问卷分割技术

——互联网背景下的新型市场调查设计研究

摘 要 市场调查中的问卷设计由于沿袭传统调查冗长而复杂的方式，已成为当前互联网和移动互联网背景下市场调查的突出矛盾，造成受访者合作率下降、调查中断、数据异常等不良影响，从而导致调查质量下降，消费者洞察出现偏差。本文提出问卷分割技术，包含从问卷如何分解到调查数据如何处理及研究，是技术流程环环相扣的完整的技术体系。本文结合实证阐述了问卷分割技术过程，该技术可以作为互联网市场调查设计和固定样本管理的一个新兴工具，在改善项目回答率和提高数据质量方面加以应用，成为深刻洞察消费者的一把利器。

关键词： 聚类算法；最优样本量设计；缺失数据处理；参数估计

Abstract Under the background of Internet and mobile Internet, the design of questionnaire in market survey has become a prominent contradiction because of the long and complicated way. It causes the adverse effects such as the decline of cooperation rate of respondents, the interruption of survey and the abnormal survey data. In this paper, a questionnaire segmentation technique is proposed. Questionnaire segmentation technology is a complete technical system. This paper elaborates the specific process of questionnaire segmentation technology based on empirical research. This technology can be used as a new tool for market survey design and fixed sample management. It can be applied in improving project response rate and data quality, and become a sharp tool for deep insight into consumers.

Key words: clustering algorithm; optimal sample size design; missing data imputation; parameter estimation

一、引言

1.1 问卷分割技术的研究背景

大数据与人工智能技术突飞猛进的发展潮流下，市场调研数据在消费者洞察领域依然具有自身独特的优势和不可替代的价值，只有代表人“态度”的调研数据和代表人“行为”的大数据的完美融合才能更深刻更全面地洞察消费者。市场调查设计是消费者洞察背后看不见的“触手”，对调研结果的影响不容小觑。然而，它与当前调研方式逐步向互联网及移动互联网调查多元、飞速的变革相比则显得相对滞后，甚至依旧沿袭传统的调查问卷设计，冗长而复杂，导致调查数据存在许多潜在问题，消费者洞察出现偏差。因此，调查问卷设计是调查行业目前面临的主要挑战和亟待解决的难题。

市场调查设计一般要综合平衡成本效率、准确和速度三个目标之间的关系。为了实现成本效率目标，市场调查机构通常会最大限度地通过调查问卷来实现客户对信息“量”的需求。因此，大多数调查往往包含人口特征、消费行为、媒体接触等多个主题，相应的问卷必然冗长而

复杂。这种设计通常会严重破坏市场调查目标间的权衡关系，特别是在网络调查中尤为突出，主要表现在(1)**调查项目响应率下降**。受访者合作的积极性下降，严重制约了调查项目的进度，降低效率、抬高成本，导致调查机构负担加重；(2) **调查中断。**以往的调查经验显示，问卷长度每增加 5 分钟，受访者中途放弃率便会上升 5%左右，而网络调查中途放弃的可能性比传统面访要大得多；(3)**增加数据异常值比率**。受访者的厌倦情绪将会增加数据未在估计范围内的可能性，要么夸大或降低回答的真实得分，要么虚增或弱化变量间的相关性，这将直接影响对调查目标的推断；(4)**增加受访者与调查项目间的摩擦**。特别是对固定样本，受访者合作态度消极，对后续更多调查项目的正常开展不利[1]。可见，调查问卷冗长而复杂已经成为市场调查设计不能承受之重，为了从根本上提高市场调查质量，更深刻地洞悉消费者，问卷分割技术的研究与应用迫在眉睫、呼之欲出。

1.2 问卷分割技术研究综述

问卷分割技术是从“矩阵抽样”发展而来，最初适用于教育服务的评估上，并且已经在国外的市场调查领域得到一定应用。例如，德国媒体和消费调查采用问卷分割技术来设计和分析样本，最终证实了问卷分割技术和数据插补能够再现原始数据 [2]。美国劳工部将问卷分割技术应用到消费者调查中，Gonzalez 和 Eltinge[3]考虑了在问卷分割条件下基于插补的样本均值和总体的估计方法。Adigüzel 和 Wedel[4]、Chipperfield 和 Steel[5]详细论述了问卷分割技术结合数据插补是提升抽样效率的有效方法。

本文研究的问卷分割技术是现有研究成果的延伸和突破，增加了问卷聚类分割设计、分割模式选择、最优样本量设计等环节，是包含从问卷如何分解到调查数据如何处理及研究的完整的技术体系，技术流程环环相扣，有别于以往的调查设计片段式孤立地解决某些单独环节的概念，更加具有全局效应。

二、复杂调查中的问卷分割技术研究

2.1 问卷分割技术概述

问卷分割技术，简而言之是指将长问卷分割成若干个子问卷，每个受访者只需回答其中部分问题即可。由于每个人回答的问题数量减少，这也在一定程度上提高了问卷的响应率和完成率，再通过最优样本量设计、多重插补等技术手段保证调查成本效率和调查数据质量。

问卷分割技术主要解决两个问题，一个是如何分割问卷，一个是如何处理实施问卷分割后得到的调查数据。对于第一个问题，首先，我们在做问卷分割前要明确我们的调查目的和需求，结合项目预算和问卷架构对问卷问题进行分组形成若干个数据项，然后进行分割模式的划分，根据总的模式个数、各个数据项的重要性、费用和精确度的约束来计算不同分割模式对应的样本量；对于第二个问题，一种方式是直接根据现有采集的调查数据进行参数估计而不做其他处理，另一种方式是针对缺失数据利用数据插补技术进行插补，然后根据插补后的完全数据进行参数估计和预测等；最后汇总分析，撰写研究报告。问卷分割技术涉及问卷调查设计的各个方面，每一环节对应不同的技术手段和处理方式。

2.2 问卷分割技术——变量聚类分割

2.2.1 基于相关矩阵的聚类变量分割

问卷分割一般是将问卷分割成一个核心部分和若干附属部分，核心部分的问题一般是指受访者的个人属性特征，如受访者的年龄、性别、学历、职业、收入状况等，而附属部分主要包含问卷主体调查内容。每个受访者则只需回答核心部分和其中几个附属部分（或称数据项），从而有效地减少了问卷长度。问卷分割技术的思想非常相似于在评价考核时常用到的“矩阵抽样”方法，可以说是矩阵抽样方法的延伸和改进。表 1 就是将长问卷分为 1 个核心部分和 n 个附属部分，每个附属部分中对应的问题数为$n_1,n_2,n_3,\ldots,n_n$

表 1：长问卷分割的结构特征

核心部分	附属部分				
	第 1 部分	第 2 部分	第 3 部分	第…部分	第 n 部分
年龄、性别、学历、职业、收入等个人属性特征	问题n_1	问题n_2	问题n_3	问题…	问题n_n

具体分割方法：首先要计算问卷中变量的相关性，通过对相关系数的聚类将变量聚成不同的数据项，分类遵循的原则是“类内差异小、类间差异大”的原则，即相关性高的变量聚为一类，相关性低的变量聚为不同的类。

长问卷往往为了达到多个不同的调查目的而被设计成若干个主题组成，如媒体接触、消费行为，生活态度等，各主题间并没有明显的相关关系，可视为独立的调查体系，因此直接按问卷中原有主题设计分割问卷也不失为一种高效简单的分割方法。

2.2.2 数据实证

以某电视频道受众调查项目为例，除背景信息外，问卷包括 A.媒介接触习惯与收视需求、B.频道评价、C.节目评价、D.价值观和生活态度 4 个部分，问题数量多达 60 道，调查时长将近半小时。基于变量相关矩阵的聚类分割将长问卷分割为 5 个附属部分，其中 A 部分问题多达 36 道，根据调查时长可以将其分为 2 个部分。下表 2 为该问卷聚类分割后的结构特征以及每个附属部分对应的问题数量和时长分配。

表 2：某电视频道受众调查问卷分割后的结构特征

核心部分	附属部分				
	第 1 部分	第 2 部分	第 3 部分	第 4 部分	第 5 部分

	媒介接触习惯与收视需求 1	媒介接触习惯与收视需求 2	频道评价	节目评价	价值观和生活态度
年龄、性别、学历、职业、收入（1 分钟）	18 道题（8 分钟）	18 道题（8 分钟）	11 道题（5 分钟）	7 道题（3 分钟）	1 道矩阵题（5 分钟）

2.3 问卷分割技术——分割模式选择

一种是利用单调的累积模式形成分割后的问卷，简称为**累积模式**；

一种是从总的数据项中选择固定数目的数据项，进行排列组合得到不同的模式形成分割后的问卷，简称为**排列组合模式**；

2.3.1 累计模式

累计模式是一种最容易理解的模式，此模式主要是在问卷变量重要性程度分别比较明显的情况下使用，比如调查一些综合性问卷，根据研究的需求对各个分割的数据项或问卷模块的重视程度进行等级划分。对于特别重要的数据项，让所有受访者回答或大部分受访者回答；而对于较为不重要的数据项，按照一个较低的比例来进行分配。累计模式相对来说是一种比较主观的分割模式，以 $K=5$ 为例具体的问卷分割模式如表 3 所示：

表 3　$K=5$ 问卷分割的累计模式

问卷分割编号	样本量分布	核心部分	分离变量				
			数据项 1	数据项 2	数据项 3	数据项 4	数据项 5
1	n_1	√	√				
2	n_2	√	√	√			
3	n_3	√	√	√	√		
4	n_4	√	√	√	√	√	
5	n_5	√	√	√	√	√	√

2.3.2 排列组合模式

排列组合模式主要是在问卷数据项重要性较为均衡的情况下，根据预期的问卷调查时长进行设定的模式。具体是指，在分割后的 K 个数据项中随机抽取其中 m（m 为整数，且 $1 \leq m \leq K$）项来构成新的问卷，可得到总共 C_K^m 个模式。选择组合的方式也比较灵活，如果想把调查时间控制在一个时间段（如 15 分钟）内，则需要保证分割问卷的组合方式尽可能在此范围内，如从原来的 5 组选择其中的 2 组或 3 组，组成C_5^2 或C_5^3 的形式，以 C_5^2 为例，问卷的组合形式

可用下表 4 表示：

表 4 $K=5$ 问卷分割的排列组合模式（C_5^2）

问卷分割编号	样本量分布	核心部分	分 离 变 量				
			数据项 1	数据项 2	数据项 3	数据项 4	数据项 5
1	n_1	√	√	√			
2	n_2	√		√	√		
3	n_3	√			√	√	
4	n_4	√				√	√
5	n_5	√	√		√		
6	n_6	√		√		√	
7	n_7	√			√		√
8	n_8	√	√			√	
9	n_9	√		√			√
10	n_{10}	√	√				√

表 4 中打“√”部分表示受访者回答部分，空白部分表示未回答部分。

这种方式的一个优势就是从整体上来看各个模式有相同的缺失比例，数据的缺失情况可以看成是随机缺失，这样在后续数据插补时可以选择已经成熟的插补技术。

2.3.3 数据实证

仍以 2.2 节电视频道受众调查项目为例，聚类分割将长问卷分割为 5 个数据项，根据排列组合方式，问卷调查时长设定为大约 10 -15 分钟的范围，因此应选择 C_5^2的模式，分割后短问卷共计 10 种模式，调查时长最短 9 分钟（问卷 3 、问卷 4），最长 17 分钟（问卷 1）。

表 5 某电视频道受众调查项目问卷分割的可能形式（C_5^2）

问卷编号	样本量分布	核心部分	分 离 变 量				
			第 1 部分 媒介接触习惯与收视需求 1	第 2 部分 媒介接触习惯与收视需求 2	第 3 部分 频道评价	第 4 部分 节目评价	第 5 部分 价值观和生活态度
		年龄、性别、学历、职业、收入（1 分钟）	18 道题（8 分钟）	18 道题（8 分钟）	11 道题（5 分钟）	7 道题（3 分钟）	1 道矩阵题（5 分钟）

1	n_1	√	√	√			
2	n_2	√		√	√		
3	n_3	√			√	√	
4	n_4	√				√	√
5	n_5	√	√		√		
6	n_6	√		√		√	
7	n_7	√			√		√
8	n_8	√	√			√	
9	n_9	√		√			√
10	n_{10}	√	√				√

表 5 中打“√”且为灰色填充部分表示受访者回答部分，空白部分表示未回答部分。

经过分割及模式选择后，短问卷共 10 种模式，即 10 份短问卷，通过随机发放给受访者的形式进行调查数据采集，对应的样本量 *n1*、*n2*、*n3…n10* 的设定，将在后续 2.4 节中具体介绍。

2.4 问卷分割技术——最优样本量设计

确定样本量是设计市场调查方案的一个重要内容，需要综合考虑精度、费用、抽样方式等多方面因素。问卷分割技术将原来的长问卷分割为短问卷，问卷模式、抽样方式、费用等均发生了变化，因此，分割后短问卷的样本量需要重新设计。如果仅是简单地按照原有长问卷的样本量来确定短问卷的数量，显然缺乏理论依据和合理解释。在此，我们考虑在**固定费用下精度最高**或**固定精度下费用最小**两种框架下探讨问卷的最优样本量设计。实际项目中，具体应用哪一种框架则要由市场研究决策者根据实际情况确定。

这里对需要用到的一些符号给出定义，用C_B表示调查所需要的总花费， c_0 表示与样本量无关的固定花费，c_j 表示第 j 个模式（问卷）每个样本的花费；W_k 表示不同数据项的重要性，且 $\sum_{k=1}^{K} W_k = 1$，估计精度为 μ_k。

2.4.1 固定费用下精度最高

在此约束条件下，精度使用如下最小化距离函数D^{sq} 表示：

$$arg\min_{\boldsymbol{n}} D^{sq}(\boldsymbol{n}) = \sum_{k=1}^{K} W_k CV(\widehat{Y}_k)^2 \qquad \text{公式 (3)}$$

使得 $c_0 n + \sum_{j=1}^{J} c_j n_j \le C_B$, $\sum_{k=1}^{K} W_k = 1$

其中，$\boldsymbol{n} = (n_1, n_2, \ldots, n_J)$ 表示 J 个不同模式（问卷）的最优样本量分配组合，n 表示样本

总量。$CV\left(\widehat{Y}_k\right)^2$表示第 k 个数据项的误差，在这里用 $CV(\widehat{Y}_k)^2 = \mathrm{Var}(\widehat{Y}_k)/\widehat{Y}_k^2$这个变异系数来衡量，将最小化距离函数和约束条件转化为拉格朗日乘子法，可求出各问卷模式的样本量分配。

2.4.2 固定精度下费用最小

固定精度情况下费用最小化是为了保证我们调查数据的质量，我们用另外一种方式来表达，如下最小化距离函数C^{sq}：

$$\arg\min_{\boldsymbol{C}} C^{sq}(\boldsymbol{n}) = c_0 n^* + \sum_{j=1}^{J} c_j n_j^* \qquad \text{公式（4）}$$

使得 $CV\left(\widehat{Y}_k\right)^2 \le \mu_k, k = 1,2,\cdots,K$

其中，$\boldsymbol{n}^* = \left(n_1^*, n_2^*, \ldots, n_J^*\right)$ 表示 J 个不同模式（问卷）的最优样本量分配组合，n^* 表示样本总量。$CV\left(\widehat{Y}_k\right)^2$表示第 k 个数据项的误差，$\mu_k$是第 k 个数据项的误差上确界，求解方案同2.4.1。

2.4.3 数据实证

实际数据的误差计算涉及到多变量协方差矩阵的计算，利用程序可以实现，实现结果可以作为高精度的样本量理论值，其价值在于抽样设计阶段作为参考和指导。但是，由于实际项目的调查过程较为复杂，样本的实际花费与项目执行进展关系密切，可能出现波动变化的情况，不能完全依照公式推导，而根据项目实际情况来设计样本量更为合理。另外，综合考虑成本花费、回答率、数据质量，市场调查经验告诉我们 10-15 分钟的问卷性价比最高。因此，综合项目实际执行情况和行业经验，借助样本量设计理论框架，在费用和精度的约束下进行样本量设计是最佳的解决方案。

这里推荐一个简单方便的方式，即采用根据经费划拨和精度双项验证的方式，即根据经费划拨大体情况，和样本量计算公式计算出一个大体值，然后看此种情况下模拟是否能达到最优值，如果不能达到需要增加或减少花费直到达到为止。

2.5 问卷分割技术——多重插补技术探索

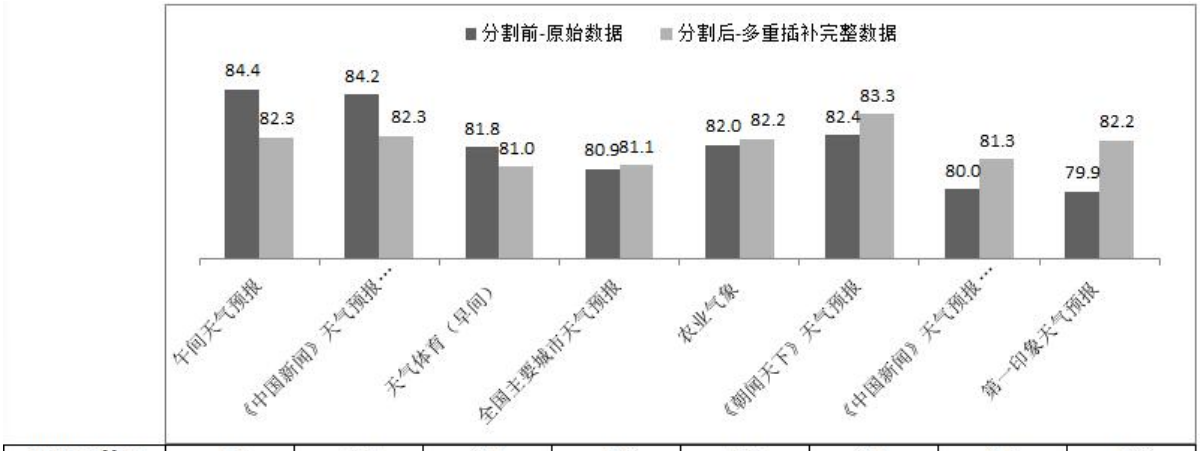
2.5.1 缺失数据处理的可行性分析

针对调查数据采集后的数据处理，一种方式是可直接根据现有调查数据进行各数据项的汇总分析和参数估计；另一种方式是针对样本量重新设计后可能出现的调查数据精度损失的情况，以及相对于原来的完整问卷而言，将受访者只回答了部分题目的短问卷视为包含“缺失数据”的情况，我们探索利用多重插补技术进行数据插补，根据插补后数据进行参数估计的可行性。近年来，处理无回答的插补技术日趋成熟，尤其是多重插补技术，完全能够应用在分割问卷中对缺失部分进行插补以构成完整的数据集。

2.5.2 数据实证

为验证多重插补技术的效果，我们需比较问卷分割前后的调查数据间有无统计学显著性差异，以及需要探讨不同数据缺失比例下多重插补的效果。调查问卷的数据类型主要分为离散型和连续型，因此我们针对不同的数据类型进行多重插补的效果分析。我们已通过电视观众满意度调查、消费者态度和行为研究、新产品上市研究等若干项目的模拟和实际测试，验证了基于问卷分割的多重插补技术能够再现原始数据，弥补数据缺失造成的精度损失。

首先，针对连续型数据验证多重插补效果，数据来源：2.2 节中某电视栏目受众调查。问卷要求受访者针对电视频道、电视栏目的满意度进行评分。



平均值差异	-2.1	-1.8	-0.8	0.2	0.2	0.9	1.3	2.2
标准差差异	1.7	0.9	-0.9	-0.4	-3.2	0.2	-1.0	-1.8

图 1 问卷分割前后，不同频道观众满意度打分平均分、标准差差异

由上图可见，问卷分割前后不同栏目观众满意度平均分差异、标准差差异均控制在微小范围内，进一步对问卷分割前后数据进行假设检验，概率 P 值均>0.05,无统计学显著性差异，说明多重插补的效果可以保障问卷分割后数据质量的稳定和可靠。

表 6 问卷分割前后，不同数据缺失比例，不同频道观众满意度打分平均分差异

	午间天气预报	《中国新闻》天气预报（午间）	天气体育	全国主要城市天气预报	农业气象	《朝闻天下》天气预报	《中国新闻》天气预报（早间）	第一印象天气预报
10%缺失	-2.1	-1.8	-0.8	0.2	0.2	0.9	1.3	2.2
30%缺失	1.2	3.1	1.2	1.5	3.5	1.5	2.9	1.3
50%缺失	3.5	5.5	2.9	2.3	3.3	-5.1	5.0	7.3

上表表示，不同数据缺失比例下不同栏目观众满意度平均分差异。可见，随着缺失比例的增加，该差异呈现逐渐增长的趋势，但是绝对值差异均在很小的范围内，并满足假设检验概率

P值>0.05，无统计学显著性差异。因此，多重插补仍能达到预期效果。

接着，针对离散型数据验证多重插补效果，数据来源：某进口鲜奶消费行为调查。

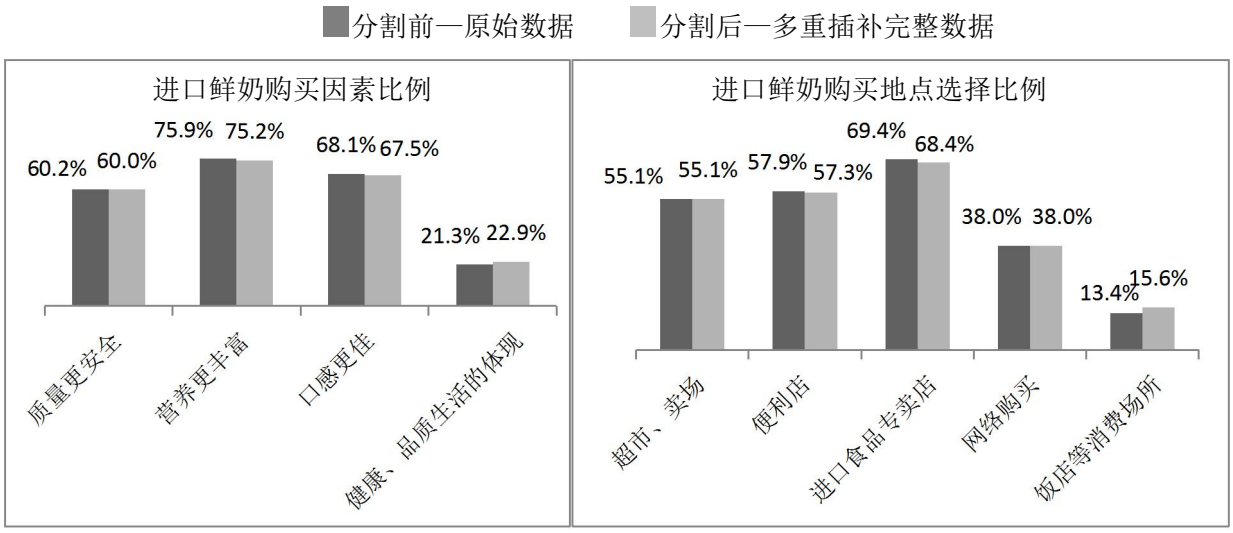


图 2 问卷分割前后，进口鲜奶的购买因素和地点选择比例差异

由上图可见，问卷分割前后不同选项的选择比例差异很小，进一步对问卷分割前后数据进行假设检验，概率 P 值均>0.05,无统计学显著性差异，说明多重插补的效果可以保证离散型数据的稳定和可靠。

表 7 问卷分割前后，不同数据缺失比例，进口鲜奶的购买因素和地点选择比例差异

	进口鲜奶购买因素				进口鲜奶购买地点				
	质量更安全	营养更丰富	口感更佳	健康、品质生活的体现	超市、卖场	便利店	进口食品专卖店	网络购买	饭店等消费场所
10%缺失	0.2%	0.7%	0.6%	-1.6%	0.0%	0.6%	1.0%	0.0%	-2.1%
30%缺失	-4.3%	3.8%	-1.2%	-3.6%	-2.9%	0.1%	2.7%	0.6%	-3.2%
50%缺失	-1.4%	1.9%	2.8%	-5.6%	-0.4%	1.4%	1.9%	-4.4%	-6.6%

上表表示，不同数据缺失比例下，不同问题选项的选择比例差异。由于离散型变量的变异比较微妙，插补的难度较连续性变量稍大，随着缺失比例的增加，差异也比连续性变量要大。因此，离散型数据进行多重插补时，尽量将数据缺失比例控制在较小的范围内（<30%）为宜。

通过上述分析，可得以下结论：数据缺失比例越小，多重插补效果越好。连续型变量的插补效果非常完美，二元离散型变量的插补效果略逊色于连续型变量，但在实际项目中能大概率地通过显著性差异检验，达到预期效果。并且，在较高的缺失比例情况下，多重插补与其他方法相比表现出更好的效果，我们也已通过若干项目的对比分析证明此结论，在此不再赘述。因

此，对于多变量的数据缺失，多重插补可以说是插补效果最好的缺失数据处理技术。

本章小结：问卷分割技术的实际项目应用，不宜过度复杂，应按照调查时长的要求，尽量选择简单的问卷模式，既要考虑到样本量设计的合理和高效，又要考虑到后期数据缺失比例尽可能控制在较小的范围内。技术流程中很多细节的把控需要丰富的市场调查经验的辅助和指导。

三、总结和展望

短小而精炼的问卷设计必将更好地适应互联网市场调查的发展趋势，更精准更透彻地洞悉消费者背后的真相。问卷分割技术的研究及应用，主要包括利用聚类算法将长问卷合理分解为短问卷，在费用和精度的约束下进行样本量的设计优化，以及采用多重插补技术来提高调查数据精度，从而保证调查数据质量。本文不仅完整搭建了问卷分割技术的理论模型，并且结合实际案例详细阐述了问卷分割技术的具体实施过程。

在当前市场调查研究复杂性越来越高，获取调查数据途径越来越广的形势下，客户和受访者对问卷调查设计的要求也越来越高，各个调查机构对长问卷进行适当分割再进行相应调查在未来势必将成为一种趋势。因此，我公司研究问卷分割技术的系统应用是十分必要且具有现实指导意义，丰富的案例讨论和经验的积累将对以后实际调查设计中实施问卷分割技术具有较大的借鉴价值，我们将在实践中不断对其进行完善和优化。

参考文献

[1] 赵雪慧.问卷分割方法在抽样调查中的应用[J].统计应用,2004，58 期

[2] Rässler, Susanne, Florian Koller, and Christine Mäenpää. *A split questionnaire survey design applied to German media and consumer surveys*. No. 42b/2002. Diskussionspapiere//Friedrich-Alexander-Universität Erlangen-Nürnberg, Lehrstuhl für Statistik und Ökonometrie, 2002.Chipperfield J O, Steel D G. Efficiency of split questionnaire surveys[J]. Journal of Statistical Planning & Inference, 2011, 141(5):1925-1932.

[3] Gonzalez, J. M. and Eltinge, J. L. (2007). Properties of Alternative Sample Design and Estimation Methods for the Consumer Expenditure Surveys. Paper presented at the 2007 Research Conference of the Federal Committee on Statistical Methodology, Arlington, VA, November, 2007.

[4] Feray Adigüzel and Michel Wedel (2008), Split-survey Design for Massive Surveys, Journal of Marketing Research, Vol. XLV, pp. 608-617.

[5] Chipperfield J O, Steel D G. Efficiency of split questionnaire surveys[J]. Journal of Statistical Planning & Inference, 2011, 141(5):1925-1932.

[6] Raghunathan T E, Grizzle J E. A Split Questionnaire Survey Design. Journal of the American Statistical Association, 1995, 90(429):54-63.

[7] Gonzalez J M . The Use of Responsive Split Questionnaires in a Panel Survey[J]. Dissertations & Theses - Gradworks, 2012.

[8] Rao J N K , Molina I . Small Area Estimation, 2nd Edition[M]., New York: John Wiley&Sons, 2015.

[9] Rubin D. B.(1987) Multiple Imputation For Nonresponse In Surveys[M]. New York:John wiley.

[10] Schenker N , Taylor J M G . Partially parametric techniques for multiple imputation[J]. Computational Statistics & Data Analysis, 1996, 22(4):425-446.