancy at birth interpolation results MSE

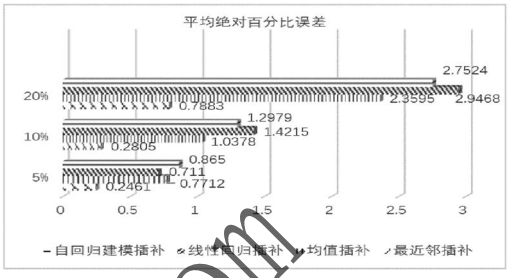

图6 出生时预期寿命插补结果MAPE

Fig.6 Life expectancy at birth interpolation results MAPE

由以上插补统计结果可看出，最近邻插补方法在实际数据中均方误差和平均绝对百分比误差数值明显小于其他三种插补方法。但值得注意的是，模拟数据的预测插补过程中我们最终选用*K*的数值为5，而在实际数据的预测插补中*K*的最终值却因缺失比例不同而有所变化。究其原因是在最近邻插补方法的计算过程中，当缺失比例较大时，较容易出现大量数据连续缺失的情况，而此时计算机软件无法对距离做出精确的测算，也就是说无法得出具体的插补结果，而人工计算对于如此容量的数据测算更是难上加难，因此在本部分实际数据的插补效果比较中，当缺失比例高达20%时，最近邻插补方法可暂时退出比较，同时这也从一定程度上检验出了该方法在实际缺失数据的预测插补等应用上的局限性。

另外，其他三种插补方法的效果按整体趋势从好到坏依次可排序为均值插补、自回归建模插补和线性回归插补。其中线性回归插补除在缺失比例为5%的情况下统计结果略低于均值插补和自回归建模插补外(MSE为4.0008，MAPE为0.7110)，均明显劣于前两者，这可能说明线性回归插补在实际缺失数据的预测中较适用于缺失比例低的情况，而当缺失数据量较大时则不适用。

4 结论(Conclusion)

本文通过运用四种插补方法对两组模拟数据和一组实际数据进行缺失数据插补，对比统计指标均方误差和平均绝对百分比误差数值后，得出不同分布数据和不同缺失比例下的适用插补方法，主要有以下结论：无论是模拟数据还是实际数据，以整体插补效果来看(考虑不同缺失比例)，自回归建模插补和均值插补略优于最近邻插补和线性回归插补；而在实际数据出生时预期寿命这一变量的预测插补过程中，若缺失样本量较少即缺失比例较低时，最近邻插补和线性回归插补

(下转第10页)