様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月　 ７ 日現在

機関番号：１２６０１
研究種目：挑戦的萌芽研究
研究期間：2014～2015
課題番号：２６５６０１６３

研究課題名（和文）データ駆動型統計学の可能性：動的で複雑な経済・社会現象の帰納的理解に挑む

研究課題名（英文）Possibility of data-driven statistics: Challenge to inductive understanding of dynamic and complex socio-economic phenomena

研究代表者
大西　立顕（Ohnishi, Takaaki）

東京大学・情報理工学（系）研究科・准教授

研究者番号：１０３７６３８７

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：データ駆動の視点から，動的で複雑な経済・社会現象に関連したデータを解析した．k-近傍法を用いて，データから自動的に不動産価格を予測する方法を提案した．予測誤差を最小にする近傍数が存在し，説明変数が増えるほど誤差が小さくなることを見出した．相関の強さから面積と場所が説明変数として重要であることを明らかにした．貿易品別の国際貿易ネットワークから統計的有意に出現しやすいネットワークモチーフを抽出し，国の特徴や経済学的機能を反映したネットワーク構造を明らかにした．

研究成果の概要（英文）：Dynamic and complex data observed in socio-economic phenomena are studied based on inductive approach. We proposed the use of k-nearest neighbor regression to automatically value real estate property. We showed that there is an optimal number of nearest neighbors minimizing the prediction error. As the number of explanation variables increases, the error becomes small. We estimated strength of the correlation and found its size and location are important to predict house price. We detected significant three-node motifs, which are significantly more abundant than expected by chance, observed in the world trade network of multiproduct world trade between countries. The obtained motifs characterize the country and reflect particular economic functions.

研究分野：数理工学

キーワード：経済物理学　超並列計算　不動産市場　価格予測　非線形　k-近傍法　国際貿易ネットワーク　ネットワークモチーフ

１．研究開始当初の背景

情報通信技術と計算機性能の向上により，経済・社会システムは情報通信技術を基盤としたものになり，我々が日々行っている経済・社会活動に関する多様で詳細な情報が高頻度に記録されるビッグデータ時代になってきている．これまでの人文・社会科学はこのような詳細なデータが観測できなかったために，概念・理論を重要視して理論構築されてきたが，今ではこれらの膨大なデータに基づいて実証科学の視点から人文・社会科学を再構築することが可能になってきている．本課題では，スーパーコンピュータを活用して現実の経済・社会に関するビッグデータをデータ駆動の視点から実証分析する．

２．研究の目的

経済・社会活動は非常に複雑なシステムであり，理論・仮説（第一原理）に基づくアプローチよりも，データに基づいてモデルを構築するデータ駆動型のアプローチの方が有効である．特にビッグデータ時代において，その有効性は増している．このような視点から経済・社会データの実証分析を行うことで，動的で複雑な経済・社会現象を帰納的に理解することを目的とする．

３．研究の方法

(1) 首都圏中古マンション売買データ

過去27年間の首都圏の中古マンション売買の取引データ約 100 万件を用いて物件価格を予測する問題を考えた．金融市場では同一銘柄の取引が頻繁に行われているため，価格を予測する際にはその銘柄の直近過去の取引価格を参考にすれば良い．しかし，不動産市場ではまったく同一の物件が頻繁に取引されることはほとんどなく，直近過去の取引が10年以上も昔になってしまうこともあり，同一物件の過去の取引事例は必ずしも参考にならない．したがって，実務上は勘と経験に基づいてその物件と同じような特徴を持つ物件を何件か取り出し，価格を外挿して手動で予測を行うのが典型的な方法になっている．そこで，データを十分に活用して価格を自動的に決定する方法を開発するために，直近過去の取引事例の中から予測したい物件に類似した属性（面積，緯度，経度，築年数）を有する物件を k 件取り出し，これらの平均価格を予測値とする重みつき k-近傍法のモデルを用いて解析する．

(2) 貿易品別の国際貿易ネットワーク

1962年から2000年についての国際貿易ネットワークの分析を行った．多くのネットワーク指標は，単にその値を求めるだけでは意味をなさず，適切にランダマイズされたネットワークと比較することではじめて，統計的有意なネットワークの性質を明らかにすることができる．そこで，次数を保存してランダムにつなぎ替えたランダマイズドネットワークを作成し， 3 カ国間の貿易関係（部分グラフ）に注目した．注目している部分グラフについて，実ネットワークでの出現回数とランダマイズドネットワークでの出現回数とで統計的有意に差があるかどうかを分析し， 3 カ国間の貿易関係において統計的有意に出現しやすい（しにくい）構造を抽出する．

４．研究成果

(1) 首都圏中古マンション売買データ

k-近傍法により物件価格を予測する分析を行った．物件価格は専有面積だけでなく，緯度（図1），経度，築年数など様々な変量と複雑に相関している．このような複雑で非線形な相関構造を明らかにし，この相関を考慮した上で物件価格を精度高く予測するために k-近傍法によ

る物件価格予測を検討した．面積　A，緯度 Long，経度 Lat，築年数 Y のみを用いて各物件の属性は４次元ユークリッド空間上の一点として表現できるとする．ただし，各変量は平均 0，分散 1 に標準化するものとする．直近過去１ヶ月間の取引事例の中から予測したい物件に類似した属性を有する物件（４次元空間上の最近傍の標本　）を k 個取り出し，これらの平均対数価格を予測値とする k-近傍法による手法：

$$\hat{\log P} = \frac{1}{k} \sum_{i \in knn} \log P_i$$

を検討した．この手法は，非線形でノンパラメトリックな予測であり，局所的な情報のみを使った予測になる．一般に，パラメータ k は平均二乗予測誤差を最小にするように定める．　k が非常に大きくなるとほぼすべての標本を参照するようになるため予測値が全データの平均値に近づいて誤差が大きくなり，逆に，　k が非常に小さくなると参照する標本が少な過ぎるために誤差が大きくなる．その間の値に，誤差を最小にする k が存在すると考えられる．

２年間の期間のデータを用いて分析を行った結果，　k=8 のときに誤差が最小になることが分かった（図 2 の黄）．つまり，属性の近い物件 8 件を参照して予測するのが良いことになる．さらに，　4 変量すべてを用いるのではなく，　3 変量のみを用いた予測，　2 変量のみを用いた予測，　1 変量のみを用いた予測も行った（図2)．3 変量のみを用いた場合は（A，Long，Lat），（A，Long，Y），（A，Lat，Y），　（Long，Lat，Y），　2 変量のみを用いた場合は　（A，Long），（A，Lat），（A，Y），（Lat，Y），（Long，Lat），（long，Y）の順で誤差が大きくなり，価格予測において，面積　A と緯度Longが重要であることを明らかにした．

（面積，経度，緯度，築後日数，都心までの時間，階，構造，向き，交通手段　）を

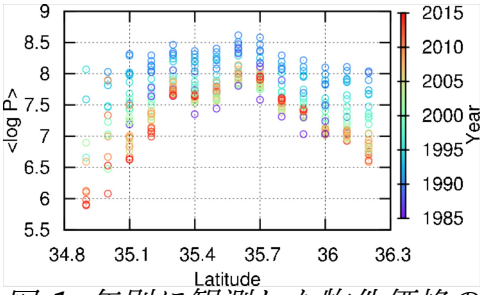

図 1: 年別に観測した物件価格の緯度依存性

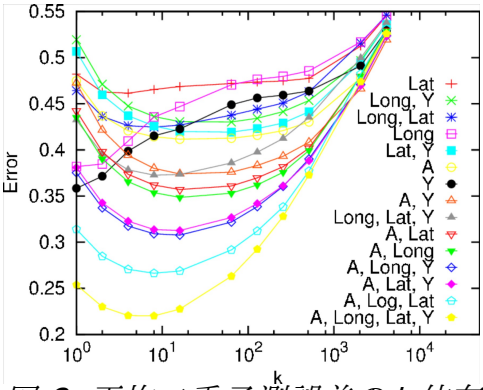

図 2: 平均二乗予測誤差の k 依存性

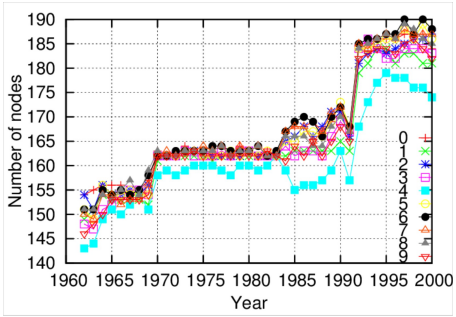

図 3: ノード数の年次推移

説明変数として価格を分類（予測）する決定木の解析でも，価格を予測するためには面積と都心までの時間が重要であることが判明した．したがって，面積と位置が価格を決める重要な要因になっていると考えられる．

(2) 貿易品別の国際貿易ネットワーク

国際貿易ネットワークのデータは数千の貿易品別に分かれているが，それらを10品目（図 3 では 0 ～ 9 と表示する ）にまとめて分類して分析した．ノード（国）の個数は 140 ～ 190 カ国である（図3)．リンク（貿易関係 ）の個数は数

1000程度であり，1980年代前半に大きく減少している（図4)．ネットワークの構造・機能においてはリンク数の分布（次数分布）が重要な影響を及ぼしている．多くの経済・社会ネットーワークでは次数がベキ分布に従うこと（スケールフリーネットワーク）が知られている．国際貿易ネットワークも次数はベキに近い分布に従っているが，ノード（国）数に上限（約200)があるために明確ではない．

多くのネットワーク指標は，単にその値を求めるだけでは意味をなさず，適切にランダマイズされたネットワークと比較することではじめて，統計的有意なネットワークの性質を明らかにすることができる．そこで，バイアスが生じるのを避けるためMCMC switching algorithmを用いて，次数を保存してランダムにつなぎ替えたランダマイズドネットワークを作成した．そして，３カ国間の貿易関係（部分グラフ）に注目した．注目している部分グラフについて，実ネットワークでの出現回数　Nとランダマイズドネットワークでの出現回数　Mとで統計的有意に差があるかどうかを

$$Z = \left(N - < M >\right) / \sqrt{< \left(N - < M >\right)^2 >}$$

を用いて分析し（図5)，３カ国間の貿易関係において統計的有意に出現しやすい（しにくい）構造を明らかにした（図6)．

次に，貿易金額を考慮した詳細な解析を行うために，三角貿易に注目した．国　iが国 jから貿易品 lを輸入するときの輸入金額を$w_{ij}^{(l)}$とする$w_{ij}^{(l)} > 0$，$w_{jk}^{(m)} > 0$，

$w_{ki}^{(n)} > 0$である国 i，j，k，貿易品 l，m，nについて，任意の貿易品　pに対して

$$w_{ij}^{(p)} = 0 (p \neq l), w_{ik}^{(p)} = 0,$$

$$w_{jk}^{(p)} = 0 (p \neq m), w_{ji}^{(p)} = 0,$$

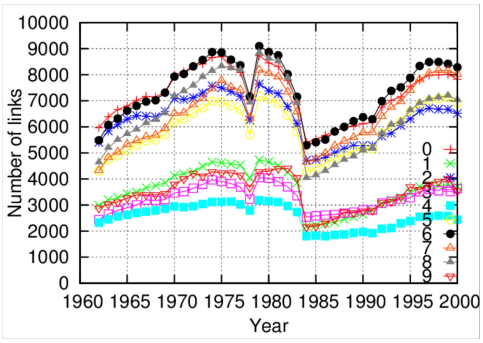

図 4: リンク数の年次推移

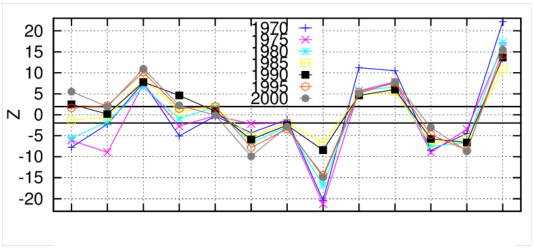

図　5: 1970，1975，1980，1985，1990，1995，2000 年についての各部分グラフのZ値

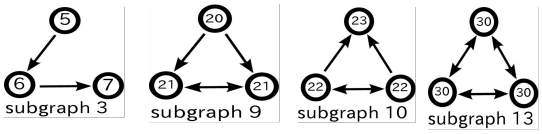

図 6: 統計的有意に出現する部分グラフ（モチーフ）

$$w_{ki}^{(p)} = 0 (p \neq n), w_{kj}^{(p)} = 0$$

を満たすとき，国　i，j，k の三角貿易の重要性を

$$w_{ij}^{(l)} + w_{jk}^{(m)} + w_{ki}^{(n)}$$

により定量化し，潜在的な三角貿易の構造を明らかにした．

５．主な発表論文等

〔学会発表〕（計 6 件）

1_ Takaaki Ohnishi, Takayuki Mizuno, Chihiro Shimizu, Hiroshi Iyetomi, Tsutomu Watanabe, ”Real estate valuation using k-nearest neighbor regression”,

Econophysics-2015, 2015年11月27日 ～ 2015 年 12 月 01 日 ， New Delhi(India)

2_ 大西立顕，水野貴之，清水千弘，家富洋，渡辺努， ”k- 近傍回帰法による経済データの実証分析 ”，平成 27年度統数研共同研究集会「経済物理学とその周辺」第 1回研究会， 2015年09月24日～2015年09月25日，鳥取大学（ 鳥取県鳥取市 ）

3_ Takaaki Ohnishi, Takayuki Mizuno, Chihiro Shimizu, Hiroshi Iyetomi, Tsutomu Watanabe, ”Using k-nearest neighbor method to estimate real estate prices”, Econophysics Colloquium 2015, 2015年09月14日～2015年09月16日 , Prague(Czech Republic)

4_ Takaaki Ohnishi, Takayuki Mizuno, Yuichi Ikeda, Hiroshi Iyetomi, Tsutomu Watanabe, ”Network Motifs in the World Trade Network”, NetSci2015, 2015年06月01日～2015年06月05日 , Zaragoza(Spain)

5 小林秀輔，大西 立顕，”ベイジアンネットワークの日経平均推定への応用 ”，平成 26年度統数研共同研究集会「経済物理学とその周辺」第 2 回研究会， 2015年03月26日～2015年03月27日，統計数理研究所（ 東京都立川市 ）

6_ 大西立顕，水野貴之，清水千弘，家富洋，渡辺努， ”k近傍法による不動産価格の予測 ”，平成 26年度統数研共同研究集会「経済物理学とその周辺」第 1 回研究会， 2014年09月11日～2014年09月12日，キヤノングローバル戦略研究所 （ 東京都千代田区 ）

６．研究組織

(1) 研究代表者

大西 立顕（ OHNISHI, Takaaki ）
東京大学・情報理工学系研究科・准教授
研究者番号：１０３７６３８７