様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ５ 月 ２８ 日現在

機関番号：１２６１３
研究種目：研究活動スタート支援
研究期間：2019～2023
課題番号：１９Ｋ２３２４３

研究課題名（和文）GPSデータの個人特定化リスクに対する統計手法の開発

研究課題名（英文）Development of Statistical Methods for the Risk of Personal Identification from GPS Data

研究代表者
城田　慎一郎（Shirota, Shinichiro）

一橋大学・大学院ソーシャル・データサイエンス研究科・准教授

研究者番号：９０８４５９１８

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：　位置情報の社会科学への応用を考える際には、位置情報に紐づいた各個人情報とセットで提供することが重要であるが、個人特定リスクが内在する。個人情報の中でも、とりわけ位置情報は特定化リスクが高い。データ合成手法は、各変数の依存構造を保持したまま、対象となるデータ自体をシミュレートすることで、データの分布構造を保持した上での擬似データにより元データを代替するものである。本研究では、既存研究より柔軟性の高いアプローチとしてガウス過程を用いたデータ合成手法の開発を行った。これにより、個人特定化リスクを抑えたまま変数間の依存構造を保持したデータ合成手法の発展に貢献している。

研究成果の学術的意義や社会的意義
位置情報の社会科学への応用を考える際には、位置情報に紐づいた各個人情報とセットで提供することが重要であるが、個人特定リスクが内在する。個人情報の中でも、とりわけ位置情報は特定化リスクが高い。データ合成手法は、各変数の依存構造を保持したまま、対象となるデータ自体をシミュレートすることで、データの分布構造を保持した上での擬似データにより元データを代替するものである。本研究では、既存研究より柔軟性の高いアプローチとしてガウス過程を用いたデータ合成手法の開発を行った。これにより、個人特定化リスクを抑えたまま変数間の依存構造を保持したデータ合成手法の発展に貢献している。

研究成果の概要（英文）：　When considering the application of location information to social sciences, it is important to provide it along with individual information linked to the location. However, there is an inherent risk of personal identification. Among personal information, location information carries a particularly high risk of identification. Data synthesis methods simulate the target data itself while maintaining the dependency structure of each variable, thus replacing the original data with pseudo-data that retains the distribution structure of the data. In this study, we developed a data synthesis method using Gaussian processes as a more flexible approach than existing studies. This contributes to the development of data synthesis methods that maintain the dependency structure between variables while suppressing the risk of personal identification.

研究分野：統計学

キーワード：データ合成　空間統計　ガウス過程

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

(1)近年のデータサイエンスの流行を背景に、企業や官公庁では大規模かつ情報量の豊富なデータが蓄積しており、これらのデータの公開需要が高まっている。とりわけ経済学・マーケティングなどの分野では、個人の属性を含むデータの集積が多く、その公開の需要が高い。データを公開する際には、個人が特定できないように匿名化の処理を適切に行う必要がある。とりわけ、位置情報は個人特定リスクが高く、何らかの匿名化処置が必須となる。

(2)位置情報の社会科学への応用を考える際には、位置情報に紐づいた各個人情報とセットで提供することが重要である。これにより、位置情報を考慮した上での各個人情報を用いた分析が可能となる。一方で、位置情報をそのまま公開すると。個人特定リスクが内在する。とりわけ位置情報は個人特定リスクが高く、簡便な方法として空間的に粗いメッシュを作成することで位置情報を特定しにくくする方法が考えられる。この方法は扱いやすい反面、位置情報と個人情報の依存構造の情報が大幅に失われることに加え、データによってどの程度メッシュを粗くすれば十分なのかなどの課題がある。実際の運用においては、個人特定を避けるため、必要以上に粗いメッシュなどの処理をすることが考えられる。

(3)データ合成手法は、各変数の依存構造を保持したまま、対象となるデータ自体をシミュレートすることで、データの分布構造を保持した上で、生成した擬似データにより元データを代替するものである。データのプライバシーを扱う重要な手法の一つであり、データ合成手法によって生成した位置情報を用いることで、個人特定のリスクを回避することができる。また、メッシュで粗くすることによる位置情報に関する情報を失われることも少ないため、社会科学の分析に有用なデータを提供できる。

(4)一方で、位置情報のシミュレーションに関しては、変数の依存構造を十分に反映したものが研究開始当初は開発されていなかった。また、付随する変数のみでは、位置情報の空間パターンを十分に説明しきれないため、空間パターンを補完する手法の開発が必要であると考えられていた。データのプライバシーに関する研究を行なっている海外の研究者に相談を受けて、本研究が始まったという経緯がある。

２．研究の目的

(1)既存のデータ合成手法において、付随するデータ間の依存構造を十分に保持した上で、位置情報を生成することは難しかった。本研究では、データ間の依存構造をより柔軟に表現する手法を開発することで、より元データに近い形でデータ(とりわけ位置情報)を生成することを目指す。

(2)元データに近い性質を持つ位置情報を生成することができれば、生成したデータを利用した広範な分析が可能となり、社会や学術領域におけるデータサイエンスのさらなる発展への寄与を目的としている。とりわけ、位置情報を積極的に活用したマーケティングや経済分析がより進んでいくことを期待している。

３．研究の方法

(1)位置情報は点パターンデータと解釈できる。点パターンデータに対する統計的確率モデルとして空間点過程モデルが知られている。本研究でも、空間点過程をベースとしたモデルを構築し、位置情報を生成する。

(2)変数間の構造に加えて、内在する空間的な相関を考慮するためにガウス過程を用いる。通常、付随する情報だけでは、位置情報が持つ空間的な構造を説明しきれない場合が多い。ガウス過程を導入することで、これらの構造を捉えた上で、より元データに近い形で位置情報を生成する。

(3)既存手法によっても位置情報も生成し、比較検証を行う。

４．研究成果

(1)位置情報に関する空間展過程モデルを開発した。対数ガウスコックス過程と呼ばれ、対数強度関数にガウス過程を含むことで、モデルに柔軟性を与えている。これにより変数間の依存構造を正確に反映し、説明しきれない空間相関を反映した上での位置情報の生成が可能となる。

(2)開発したプログラムコード類は公開予定であり、現在整備中である。

(3)これにより、元データの性質を反映したデータ公開に向けての手法開発が進んだことになり、とりわけ社会科学系でのデータ利活用に貢献できたと考えている。

(4)一方で、非常に大規模なデータに関しては、計算コストが高いなどの課題も残った。これについては、高速かつ効率的な計算手法が見つかっており、本研究でも適応可能と考えており、今後の研究として取り組む予定である。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関