摘要： 本文通过分析原始数据和“千车故障数”计算的不合理性，提出了改进的千车故障数的计算方法，并对原始数据进行变换得到了合理的数据表。通过假设检验得出瞬时故障率是服从 Weibull 分布，在此基础上提出了单批次预测的概率模型。进一步，利用数据横向和纵向的相关性，在单批次概率模型的基础上，提出了基于样条拟合的多批次双向联合预测模型，这个模型保证了横向和纵向的有机结合，通过拟合汽车的质量参数，来进行联合预测，维护了数据的统一性。

关键词： 千车故障数；瞬时故障率；相关分析；Weibull 分布；质量参数；样条模型

1 引言

近些年来，为了提高汽车产品的可靠性水平，我国的科研工作者对利用汽车售后故障信息评价汽车的可靠性做了大量的研究，取得了一定的成绩[1-3]。这些研究基本上都是利用故障数据建立汽车故障的概率分布模型，一定程度上能反映问题所在，但还不足以能完全反映汽车的可靠度水平，进一步可以从汽车的生产批次来考察汽车的可靠性，提供更小粒度的考察。这里我们根据国内某一轿车公司的售后服务数据进行了处理，除了运用传统的概率方法建模外，还充分利用了不同轿车批次间的相关关系，运用样条的方法进行双向联合统一建模，并进行故障预测，预测达到了较好的效果，有利于汽车生产厂家及时调整生产对策，提高产品的市场竞争力。

2 数据分析

2．1 原数据的不合理性

现有某轿车厂家 2004 年 4 月 1 日从数据库中整理出来的某个部件的故障数据表（见原题），该表是工厂的真实数据，基本上能反映出零件随时间出现故障的情况，但制表方式存在不合理之处，还有待改进。不合理性主要来自两方面：计算公式的不合理性和数据时滞带来的不合理性。原表中“千车故障数”的计算公式为：$w_i = \dfrac{n_i \times 1000}{N}$。这里 w_i 为某批次车 i 个月以内的“千车故障数”，n_i 为该批次车中使用了 i 个月累计出现了故障的次数，N 为到制表时刻为止该批次车销售的总数。这个计算公式存在问题，主要是分母问题（车的销售总数）。因为车若要使用了 i 个月，则它一定是在当前时刻（制表时刻）前 i 个月以前买的，而销售是在不断进行的，在这 i 个月以内售出的车数 N_i' 还没有使用到 i 个月，因此在计算上时分母实际上是被扩大了，而且由于累积效果，导致使用月份越多，误差也越来越大。

为此本文提出了瞬时故障率的概念，定义 $\lambda_i = \dfrac{n_i - n_{i-1}}{N - N_i'}$ 为第 i 个月的瞬时故障率，显然 $0 \le \lambda_i \le 1$。瞬时故障率的好处是既保证了样本空间的准确性，又避免了累积所带来的问题。

另一方面，对于这类用于指导生产的数据，时效性是很强的，但如问题中所言，数据显得滞后很多，对于预测是很不利的。越靠近制表时刻，知道的数据越少，用这些数据来进行预测的准确性就越差，但同时这些数据又是非常重要的，这是一对矛盾体。在处理上靠近制表时刻的数据
重要性的权值可以设的小一些。

2．2 转换数据

在分析了原表的不合理性外，可以根据不合理数据反推出合理的千车故障数，但是因为要扣除 $t-i$ 到 t 时刻间的销售量 N_i'，而这个量到目前为止我们并不是知道的，只知道到制表日为止的销售总量，因此给转换数据带来一定困难。这里本文基于一种简单假设，认为汽车的销售量随时间是呈均匀分布的，当然这并不一定合理，因为汽车的销售有很大的随机性，有可能还要受到季节性的影响，但是在缺少销售信息的前提下，可以近似认为其服从均匀分布。

视销售量为均匀分布，利用已知数据，根据瞬时故障率的定义从 0 个月到 12 个月依次进行数据反推，本文得出了合理的千车故障数分布表。与原千车故障数表相比较发现，数据基本上比原来扩大了，这是因为

分母缩小的结果。改进的数据更能体现出“千车故障数”的含义， 图 1 是部分批次的数据对照图。

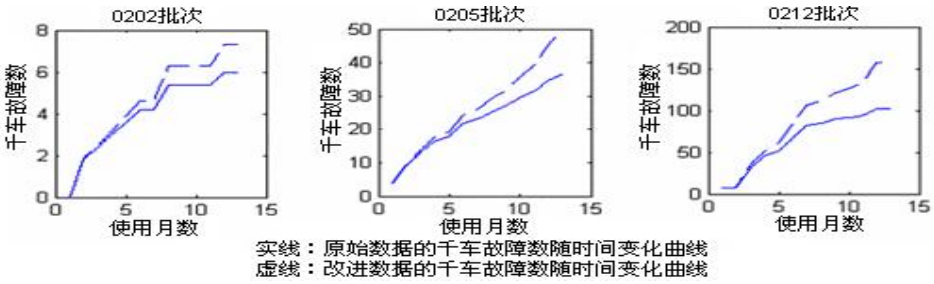

图 1：部分批次改进后千车故障数与原始千车故障数对比图

从图中也可发现，对于原来计算方法，随着月份的增加，误差会越来越大。

2．3 数据的相关分析

由于数据本身的累积特性，各批次间（纵向）以及不同月份间（横向）都有可能存在相关关系，利用好这些相关关系无论对于建立模型或者进行预测都是非常重要的，尤其是数据量比较小的时候，这种相关关系显得尤为重要。利用多元统计分析中的相关性检验[4] ，发现横向和纵向数据相关矩阵的最大特征根都非常大，结果如下：

$$\lambda_1=(0\quad 0\quad 0\quad 0\quad 1\quad 1\quad 4\quad 4\quad 5\quad 26\quad 82\quad 657\quad 34827)$$

$$\lambda_2=(0\quad 0\quad 0\quad 0\quad 0\quad 1\quad 1\quad 2\quad 4\quad 5\quad 25\quad 71\quad 597\quad 23138)$$

其累积贡献率都超过 99%，因此可以认为横向数据和纵向数据具有很强的相关性，它们之间并不是相互独立的。推究其深层原因，可以这样认为，横向数据的相关性主要是由于汽车的故障率随时间是连续变化的，而纵向的相关性是由于厂家部分的改进了技术，而这种技术的改进也是渐进变化的。

3 基于单批次的概率模型

前面分析了原数据的不合理性，并得出了改进后的数据表。我们可以仅仅从某个单一批次出发，利用汽车发生故障的物理规律建立单批次预测的概率模型。根据改进后的数据，我们得到的仍是 i 个月以内的千车故障数 w_i，$i=0,1,\cdots 12$，这是一个简单的时间序列，可以考虑用时间序列中的平滑方法或 ARMA 模型来处理，但由于数据太少（只有 13 个），且还要进行平稳行检验，利用时间序列显得力不从心。因此本文从汽车可靠性理论建立模型[5]，这里选用威布尔分布（Weibull Distribution）来建模，主要是考虑它是工程上常用的分布同时又涵盖了指数分布、正态分布等特殊形式的分布。

威布尔分布 $W(m,t_0)$ 中参数 m 称为形状参数，参数 t_0 称为尺度参数，又称为零件的特征寿命参数。$m=1$ 退化为指数分布 。我们已知单一批次 i 个月以内轿车的千车故障数 w_i，$i=0,1,\cdots 12$，至于这些数据是否服从威布尔分布，还要进行假设检验。这里采用基于误差的 Weibull 检验，即先用威布尔概率纸的方法拟合出质量参数 $\{\hat{m},\hat{t}_0\}$，然后回代入分布函数，判别系统误差 $\varepsilon(t)=F(t)-\hat{F}(t)$ 是否服从零均值的正态分布 $\varepsilon\sim N(u,\sigma^2)$，其中 $u=0$，σ 未知。即检验 $H_0:u=0,\quad H_1:u\neq 0$，取显著性水平 $\alpha=0.1$，则拒绝域为

$$|t|=\left|\frac{\bar{\varepsilon}-0}{S/\sqrt{n-1}}\right|>t_{\frac{\alpha}{2}}(n-1)\,, \qquad (3.1)$$

这里 $\bar{\varepsilon}$ 为 $\{F(t_i)-\hat{F}(t_i)\},i=0,1,\cdots 12$ 的均值，S 为标准差，$n=13$。查表得 $t_{0.05}12=1.3502$，因此如果算出 $|t|$ 值小于 1.3502 就可以认为原数据服从威布尔分布。经过分布假设检验，所有批次均不拒绝原假设，因此可以认为轿车的故障服从威布尔分布。图 2 以 0205 批次为例显示了威布尔分布的拟合结果，左图为威布尔概率纸，右图为拟合前后的比较。

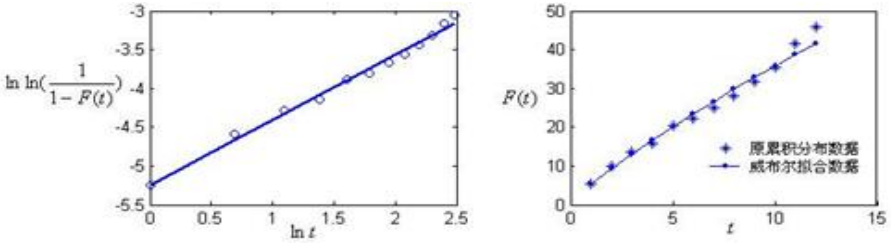

图 2 ：0205 批次的故障累积分布威布尔拟合图

根据前面建立的威布尔分布模型，可以仅从单一批次得到的汽车质量参数出发进行预测，预测结果为：0205 批次使用 18 个月的“千车故障数”为 40.259，0306 批次使用 9 个月的“千车故障数”为 6.0819。0310 批次使用月数为 12 的“千车故障数”由于数据的特殊性（都为 0），无法仅仅根据横向数据做出相应的预测。

4 基于样条模型的多批次双向联合预测模型

在单批次预测模型中，用威布尔分布函数来拟合单批次轿车的故障累积分布，很好地刻画了横向数据的规律性，但单批次预测模型无法刻画纵向数据的变化规律。如何综合利用数据的横向和纵向相关性来进行预测呢？本文认为数据的横向相关性体现在每个批次 i 的数据服从参数为（m_i, t_0^i）的威布尔分布，数据的纵向相关性体现在不同批次的分布参数（m_i, t_0^i）的连续变化。因此根据函数逼近理论，可以在纵向一段批次内用基函数建立威布尔分布参数（m_i, t_0^i）的参数表示模型。本文选用样条函数作为基函数，提出基于样条模型的多批次双向联合预测模型。这里需要强调的是横向和纵向数据的相关性（连续性）是我们统一建模的基础。

有两种样条处理方法：一种是先用单批次模型对横向数据进行处理，得到每个批次 i 对应的分布参数（m_i, t_0^i），然后将得到的（m_i, t_0^i）在纵向批次间进行样条拟合平滑处理。这种处理方法虽然考虑了纵向数据的相关性，但是将横向相关与纵向相关分开处理，割裂了二者的整体性。特别在后面的批次中，横向数据量较少，单批次模型的估计信息不足，导致（m_i, t_0^i）的估计精度较差，甚至根本无法估计，在此基础上再对（m_i, t_0^i）进行样条拟合平滑处理，效果较差。

本文采用第二种样条处理方法：先将不同批次的分布参数（m_i, t_0^i）用样条模型表示后，再代入单批次模型的威布尔分布表达式中，对所有批次的双向数据统一建模。模型由原来分别估计每个批次的威布尔分布参数（m_i, t_0^i），转化为统一估计（m_i, t_0^i）的样条表示参数，在数据处理上不再区分横向和纵向的先后，对数据进行双向联合处理，反映数据的双向相关性，故称为双向联合预测模型。下面以等距节点的 3 次 B 样条函数[7]为例，阐述基于样条模型的多批次双向联合预测模型原理：

在连续区间 $i \in [a,b]$ 内，用 $m(i)$，$t_0(i)$ 来表示批次 i 的威布尔分布参数。$m(i)$，$t_0(i)$ 在 $i \in [a,b]$ 区间内连续，于是可用 N+3 个等距节点的 3 次 B 样条函数来拟合表示：

$$m(i) = \sum_{j=-1}^{N+1} b_{1j} B\left(\frac{i-\tau_j}{h}\right) \qquad j = -1,0,1,\cdots,N+1 \qquad (4.1)$$

$$t_0(i) = \sum_{j=-1}^{N+1} b_{2j} B\left(\frac{i-\tau_j}{h}\right) \qquad j = -1,0,1,\cdots,N+1 \qquad (4.2)$$

其中 τ_j 为 B 样条函数的节点，节点间距 $h = \tau_{i+1} - \tau_i$，$B(\bullet)$ 是 B 样条的基函数，b_{1j}, b_{2j} 都是样条系数。用向量 β_1 表示 $\left(b_{1,-1},\cdots,b_{1,j}\cdots b_{1,N+1}\right)$，$\beta_2$ 表示 $\left(b_{2,-1},\cdots,b_{2,j}\cdots b_{2,N+1}\right)$。

将(4.1),(4.2)代入威布尔分布函数中：

$$F_i(t) = \begin{cases} 1 - e^{-\frac{t^{m(i)}}{t_0(i)}}, & t > 0 \\ 0, & t \le 0 \end{cases}$$

对其进行对数变换得到：

$$\ln\ln\frac{1}{1-F_i(t)} = m(i)\ln t - \ln t_0(i) \qquad (4.3)$$

记等式的左边 $\ln\ln\frac{1}{1-F_i(t)}$ 为 $Y_i(t)$。而由(4.1),(4.2)知，$m(i)$、$t_0(i)$ 可分别用可用样条参数 β_1，β_2 拟合表示。因此不妨记右边 $m(i)\ln t - \ln t_0(i)$ 为：$G(\beta_1,\beta_2)$。于是得到：

$$Y_i(t) = G(\beta_1,\beta_2) + \varepsilon_i(t) \qquad (4.4)$$

这里 i 表示某一批次，$\varepsilon_i(t)$ 为观测噪声。于是模型由原来估计单批次的威布尔分布参数（m_i, t_0^i），转化到估计威布尔分布参数（m_i, t_0^i）的样条表示参数 β_1，β_2。

$G(\beta_1,\beta_2)$ 为非线性函数，用 Gauss-Newton 方法迭代求解，需将 $G(\beta_1,\beta_2)$ 在概略点 (β_1^*,β_2^*) 处线

性展开：

$$Y_i(t)=G(\beta_1^*,\beta_2^*)+\left.\frac{\partial G(\beta_1,\beta_2)}{\partial\beta_1}\right|_{\beta_1=\beta_1^*}(\beta_1-\beta_1^*)^T+\left.\frac{\partial G(\beta_1,\beta_2)}{\partial\beta_2}\right|_{\beta_1=\beta_2^*}(\beta_2-\beta_2^*)^T+\varepsilon_i(t)$$

上式可以简记为：$Y_i(t)-G(\beta_1^*,\beta_2^*)=A(\beta_1-\beta_1^*)^T+B(\beta_2-\beta_2^*)^T+\varepsilon_i(t)$ (4.5)

其中：$A=\left.\frac{\partial G(\beta_1,\beta_2)}{\partial\beta_1}\right|_{\beta_1=\beta_1^*}$，$B=\left.\frac{\partial G(\beta_1,\beta_2)}{\partial\beta_2}\right|_{\beta_1=\beta_2^*}$。

(β_1^*,β_2^*) 是通过拟合单批次模型的粗解获得的。向量 A,B 中的分量可采用复合函数求导规则求得：

$$\frac{\partial G(\beta_1,\beta_2)}{\partial b_{1j}}=\ln t\cdot\frac{\partial m(i)}{\partial b_{1j}};\quad\frac{\partial G(\beta_1,\beta_2)}{\partial b_{2j}}=\frac{1}{t_0(i)}\cdot\frac{\partial t_0(i)}{\partial b_{2j}}$$

于是参数 (β_1,β_2) 的估计过程，可归结为寻找 $(\hat\beta_1,\hat\beta_2)$，使

$$\min_{\beta_1,\beta_2}\left\|Y_i(t)-G(\beta_1^*,\beta_2^*)-A(\beta_1-\beta_1^*)^T-B(\beta_2-\beta_2^*)^T\right\|^2 \quad (4.6)$$

用估计得到的参数 $(\hat\beta_1,\hat\beta_2)$ 代替 (β_1^*,β_2^*)，采用 Guass-Newton 叠代方法进行叠代计算，以满足叠代收敛条件的 $(\hat\beta_1^l,\hat\beta_1^l)$ 作为 $(\hat\beta_1,\hat\beta_2)$ 的近似解，代入（4.1）、（4.2）式即可求得每个批次的威布尔分布参数 $m(i)$，$t_0(i)$。

下面是模型计算结果：在批次 $i\in[0\quad 23]$ 内，模型采用两个区间的 3 次 B 样条函数。经过求解得到样条系数分别为：

$$\beta_1=(21815\quad 13.225\quad 82.524\quad -15.985\quad 529.38)$$
$$\beta_2=(5.3681\quad 0.016868\quad 1.7117\quad 0.45216\quad 3.3923)$$

表 1：双向联合预测模型解算出的不同批次的威布尔分布参数

批次	曲线形状 m	特征寿命 t_0	批次	曲线形状 m	特征寿命 t_0
0201	1.19121	3658.4	0301	1.07565	44.8081
0202	1.00613	2395.28	0302	1.05474	52.0745
0203	0.922612	1468.72	0303	1.07457	66.9396
0204	0.916946	826.513	0304	1.15211	91.3273
0205	0.965428	416.4	0305	1.29732	126.162
0206	1.04436	186.144	0306	1.49222	168.37
0207	1.13002	83.506	0307	1.71186	213.877
0208	1.19872	56.2451	0308	1.93126	258.609
0209	1.2276	53.2495	0309	2.12546	298.492
0210	1.21329	49.3611	0310	2.2695	329.454
0211	1.17188	45.3754	0311	2.33841	347.42
0212	1.12035	43.2164	0312	2.30823	348.48

表 1 是利用联合模型解算出的不同批次的威布尔分布参数，根据上面的参数画出千车故障数的空间分布，见图 3：

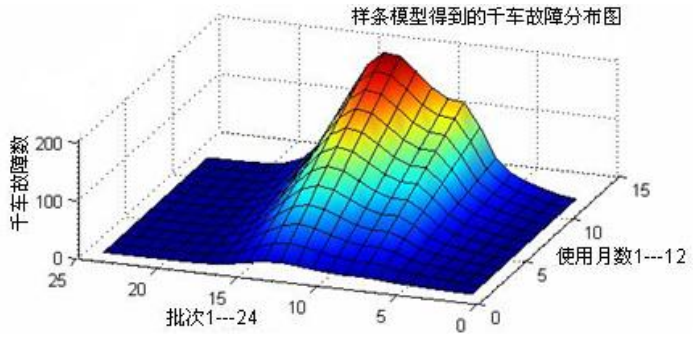

图 3　多批次双向联合预测模型还原的千车故障数的空间分布图

图 3 很好的反映了原有数据的走向规律，中间隆起，反映了原数据中，中间部分的千车故障数比较高；越往两端，故障数越低，同时很好的预测了原表中没有的数据。表 2 和表 3 是单批次模型和联合模型的预测结果比较。

表 2：本文定义的千车故障数得出的结果

预测点	0205 使用 18 个月时	0306 使用 9 个月时	0310 使用 12 个月时
单批次预测	58.037	9.6364	—
多批次联合预测	58.037	12.564	0.54319

表 3：换算成原题“千车故障数”定义得出的结果

预测点	0205 使用 18 个月时	0306 使用 9 个月时	0310 使用 12 个月时
单批次预测	40.259	6.0819	—
多批次联合预测	40.259	6.3746	0．020396

[注意] 表 2，表 3 中的折线表示这个点不能预测，原因是 0310 这个批次的已知数据都是零，用单批次模型无法拟和其威布尔分布曲线。但是多批次样条模型因为考虑了纵向相关，是可以预测的。

5 模型的评价

前面根据售后数据，建立了汽车故障分布的单批次概率模型，在这基础上又利用纵向间的相关信息进行样条统一建模，并进行预测，达到了较好的效果。统一建模具有以下优点：

（1）用样条模型对双向关联数据统一建模，将横向和纵向有机的结合起来，充分利用了数据的双向连续性，保持了数据整体统一性，提高了预测精度。另外基于数据时滞性的考虑，还可以在样条模型中进行加权处理，靠近售出点数据的权值可以取得小一些，而其他部分的权值可以取得大一些。

（2）样条模型可以节省估计参数个数。设共有 n 批次的轿车，用单批次解算的方法，待估参数个数为 $2n$，而使用 B 样条表示，假设用 N 个区间进行拟合时，只需要 $2*(N+3)$ 个参数（ N 一般较小），两者相差 $2*(n-N-3)$ 个参数。以本题为例： $n=24$， $N=2$，相差 38 个，因此大大减少待估参数的个数。

（3）样条模型可以平滑噪声。通过拟合汽车的质量参数来进行建模，避免了直接拟合数据带来的噪声影响。就其频域特性而言，由于样条拟合相当于一个低通滤波器，可以有效地保留低频状态的信号，而滤掉高频状态的噪声，从而大大提高估计精度。

（4）对初值不敏感。样条模型利用单批次的粗解进行拟合得到的参数值作为初值，个别采样时刻初值的较大扰动一般不会影响到样条模型的收敛。

参考文献

[1] 祁型红，严运兵． 基于售后故障信息的汽车可靠性评价方法． 机电产品开发与创新，2004.3，17（2）：49-51

[2] 阎勇，曹正清． 利用三包服务期内故障数据评估汽车的可靠性的方法.农业机械学报 1999， 30（3）

[3] 徐坤，赵致宽 .中型载重汽车可靠性与维修性指标评定方法.交通科技与经济.2000，（1）

[4] 李排昌. 在相关变量间寻找主因素的一种方法.吉林大学学报，Vol 40

[5] 王秉刚. 汽车可靠性工程方法.北京：机械工业出版社，1991.11

[6] 张湘伟. 结构分析中的概率方法.科学出版社

[7] 王正明，易东云.测量数据建模与参数估计.国防科技大学出版社

The combine forecasting of the reliability of auto based on spline model

Wang Dan Gu De-feng Rao Bin

Advisor: Wu Meng-da

(Dept. of Mathematic and System Science, National University of Defense Technology, Changsha 410073, China)

Abstract: the paper analyzes the inconsequence of the primitive data and “failure number per thousands auto ” and presents the computing method of correctional “failure number per thousands auto ”. Based on new method, the reasonable data table is gained. Meanwhile, by test of hypothesis, we conclude that the failure rate submits to WEIBULL distribution. Based on the conclusion, the probability model of single batch forecasting is presented. Further, by the correlation of horizontal data and vertical data, multi-batch bidirectional forecasting model based on spline fitting is presented. The model guarantees the unification of horizontal data and vertical data.

Key words: failure number per thousands, instantaneous failure rate, correlation analysis, WEIBULL distribution