# 擬似ミクロデータ作成についての一考察

統計数理研究所　馬場康維

## １．はじめに

データ解析のためのプログラムの作成や統計教育のためには個票の利用が望ましい。しかし、様々な制約があるために実際の個票の利用は難しい。このような場合、集計結果がオリジナルによるものに近い結果となるようなデータがあれば有用である。

分析の多くは、変数間の関係を求めることにある。このため、しばしば、多変量解析法が手法として用いられる。馬場の一連の研究では、主成分分析等の線形変換に基づく多変量解析の諸手法では、連続量データをカテゴリーに変換したデータを用いてもオリジナルデータの結果に近い結果が得られることが示されている。これは、多変量解析の多くの方法が、データの分散に基づく方法であることに由来する。

連続・離散変換のこの性質を利用して擬似データを作成することができる。連続量をカテゴリー化しそれに雑音を加えてもとの分散・共分散構造を保ちながらオリジナルではないが本物に似せた擬似データの作成が可能である。このプロセスは、連続→離散→連続という変換の過程と考えることができる。このプロセスの中の離散→連続のなかに、カテゴリー⇒数量化というプロセスを挿入することにより、オリジナルデータ→擬似データ、という過程を少し複雑化して個票情報の秘匿に使えないかと言うのがこの研究の発端である。

## ２．数量化

数量化はカテゴリーそれぞれに数値を付与する方法である。変数間の相関が高くなるように数値が付与されるが、これは変数間に非線形の関係があったとしてもそれを線形で表現するように空間を変形する変換と考えることができる。この変換の過程で１次元尺度が得られれば、それに雑音を加えることによって擬似データの作成ができる。こうして作られた擬似データは、単純にカテゴリーに雑音を加えたものより複雑であるが故に単純な方法よりは個票情報にたどり着く危険性は遠い物になっている可能性がある。

## 参考文献

馬場康維（2010）．連続・離散変換による情報の保持と秘匿、日本計算機統計学会第 24 回大会予稿集，pp41-42

馬場康維, 岡本基, 野呂竜夫, 加藤真二 (2017). 研修教材としての擬似データの作成と利用, 2017 年度統計関連学会連合大会講演報告集、pp84

馬場康維（2017）．連続・離散変換の影響評価－線形と非線形－，日本行動計量学会第 45 回大会抄録集、pp158-161.